IMMM 2016

The Sixth International Conference on Advances in Information Mining and Management

ISBN: 978-1-61208-477-0

DATASETS 2016

The International Symposium on Challenges for Designing and Using Datasets

May 22 - 26, 2016

Valencia, Spain

IMMM 2016 Editors

Elsa Macias-López, Las Palmas de Gran Canaria University, Spain
Alvaro Suarez, Las Palmas de Gran Canaria University, Spain

IMMM 2016

Foreword

The Sixth International Conference on Advances in Information Mining and Management (IMMM 2016), held between May 22-26, 2016, in Valencia, Spain, continued a series of academic and industrial events focusing on advances in all aspects related to information mining, management, and use.

The amount of information and its complexity makes it difficult for our society to take advantage of the distributed knowledge value. Knowledge, text, speech, picture, data, opinion, and other forms of information representation, as well as the large spectrum of different potential sources (sensors, bio, geographic, health, etc.) led to the development of special mining techniques, mechanisms support, applications and enabling tools. However, the variety of information semantics, the dynamic of information update and the rapid change in user needs are challenging aspects when gathering and analyzing information.

IMMM 2016 also featured the following Symposium:

- DATASETS 2016: The International Symposium on Challenges for Designing and Using Datasets

We take here the opportunity to warmly thank all the members of the IMMM 2016 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to IMMM 2016. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the IMMM 2016 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that IMMM 2016 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of information mining and management.

We are convinced that the participants found the event useful and communications very open. We hope that Valencia provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

IMMM 2016 Chairs:

IMMM General Chairs

Elsa Macias-López, Las Palmas de Gran Canaria University, Spain
Alvaro Suarez, Las Palmas de Gran Canaria University, Spain

IMMM Advisory Committee

Andreas Schmidt, Karlsruhe University of Applied Sciences, Germany
Kuan-Ching Li, Providence University, Taiwan

Abdulrahman Yarali, Murray State University, USA
Alain Casali, Aix Marseille Université, France
Ingrid Fischer, Universität Konstanz, Germany
Christian Ernst, Ecole des Mines de St Etienne - Gardanne, France
Paolo Garza, Politecnico di Torino, Italy
Bartlomiej Jefmanski, Wroclaw University of Economics, Poland
Nathalie Pernelle, Université Paris-Sud, France
Jürgen Pfeffer, Carnegie Mellon University, USA
Jörg Scheidt, University of Applied Sciences Hof, Germany
Ariella Richardson, Jerusalem College of Technology, Israel
Lorna Uden, Staffordshire University, UK
Eli Upfal, Brown University - Providence USA
Domenico Ursino, Università Mediterranea di Reggio Calabria, Italy
Jan Zizka, Mendel University - Brno, Czech Republic

IMMM Industry/Research Liaison Committee
Stefan Brüggemann, Astrium GmbH - Bremen, Germany
Olivier Caelen, Atos Worldline, Belgium
Feng Yan, Facebook Inc., USA
Katja Pfeifer, SAP AG, Germany
Arno H.P. Reuser, Reuser's Information Services, The Netherlands
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Wei Jin, Amazon.com, Seattle, USA
Olivier Caelen, Atos Worldline, Belgium
Yili Chen, Monsanto Company, USA
Gloria Bordogna, CNR - National Research Council / IDPA - Institute for the Dynamics of Environmental Processes - Dalmine, Italy
Josiane Mothe, IRIT, France
Dirk Labudde, Hochschule Mittweida, Germany
Yoseba Penya Landaburu, University of Deusto - Basque Country, Spain
Robert Wrembel, Poznan University of Technology, Poland

IMMM Publicity Chairs
Alessia Saggese, University of Salerno, Italy
Ludovico Boratto, Università di Cagliari, Italy
Toshio Kodama, University of Tokyo, Japan

DATASETS General Chairs
Elsa Macias-López, Las Palmas de Gran Canaria University, Spain
Alvaro Suarez, Las Palmas de Gran Canaria University, Spain

DATASETS 2016 Advisory Committee
Sung-Bae Cho (Chair), Yonsei University, Korea
Kyung-Joong Kim, Sejong University, Korea

IMMM 2016

COMMITTEE

IMMM General Chairs
Elsa Macias-López, Las Palmas de Gran Canaria University, Spain
Alvaro Suarez, Las Palmas de Gran Canaria University, Spain

IMMM Advisory Committee

Andreas Schmidt, Karlsruhe University of Applied Sciences, Germany
Kuan-Ching Li, Providence University, Taiwan
Abdulrahman Yarali, Murray State University, USA
Alain Casali, Aix Marseille Université, France
Ingrid Fischer, Universität Konstanz, Germany
Christian Ernst, Ecole des Mines de St Etienne - Gardanne, France
Paolo Garza, Politecnico di Torino, Italy
Bartlomiej Jefmanski, Wroclaw University of Economics, Poland
Nathalie Pernelle, Université Paris-Sud, France
Jürgen Pfeffer, Carnegie Mellon University, USA
Jörg Scheidt, University of Applied Sciences Hof, Germany
Ariella Richardson, Jerusalem College of Technology, Israel
Lorna Uden, Staffordshire University, UK
Eli Upfal, Brown University - Providence USA
Domenico Ursino, Università Mediterranea di Reggio Calabria, Italy
Jan Zizka, Mendel University - Brno, Czech Republic

IMMM Industry/Research Liaison Committee

Stefan Brüggemann, Astrium GmbH - Bremen, Germany
Olivier Caelen, Atos Worldline, Belgium
Feng Yan, Facebook Inc., USA
Katja Pfeifer, SAP AG, Germany
Arno H.P. Reuser, Reuser's Information Services, The Netherlands
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Wei Jin, Amazon.com, Seattle, USA
Olivier Caelen, Atos Worldline, Belgium
Yili Chen, Monsanto Company, USA
Gloria Bordogna, CNR - National Research Council / IDPA - Institute for the Dynamics of Environmental Processes - Dalmine, Italy
Josiane Mothe, IRIT, France
Dirk Labudde, Hochschule Mittweida, Germany
Yoseba Penya Landaburu, University of Deusto - Basque Country, Spain
Robert Wrembel, Poznan University of Technology, Poland

IMMM Publicity Chairs

Alessia Saggese, University of Salerno, Italy
Ludovico Boratto, Università di Cagliari, Italy
Toshio Kodama, University of Tokyo, Japan

IMMM 2016 Technical Program Committee

Aseel Addawood, Cornell University, USA
Zaher Al Aghbari, University of Sharjah, UAE
Riccardo Albertoni, Consiglio Nazionale delle Ricerche - Genova, Italy
César Andrés Sanchez, Universidad Complutense de Madrid, Spain
Annalisa Appice, Università degli Studi di Bari Aldo Moro, Italy
Avi Arampatzis, Democritus University of Thrace, Greece
Liliana Ibeth Barbosa Santillán, University of Guadalajara, Mexico
Shariq Bashir, National University of Computer and Emerging Sciences, Pakistan
Bernhard Bauer, University of Augsburg, Germany
Grigorios N. Beligiannis, University of Western Greece - Agrinio, Greece
Jorge Bernardino, ISEC - Polytechnic Institute of Coimbra, Portugal
Konstantinos Blekas, University of Ioannina, Greece
Jacek Blazewicz, Poznan University of Technology, Poland
Ludovico Boratto, Università di Cagliari, Italy
Gloria Bordogna, CNR - National Research Council / IDPA - Institute for the Dynamics of Environmental Processes - Dalmine, Italy
Stefan Brüggemann, Astrium GmbH - Bremen, Germany
Olivier Caelen, Atos Worldline, Belgium
Alain Casali, Aix Marseille Université, France
Mirko Cesarini, University of Milano Bicocca, Italy
Nadezda Chalupova, Mendel University - Brno, Czech Republic
Chi-Hua Chen, National Chiao Tung University, Taiwan R.O.C.
Weifeng Chen, California University of Pennsylvania, USA
Yili Chen, Monsanto Company, USA
Been-Chian Chien, University of Tainan, Taiwan
Sung-Bae Cho, Yonsei University, Korea
Kendra Cooper, University of Texas at Dallas, USA
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Lois Delcambre, Portland State University, USA
Frantisek Darena, Mendel University - Brno, Czech Republic
Sébastien Déjean, Université de Toulouse & CNRS, France
Mustafa Mat Deris, University of Tun Hussein Onn, Malaysia
Emanuele Di Buccio, University of Padua, Italy
Qin Ding, East Carolina University - Greenville, USA
Mario Döller, University of Passau, Germany
Aijuan Dong, Hood College - Frederick, USA
Nikolaos Doulamis, National Technical University of Athens, Greece
Anass Elhaddadi, University of Paul Sabatier - Toulouse, France
Christian Ernst, Ecole des Mines de St Etienne - Gardanne, France
Manuel Filipe Santos, University of Minho, Portugal
Ingrid Fischer, Universität Konstanz, Germany
Rita Francese, Università degli studi di Salerno, Italy

Vadim Strijov, Computing Centre of the Russian Academy of Sciences, Russia
Tatiana Tambouratzis, University of Piraeus, Greece
Tõnu Tamme, University of Tartu, Estonia
Mehmet Tan, TOBB University of Economics and Technology, Turkey
Yi Tang, Chinese Academy of Sciences, China
Xiaohui (Daniel) Tao, The University of Southern Queensland, Australia
Olivier Teste, Université de Toulouse, France
Daniel Thalmann, Institute for Media Innovation (IMI) - Nanyang Technological University, Singapore
Alberto Tonda, UMR 782 GMPA - INRA, France
Duarte Trigueiros, University Institute of Lisbon, Portugal
Michael Tschuggnall, University of Innsbruck, Austria
Vincent S. Tseng, National Cheng Kung University, Taiwan, R.O.C.
Chrisa Tsinaraki, European Union - Joint Research Center (JRC), Italy
Pavel Turcinek, Mendel University - Brno, Czech Republic
Franco Turini, University of Pisa, Italy
Lorna Uden, Staffordshire University, UK
Eli Upfal, Brown University - Providence USA
Domenico Ursino, Università Mediterranea di Reggio Calabria, Italy
Fabio Vandin, University of Padova, Italy
Julien Velcin, Université de Lyon 2, France
Corrado Aaron Visaggio, University of Sannio, Italy
Zeev Volkovich, ORT Braude College Karmiel, Israel
Stefanos Vrochidis, Information Technologies Institute - Centre for Research and Technology Hellas, Greece
Baoying (Elizabeth) Wang, Waynesburg University, USA
Qi Wang, Northwestern Polytechnical University, China
Alexander Wöhrer, Vienna Science and Technology Fund, Austria
Hao Wu, Yunnan University - Kunming, P.R.China
Feng Yan, Facebook Inc., USA
Chao-Tung Yang, Tunghai University, Taiwan
Zhenglu Yang, University of Tokyo, Japan
Kui Yu, School of Computing Science - Simon Fraser University, Canada
Jan Zizka, Mendel University - Brno, Czech Republic

DATASETS General Chairs
Elsa Macias-López, Las Palmas de Gran Canaria University, Spain
Alvaro Suarez, Las Palmas de Gran Canaria University, Spain

DATASETS 2016 Advisory Committee

Sung-Bae Cho (Chair), Yonsei University, Korea
Kyung-Joong Kim, Sejong University, Korea

DATASETS 2016 Program Committee Members

Walid G. Aref, Purdue University, USA
Francisco Henrique Cerdeira Ferreira, Universidade Federal de Juiz de Fora, Brazil
Yun-Maw Kevin Cheng, Tatung University, Taiwan

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Semi-supervised Learning in the Framework of Data Multiple 1-D Representation 1
Jianzhong Wang

Onto-Traffic: A Semantic Traffic Analysis Tool based on GPS Data for Smart-Cities 5
Juan J. Sanchez-Escobar, Liliana I. Barbosa-Santillan, Luis F. Barbosa-Santillan, Fatemeh G. Nasri-Shandiz, and Gabriel A. Leon-Paredes

Using Recurrent Boltzmann Machines in the Detection of Process Drifts Using Several Process Mining Perspectives 9
Alessandro Berti

Machine Learning in Cloud Environments Considering External Information 11
Matthias Lermer, Stefan Frey, and Christoph Reich

Automatic City Knowledge Discovery from Internet Resources 18
Nazanin Firoozeh

Log-Modulus for Knowledge Discovery in Databases of Financial Reports 26
Duarte Trigueiros and Carolina Sam

The Mining and Analysis of Data with Mixed Attribute Types 32
Edward Wakelam, Neil Davey, Yi Sun, Amanda Jefferies, Parimala Alva, and Alexander Hocking

Practical Application of the Data Preprocessing Method for Kohonen Neural Networks in Pattern Recognition Tasks 38
El Khatir Haimoudi, Loubna Cherrat, Otman Abdoun, and Mostafa Ezziyyani

Semi-Supervised Learning in the Framework of Data Multiple 1-D Representation

Jianzhong Wang
College of Sciences
Sam Houston State University
Huntsville, Texas 77341–2206, USA
e-mail: jzwang@shsu.edu

***Abstract*—The paper develops 1D-based ensemble method for semi-supervised learning (SSL). The method integrates the classifier based on *data 1-D representations* and *label boosting* in a serial ensemble. In each stage, the data set is first represented by several 1-D stacks, which preserve the local similarity between data samples. Then, a *1-D ensemble labeler* (1DEL) is constructed and used to create a *newborn labeled subset* from the unlabeled set. United with the subset, the original labeled is boosted for the next learning stage. The boosting process is repeated till the updated labeled set reaches a certain size. Finally, a 1DEL is applied again to build the classifier. The validity and effectiveness of the method are confirmed by experiments. Comparing to several other popular SSL methods, the results of the proposed method are very promising.**

Keywords–Data 1-D representation; regularization; label boosting; ensemble; semi-supervised learning.

I. INTRODUCTION

In this paper, we introduce a novel ensemble method for SSL based on *data 1-D representation*. In SSL, the essential problem is data binary classification, which can be briefly described as follows: Assume that the samples (or members, points) of a given data set $X = \{\vec{x}_i\}_{i=1}^n \subset \mathbf{R}^m$ belong to two classes *A* and *B*, labeled by 1 and -1, respectively. Denote by y_j the label of the sample $\vec{x}_j$, where $y_j \in \{1, -1\}, 1 \leq j \leq n$. In a SSL problem, X is divided into two disjoint subsets: $X = X_\ell \cup X_u, X_\ell \cap X_u = \emptyset$, where the members in $X_\ell = \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_{n_0}\}$ have known labels $Y_\ell = \{y_1, y_2, \cdots, y_{n_0}\}$, while the labels for the members in $X_u = \{\vec{x}_{n_0+1}, \vec{x}_{n_0+2}, \cdots, \vec{x}_n\}$ are unknown. We often call a function $f : X \to \{1, -1\}$ a *classifier* (or *labeler*) on X. The *classification error* is measured by the misclassified number:

$$E(f) = |\{\vec{x} \in X |\ f(\vec{x}_i) \neq y_i\ 1 \leq i \leq n\}|,$$

where $|S|$ denotes the cardinality of a set S. Then, the quality of a classifier is measured by the *error rate* $E(f)/|X|$. The task of SSL is to find a classifier f with the error rate as small as possible.

The monograph [1] and the survey paper [2] gave a comprehensive review of various SSL methods, among which the popular ones are based on kernel technique such as transductive support vector machines, manifold regularization, and other graph-based methods [3] [4]. In these methods, using *kernel trick*, people construct a kernel function to map original samples onto a reproducing kernel Hilbert space (RKHS) [5], where the non-linear decision boundary in the raw data space becomes nearly linear. Thus, people can construct classifiers in the RKHS using regularization methods. The success of a kernel-based method strongly depends on the exploration of data structure by kernels. However, it is often difficult to design suitable kernels, which precisely explore the feature spaces. Recently, researchers have developed new SSL models, which construct classifiers without adopting kernel technique, for instance, the data-tree based method [6] [7] constructs the classifier based on the data multi-layer structure.

In all of the models above, a single classifier is employed to label unlabeled points. However, when a data set has a complicate intrinsic structure and high-dimensionality, a single classifier usually cannot complete the task satisfactorily. The proposed method takes the idea of ensemble methodology in the *multiple classifier systems* (MCSs) [8]: It build a final classifier by integrating multiple pre-classifiers. Since MCSs perform information fusion at different levels, they overcome the limitations of the traditional approaches [9]–[11].

The novelty of the introduced ensemble SSL method is the following: It adopts the framework of data 1-D representation, in which the data set is represented by several different 1-D sequences, then a classifier is constructed as an ensemble of pre-classifiers built on these sequences. Here, we choose 1-D models because 1-D decision boundary is a set of points on a line, which has the simplest topological structure. As a result, the pre-classifiers can be easily constructed by classical 1-D regularization methods without using kernel trick or data trees. Furthermore, the simplicity of 1-D models makes the algorithm for building the final classifier relatively reliable and stable. We new describe the architecture and technological process of our method in the following.

1) The data set X is first mapped to several 1-D sets $\{T^i\}_{i=1}^k$, which preserve the local similarity of members in X. Correspondingly, the couple $\{X_\ell, X_u\}$ is mapped to $\{T_\ell^i, T_u^i\}$ for each 1-D set T^i.
2) A pre-classifier g^i on X is constructed based on T^i by a 1-D regularization method. Then an ensemble labeler g on X is assembled from $\{g^i\}_{i=1}^k$ to label all members of X.
3) A *feasibly confident subset* $L \subset X_u$ is produced by g. According to the *class weights* of the members of L, a half of members in L is chosen into the *newborn labeled subset* S. Then, the initial labeled set X_ℓ is boosted to $X_\ell^{new} = X_\ell \cup S$.
4) The procedure above is repeated till the updated labeled set X_ℓ^{new} reaches a certain size. Finally, the classifier f is obtained by applying the ensemble labeler g on the newest couple $\{X_\ell^{new}, X_u^{new}\}$.

Our strategy adopts *Model-guided Instance Selection* (MIS) approach [9], but is slightly different from AdaBoost algorithm [12] in the sense that AdaBoost updates the misclassified weights on X_u, while our method updates the set X_u itself.

The paper is organized as follows: In Section II, we develop the 1-D based ensemble SSL method. In Section III, we

 ISBN: 978-1-61208-477-0

demonstrate the validity of our method in two examples and give the comparison of our results with other methods. The conclusion is given in the last section.

II. The 1-D based ensemble SSL method

In this section, we introduce the novel SSL method based on data 1-D representation.

A. Data 1-D Representations

Assume that the data set X is initially arranged in a stack $\mathbf{x} = [\vec{x}_1, \cdots, \vec{x}_n]$, where the first n_0 members are in Class A and others are in Class B. Let $d(\vec{x}, \vec{y})$ be a metric on X that measures the dissimilarity between the points of X. Let π be an index permutation of the index sequence $[1, 2, \cdots, n]$, which induces a permutation P_π on the initial stack $\mathbf{x}$, yielding a stack of X headed by $\vec{x}_{\pi(1)}$: $\mathbf{x}_\pi = P_\pi \mathbf{x} = [\vec{x}_{\pi(1)}, \cdots, \vec{x}_{\pi(n)}]$. We define the set of all permutations of X headed by $\vec{x}_\ell$ by

$$\mathcal{P}_\ell = \{P_\pi; \quad \pi(1) = \ell\}.$$

According to [13], the *shortest-path sorting of* X headed by $\vec{x}_\ell$ is the stack $\mathbf{x}_\pi$ that minimizes the path starting from $\vec{x}_\ell$ and though all points in X, i.e., $\mathbf{x}_\pi = P_\pi \mathbf{x}$, where P_π is given by

$$P_\pi = \arg\min_{P \in \mathcal{P}_\ell} \sum_{j=1}^{n-1} d((P\mathbf{x})_j, (P\mathbf{x})_{j+1}). \tag{1}$$

Let the stack $\mathbf{x}_\pi$ be the shortest-path sorting of X headed by $\vec{x}_\ell$. Set

$$t_1 = 0, \; t_{j+1} - t_j = \frac{d(\vec{x}_{\pi(j)}, \vec{x}_{\pi(j+1)})}{\sum_{k=1}^{n-1} d(\vec{x}_{\pi(k)}, \vec{x}_{\pi(k+1)})}. \tag{2}$$

Then, the stack $\mathbf{t} = [t_1, \cdots, t_n]$ is called the 1-D (shortest-path) representation of X headed by $\vec{x}_\ell$.

The problem (1) has NP computational complexity. A greedy algorithm to find an approximation of P_π in (1) is referred to [13]. Once, P_π is found, the corresponding 1-D representation is obtained by (2).

Denote by T the set of the components of $\mathbf{t}$. The bijective mapping $h : T = h(X_\ell)$ is called a 1-D (shortest-path) embedding of X headed by $\vec{x}_\ell$, which also map the unlabeled set X_u onto $T_u = h(X_u) \subset T$. Then, a classifier on T induces a classifier on X. Since T is a 1-D set, its class decision boundary is reduced to a discrete set in $[0, 1]$.

B. The 1-D based ensemble labeler

Although the simplest topological structure of data 1-D representation reduces the decision boundary to a discrete set in $[0, 1]$ points, a single 1-D representation cannot truly preserve the data similarity because the sorting is a serial process that makes earlier selected adjacent pairs are more similar than the later selected ones. To overcome the drawback of a single 1-D embedding, we employ the *spinning technique* to build several 1-D representations. Based on each of them, we first construct a pre-classifier, then assemble an ensemble labeler from them. The following is the details.

Let $\vec{h} = [h_1, \cdots, h_k]$ be a k-ple 1D-embedding and P_i be the permutation operator on X corresponding to h_i such that the stack $\mathbf{x}_{\pi_i} = P_i \mathbf{x}$ is headed by a randomly selected point $\vec{x}_{\pi_i(1)}$. The embedding h_i produces a 1-D representation of X: $\mathbf{t}^i = h_i(\mathbf{x}_{\pi_i})$. For a function f on X, $s^i = f \circ h_i^{-1}$ is a function on $\mathbf{t}^i$. We now represent a function f on X by its vector form $\mathbf{f} = [f_1, \cdots, f_n]$, $f_j = f(\vec{x}_j)$, and a function s on $\mathbf{t}^i$ by the vector $\mathbf{s} = [s_1^i, \cdots, s_n^i]$, $s_j^i = s(t_j^i)$.

Let $T_\ell^i = h_i(X_\ell)$ and $T_u^i = h_i(X_u)$. Using a classical regularization method, we construct a pre-classifier g_i for X based on the couple $\{T_\ell^i, T_u^i\}$. For instance, denote by $C^1[0, 1]$ the space of smooth functions on $[0, 1]$ and by $Ds_j = (s(t_{j+1}^i) - s(t_j^i))/(t_{j+1}^i - t_j^i)$ the difference quotient of $s \in C^1[0, 1]$ on the stack $\mathbf{t}^i$ at t_j^i. Let q^i be the solution of the following constrained minimization problem:

$$q^i = \arg\min_{s \in C^1[0,1]} \frac{1}{n_0} \sum_{j=1}^{n_0} \left(s(h^i(\vec{x}_j)) - y_j\right)^2 + \frac{\lambda}{2} \sum_{j=1}^{n-1} (Ds_i)^2, \tag{3}$$

subject to the constraint

$$\frac{1}{n} \sum_{j=1}^{n} s(t_j^i) = M,$$

where M can be chosen to $M = \frac{1}{n_0} \sum_{j=1}^{n_0} y_j$. We denote by $\vec{1}$ the vector whose all entries are 1, denote by I_{n_0} the $n \times n$ diagonal matrix, in which only $(\pi^i(j), \pi^i(j))$-entries are 1, $1 \le j \le n_0$, but others are 0. Set $w_0 = w_n = 0, w_j = 1/(t_{j+1}^i - t_j^i)^2$, and denote by $D = [D_{i,j}]$ the $n \times n$ three-diagonal matrix, in which

$$\begin{cases} D_{j,j} = w_{j-1} + w_j & 1 \le j \le n, \\ D_{j,j+1} = D_{j+1,j} = -w_j & 1 \le j \le n-1, \end{cases}$$

Then, the vector representation of q^i on the stack $\mathbf{t}^i$ is the following solution

$$\mathbf{q}^i = (I_{n_0} + n_0 \lambda D)^{-1} (\vec{y} + \mu \vec{1}), \tag{4}$$

with

$$\mu = \frac{M - \mathcal{E}\left((I_{n_0} + n_0 \lambda D)^{-1} \vec{y}\right)}{\mathcal{E}\left((I_{n_0} + n_0 \lambda D)^{-1} \vec{1}\right)},$$

where $\mathcal{E}(\vec{v})$ denotes the mean value of the vector $\vec{v}$. We define the pre-classifier on X associated with the 1-D embedding h_i by $g^i = q^i \circ h_i^{-1}$. Finally, we define 1DEL on X by

$$g(\vec{x}) = \frac{1}{k} \sum_{i=1}^{k} \text{sign}(g^i(\vec{x})), \quad x \in X. \tag{5}$$

C. The newborn labeled subset selector

Using the 1-D ensemble labeler g in (5), we construct

$$L^+ = \{\vec{x} \in X_u; \; g(\vec{x}) = 1\}, \quad L^- = \{\vec{x} \in X_u; \; g(\vec{x}) = -1\}.$$

In a great chance, L^+ contains the members in Class A, while L^- contains the members in Class B. We call $L = L^+ \cup L^-$ the *feasibly confident subset* created by g. For convenience, we denote the set operator that create the feasibly confident subset L from X_u by $\mathbf{G} : \mathbf{G}(X_u) = L$.

We now select a half of the members in L to form a *newborn labeled subset* $S = S^+ \cup S^-$, where S^+ contains all Class-A members in L^+ and S^- contains all Class-B members in L^-. They are constructed as follows. Let X_ℓ^+ contain all Class-A members of X_ℓ and X_ℓ^- contain all Class-B members of X_ℓ. For each $\vec{x} \in L$, define $d(\vec{x}, X_\ell^+) = \min_{\vec{y} \in X_\ell^+} d(\vec{x}, \vec{y})$

and $d(\vec{x}, X_\ell^-) = \min_{\vec{y}\in X_\ell^-} d(\vec{x}, \vec{y})$. We now associate $\vec{x}$ with the *class weight*

$$w(\vec{x}) = \frac{d(\vec{x}, X_\ell^-)}{d(\vec{x}, X_\ell^-) + d(\vec{x}, X_\ell^+)}.$$

Finally, let the set S^+ contain the half of members of L^+ with the greatest class weights and S^- contain the half of members in L^- with the smallest class weights. We call the operator $\mathbf{S} : \mathbf{S}(L) = S$ a *newborn labeled subset selector* and call the composition $\mathbf{M}= \mathbf{S} \circ \mathbf{G}$ a *newborn labeled subset creator* because the newborn labeled subset $S = M(X_u)$.

D. Construction of the final classifier

We now build the (final) classifier by a serial ensemble, in which the labeled set is cumulatively boosted. Let the initial labeled set be equipped with the index 0: $X_\ell^0 = X_\ell$. Starting from X_ℓ^0, we apply the newborn labeled subset creator $\mathbf{M}_1$ to create a newborn labeled set S^1, which is united with X_ℓ^0 to produce $X_\ell^1 = X_\ell^0 \cup S^1$. Repeating the procedure n times, the labeled set will be cumulatively boosted to a labeled set X_ℓ^n:

$$X_\ell^0 \Longrightarrow X_\ell^1 \Longrightarrow \cdots \Longrightarrow X_\ell^n.$$

We set a *boosting-stop parameter* $p, 0 < p < 1$. The process will not be terminated until the labeled set X_ℓ^n reaches the size $|X_\ell^n| \geq p|X|$. Finally, we apply 1DEL on the couple $\{X_\ell^n, X_u^n\}$ to construct the final classifier f on X, which labels each $\vec{x} \in X$ by $\operatorname{sign} f(\vec{x})$.

III. EXPERIMENTS

We use two benchmark databases of handwritten digits, MNIST [21] and USPS [22] in the experiments to present the validity and effectiveness of the proposed method. In the literature of machine learning, MNIST is often used to test the error rate of classifiers obtained by supervised learning. The best result for the error rate up to 2012 was 0.23%, reported in [14] by using the convolutional neural network technique. In 2013, the authors of [15] claimed to achieve 0.21% error rate using DropConnect, which is based on regularization of neural networks. Because in SSL no large training set is available for producing classifiers, the error rates obtained by SSL methods usually are much higher than the claimed error rates obtained by supervised learning. Besides, the error rates of SSL are strongly dependent the size of the initial label set X_ℓ. In general, the smaller the size of X_ℓ, the higher the error rate. Hence, it is unfair to compare the error rates obtained by SSL methods to the above recorded ones.

In all of our experiments, the spin number 3 is used for constructing 1DEL while 20 for building the final classifier, and the boosting-stop parameter p is set to 0.7.

For comparison, we choose the same data setting as in [7]: In MINST, for each of the digits $\{3, 4, 5, 7, 8\}$, 200 samples were selected at random so that the cardinality of the data set is $|X| = 1000$, where the digit 8 is assigned to Class B, and others belong to Class A. In USPS, for each of the digits $0-9$, 150 samples are selected at random so that $|X| = 1500$, where the digits 2 and 5 are assigned to Class B, and others belong to Class A. In all experiments, the initial labeled set X_0 is preset to 10 various sizes of $10, 20, \cdots, 100$, respectively, and the labeled digits are distributed evenly on each chosen digit.

TABLE I. ERROR RATE OF THE PROPOSED 1-D BASED ENSEMBLE SSL METHOD FOR 50 RANDOMLY SELECTED SUBSETS FROM MNIST WITH $|X| = 1000$.

$\lvert X_0\rvert$	10	20	30	40	50	60	70	80	90	100
Mean%	7.8	7.9	4.6	2.5	2.1	1.9	1.9	1.9	1.2	1.2
Min%	7.6	7.9	4.6	1.9	2.1	1.9	1.9	1.9	1.2	1.2
Max%	19.4	7.9	4.6	3.5	2.1	1.9	1.9	1.9	1.2	1.2
STD%	1.7	0	0	0.7	0	0	0	0	0	0

TABLE II. ERROR RATE OF THE PROPOSED 1-D BASED ENSEMBLE SSL METHOD FOR 50 RANDOMLY SELECTED SUBSETS FROM USPS WITH $|X| = 1500$.

$\lvert X_0\rvert$	10	20	30	40	50	60	70	80	90	100
Mean%	3.3	2.1	1.5	1.5	1.3	1.4	1.4	1.4	1.4	1.2
Min%	2.0	1.3	1.5	1.5	1.3	1.4	1.4	1.4	1.4	1.2
Max%	16.8	2.9	1.7	1.5	1.3	1.4	1.4	1.4	1.4	1.2
STD%	1.99	0.8	0.02	0	0	0	0	0	0	0

Note that a vector $\vec{x} \in X$ is originally represented by a $c \times c$ matrix $[x_{i,j}]_{i,j=1}^c$, where $c = 20$ for MNIST and $c = 16$ for USPS. To reduce the shift-variance, we define the 1-shift distance between two digit images [16]:

$$d(\vec{x}, \vec{y}) = \min_{\substack{|i'-i|\leq 1 \\ |j'-j|\leq 1}} \sqrt{\sum_{i=2}^{c-1}\sum_{j=2}^{c-1} (x_{i,j} - y_{i',j'})^2}.$$

We first run our algorithm on 50 subsets (with 1000 members) randomly chosen from the MNIST database and show the test results in Table I, where the first row is the number of samples in X_ℓ, and the $2^{nd}-5^{th}$ rows are the mean, minimum, maximum, and standard deviation of the error rates of the 50 tests, respectively. In the second experiment, we run our algorithm for USPS is a similar way: 50 subsets with 1500 members are randomly chosen from USPS database. The test results are shown in Table II, where the setting for the rows is the same as in Table I. The Tables I and II show that the standard deviations of the error rates are quite small, particular when the known labeled members are more than 1%. This indicates the high stability of the proposed SSL algorithm.

In Figure 1, we give the comparison of the average error rates (of 50 tests) of our 1-D based ensemble method to Laplacian Eigenmaps (Belkin & Niyogi, 2003 [3]), Laplacian Regularization (Zhu et al., 2003 [17]), Laplacian Regularization with Adaptive Threshold (Zhou and Belkin, 2011 [18]), and Haar-Like Multiscale Wavelets on Data Trees (Gavish et al., 2011 [7]) on the subsets randomly chosen from both MNIST and USPS databases. The results show that our method achieves competitive results comparing to other SSL methods.

We have also applied the proposed method on the real-world applications, such as the classification of hyperspectral images [19] and the face recognition [20]. In these experiments, we have even adopted a much simpler label boosting method: Choosing the newborn labeled subset at random. The obtained results are still very promising and superior over many other popular methods. It is also worth to point out that the method is not very sensitive to the parameters. For instance, in our experiments, if spinning numbers are set to 3–5, and the boosting-stop parameter is set in the range of

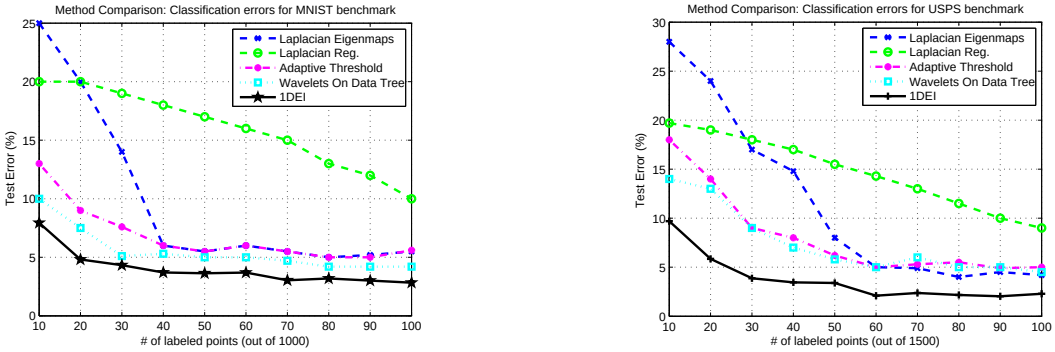

Figure 1. RESULT COMPARISON WITH DIFFERENT SSL MODELS.

0.6–0.8, the results are similar. The detailed discussion on the parameter tuning can be found [19] [20].

IV. CONCLUSION

We proposed a new ensemble method for SSL based on data 1-D representations, which enable us to construct ensemble classifiers assembled from several pre-classifiers for the same data set using classical 1-D regularization technique. Furthermore, a label boosting technique is applied for robustly enlarging the labeled set to a certain size so that the final classifier is built based on the boosted labeled set. The experiments show that the performance of the proposed method is superior to many popular methods in SSL. The new method also exhibits a clear advantage for learning the classifier when only a small labeled set is given. Because the method is independent of the data dimensionality, it can also be applied to various types of data. Since the algorithms to construct the classifiers in the proposed method only employ 1-D regularization technique, avoiding the complicate kernel trick, they are simple and stable. It can be expected that the created 1-D framework in this paper will be applied to the development of more machine learning methods for different purposes. In the future work, we will study how to accelerate the sorting algorithm in 1-D embedding and consider to integrate the data-driven wavelets with the proposed method.

ACKNOWLEDGMENT

This work is supported by SHSU-2015 ERG Grant no. 250711. The authors would like to thank the anonymous reviewers for their highly insights and helpful suggestions.

REFERENCES

[1] O. Chapelle, A. Zien, and B. Schölkopf, Semi-supervised Learning. MIT Press, 2006.

[2] X. Zhu, “Semi-supervised learning literature survey,” University of Wisconsin-Madison, Computer Sciences TR-1530, July 2008.

[3] M. Belkin and P. Niyogi, “Using manifold structure for partially labeled classification,” Advances in Neural Information Processing Systems, vol. 15, 2003, pp. 929–936.

[4] T. Joachims, “Transductive learning via spectral graph partitioning,” in Proceedings of the 20th International Conference on Machine Learning, 2003, pp. 290–297.

[5] V. Vapnik, Statistical Learning Theory. Wiley-Interscience, New York, 1998.

[6] R. Coifman and M. Gavish, “Harmonic analysis of digital data bases,” Applied and Numerical Harmonic Analysis. Special issue: Wavelets and Multiscale Analysis, 2011, pp. 161–197.

[7] M. Gavish, B. Nadler, and R. Coifman, “Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning,” in Proceedings of the 27th International Conference on Machine Learning, 2010, pp. 367–374.

[8] T. K. Ho, J. J. Hull, and S. N. Srihari, “Decision Combination in Multiple Classifier Systems,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 1, Jan. 1994, pp. 66–75.

[9] L. Rokach, “Ensemble-based classifiers,” Artif. Intell. Rev., vol. 33, 2010, pp. 1–39.

[10] M. Wozniak, M. Grana, and E. Corchado, “A survey of multiple classifier systems as hybrid systems,” Information Fussion, vol. 16, 2014, pp. 3–17.

[11] Z-H. Zhou, J. Wu, and W. Tang, “Ensembling neural networks: Many could be better than all,” Artif. Intell. vol. 137, 2002, pp. 239–263.

[12] D. Z. Li, W. Wang, and F. Ismail, “A selective boosting technique for pattern classification,” Neurocomputing, vol. 156, 2015, pp. 186–192.

[13] I. Ram, M. Elad, and I. Cohen, “Image processing using smooth ordering of its patches,” IEEE Trans. on Image Processing, vol. 22, no. 7, July 2013, pp. 2764–2774.

[14] C. Dan, U. Meier, and J. Schmidhuber, “Multi-column deep neural network for image classification,” IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3642–3649.

[15] W. Li, M. Zeiler, S. Zhang, Y. LeCun, and R. Fergus, “Regularization of neural network using dropconnect,” Journal of Machine Learning Research, vol. 28, no. 3, 2013, pp. 1058–1066.

[16] J. Z. Wang, “Semi-supervised learning using multiple one-dimensional embedding based adaptive interpolation,” International Journal of Wavelets, Multiresolution and Information Processing, vol. 14, no. 2, 2016, pp. 1640002: 1–11.

[17] X. Zhu, Z. Ghahramani, and J. Lafferty, “Semi-supervised learning using gaussian fields and harmonic functions,” in Proceedings of the 20th International Conference on Machine Learning, vol. 3, 2003, pp. 912–919.

[18] X. Zhou and M. Belkin, “Semi-supervised learning by higher order regularization,” in Proceedings of the 14th International Cofference on Artificial Intelligence and Statistics, 2011, pp. 892–900.

[19] H. Luo et al., “Hyperspectral image classification based on spectral-spatial 1-dimensional manifold,” IEEE Trans. Geosci. d Remote Sens., in press.

[20] Y. Wang, Y. Y. Tang, L. Li, and J. Z. Wang, “Face recognition via collaborative representation based multiple one-dimensional embedding,” International Journal of Wavelets, Multiresolution and Information Processing, vol. 14, no. 2, 2016, pp. 1640003:1–15.

[21] Y. LeCun, C. Cortes, and C. J. C. Burges, “THE MNIST DATABASE of handwritten digits,” http://yann.lecun.com/exdb/mnist/, accepted March 17, 2016.

[22] “USPS handwritten digit data,” http://www.gaussianprocess.org/gpml/data/, accepted March 17, 2016.

 ISBN: 978-1-61208-477-0

Onto-Traffic: A Semantic Traffic Analysis Tool based on GPS Data for Smart-Cities

Juan J. Sánchez-Escobar[‡], Liliana I. Barbosa-Santillán[*],
Luis F. Barbosa-Santillán[*], Fatemeh G. Nasri-Shandiz[*] and Gabriel A. León-Paredes[* †]
[*]University of Guadalajara
[†]Politécnica Salesiana University
[‡]Technical and Industrial Teaching Center
email: jjsanchez@ceti.mx, ibarbosa@cucea.udg.mx,
francisco.barbosa@alumno.udg.mx, ghazalnasri@hotmail.com and gleon@ups.edu.ec

***Abstract*—Nowadays by growing usage of vehicles, control and monitor the traffic is a promise for smart cities. Recommendation Systems are one of the solutions provided by smart cities for such issues. These kind of systems are trend in Information Technologies and they are acquiring acceptance among scientific communities and enterprises as well. Such systems feed users with valuable information about their environment and help them to take better decisions related to traffic jam. This paper presents OntoTraffic, a recommendation system based on ontologies, which supports the user to obtain information related to urban traffic in the city of Beijing. The information is provided according to the user's requests. This recommendation system works with Global Positioning System (GPS) traces taken from the Beijing data set, which has twenty one millions traces. In addition, 5888 traces were gathered through smartphones of Guadalajara citizens. In OntoTraffic system, data are collected by certain algorithms that extract the data from a reliable databases. Produced information from such system, which depends on the other recommendation, allows users being more productive and efficient in urban issues. This information also improves user's life quality by spending less time and money on their daily activities. OntoTraffic also provides recommendations to obtain: (1) average traffic on a specific street, (2) find rush hours for each street, (3) discover the most crowded streets in a certain time, (4) demonstrate the distance between two points on the map, and (5) show the bustle streets between 7:00 pm to 3:00 am. Finally this paper suggests some activities for future work such as: (a) build new models to support smart cities, (b) supporting recommended system developments for decision makers, (c) valuing data analysis and ontologies for future High Performance Computing (HPC) center.**

***Keywords*—Recommendation systems; urban traffic; smart cites; OntoTraffic system; smart applications; GPS data analysis.**

I. INTRODUCTION

Over the last five years, urban density in big cities has been increased constantly and in some cases, such as less-developed regions it has expanded dramatically. As a result, the increase of urban density generated many issues like healthcare, public security, pollution, etc. Moreover, urban mobility has been affected drastically by the increase of population in major cities. This situation has a tendency to get worse due to increase of migration from rural areas to urban areas in the coming years [1]. These issues can be more remarkable on travel seasons, because there are more people and more vehicles traveling at the same time to the same place. In addition, road infrastructures cannot expand at the same pace to solve the problem. As a consequence, the inhabitants of big cities are the ones who suffer from daily troubles such as traffic jam, delays and high fuel consumption [2]; in addition, there are some other intangible problems such as fatigue, stress, air pollution, higher rate of accidents and diseases, that are caused by urban growth and traffic. The aforementioned issues are very complex and involve multiple factors. Hence, these significant problems have been analysed and treated from different perspectives and via different entities. Finding a solution for cited problems is not easy because there are also some governmental issues involved, such as politics, guidelines, infrastructure investments and lack of using intelligent applications. However, systems like mobile sensors that employ Global Positioning System (GPS) are growing so fast and almost all of the smartphones are using GPS benefits, but still there are issues to provide the accurate users location. Calabrese et al., [3] mention in their work that use of smartphones for geo-location is only useful when traffic routes have been defined and identified. Herring et al., [4] have utilized gadgets with GPS to improve the accuracy of user's location. Disadvantage of this phenomena is that data acquisition depends on people participation by downloading and executing the application on their smartphones. Incursion of social networks and massive spread of smartphones have contribute to the develop of various projects where GPS are utilized as data collectors systems, like tracking GPS trajectories projects. For example, the GeoLife project [5] identifies the relationship between people and places, enabling people to share life experiences and build connections among each other using location history. Users share travel experience using GPS trajectories. Also, they obtain information from the GPS traces with the objective to recommend sites where other users and travellers visited. In other project presented by Microsoft Research Asia [5], they have focused on characterizing and comprehending people behavior, based on supervised learning through GPS traces to infer and predict their mobility behavior. Another ambitious project called T-Drive [6] Yuan et al., have utilized taxis GPS as mobile sensor to register the trajectory. Hence, they use the experience and ability of taxi drivers to locate the best route for the user's destiny. In this project, a database has been created with most used routes trajectories of Beijing city using GPS traces of 33,000 taxis during 3 months. The authors could identify each segment (every block in a trajectory) and time estimation between each node (an intersection between streets). Thus, the fastest route from one point to another could be calculate and demonstrate. Subsequently, users can choose the best route to their selected destination [7]. Many developing countries have witnessed an explosive vehicular growth. Our motivation is to improve people's lives through the knowledge of a recommender, by using an open source standard-based learning management system, that allows them to choose the better available option.

 ISBN: 978-1-61208-477-0

It is essential to have a recommendation system that could predict the behaviour of certain routes based on user's previous queries. Therefore, the use of recommendation systems could reduce the travel in time and costs, and by this way users can reach their destination with less stress.
One of the challenges is the architecture proposal, which transforms the input using a data structure (ontologies) with a huge datasets.

A. Contributions

Our contributions are as follows: 1) automatically identifying traffic interests from a user's history, 2) alert techniques that can be used to better determine whether a new recommendation is interesting, and 3) present a sample of specific SPARQL queries in order to evaluate system's answers.

B. Structure of the paper

This paper is organized as following: in Section 2, methodology is introduced. In Section 3, experiment results are presented. Finally, Section 4, presents conclusions.

II. METHODOLOGY

A description of our recommender system based on Ontologies is shown in Figure 1. In the first level, dataset are grouped into two sets (a) 21 millions of Beijing GPS records and (b) 5888 Guadalajara GPS records. On the second level, the GPS traces are transformed into Ontology Web Language (OWL) Lite. Then, OntoTraffic is populated and its consistency is verified with Protege. In the fourth level, valid questions for OntoTraffic were identified and agreed. Next, competence questions are translated to SPARQL language. We use ARQ and JENA engines in order to run the SPARQL expressions into OntoTraffic. Results are validated by a focus group, and translated to XML format to integrate with the project TUI-TRAFFIC. Finally, all files transformed to CAP-XML file format.

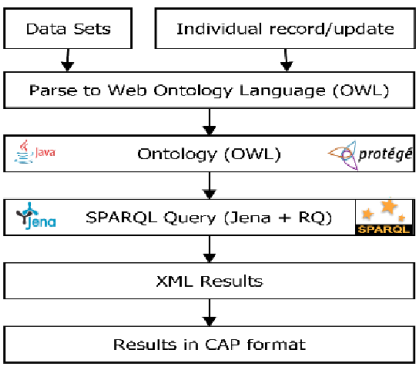

Figure 1. Recommender system architecture based on OntoTraffic

Some of the objectives that OntoTraffic covers are:

- Average traffic flow on a specific street.
- Obtain street rush hours.
- Identify the most crowded streets in a certain time.
- Measure distance between two particular points on the map.
- Recognize the most crowded streets between 7:00 pm to 3:00 A.M.

On the other hand, GPS Taxi Cabs files contain data collected by monitoring local taxicabs GPS in the city of Beijing, China. This dataset was obtained from [5] where 10383 files have been collected. Each file contains the GPS lecture of a single taxicab. This means that 10383 taxicabs were tracked. Figure 2 shows a fragment of a GPS trace file content. Each file has a ".txt" extension and the information stored in all of them is plain text.

```
1131,2008-02-02 13:30:59,116.45847,39.86964
1131,2008-02-02 13:31:04,116.45852,39.86954
1131,2008-02-02 13:31:09,116.45851,39.86955
1131,2008-02-02 13:31:14,116.45851,39.86956
1131,2008-02-02 13:31:24,116.45859,39.86954
1131,2008-02-02 13:31:29,116.45867,39.8695
1131,2008-02-02 13:31:34,116.45889,39.86946
1131,2008-02-02 13:31:39,116.45879,39.86946
1131,2008-02-02 13:31:44,116.4588,39.86943
1131,2008-02-02 13:31:49,116.45897,39.86941
1131,2008-02-02 13:31:54,116.45925,39.8694
1131,2008-02-02 13:31:59,116.45954,39.86939
1131,2008-02-02 13:32:04,116.46003,39.86935
1131,2008-02-02 13:32:09,116.46044,39.86934
1131,2008-02-02 13:32:14,116.46081,39.86941
1131,2008-02-02 13:32:19,116.46122,39.86951
1131,2008-02-02 13:32:24,116.46152,39.86951
1131,2008-02-02 13:32:29,116.46157,39.86959
1131,2008-02-02 13:32:34,116.46136,39.87011
1131,2008-02-02 13:32:39,116.46127,39.8703
1131,2008-02-02 13:32:44,116.46127,39.8703
1131,2008-02-02 13:32:49,116.46127,39.8703
1131,2008-02-02 13:32:54,116.46127,39.87029
1131,2008-02-02 13:32:59,116.4615,39.87025
```

Figure 2. GPS Traces located in the txt files.

It can be noticed that every taking reading is represented as a single line. The name of the fields is as follows: id_taxi, date, hour, latitude and longitude.
Additionally, the first column of Table 1 shows the name of the fields from the collected data of GPS lectures, while on the second column we present an example of values. This was done for clarification purposes and all parameters are self-explanatory.

TABLE I. SCHEMA AND EXAMPLE OF A GPS TRACE FILE

avgflow

Field	Example
id_taxi	9999
date	2008-02-02
hour	13:39:30
latitude	40.03238
longitude	116.28176

In order to process all the instances of OntoTraffic we convert the CSV files to turtle files as shown in Figure 3.

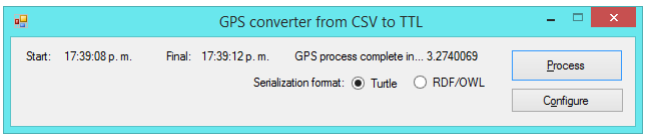

Figure 3. Convert the CSV files to turtle files.

III. RESULTS

This system can also give recommendations on (1) average traffic on a specific street, (2) rush hours of streets, (3) most crowded streets in a certain time, (4) distance between two points on the map, and (5) most crowded streets between 7:00 pm to 3:00 am, such as explained form A to E.

A. Average traffic flow on a specific street

The first test aim was to find the average traffic flow on a specific street. Our dataset has been used for this test. Figure 4 shows results of the SPARQL query:

```
SELECT (AVG(?Flujo) AS ?Avg_flow)
WHERE {
 ?IDStreet s:flujo?Flujo;
 s:name "DKMLTADKHFYLRKHQRMPO";
 s:col "SZMAXRLJMHFKURUJYYOG".
}
```

Avg_flow
98.190623122084040777777777

Figure 4. Result of average traffic flow on a specific street by using Beijing's dataset

B. Streets rush hours

Second test purpose was to find the busiest points in a certain time (rush hours). To fulfill this test aim, we used Beijing's dataset.
Figure 5 shows results in a statistical graph and Figure 6 shows the result of the SPARQL query:

```
SELECT ?latitude ?longitude ?Hour
       (COUNT(?gps) AS ?Amount)
WHERE
{
?gps gps:latitude ?latitude;
gps:longitude ?longitude;
gps:hour?Hour

FILTER(?latitude != 0.0).
FILTER(?longitude != 0.0).
}
GROUP BY ?latitude ?longitude ?Hour
HAVING(COUNT(?gps) > 1)
ORDER BY DESC (?Amount)
LIMIT 20
```

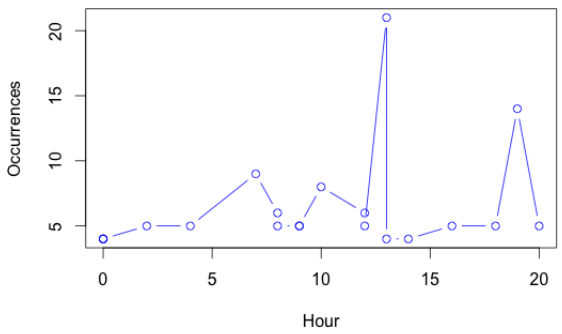

Figure 5. Results of busiest points in a certain time on Beijing

latitude	longitude	Hour	Amount
116.39696	39.81439	13	21
116.49653	39.9659	19	14
116.49711	39.95284	07	9
116.4436	39.82555	10	8
116.01101	40.35158	08	6
119.11988	40.06314	12	6
116.50432	40.00928	02	5
116.3726	39.96738	04	5
116.2737	39.9397	08	5
115.46215	41.23102	09	5
117.0063	40.14755	09	5
117.18312	40.18602	12	5
116.16002	39.67661	16	5
116.48792	40.01089	18	5
116.31802	39.93755	20	5
116.39075	39.84218	00	4
116.41175	39.93947	00	4
116.51686	40.23957	00	4
116.49892	39.723	13	4
116.46863	39.97601	14	4

Figure 6. Results of busiest points in a certain time on Beijing

C. The most crowded streets in a certain time

The third test aim was to find certain points with more occurrences regardless of day and time. To achieve results for this test, our dataset has been used. Figure 7 shows the results of SPARQL query:

```
SELECT ?latitude ?longitude
       (COUNT(?gps) AS ?Amount)
WHERE
{
  ?gps m:Latitude_from ?latitude;
m:Longitude_from ?longitude.

FILTER(?latitude != 0.0).
FILTER(?longitude != 0.0).
}
GROUP BY ?latitude ?longitude
HAVING(COUNT(?gps) > 1)
ORDER BY DESC(?Amount)
```

latitude	longitude	Amount
20.6724837	-103.3305215	3
20.671611	-103.3463266	2
20.6748546	-103.3568629	2
20.6750575	-103.3401175	2
20.6973759	-103.3100631	2
20.6980102	-103.3116547	2
20.6982709	-103.3121455	2
20.6992375	-103.3139986	2
20.6992935	-103.3115487	2
20.6997488	-103.3136851	2
20.6999986	-103.3154296	2
20.7001552	-103.3110331	2
20.7002576	-103.3133749	2
20.7005223	-103.3150944	2
20.7015567	-103.3181134	2
20.7025497	-103.3118661	2
20.7031258	-103.3114649	2
20.7056635	-103.3100847	2
20.7061499	-103.3116284	2
20.706695	-103.3133045	2
20.7067842	-103.3096909	2
20.7070555	-103.3144593	2
20.7072601	-103.3112226	2
20.7097507	-103.3230464	2
20.7108174	-103.3224945	2
20.7137486	-103.30201	2
20.7175645	-103.2967534	2
20.7203743	-103.2957234	2

Figure 7. Results of the most crowded streets in Beijing

D. Distance between two particular points on the map

Fourth test target was to measure and demonstrate the distance between two particular points on the map. For this result, Beijing dataset has been used. Figure 8 shows the result to SPARQL query:

 ISBN: 978-1-61208-477-0

```
SELECT (SUM(?Distancia_Seg) AS
       ?Distance_mts)
WHERE
{
  ?gps1 m:IDSegmento  ?Segmento;
        m:Distancia_Seg ?Distancia_Seg .

   FILTER(?Segmento > A ).
   FILTER(?Segmento < B ).
}
```

Distance_mts
9822.914631589

Figure 8. Results of distance between point A to B on Beijing dataset

E. The most crowded streets between 7:00 PM to 3:00 AM

The aim of the fifth test was to find the crowded points between 7:00 PM to 3:00 AM. For obtain results of this test, Beijin dataset has been used. The following query was used in this aim.

```
SELECT ?longitude  ?latitude ?t_hour
       (COUNT(?gps) as ?Amount )
WHERE
{  {?gps gps:hour ?hour;
          gps:longitude ?longitude;
          gps:latitude ?latitude .
    BIND (SUBSTR(?hour,1,2) as ?t_hour)
    FILTER(?latitude != 0.0).
    FILTER(?longitude != 0.0).
    FILTER(?hour>="19:00:01").
    FILTER(?hour<="23:59:59").
    FILTER(?latitude=?latitude).
    FILTER(?longitude=?longitude)
    }
UNION
    {?gps gps:hour ?hour;
          gps:longitude ?longitude;
          gps:latitude ?latitude .
    BIND (SUBSTR(?hour,1,2) as ?t_hour)
    FILTER(?latitude != 0.0).
    FILTER(?longitude != 0.0).
    FILTER(?hour>="00:00:01").
    FILTER(?hour<="03:00:00") .
    FILTER(?latitude=?latitude).
    FILTER(?longitude=?longitude)
    }
}
GROUP BY ?longitude ?latitude ?t_hour
HAVING (COUNT(?gps)>2)
ORDER BY DESC (?Amount)
```

Figure 9 shows result to the SPARQL query:

longitude	latitude	t_hour	Amount
39.9659	116.49653	"19"	14
39.93755	116.31802	"20"	5
40.00928	116.50432	"02"	5
39.84218	116.39075	"00"	4
39.93947	116.41175	"00"	4
39.90652	116.41823	"20"	4
40.23957	116.51686	"00"	4
39.76609	116.2228	"21"	3
39.94916	116.37407	"21"	3
39.96749	116.40391	"21"	3
39.69164	116.42191	"20"	3
39.93284	116.45692	"23"	3
39.98753	116.48178	"02"	3
40.0798	116.58538	"22"	3
40.13195	116.66135	"02"	3
40.11647	116.89351	"20"	3
39.78471	117.0719	"21"	3

Figure 9. Results of the most crowded streets between 7:00PM and 3:00AM of Beijing dataset

IV. CONCLUSION

OntoTraffic was developed as a basis for creating a recommendation system. Our system will recommend the best traffic route based on five solutions. The present paper shows that, by using GPS traces from citizens mobile devices, can recommend the best route in order to travel from point A to point B in a specific time of day or under certain conditions in a moment. It also helps to know the most frequent route, the most utilized roads, time with the most heavy traffic load, city areas with the biggest traffic jam, the best places to take a Taxi, the preferred route for drivers, the different routes available to reach a place or the alternative routes according to a specific daytime, in a reasonable execution time. It was observed that the processing time increases by the quantity of GPS traces to be analysed. Even in the first experiment, the quantity of traces was relevant (more than 1,000,000). Our main contribution is to have a recommendation system that could predict the behaviour of the traffic based on five questions and answers of certain routes depending on different user's search criteria. Besides, our methodology is simple because based on ontologies and SPARQL language is possible to discover data in a huge dataset as Beijing. Above all, it is important to highlight that this work is in development process. For future work, we propose: (a) Build new models to support smart cities. (b) Support recommender systems development for decision makers. (c) Use data analytic and ontologies for a future HPC center.

ACKNOWLEDGMENT

We thank Sciences Research Council (CONACyT) for funding this research project.

REFERENCES

[1] UNPD, "Unpd. 2012. world urbanization prospects: The 2011 revision," tech. rep., 2011.

[2] D. Schrank, T. Lomax, and B. Eisele, "Urban mobility report 2011. report for the texas transportation institute," tech. rep., 2011.

[3] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in rome," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, pp. 141–151, March 2011.

[4] R. Herring *et al.*, "Using mobile phones to forecast arterial traffic through statistical learning," *89th Transportation Research Board Annual Meeting, Washington DC*, 2010.

[5] Y. Zheng, X. Xie, and W.-Y. Ma, "Geolife: A collaborative social networking service among user, location and trajectory.," *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, 2010.

[6] J. Yuan *et al.*, "T-drive: driving directions based on taxi trajectories.," in *GIS* (D. Agrawal, P. Zhang, A. E. Abbadi, and M. F. Mokbel, eds.), pp. 99–108, ACM, 2010.

[7] K. Sia-Nowicka, J. Vandrol, T. Oshan, J. A. Long, U. Demar, and A. S. Fotheringham, "Analysis of human mobility patterns from gps trajectories and contextual information," *International Journal of Geographical Information Science*, vol. 30, no. 5, pp. 881–906, 2016.

 ISBN: 978-1-61208-477-0

Using Recurrent Boltzmann Machines in the Detection of Process Drifts Using Several Process Mining Perspectives

Alessandro Berti

SIAV
35030 Rubano PD
Email: alessandro.berti89@gmail.com

***Abstract*—In this paper is proposed an algorithm that uses Recurrent Boltzmann Machines to detect concept drifts in Process Mining event logs. The method is able to use several perspectives (Control Flow, Resources, Data) in comparison to existing methods that are conceived to use a single perspective (Control Flow). The approach has been tested on some artificial event logs and on a real-life log.**

Keywords–Concept Drift; Process Mining; Boltzmann Machines.

I. INTRODUCTION AND BACKGROUND

Process Mining [1] is a relatively new discipline related to the discovery and the analysis of business processes starting from event logs. These logs are organised in traces, that correspond to single executions of the process (a process instance). Each trace can contain several events, that can be described by some attributes, like the organizational resource that has performed the event, the activity that has been performed and the timestamp. In most cases, events are instantaneous (so, only the completion time of an activity is recorded), and a trace could be briefly described (the Control-Flow perspective) by a list of activities. A path is a succession of activities that can be observed in traces. For example, if a trace can be described by this list of activities: A,B,C,D,E; then the paths are AB, BC, CD, DE. The start timestamp of a trace is the minimum of the timestamps of its events; the end timestamp is the maximum of the timestamps.

Process Mining algorithms are often hampered by concept drifts in the underlying process, that are changes in the process during the analyzed time interval. Some papers (like [2], [3], [4]) have analyzed a way to cope with the drift in processes, but the analysis is restricted to the Control-Flow perspective, and ignore other information related to the process instance. The Control Flow perspective is the list of activities performed in order to complete a single business process instance. Other perspectives are the data perspective and the resource perspective (people that are involved in the completion of the instance). In this paper is described a way to detect the drift in the underlying process that takes into account also the other information related to business instances. The method is based on the estimation, for each of the business instances, of the probability that a concept drift has actually happened. The algorithm is based on Recurrent Boltzmann Machines [5] that are useful to model high dimensional sequences. Other methods (like NADE [6]) are known to be able to detect the probability to observe a given point; however, their application is usually done point-by-point and ignores the the history of the observed points (that are business instances). The proposed approach is a necessary extension of the methods that take in account only the Control Flow because they cannot detect changes happening in other perspectives. In the Background Section of this paper, Boltzmann Machines are presented. In the Method Section, the proposed approach to detect the concept drift is analysed. In the Results Section, some results related to artificial and real event logs are presented.

II. BACKGROUND

Restricted Boltzmann Machines (RBM) are useful to learn a probability distribution over a set of inputs and are based on the concepts of visible and hidden binary units. Hidden units are activated by the RBM, that works taking in input a weighted sum of the visible units, applying a $[0, 1]$-valued function and activating the hidden unit with a probability equal to the function value. An energy function can be associated with RBM, that is based on the visible and hidden units value; a low energy configuration is preferred as the definition of the energy function makes RBM useful for classification purposes [7]. A well-known method for RBM training is Contrastive Divergence [8]; basically, it is an iterative method for RBM weights discovery that tries to minimise the difference between the visible units and some temporary (binary) units whose value is found by the inverse application of the RBM (in this step, the hidden units become the visible units). RBM, however, do not handle sequence of points as the hidden units' activation depend only on the current iteration of visible units (that are single points), and do not handle history. Recurrent Boltzmann Machines are conceived to use the history of the sequence, as the hidden units activation do not depend only on the visible units but also on the previous states of the hidden units. However, other papers (like [5]) can be relevant for further information.

III. METHOD

The method is based on the construction of a sequence of binary points (each one corresponding to an event log trace) that is provided to a Recurrent Boltzmann Machine in order to learn a meaningful representation of the sequence. The number of hidden units has been set to be equal to the number of visible units. The trace is being described in both the Control-Flow perspective and the other perspectives:

1) The Control-Flow is described by the paths followed in the trace.
2) Other perspectives are described by recording all the different values for an attribute that can be seen in the various events of the trace.

A binary representation, whose length is equal to the sum of the number of different paths in all the traces of the

 ISBN: 978-1-61208-477-0

log and the number of different values for the considered attributes in all the events of the log, can be obtained by giving value 1 in a position that describes a path / attribute value that is contained in the trace, and giving 0 otherwise. For example, if there are the following two traces: Trace 1 (events: A(Mike),B(Tom),C(Mike); paths: AB, BC; resources: Mike, Tom), Trace 2 (events: A(Alex),D(Maria),E(Maria); paths: AD, DE; resources: Alex, Maria); the binary representation could be as follow: position 1 is relative to the presence of the path AB, pos. 2 is relative to BC, pos. 3 is relative to AD, pos. 4 is relative to DE, pos. 5 is relative to the resource Mike, pos. 6 is relative to Tom, pos. 7 is relative to Alex, pos. 8 is relative to Maria; the eventual representation is (1,1,0,0,1,1,0,0) for Trace 1 and (0,0,1,1,0,0,1,1) for Trace 2. The traces are considered to be ordered by their start timestamp.

After building the sequence of binary points, the RBM could be trained. The results of the training are then used by "stopping before" the activation of the hidden units and recording the activation function (probability) values. So, a $[0, 1]$-valued vector can be obtained for each trace, that refers to the probability of activation of the hidden units. From these vectors, the maximum value is taken; in doing so, each trace is described by a single probability value. Traces with a lower value of probability are likely to show a concept drift in the underlying process. This is because the RBM tries to learn a representation that maximises the probability of a given sequence, and at least one hidden unit for trace should be activated with high probability. Considering also the previous states of the sequence, a trace that follows the previous schema shows usually a high value in the defined probability, while traces that show a different schema produce lower values of probability.

IV. Results

The following artificial logs have been used in order to evaluate the effectiveness of the method:

1) An event log that contains 1000 equal traces (events: A(Mike),B(Tom),C(Mike); paths: AB, BC; resources: Mike, Tom), and other 1000 equal traces (events: A(Alex),D(Maria),E(Maria); paths: AD, DE; resources: Alex, Maria).
2) An event log that contains 500 traces for each of the following schemas:
   - events: A(Mike),B(Tom),C(Mike); paths: AB, BC; resources: Mike, Tom
   - events: A(Alex),D(Maria),E(Maria); paths: AD, DE; resources: Alex, Maria
   - events: F(Billie),G(Louise),H(Billie); paths: FG, GH; resources: Billie, Louise
   - events: I(Barack),L(Francois),M(Barack); paths: IL, LM; resources: Barack, Francois
3) An event log that contains 500 traces for each of the following schemas:
   - events: A(Mike),B(Tom),C(Mike); paths: AB, BC; resources: Mike, Tom
   - events: A(Alex),B(Maria),C(Maria); paths: AB, BC; resources: Alex, Maria
   - events: F(Billie),G(Louise),H(Billie); paths: FG, GH; resources: Billie, Louise
   - events: F(Barack),G(Francois),H(Barack); paths: FG, GH; resources: Barack, Francois

The method is able to correctly identify concept drifts in all the cases. The first two logs are very simple, as there is a drift both in the Control Flow and the resources. The third log shows changes in the resource set; existing methods for concept drift in Process Mining (as [2], [3], [4]) would not have been able to identify changes in this log.

The method has been tested also on a real-life event log, that is "Receipt phase of an environmental permit application process" [9] containing an interesting shift in the process. Using Dotted Chart feature in ProM framework [10] you can identify a change in the underlying process after the timestamp 31/03/2011, when some activities (T06 and T10) became slightly less frequent. This shift has not been identified by the method described in [2] as the change in the Control Flow is not so great, but is identified by the proposed method taking into account the other perspectives. Also the methods described in [3] based on SVM, and [4] which is an improvement of [2], fail to identify this change.

V. Conclusion and Future Work

The proposed approach for the detection of concept drifts takes into account several Process Mining perspectives and correctly identifies process changes in the examined logs. Further work is needed to classify the change points, as there can be several drifts (reported in [2]): sudden drifts, gradual drifts, seasonal drifts; also, other types of hidden units (as [11]) might produce better results.

References

[1] W. Van Der Aalst, Process mining: discovery, conformance and enhancement of business processes. Springer Science & Business Media, 2011.

[2] R. J. C. Bose, W. M. van der Aalst, I. Žliobaitė, and M. Pechenizkiy, "Handling concept drift in process mining," in Advanced Information Systems Engineering. Springer, 2011, pp. 391–405.

[3] R. Klinkenberg and T. Joachims, "Detecting concept drift with support vector machines." in ICML, 2000, pp. 487–494.

[4] D. Luengo and M. Sepúlveda, "Applying clustering in process mining to find different versions of a business process that changes over time," in Business Process Management Workshops. Springer, 2011, pp. 153–158.

[5] I. Sutskever, G. E. Hinton, and G. W. Taylor, "The recurrent temporal restricted boltzmann machine," in Advances in Neural Information Processing Systems, 2009, pp. 1601–1608.

[6] H. Larochelle and I. Murray, "The neural autoregressive distribution estimator," in International Conference on Artificial Intelligence and Statistics, 2011, pp. 29–37.

[7] H. Larochelle and Y. Bengio, "Classification using discriminative restricted boltzmann machines," in Proceedings of the 25th international conference on Machine learning. ACM, 2008, pp. 536–543.

[8] M. A. Carreira-Perpinan and G. E. Hinton, "On contrastive divergence learning," in Proceedings of the tenth international workshop on artificial intelligence and statistics. Citeseer, 2005, pp. 33–40.

[9] J. Buijs, "Receipt phase of an environmental permit application process (wabo), coselog project," 2014. [Online]. Available: http://dx.doi.org/10.4121/uuid:a07386a5-7be3-4367-9535-70bc9e77dbe6

[10] B. F. van Dongen, A. K. A. de Medeiros, H. Verbeek, A. Weijters, and W. M. Van Der Aalst, "The prom framework: A new era in process mining tool support," in Applications and Theory of Petri Nets 2005. Springer, 2005, pp. 444–454.

[11] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814.

 ISBN: 978-1-61208-477-0

Machine Learning in Cloud Environments Considering External Information

Matthias Lermer, Stefan Frey, Christoph Reich
Cloud Research Lab
Furtwangen University
Furtwangen, Germany
Email: {matthias.lermer, stefan.frey, christoph.reich}@hs-furtwangen.de

***Abstract*—Machine learning applied to cloud environments can lead to many advantages. One example is the possibility of improved Quality of Service (QoS) by predicting future workloads and reacting dynamically with automated scaling. In reality however, there are cases where the use of machine learning algorithms is not as efficient as imagined. One current problem is the disregard of external information, whose inclusion could help to create better models of the reality. The approach of this paper shows that different machine learning algorithms like Neural Networks (NN), Support Vector Machines (SVM) and Linear Regression can be successfully used to predict the response time of Virtual Machines (VM) within cloud environments. The performed application of those algorithms to different cloud usage scenarios and following evaluation enables to gain insight into the strengths and weaknesses of each algorithm. Furthermore, a work in progress architecture is proposed to deal with the two big challenges, inclusion of external information and handling live data streams.**

Keywords–Machine Learning; Support Vector Machines; Neural Network; Linear Regression; Cloud Computing; SLA.

I. INTRODUCTION

The use of machine learning, a relatively mature and established discipline of computer science, has become more important than ever. Various challenges in the area of Cloud Computing, like efficiently handling big data or the realization of green IT, can be tackled by applying machine learning algorithms. This paper looks at the specific application of response time prediction of Virtual Machines (VM), in order to improve scaling functionality and prevent Service Level Agreement (SLA) violations. The associated implication on the utilization of resources when using different machine learning algorithms is not covered in this evaluation. The remainder of the paper is organized as follows: In Section II a short explanation about the used CloudSim framework and created scenarios is given. Section III describes the application and evaluation of various machine learning algorithms. In Section IV an architectural approach to include external informations during the learning process is presented. Section V completes the paper by drawing a conclusion and suggesting future work.

II. RELATED WORK

Similar research with different focus has been conducted in the past for the use of machine learning in cloud environments. Prevost et al. used a Neural Network (NN), as well as a Linear Predictor [1] to anticipate future workloads by learning from historical URL requests. Although both models were able to give efficient predictions, the Linear Predictor was able to predict more accurately. Li and Wang proposed their modified Neural Network algorithm nn-dwrr in [2]. The application of this algorithm led to a lowered average response time compared to application of traditional capacity based algorithms for scheduling incoming requests to VMs. In similar research Hu et al. [3] have shown that their modification of a standard Support Vector Regression (SVR) algorithm can lead to an accurate forecasting of CPU Load what can be used to achieve a better resources utilization. Another algorithm, which is renowned for providing good results in similar scenarios, is Linear Regression. Although the results are often weaker compared to Neural Networks or Support Vector Machines (SVM) in cases of workload prediction [4] [5], the fast training and deployment time of models built with Linear Regression should not be underestimated.
Those examples show that there are a variety of optimization challenges in cloud environments which can be tackled by applying machine learning algorithms. What separates the current work from previous research is a detailed examination of specific characteristics of three different machine learning algorithms and presenting the results in a visual way. The choice to evaluate Neural Networks, Support Vector Machines and Linear Regression was made because those algorithms earned promising results in previously conducted research.

III. CLOUDSIM

The CloudSim framework [6] [7], developed by the University of Melbourne provides the means to realistically simulate Cloud Computing environments. An extended implementation of this framework was used to simulate specific scenarios in order to obtain relevant log data. This data is used to train and test models with different machine learning algorithms. Furthermore the additions made by the CloudSimEx extension [8] were used.

With the help of additional modifications of the CloudSim source code it was possible to simulate and log the 3 following Cloud usage scenarios:

1) Short bursts of peak requests in the average usage area
2) Slow ascending and descending requests with one larger peak
3) Quick changes in requests with small peaks, followed by medium and large sized peaks

IV. APPLICATION AND EVALUATION OF MACHINE LEARNING ALGORITHMS

In order to apply and evaluate different Machine Learning algorithms, the open source software RapidMiner [9] was used. The log files created by the CloudSim application were utilized as training and test sets. Furthermore, the available Series extension provided by RapidMiner was used. This extension enables an efficient way to quickly replace different

 ISBN: 978-1-61208-477-0

Machine Learning algorithms during the process of creating and evaluating a model in regard to ordered time series. With the help of a horizon of h=20 (2 seconds), it was defined that the learning algorithms gets to learn the next h time steps in order to be able to predict the value of the average response time of h+1. After the prediction, the time window is incremented by 1 and the next value gets predicted. A further advantage is that the data is transformed by the implemented Series operator in a way that enables the use of classification algorithms like Neural Networks and Support Vector Machines in the case of a numerical regression problem.

A. Configuration of the algorithms

RapidMiner provides a large number of configuration parameters which can be tuned. After the execution of test runs with different parameters the following configuration provided the best results:

1) Neural Network: |Feed forward NN, training back propagation |Hidden layers: 8 |Training cycles: 1000 |Learning rate: 0.3 |Momentum: 0.2 |Decay: true |Normalize: true |Error epsilon: 0.00001

2) Support Vector Machines: |Kernel Type: radial |Kernel Gamma: 1.0 |Kernel cache: 200 |C: 0 |Convergence epsilon: 001 |Max iterations: 10000 |Scale: true |L pos: 1.0 |L neg: 1.0

3) Linear Regression: |Feature selection: Iterative T-Test |Max iterations: 1.0 |Forward alpha: 0.05 |backward alpha: 0.05 |eliminate colinear features: true |min tolerance: 0.05 |use bias: true

B. Graphs

The following graphs show the aforementioned scenarios (see Section III). The x-axis represents the time in seconds. The y-axis to the left indicates the response time of the cloud environment with the differentiation in predicted values (red line) and actual values (black line). Whereas the y-axis to the right indicates the current number of active VMs. Furthermore, the active VMs are highlighted in a light blue in the background.

C. Scenario 1

1) Neural Network: Figure 1 shows that the NN overestimates the peak of the burst load in every case. It can be seen that the difference between predicted peak and real peak is the biggest during the first burst and that there is an improvement when predicting the later peaks of the bursts. The briefly following decline and rise after each peak, e.g., during sec 8-15 is respectively underestimated and overestimated but it can clearly be seen that there is an improvement in the last iteration. Figure 2 shows the delay characteristics and that the algorithm in general can adapt well to the problem.

2) Support Vector Machines: Figure 3 shows a contrast to the NN algorithm. In the case of SVMs, the first peak of each burst is underestimated. The following cooldown phase before the second peak of each burst is overestimated but an improvement over time can be seen, especially on the last burst. Figure 4 looks specifically at the first burst and a comparison to the NN 2 makes it clear that SVMs predict a more smooth curve. It should be kept in mind that it is realistic to assume that in real life there are scenarios with different requirements regarding the reaction to those predictions where this specific differences, smooth or rough, could be seen as either an advantage or disadvantage.

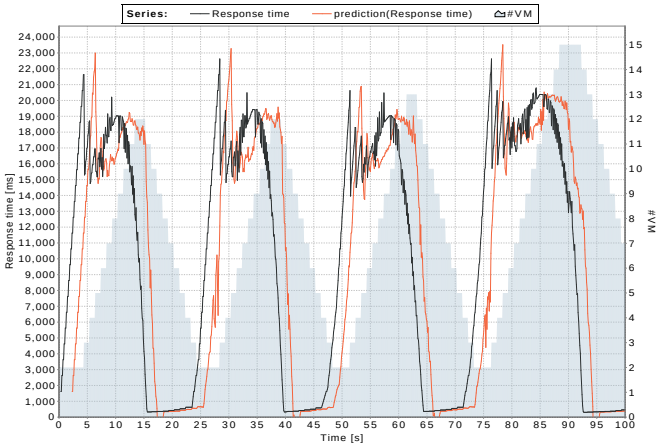

Figure 1. NN Scenario 1: 0s-100s

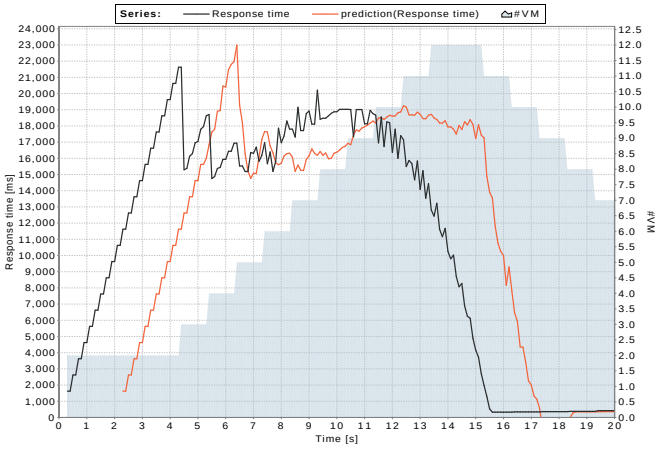

Figure 2. NN Scenario 1: 0s-20s

3) Linear Regression: The predictions made with the help of a Linear Regression model, shown in Figure 5, seem to be very similar to those predictions made by the SVM in Figure 3. This insight is further substantiated by taking a closer look at the bursts in Figure 6 and Figure 4 where a similar prediction pattern can be seen. Worth mentioning is that the predictions made by Linear Regression lead to an even smoother curve compared to the curve predicted by the other 2 algorithms.

D. Scenario 2

1) Neural Network: When looking at the overview in Figure 7 it is demonstrated that moderate changes in response times are learned rather well. The interesting part, shown in more detail in Figure 8, showcases the nature of overestimation. Again, the peak is overestimated, but the predicted curve recovers very fast and yet this issue occurs again after the second plateau. It should be noted that with a different configuration of the NN algorithm a very different curve can be predicted. For this paper we looked at a specific configuration of the algorithm because this characteristic can be utilized and will be explained during the comparison of the algorithms.

2) Support Vector Machines: The overview shown in Figure 9 displays the capability of the algorithm to be able to adapt to a singular, steadily climbing response time. The prediction

 ISBN: 978-1-61208-477-0

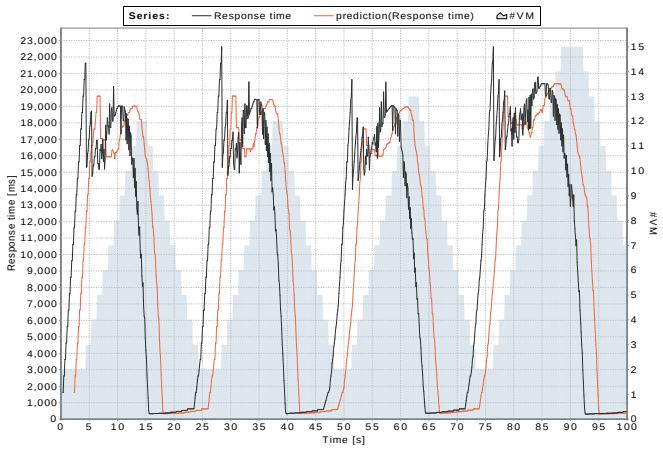

Figure 3. SVM Scenario 1: 0s-100s

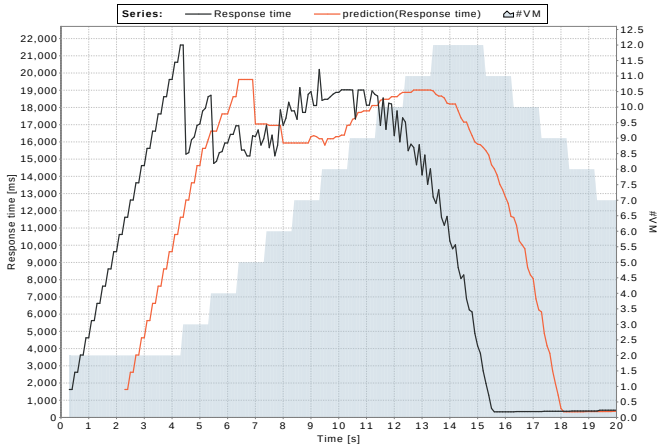

Figure 4. SVM Scenario 1: 0s-20s

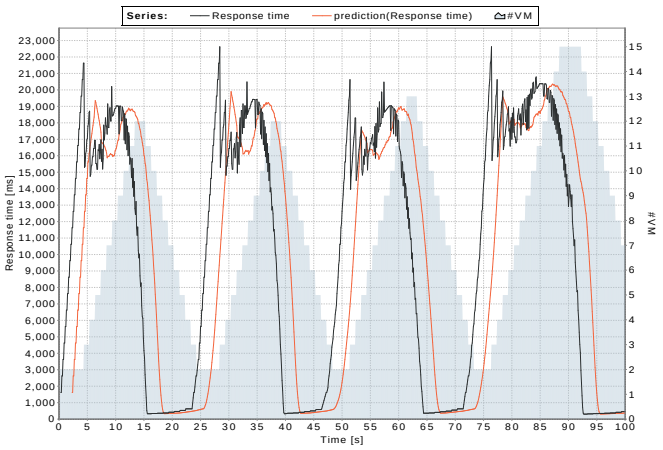

Figure 5. Linear Regression Scenario 1: 0s-100s

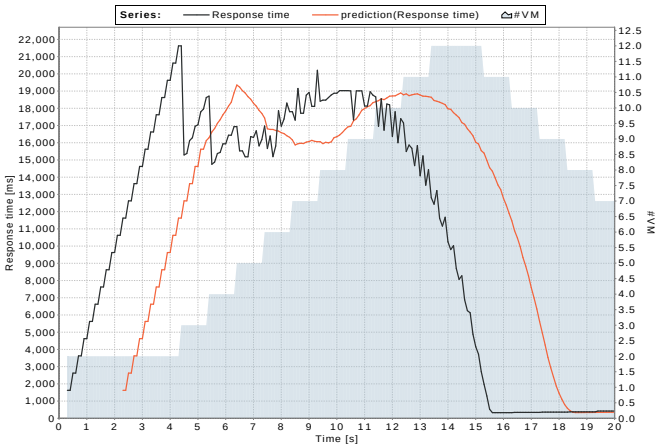

Figure 6. Linear Regression Scenario 1: 0s-20s

during the first phase (0-60s) is handled well by the algorithm. Figure 10 illustrates that the spontaneous and large decline in response time is learned very well. This is an important characteristic as predictions based on those quick changes could be the focus during the application in real-time scenarios. None of the other algorithms is able to predict scenario 2 this precisely.

3) Linear Regression: Figure 11 shows again great similarity between predictions made with the help of Linear Regression and SVMs. Again the difference is that the ascent of the curve is predicted in a smoother way. Additionally, the spontaneous and large decline seen in Figure 12 is not predicted very well. The same characteristic applies on the following smaller decline. As a conclusion it can be said that in this specific scenario the model trained by Linear Regression is the weakest.

E. Scenario 3

1) Neural Network: Figure 13 shows that during the first phase (0-100s) the model trained by a NN has minor problems in predicting the response time. Although the peaks are generally predicted well, sometimes they are underestimated and sometimes overestimated. But in contrast to the other algorithms the difference in error margin is very small in most cases. During the recovery times after each slope, the local minima are overestimated almost in every case. While the first big peak of a response time over 42000ms is overestimated as well, the second one is predicted almost perfectly. The relative smooth slopes before, during and after the larger peaks are predicted very well with no prominent deficit.

2) Support Vector Machines: Figure 14 shows that a model trained by SVMs can predict the response time for a varying scenario rather well. The occurring peaks during the first phase (0-100s) are underestimated in every case, but not to a large degree. This leads likewise to the underestimation of the recovery times after each peak, which are the consequence of adding and deleting Virtual Machines. The two larger peaks with a response time of over 42000ms are underestimated again by a small margin while the relative smooth slopes before, in between and after are learned well with no prominent deficit in their prediction.

3) Linear Regression: Figure 15 shows that a model trained by Linear Regression can cope well in a varying scenario. Similar to the SVM model it slightly underestimates the response time in the first phase (0-100s). In general, it can be said that those models are very similar and have only minor, negligible differences. The main difference is that the use of

 ISBN: 978-1-61208-477-0

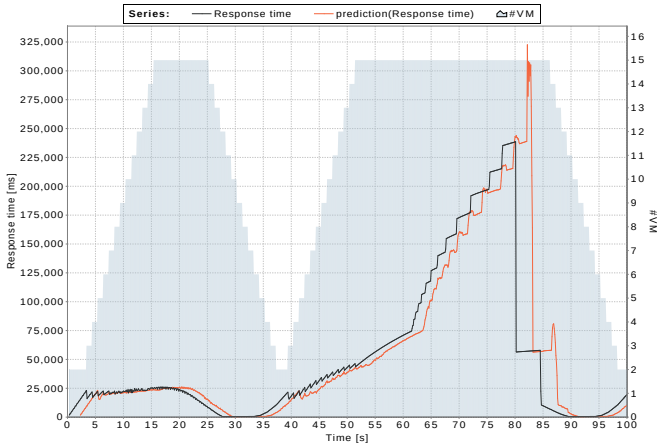

Figure 7. NN Scenario 2: 0s-100s

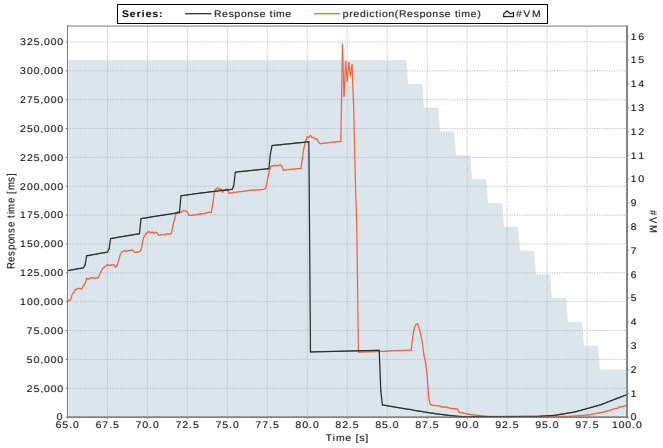

Figure 8. NN Scenario 2: 65s-100s

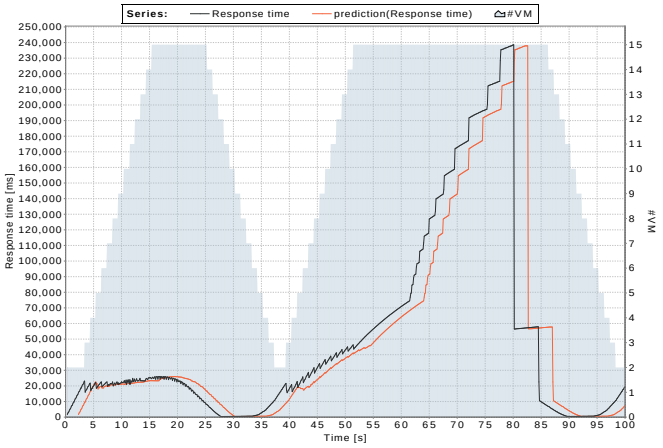

Figure 9. SVM Scenario 2: 0s-100s

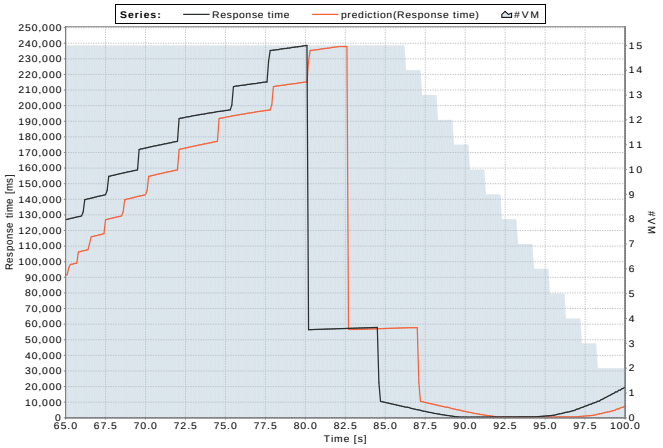

Figure 10. SVM Scenario 2: 65s-100s

Linear Regression leads to smoother slopes.

F. Comparison

While it was shown that all 3 algorithms can be effectively used for predicting the response time in different scenarios it can be said that the NN has a minor advantage over the other algorithms. The main reason is that the NN, in general, slightly overestimates and almost never underestimates the response time. The practical application of this knowledge, e.g., using those predictions in combination with a scaler who manages the quantity of VMs leads to a more assuring state that requirements like defined SLAs can be covered more carefully than with other algorithms. In less critical business-cases, where the defined SLAs and response times are not that sensitive, the other 2 algorithms, SVMs and Linear Regression, can be used despite their tendency to slightly underestimate response times. Especially the Linear Regression with its fast training and deployment times could be considered in near real-time scenarios.

G. Related Work: Fuzzy

A similar research has been conducted by Frey et al. in [10]. In their scenario the driving factor was to use predictions based on fuzzy logic to automatically scale the quantity of Virtual Machines in a Cloud Computing environment in order to be able to guarantee that a certain threshold regarding response times is not exceeded. While the paper presented here takes a more general approach, Frey et al. have successfully proven that a model trained by fuzzy logic can predict response times well and that the thereby gained knowledge can be successfully applied in a real-time scenario.

V. Addition of External Sources

In order to be able to enrich data by external information, the process and work flow has to be defined and created. The following Figure 16 provides an overview concerning that matter. In the following scenario it is presumed that all steps of the Cross Industry Standard Process for Data Mining (CRISP-DM) [11] Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment have been traversed at least once. As a result, a working system was established but after one or more evaluations it becomes clear that there are possibilities to create a better model by considering the use of appropriate external data sources. Those can enrich the existing historical training data and provide the ability to dynamically adapt the specific or general needs of a good model. This can be realized by the use of specific kinds of agents. Polling agents are responsible for the following tasks:

 ISBN: 978-1-61208-477-0

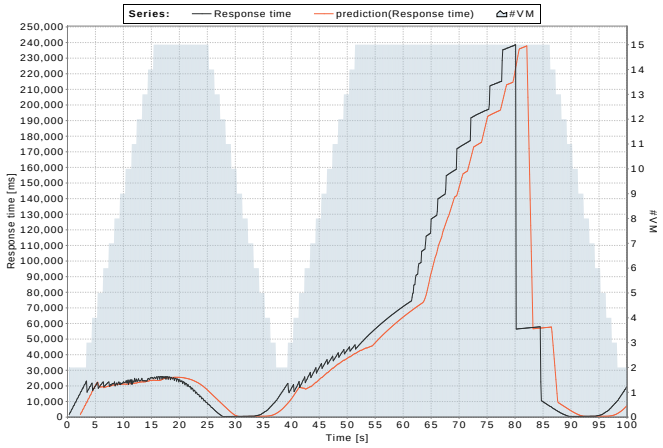

Figure 11. Linear Regression Scenario 2: 0s-100s

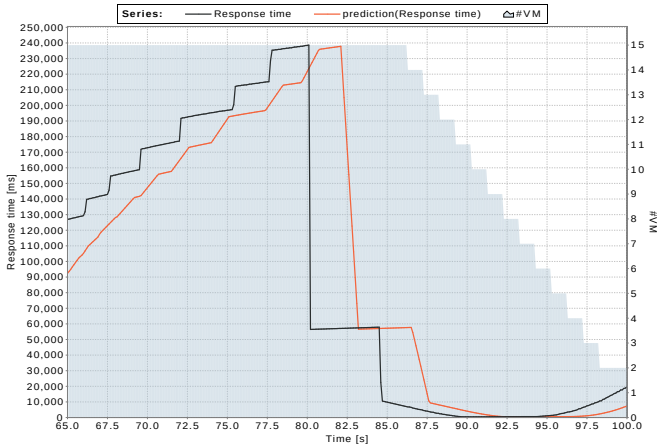

Figure 12. Linear Regression Scenario 2: 65s-100s

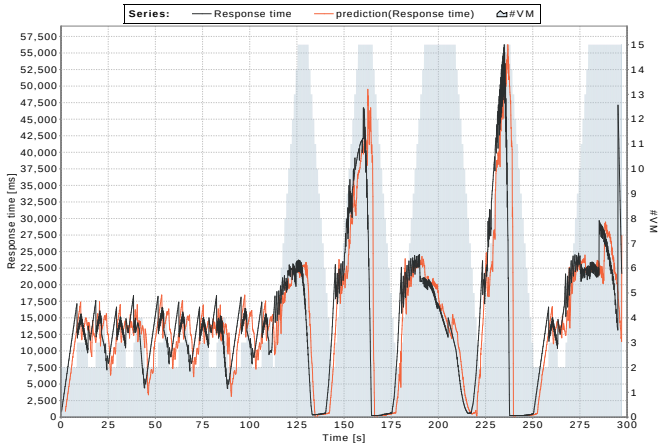

Figure 13. NN Scenario 3: 0s-300s

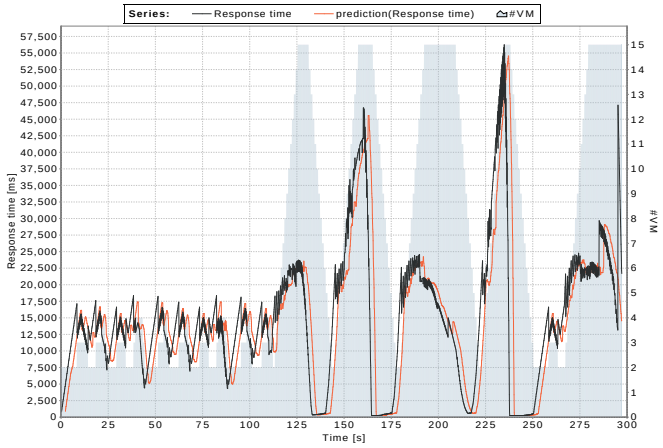

Figure 14. SVM Scenario 3: 0s-300s

- Retrieve specified information from described sources.
- Cope with missing values in a defined way, as it is not always the case that information can be gathered consistently.
- Recognize and filter wrong and erroneous data, which is especially important in cases where the external sources are filled with data collected by humans and not computers.
- Transform and correlate the gathered information and attributes to the training set, e.g., timestamp synchronization.
- Store the prepared information in the training database and thus enrich existing historical data.

Configuration agents must be able to realize the following:

- Query the polling agents about meta information
- Use this meta information to change the configuration of the training process.
- Initiate new training sessions after defined periods, as well as after enrichment of the training set.
- Initiate the application and validation of models while storing the results in a database.

If a scenario has the need for evaluating live data streams there must be a coordinating agent who has the task to react in a defined way. The application shown in Section III gives an example how a live data stream could be integrated. The knowledge gained by machine learning algorithms could be used to automatically scale an appropriate amount of VMs in order to never exceed a certain response time. Furthermore, the additional knowledge gained by training different models with various machine learning algorithms can be seen as external information. This information about the strength and weakness of each algorithm can be exploited. For example it can be declared that in critical business cases, where the transgression of response times is inevitably paired with costly SLA violations, the use of the NN algorithm, which predicts in a more cautious way by overestimating response times, could be prioritized. One possibility to fully realize this potential would be to offer classifications of individual SLAs in gold, bronze and silver. In this example the knowledge and application of the different algorithms could be used by a scaler regarding the Service Level Objective of the SLA ”Maximum Response Time of service X shall not surpass Y ms”. The use of the Neural Network could be set up by an coordination agent for gold customers whereas the use of the weaker but less expensive Linear Regression could be considered for bronze customers.

 ISBN: 978-1-61208-477-0

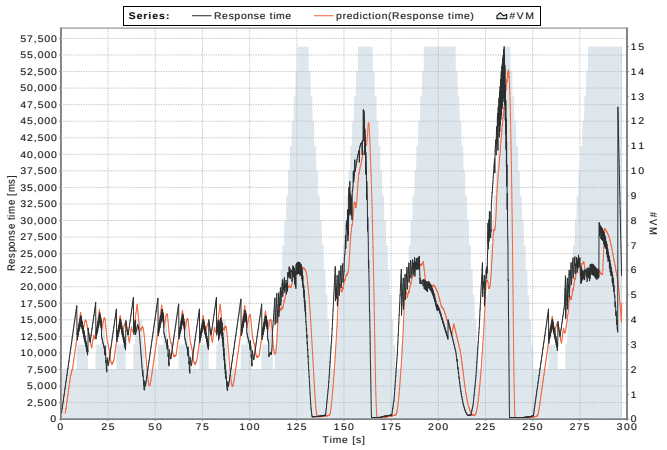

Figure 15. Linear Regression Scenario 3: 0s-300s

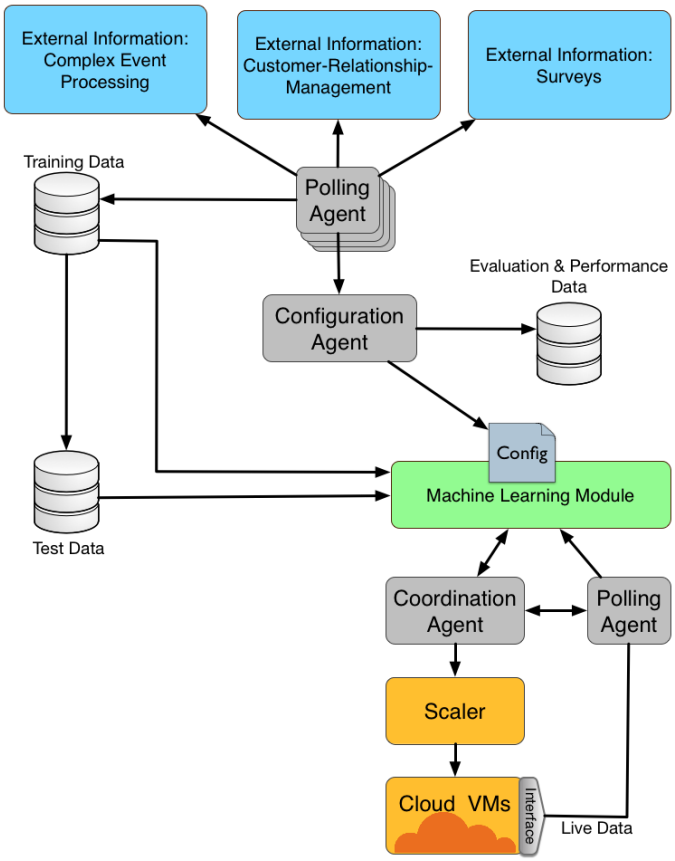

Figure 16. Architecture

This is just one example to show the synergy of the proposed architecture with the evaluated application of machine learning learning techniques.

One of the biggest challenges is without a doubt the inclusion and evaluation of external information which a model has never seen before. The correlation between historical and current information has to be established. This is no easy task as the problem starts already at the often needed transformation and preprocessing of the data in order to be able to train a model in the first place. The implementation of the architecture proposed in Figure 16 would enable a step to tackle this problem. But the next problem waits just around the corner. The evaluation of models, especially if the use of live data streams is involved. This is a current research problem and first proposals for solutions are presented by de Faria et al. in [12]. Although there is a noticeable progress in this area of expertise in general, it is still a long way from being able to provide a general approach and methodology.

VI. CONCLUSION AND FUTURE WORK

The aim of this paper was to show how selected machine learning algorithms cope with the prediction of response times in a cloud environment. Three different cloud usage scenarios were defined and three different algorithms (NN, SVM, Linear Regression) were applied. Knowledge about specific strengths and weaknesses about each algorithm was gained in the process. The general conclusion is that although each of the algorithms can be used for predicting response times effectively, some show specific characteristics which can be exploited. Additionally, an architecture was proposed in order to be able to deal with external information in an efficient way. Future work is to examine more algorithms with different configurations and scenarios in regard to response time. Furthermore, those results shall be substantiated by application on real cloud environments. Also, the creation of a framework of the architecture, proposed in Section IV is planned.

REFERENCES

[1] J. J. Prevost, K. Nagothu, B. Kelley, and M. Jamshidi, “Prediction of cloud data center networks loads using stochastic and neural models,” in System of Systems Engineering (SoSE), 2011 6th International Conference on, June 2011, pp. 276–281.

[2] C.-C. Li and K. Wang, “An SLA-aware load balancing scheme for cloud datacenters,” in Information Networking (ICOIN), 2014 International Conference on, Feb 2014, pp. 58–63.

[3] R. Hu, J. Jiang, G. Liu, and L. Wang, “KSwSVR: A New Load Forecasting Method for Efficient Resources Provisioning in Cloud,” in Services Computing (SCC), 2013 IEEE International Conference on, June 2013, pp. 120–127.

[4] A. Bankole and S. Ajila, “Predicting cloud resource provisioning using machine learning techniques,” in Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on, May 2013, pp. 1–4.

[5] M. Imam, S. Miskhat, R. Rahman, and M. Amin, “Neural network and regression based processor load prediction for efficient scaling of Grid and Cloud resources,” in Computer and Information Technology (ICCIT), 2011 14th International Conference on, Dec 2011, pp. 333–338.

[6] Cloudbus.org, “The CLOUDS Lab: Flagship Projects - Gridbus and Cloudbus,” Available: http://www.cloudbus.org/cloudsim, [retrieved: 04, 2016].

[7] R. N. Calheiros, R. Ranjan, A. Beloglazov, A. F. De Rose, and R. Buyya, “CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms,” Softw. Pract. Exper., vol. 41, no. 1, Jan. 2011, pp. 23–50, http://dx.doi.org/10.1002/spe.995, [retrieved: 04, 2016].

[8] Nikolay Grozev et al., “Cloudslab CloudSimEx,” Available: https://github.com/Cloudslab/CloudSimEx, [retrieved: 04, 2016].

[9] RapidMiner, “RapidMiner - #1 Open Source Predictive Analytics Platform,” Available: https://rapidminer.com, [retrieved: 04, 2016].

[10] S. Frey, C. Luthje, C. Reich, and N. Clarke, “Cloud QoS Scaling by Fuzzy Logic,” in Cloud Engineering (IC2E), 2014 IEEE International Conference on, March 2014, pp. 343–348.

[11] P. Chapman et al., “CRISP-DM 1.0 Step-by-step data mining guide,” August 2000, Available: https://the-modeling-agency.com/crisp-dm.pdf, [retrieved: 04, 2016].

 ISBN: 978-1-61208-477-0

[12] E. Ribeiro de Faria, I. Ribeiro Goncalves, J. Gama, and A. Carlos Ponce de Leon Ferreira Carvalho, “Evaluation of Multiclass Novelty Detection Algorithms for Data Streams,” IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 11, Nov 2015, pp. 2961–2973.

 ISBN: 978-1-61208-477-0

Automatic City Knowledge Discovery from Internet Resources

Nazanin Firoozeh

Laboratoire d'Informatique de Paris-Nord, Université Paris 13
Pixalione SAS
Paris, France
Email: nazanin.firoozeh@lipn.univ-paris13.fr

***Abstract*—Knowledge Discovery plays an important role in the Artificial Intelligence field. Due to the growing nature of the Web, using proper sources of extraction and reducing human intervention is an important step towards creating rich knowledge bases. Urban simulations are one type of Interactive Virtual Environment, which attempt to represent dynamic processes and interactions of urban development. Making these virtual environments closer to the human behaviour requires rich sources of knowledge. This paper presents a pattern-based approach for knowledge extraction. The goal is to extract the implicit knowledge behind any given city-related text. To achieve this goal, we make use of category names and infobox tables from Wikipedia. The system takes two inputs: 1. a text/Uniform Resource Locator (URL), 2. set of extraction patterns. Comparing to some of the proposed tools in the state-of-the-art, our system uses a simpler approach which reduces the human intervention. We tested the system with different text inputs and represented the results as both a text file and a set of triples. Manual evaluation of the system showed its good performance. According to the results, category names are a good resource of common sense knowledge when compared to infobox tables, which mostly contain basic knowledge.**

Keywords–Urban Simulation; Knowledge Discovery; Common Sense Knowledge

I. INTRODUCTION

Web pages contain a great amount of knowledge which is daily used by people or different systems. This knowledge plays an important role in Artificial Intelligence (AI) and Natural Language Processing (NLP). In fact, availability of large coverage and machine readable knowledge is an important step towards the goals of intelligent systems such as urban simulation systems. Urban simulation models are increasingly being used in city, country, and regional planning agencies to assess alternative transportation investments, land use regulations, and environmental protection policies. To have better simulations and also make the city agents more intelligent, using a rich knowledge base is essential.

One advantage of using large and high quality knowledge bases is to help agents to make better decisions and perform well, especially in real-time environments. In addition, having a dynamic knowledge extraction system, agents will be able to dynamically extract knowledge while facing different concepts of the real world. Moreover, as common sense knowledge (CS) affects human decision making process, providing a good resource of common sense knowledge is highly important. Having this type of knowledge, agents and humans will make closer decisions.

Wikipedia is a resource which is updated regularly and contains many statements in natural language. Due to the advantages of this resource, such as representing mostly unique information, it is considered as one of the good resources in knowledge extraction. In this work, we make use of Wikipedia and develop an automatic city knowledge discovery system. The goal of the system is to extract basic and common sense knowledge implicitly expressed in the content of any city-related text. To achieve this goal, semi-structured content of Wikipedia is used. Comparing to unstructured content, the advantage of using semi-structured content is to extract more reliable knowledge with higher accuracy rate.

The rest of the paper is organized as follows. In Section II, we briefly describe the works in the related domain. We then describe our proposed model and the implementation steps in Sections III. Evaluation of the system is shown in Section IV. We finally discuss the proposed model and conclude our work in Section V.

II. BACKGROUND

Extracting knowledge from Wikipedia has been one of the interesting domains of research. Some researchers believe that using Wikipedia as an extraction resource improves quality and size of knowledge bases. So far, different systems have been developed for extracting knowledge from Wikipedia. One instance of such systems is YAGO [1], an ontology which uses Wikipedia. In this project, instead of using information extraction methods, categories of Wikipedia articles are used as sources of knowledge. This system makes use of both Wikipedia articles and WordNet database in order to extract facts. The motivation behind combining these two resources is to take advantage of large number of individuals in Wikipedia as well as a clean taxonomy of concepts in WordNet. YAGO is able to detect both *is-a* and *not-is-a* relations such as *BornInYear*, *PoliticianOf*, etc. However, one limitation of YAGO is that it extracts only 14 types of relations and some relations cannot be detected using this system. As the second limitation, it is not able to extract facts from tables, such as infobox tables. In another work, authors extended the system to YAGO2 [2]. Comparing to YAGO, YAGO2 is able to represent the facts along dimensions of time and space. To achieve this, GeoNames resource is also used in addition to the existing resources, i.e., Wikipedia and WordNet.

In another work, WordNet database is enriched using new relations extracted from Wikipedia [3]. The work consists of four steps, which are all automatic. 1) As a pre-processing step, each Wikipedia entity is assigned to the corresponding WordNet synset. 2) First, the system looks for words which are connected to the entity through hyperlinks. WordNet is then used to see if there is any relation between the entity and any

 ISBN: 978-1-61208-477-0

of the found words. In case of having a relation, the context is analysed and a pattern is extracted for that relation. 3) In the next step, similar patterns are generalized. 4) Finally, the patterns are applied for finding new relations, which do not exist in WordNet database. Evaluation of this system shows that the precision rate of the extracted relations is not good enough, 0.61 to 0.69. Hence, one disadvantage of this system is that it produces some unreliable facts. In addition, to extract relations between different concepts, the system needs to go through definition of each entity in Wikipedia to look for hyperlinks. This increases the time complexity of the system.

Category names of Wikipedia articles have been also noted in other works. Large scale taxonomy was built from Wikipedia [4]. In this work, semantic relations between categories are found using a connectivity network and lexico-syntactic matching. This method is able to extract both *is-a* and *not-is-a* facts including relations such as of, with, contain, etc. However, although this method extracts relation between category names, it cannot find relation between a specific concept and its corresponding category names.

DBpedia [5], [6], [7] is another knowledge extraction system, which is now available on the World Wide Web. The goal of DBpedia is first to extract structured information from Wikipedia and then to allow users to ask queries against it. It also links different data sets on the Web to Wikipedia data. In the extraction process, authors make use of MediaWiki, which is a wiki software behind Wikipedia. MediaWiki enables authors to represent structured information in an "attribute-value" notation.

One drawback of the initial version of DBpedia was that its data could be based on several months old data. This problem however was solved using DBpedia-live [8]. DBpedia-live provides a live synchronization method based on the update stream of Wikipedia.

Wikipedia articles have been also used for extracting relations between different concepts of Wikipedia [9]. In this approach, an unsupervised approach is used along with linguistic analysis with web frequency information. The goal of using this analysis is to improve unsupervised classification performance. Unlike our approach, which focuses on semi-structured content of Wikipedia, here, unstructured content of articles is taken into consideration.

Similar to the described systems, we also make use of Wikipedia articles as a rich source of information. However, unlike YAGO, YAGO2 and the work done by Ruiz-Casado et al, the system does not use WordNet database. Hence, the complexity of the system is reduced. In addition, while the work done by Ponzetto et al, tries to find types of relations between different category names, our system finds relations between any given concept and its related category names. As a result, it provides various information about a specific concept rather than generating a network of category names.

DBpedia can be considered as the closest knowledge base to our proposed one. Comparing to DBpedia, our proposed approach is a simpler and less structured approach. This work is however an initial effort on knowledge extraction in order to propose a simple but efficient approach for extracting knowledge from Wikipedia. As will be seen later, as a future work, we are going to compare our extracted knowledge with DBpedia knowledge. Comparing the results, we will be then able to see if the proposed method can be considered as a complementary tool for DBpedia. In this case, using our system, some new relations can be added to DBpedia knowledge base. Specifically, in case that a concept does not have a corresponding entry in DBpedia, our system will be able to extract its basic and common sense knowledge in real time. It should be noted that although many works make use of knowledge extracted by DBpedia [10], [11], [12], according to [13] there is still a room for adding more entities and knowledge to this knowledge base.

III. A MODEL OF EXTRACTION

In this section, we explain in detail the different steps of our proposed approach.

A. Methodology

Extracting good amount of basic and common sense knowledge and using it in interactive applications is an important step in agents decision making. In particular, providing a comprehensive source of common sense knowledge enables machines to reason about everyday life. Meanwhile, developing automatic systems is highly important, as due to the growing nature of the information on the Web, it is almost impossible to manually create rich and up-to-date knowledge bases.

In this work, we make use of Wikipedia as a source of knowledge. Although Wikipedia has both unstructured and semi-structured content, we focus only on semi-structured content. The main motivation behind this choice is that common sense knowledge is a kind of knowledge that is merely expressed in unstructured content. However, as it is seen in the next sections, having extraction patterns, we will be able to extract this kind of knowledge from semi-structured content.

The two semi-structured sources used in our work are category names and infobox tables. Statistics show that from 2,390,513 available articles in Wikipedia in 2008, 1,057,563 articles (44.2%) contain infobox tables, while 1,927,525 articles (80.6%) have category names [14].

B. Model

In this section, the proposed algorithm is described. In general, our model consists of six main steps: 1. getting the Hyper Text Markup Language (HTML) source code, 2. getting the plain text of the source code, 3. parsing the plain text, 4. extracting category-based facts, 5. extracting infobox-based facts, 6. mapping the result into both a text file and a triple format. It should be noted that by category-based and infobox-based facts we respectively mean facts that are extracted from category names and infobox tables. Figure 1 illustrates the steps of the algorithm. Following is the detailed description of each step.

1) Getting the HTML Source Code: The input of the system can be either a city-related plain text or Uniform Resource Locator (URL). In case of having a plain text as an input, the system starts from the third step, i.e., parsing. Otherwise, the HTML source code of the given URL is retrieved for further processing.

 ISBN: 978-1-61208-477-0

```
Data: Text or URL of a web page, extraction patterns
Result: Extracted basic and common sense facts

initialization;
if input is a plain text then
    go to the next step;
else
    get plain text of the web page;
end
parse the text and generate concepts (based-on
  Wikipedia concepts);
while concept exists for each text do
    extract all category names from the category
      section of the corresponding Wikipedia article;
    while category name exists do
        apply extraction patterns and find the
          corresponding facts;
        if category-based facts != null then
            store the result into a text file;
            store the result as a triple
              format;
        else
            continue with the next
              category value;
        end
    end
    get the Wikitext of the concept;
    if Wikitext contains infobox table then
        apply the patterns on infobox table
          and extract the corresponding facts;
    else
        continue with the next concept;
    end
    if infobox-based facts != null then
        store the result into a text file;
        store the result as a triple format;
    else
        continue with the next concept;
    end
end
```

Figure 1. Algorithm of the developed knowledge extraction system.

2) Getting the Plain Text of the Source Code: Not all tags in the HTML code contain informative content. Hence, we apply a filtering procedure on the code to reduce its noisy content. Headers, Footers, Style, and Script tags in HTML code are examples of such noisy tags. Having the clean HTML source code, we then extract its content.

3) Parsing: The next step of the algorithm is to detect concepts of the studied text for which we want to extract facts. This is done through Parsing step. In this step, we customize the Stanford Part-Of-Speech (POS) tagger. The tagger uses Penn Treebank tag set for representing tags [15]. It makes use of different sets of models as training set of the tagging procedure. The model we use in our system is english-left3words-distsim.tagger, which is the widely used one in applications. Accuracy rate of the model is also 96.97%.

Concepts are generated by means of both POS tagger and Wikipedia articles. In fact, we assume that each Wikipedia article corresponds to one concept in the real world. Some concepts however have different representations. For these concepts, Wikipedia redirects users to the same article and shows the same page for different representations. Using the redirection feature of Wikipedia, we avoid generating concepts with the same meaning but different formulations. As an example, both UK and United Kingdom are redirected to the same URL in Wikipedia. Hence, only one of them is taken into consideration for further processing.

4) Extracting Category-Based Facts: Most of the Wikipedia articles have a category section where navigational links to other Wikipedia pages are provided. Using categories of Wikipedia, users are able to quickly find sets of pages related to any Wikipedia article. In order to reveal the semantics encoded in category names of Wikipedia articles, we develop an extraction algorithm which consists of three steps. The goal is to extract a set of triples as {*concept1, relation, concept2*}, *where concept1 is a detected concept in the given text, concept2 is a concept found in category section of Wikipedia, and relation shows the semantic relation between the two concepts.* Following is the detailed description of each step:

a) Extracting category names from HTML source code: In the first step, for any detected concept, source code of the corresponding Wikipedia article is extracted. Category section of the article is then retrieved and its category names are extracted.

b) Discarding uninformative names: Not all the extracted category names are informative. There are some general category names such as "Disambiguation pages" that appear in some of Wikipedia articles. We ignore all the categories under Wikipedia administration.

c) Extracting the facts: In order to extract the knowledge behind category names, we propose 23 extraction patterns based on structures of different category names. To generate the patterns, different criteria such as type and position of the prepositions as well as occurrences of some keywords like Type, Establishment, Disestablishment, etc. are taken into consideration. Table I shows the proposed extraction patterns. As it is seen, one or more facts along with their corresponding triples are assigned to each pattern. It should be noted that the triples represent the relation between only two concepts. Hence, in case of having more than two concepts in the extraction pattern, no triple is assigned.

For each category name, the system checks if it matches any of the patterns. If so, the associated fact is extracted. It is important to mention that in Table I, *X* refers to *concept1*, and both *Y* and *Z* refer to *concept2*. In addition, *YEAR* simply indicates a year and *Xs* shows the plural form of *X*. *X1*, *X2* and *Y1*, *Y2* also show respectively the sub-terms of *X* and *Y*. All these symbols are considered as concepts in our extraction patterns.

5) Extracting Infobox-Based Facts: Infobox is another resource that we use for the purpose of fact extraction. The triples extracted from infobox tables are in the form of {*concept, attribute, value*} which in fact represent the value of a specific attribute for the studied concept. For each detected concept in the parsing step, we do the following steps to extract infobox-based facts:

TABLE I. EXTRACTION PATTERNS AND THEIR ASSOCIATED FACTS AND TRIPLES FOR CATEGORY VALUES OF WIKIPEDIA.

	Extraction pattern	*Fact*	*Triple*
1	Y of X	Y is an attribute of X	{Y, is_attribute, X}
2	Type(s) of Y	X has a type of Y	{X, has_type, Y}
3	Y in X	X has Y	{X, has, Y}
4	Y in Z (Y contains just letters)	X is a Y in Z X is a Y X is in Z	 {X, is_a, Y} {X, is_in, Z}
5	Y of Z	X is a Y of Z (Singular Y) X is one of the Y of Z (Plural Y)	
6	Y in YEAR	In year YEAR, there was a Y of X	
7	YEAR introductions	X was introduced in YEAR	{X, was_introduced, YEAR}
8	X in YEAR	X was in YEAR	{X, happened, YEAR}
9	Y established in YEAR	X has been established in YEAR X is a Y (Y singular) X is one of the Y (Y plural)	{X, was_established, YEAR} {X, is_a, Y}
10	YEAR establishment(s)	X has been established in YEAR	{X, was_established, YEAR}
11	Y establishment(s) in Z	X has been established in Y (if Y contains digit) X is in Z	{X, was_established, Y} {X, is_in, Z}
12	Y disestablished in YEAR	X has been disestablished in YEAR X was a Y (Y singular) X was one of the Y (Y plural)	{X, was_disestablished, YEAR} {X, was_a, Y}
13	YEAR disestablishment(s)	X has been disestablished in YEAR	{X, was_disestablished, YEAR}
14	Y disestablishment(s) in Z	X has been disestablished in Y (if Y contains digit) X was in Z	{X, was_disestablished, Y} {X, was_in, Z}
15	YX (one concept)	YX is one form of X	{X, has_form, YX}
16	Y=(Y1 Y2) & X=(X1 X2) (if Y2 = X2)	X is a Y (Y is singular) X is one of the Y (Y is plural)	{X, is_a, Y}
17	Y by type	X has a type of Y	{X, has_type, Y}
18	Y invention	X is a Y invention	{X, was_invented, Y}
19	Y format(s)	Format of X is Y	{X, has_format, Y}
20	YEAR birth	X was born in YEAR	{X, was_born, YEAR}
21	YEAR death	X died in YEAR	{X, died, YEAR}
22	Xs	X is a subset of Xs	{X, is_subset, Xs}
23	If none of the above relations	X R Y (relates)	{X, relates, Y}

a) Getting the Wikitext: Wikitext is a markup language used for writing the content of wiki websites. This language is in fact a simplified alternative to HTML. As the first step of extracting infobox-based facts, we get this raw data of each Wikipedia article.

b) Finding the infobox template: As not all the Wikipedia articles contain an infobox table, for each detected concept, we should check for the existence of the infobox template. The template starts with "{{Infobox" and ends with corresponding "}}". This section is called infobox template and is used for further processing. If Wikitext of a concept does not contain this template, we stop extracting infobox-based facts for that concept.

c) Extracting attributes and values: In case of having an infobox template in the Wikipedia page, we extract the concept's attributes along with their corresponding values. However, not all content of an infobox template produces interesting facts. Hence, we first discard useless attributes including image, caption, logo, alt, coat, footnote, etc. We then define different patterns as regular expressions for extracting attribute names and values. It is important to mention that we just keep the attributes which have at least one informative value.

d) Refining the extracted facts: In some cases, in infobox template, one attribute is related to the previous one. In our work, we try to relate the dependent attributes. As an example of such an attribute we can refer to "date". In some cases, this attribute by itself represents no meaningful fact and instead shows the corresponding date of the previous attribute.

6) Writing the Facts as Text and Triple Formats: The triple format has a format of {*concept1, relation, concept2* }, which represents the relation between any two concepts. As the last step of our work, in addition of representing the results in a human-readable format, i.e., a text file, we write the facts in the triple format. As a future work, these triples can be converted into the Resource Description Framework (RDF) triples in order to make them machine-readable and applicable for further use in real systems.

IV. EVALUATION

Accuracy and reliability of the extracted knowledge is one of the main steps in knowledge discovery. Using incorrect knowledge in different applications decreases their performance. In this work, we evaluate the system by defining three labels that indicate the quality of the extracted facts. The labels are correct, incorrect and ambiguous. A correct fact is a fact with a correct meaning and a proper formulation. Incorrect fact, on the other hand, refers to the fact with an incorrect meaning. Among the extracted facts, some have correct meanings but wrong formulations. For now, these facts are labeled as ambiguous. By differently labeling these facts, we aim to differentiate them from the incorrect ones, since we believe that as a very first step of the future work, we can refine the system to get correct formulations for these cases. Hence in this work, we exclude ambiguous facts from the correct ones in the evaluation step.

In this work, we did a shallow evaluation in order to see the initial performance of the system. This evaluation was done manually by scanning all the facts and finding the ratio of the

 ISBN: 978-1-61208-477-0

correct, incorrect and ambiguous facts. To do this, 10 students were asked to label the extracted facts and the final label was assigned based on the majority vote.

Evaluation of the system is an important step as it helps us with further improvements. In this section, performance of each extraction pattern, used for extracting category-based facts, is also evaluated. The following subsection is the description of the experiment step. After, we present the results and analyse them to show the efficiency of our system.

A. Experiment

We tested the system with different city-related web pages. As a shallow and initial evaluation, we took into account the facts extracted from 10 input texts. These texts are either from news websites or city-related web pages. The extracted facts were evaluated in terms of both meaning and formulation.

It should be noted that although it is possible to run the system for any other domain, we evaluated its performance over the city-related web pages, as the global aim of our work is to apply the developed system into interactive city applications. In the following subsection, the obtained results are presented and analysed in order to show the initial performance of the system.

B. Result and Analysis

As mentioned in the previous sections, extracted facts are stored in a text file. Going through the extracted facts, we calculated the ratio of the correct, incorrect and ambiguous facts. Tables II and III show examples of the extracted facts for different labels.

Although most of the extracted facts are considered as basic knowledge, some others can be considered as common sense knowledge. As an example, “Eiffel tower is in Paris” is considered as common sense knowledge since almost all people know it. This means that while saying “I am travelling to visit the Eiffel Tower”, we are implicitly saying that “I am travelling to Paris”.

In case of having category-based facts, the extracted facts are in two types; “R-specific” that explicitly specifies types of relations and “R-generic” that just indicates that the concepts are related without explicitly showing type of the relation. In Table I, patterns 1 to 22 generate R-specific facts, whereas R-generic facts are extracted using pattern 23.

Figure 2 compares the average rates of accuracy, error and ambiguity for both category-based and infobox-based facts and over all the evaluated examples. For the former case, the evaluation metrics are calculated over R-specific facts, since this type shows the performance of the extraction patterns. Hence, in this step, by “total number of the facts” we mean total number of the R-specific facts. In calculations, accuracy, error, and ambiguity rates are obtained by respectively dividing the number of correct, incorrect, and ambiguous facts to the total number of the facts. Result of the evaluation can be also shown as a precision metric. In our evaluation, in order to calculate the precision of the system, we discard ambiguous facts by assuming that they equally affect the positive and negative examples. Having this assumption, the obtained precision values for category-based facts and infobox-based facts are respectively 90% and 91%. It should be noted that these values are related to both basic and common sense

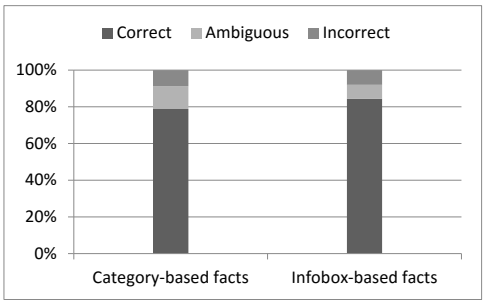

Figure 2. Comparing the average rates of correct, incorrect and ambiguous facts (common sense and basic) over 10 evaluated input texts.

facts. According to the results, the system has an acceptable precision. As an important step in the future work, we are going to use a gold standard for calculating the recall value. In our case, the gold standard may contain a set of facts extracted manually from infobox tables and category names of corresponding Wikipedia articles.

As mentioned before, in the infobox template, in order to extract value(s) of each attribute, patterns are defined using regular expressions. Due to the complexity of the infobox templates, complex patterns should be applied and this increases the complexity of the system. To overcome this issue, we tried to have a trade-off between accuracy and complexity, meaning that instead of extracting all the values, just majority of them are extracted to avoid increasing the complexity. One case that we are not able to capture in our system is the date format containing date, month and year, e.g., 13/09/1988. Although this is one of the limitations of our system, as in some cases dates can be captured using category names, we miss less information. According to our evaluations, half of the incorrect facts are related to the date format.

The next point is that in a very few cases, the extracted category-based facts are almost the same. In fact, these facts are extracted from different category names, which are close to each other. In our system, we tried to reduce the number of similar facts. As a result, in the evaluated examples, we had either no repetitive facts or just one or two cases. For instance, considering the facts “Paris is a capital in Europe” and “Paris is one of the capitals of Europe”, we represent only one of them in the final result.

According to Figure 2, the average rate of accuracy for the extracted infobox-based facts is higher than the one for the category-based facts. However, difference of their error rates is minor. This indicates that most of the infobox-based facts are labeled as correct and incorrect, whereas a higher number of the category-based facts have a degree of correctness and cannot be labeled explicitly as correct or incorrect.

Average rates of R-specific and R-generic category-based facts are shown in Figure 3. As expected, more facts are categorized as R-generic. The reason is that many of the category names have no specific structure. Retrieving facts from these category names though might contain correct facts, increases the error rate. Hence, ignoring such names is more efficient. Figure 4 also compares the amount of common sense and basic knowledge in both category-based and infobox-based facts.

 ISBN: 978-1-61208-477-0

TABLE II. EXAMPLES OF THE EXTRACTED FACTS HAVING CORRECT, INCORRECT AND AMBIGUOUS LABELS - CATEGORY-BASED FACTS.

	Extracted category-based fact	*Evaluated as*	*Basic/CS knowledge*
1	Paris is a capital in Europe	Correct	CS
2	Versailles is an art museum and gallery	Ambiguous	–
3	Eiffel Tower is a Michelin Guide	Incorrect	–

TABLE III. EXAMPLES OF THE EXTRACTED FACTS HAVING CORRECT, INCORRECT AND AMBIGUOUS LABELS - INFOBOX-BASED FACTS.

	Extracted infobox-based fact	*Evaluated as*	*Basic/CS knowledge*
1	Latitude of Paris is 48.8567	Correct	Basic
2	Roof of Eiffel Tower is, abbr=on	Incorrect	–
3	Gini year of Spain is 2005	Ambiguous	–

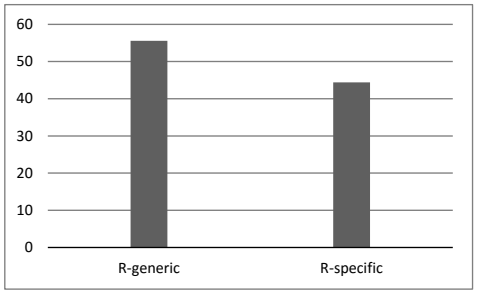

Figure 3. Comparing R-generic and R-Specific rates in category-based facts.

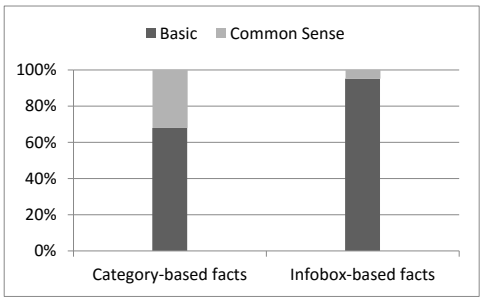

Figure 4. Comparing types of the extracted facts (Common sense vs. Basic).

According to the figure, it can be concluded that infobox template is not a good resource for extracting common sense knowledge. Instead, basic knowledge can be extracted using this table. This is due to the fact that Wikipedia has infobox tables for proper nouns such as Paris and not for general concepts such as shopping. As a result, the extracted facts in most cases cannot be considered as common sense knowledge. On the other hand, category names seem to be a better resource for extracting common sense knowledge. Results show that using category names, we are able to extract both common sense and basic knowledge.

One of the important steps in evaluating the system is to evaluate the performance of the proposed extraction patterns. In this step, pattern 23 from Table 1 is not considered, as it generates R-generic relations, while we are interested to see how efficient the extraction patterns are in extracting R-specific relations. To do this, we focused on half of the input texts (5 texts) and studied the total number of the facts extracted using each pattern (Figure 5). Performance of each pattern is also shown in Figure 6.

According to the figure, patterns 3, 10, 17, 19, 20, and 21 from Table 1 have the highest accuracy rate. However, these patterns extract a few numbers of facts. Considering the patterns with high rates of extraction, i.e., patterns 4, 5 and 22, it can be seen that pattern 22 outperforms the other two patterns due to the high rate of accuracy (96.06%). On the other hand, the result shows that pattern 15 has a poor performance when applied on the input texts, as it mostly generates incorrect facts. Hence, this pattern should be removed while improving the system.

Figure 5 also shows that patterns 11, 12, 13, and 14 extract no fact in the mentioned examples. In fact, they extract some facts but as the produced facts were similar to the previously extracted facts by the other patterns, we removed them from the final result. However, it is important to keep these patterns, as due to their structure, in some cases it might be possible to extract new facts from these patterns.

V. DISCUSSION & CONCLUSION

This paper addresses the problem of automatically extracting basic and common sense knowledge with the goal of providing rich city knowledge bases. These knowledge bases can be then used in interactive city applications to help agents to make decisions. In this work, category names and infobox tables of Wikipedia articles are used as resources of knowledge extraction. Unlike many of the systems in the state-of-the-art, our proposed model is a simple approach which reduces the human intervention. Our system just makes use of the proposed extraction patterns without having the effort of using some thesauri or ontologies. In addition, it enables agents to dynamically extract knowledge when they receive a new input text. Knowledge extracted using this approach could be considered as complementary knowledge of DBpedia.

We generated 23 extraction patterns for extracting facts from category names. Also, some complex patterns were defined to extract attribute and value pairs from infobox tables. Results of the system on 10 input texts show the average rates of correct, incorrect and ambiguous facts as 78.80%, 8.69% and 12.49% for category-based facts and as 84.36%, 8.002% and 7.62% for infobox-based facts. In terms of precision, for category-based facts and infobox-based facts values of 90%

 ISBN: 978-1-61208-477-0

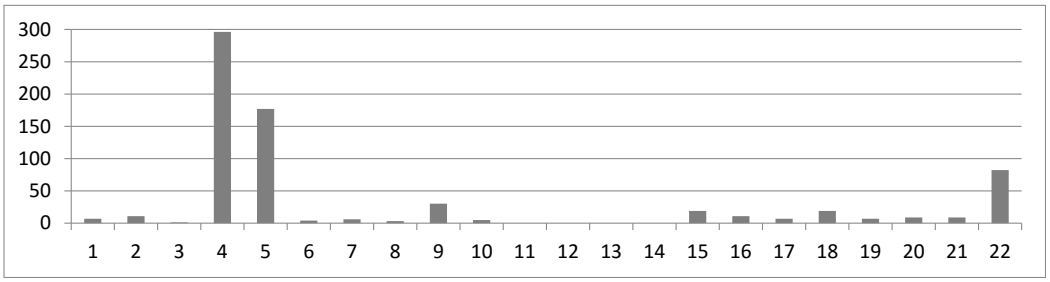

Figure 5. Total number of the facts extracted by each pattern over five input texts.

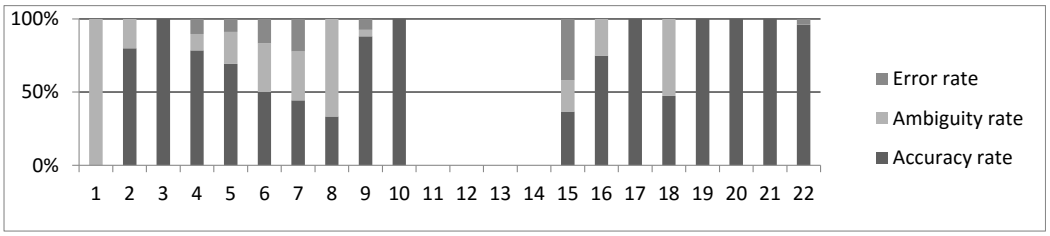

Figure 6. Performance of each extraction pattern over five input texts.

and 91% were respectively obtained. Results also show that category names are better resources for extracting common sense knowledge, while infobox tables mostly extract basic facts. According to the results, among all the category-based extraction patterns, pattern 22 has a better performance with a high accuracy rate of 96.06%.

The first step of the future work is to refine the system in order to get correct formulations for ambiguous facts. As other steps, we can make the system more automated using bootstrapping approach [16], which makes use of a training set, including pairs from infobox tables and the extracted facts, and does a recursive self-improvement. After having an automatic evaluation phase, the next step is to compare our results with the ones obtained from the previous works. To have a better evaluation, value of recall should be also calculated. Gold standard can be then generated manually. As an alternative approach, one could use the facts extracted using DBpedia. As one of the objectives of our system is to extract facts which do not exist in DBpedia, we cannot use this knowledge base for calculating the recall value.

The next step in future work is to convert the extracted triples into RDF triples in order to make them machine-readable for further use in real systems.

ACKNOWLEDGMENT

We would like to thank the Laboratoire d'informatique de Paris 6 (LIP6), where this work was conducted.

REFERENCES

[1] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A core of semantic knowledge," in Proceedings of the 16th International Conference on World Wide Web (WWW), 2007, Banff, Alberta, Canada. ACM, 2007, pp. 697–706, ISBN: 978-1-59593-654-7, URL: http://doi.acm.org/10.1145/1242572.1242667.

[2] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "Yago2: A spatially and temporally enhanced knowledge base from wikipedia," 2010.

[3] M. Ruiz-Casado, E. Alfonseca, and P. Castells, "Automatic extraction of semantic relationships for wordnet by means of pattern learning from wikipedia," in Proceedings of the 10th International Conference on Natural Language Processing and Information Systems (NLDB), 2005, Alicante, Spain. Springer-Verlag, 2005, pp. 67–79, ISBN: 3-540-26031-5, 978-3-540-26031-8, URL: http://dx.doi.org/10.1007/11428817_7.

[4] S. P. Ponzetto and M. Strube, "Deriving a large scale taxonomy from wikipedia," in Proceedings of the 22Nd National Conference on Artificial Intelligence (AAAI), 2007, Vancouver, British Columbia, Canada. AAAI Press, 2007, pp. 1440–1445, ISBN: 978-1-57735-323-2, URL: http://dl.acm.org/citation.cfm?id=1619797.1619876.

[5] S. Auer and J. Lehmann, "What have innsbruck and leipzig in common? extracting semantics from wiki content," in Proceedings of the 4th European Conference on The Semantic Web: Research and Applications (ESWC), 2007, Innsbruck, Austria. Springer-Verlag, 2007, pp. 503–517, ISBN: 978-3-540-72666-1, URL: http://dx.doi.org/10.1007/978-3-540-72667-8_36.

[6] C. Bizer et al., "DBpedia - A Crystallization Point for the Web of Data," Web Semant., vol. 7, 2009, pp. 154–165, ISSN: 1570-8268, URL: http://dx.doi.org/10.1016/j.websem.2009.07.002.

[7] J. Lehmann et al., "DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia," Semantic Web Journal, vol. 6, 2015, pp. 167–195.

[8] M. Morsey, J. Lehmann, S. Auer, C. Stadler, and S. Hellmann, "DBpedia and the Live Extraction of Structured Data from Wikipedia," Program: electronic library and information systems, vol. 46, 2012, p. 27.

[9] Y. Yan, N. Okazaki, Y. Matsuo, Z. Yang, and M. Ishizuka, "Unsupervised relation extraction by mining wikipedia texts using information from the web," in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009, Suntec, Singapore. Association for Computational Linguistics, 2009, pp. 1021–1029, ISBN: 978-1-932432-46-6, URL: http://dl.acm.org/citation.cfm?id=1690219.1690289.

[10] M. Héder and P. N. Mendes, "Round-trip semantics with sztakipedia and dbpedia spotlight," in Proceedings of the 21st In-

ternational Conference on World Wide Web (WWW), 2012, Lyon, France. ACM, 2012, pp. 357–360, ISBN: 978-1-4503-1230-1, URL: http://doi.acm.org/10.1145/2187980.2188048.

[11] M. Szczuka, A. Janusz, and K. Herba, "Clustering of rough set related documents with use of knowledge from dbpedia," in Proceedings of the 6th International Conference on Rough Sets and Knowledge Technology (RSKT), 2011, Banff, Canada. Springer-Verlag, 2011, pp. 394–403, ISBN: 978-3-642-24424-7, URL: http://dl.acm.org/citation.cfm?id=2050461.2050520.

[12] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, "Unsupervised graph-based topic labelling using dbpedia," in Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM), 2013, Rome, Italy. ACM, 2013, pp. 465–474, ISBN: 978-1-4503-1869-3, URL: http://doi.acm.org/10.1145/2433396.2433454.

[13] G. Quercini and C. Reynaud, "Entity discovery and annotation in tables," in Proceedings of the 16th International Conference on Extending Database Technology (EDBT), 2013, Genoa, Italy. ACM, 2013, pp. 693–704, ISBN: 978-1-4503-1597-5, URL: http://doi.acm.org/10.1145/2452376.2452457.

[14] Q. Liu et al., "Catriple: Extracting triples from wikipedia categories," in The Semantic Web, 2008, URL: http://dx.doi.org/10.1007/978-3-540-89704-0_23.

[15] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," Comput. Linguist., vol. 19, 1993, pp. 313–330, ISSN: 0891-2017, URL: http://dl.acm.org/citation.cfm?id=972470.972475.

[16] S. Zhao and J. Betz, "Corroborate and learn facts from the web," in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2007, San Jose, California, USA. ACM, 2007, pp. 995–1003, ISBN: 978-1-59593-609-7, URL: http://doi.acm.org/10.1145/1281192.1281299.

 ISBN: 978-1-61208-477-0

Log-Modulus for Knowledge Discovery in Databases of Financial Reports

Duarte Trigueiros
University of Macau, University Institute of Lisbon
Lisbon, Portugal
Email: dmt@iscte.pt

Carolina Sam
Master of European Studies Alumni Association
Macau, China
Email: kasm@customs.gov.mo

***Abstract*—An alternative is proposed to the use of ratios in financial predictive modelling. Such alternative, the "log-modulus", overcomes limitations, which have hitherto thwarted most of the previous attempts to predict financial attributes from data. Moreover, the use of log-modulus opens-up the prospect of performing Knowledge Discovery in Databases (KDD) of financial reports. Using controlled experiments, the paper shows that models using log-modulus are accurate, robust and balanced in cases where ratios fail to deliver feasible results. The paper also provides a theoretical basis supporting the observed ability of log-modulus to allow knowledge discovery of financial statements.**

Keywords–Type of Information Mining; Knowledge Discovery in Databases; Predictive Modelling; Financial Reports.

I. INTRODUCTION

Business companies, namely those listed in stock markets, are required to prepare annual reports reflecting their financial activity and position at the end of each year. Large databases containing these reports are routinely scrutinised by investors, banks, regulators and other parties with the object of taking decisions regarding individual companies and industrial sectors. Such scrutiny, and the corresponding diagnostic, is known as "Financial Analysis".

Financial Analysis aims to diagnose the financial outlook of a company. The major source of data for such diagnostic is the set of financial reports regularly made public by the company and by other companies in the same industrial sector. The diagnostic itself consists of identifying and in some cases measuring the state of financial attributes, such as Manipulation, Going Concern, Solvency, Profitability and others. The tool used by analysts to assess such attributes is the "ratio", a quotient of two monetary amounts appropriately chosen. After being identified and measured, financial attributes convey a clear picture of a company's future economic prospects and may support the taking of momentous decisions, such as to buy or not to buy shares, to lend money and others. Financial attributes, therefore, are the knowledge set where investing, lending and other decisions are based.

The paper is about the discovery and assessment of underlying attributes in databases of financial reports. It describes a methodology capable of reliably producing, from such databases, knowledge represented so as to allow inferencing.

Attempts to perform analytical modelling of financial attributes have largely failed except in one instance, bankruptcy prediction [1]. Other, equally vital attributes, such as the trustworthiness as opposed to fraudulent reports, have resisted attempts to be reliably predicted [2]. Such failure is largely due to difficulties posed by ratios when used as predictors but, hitherto, no attempt has been made to find alternatives. The objective of the paper is to overcome the current stalemate by proposing a type of predictor, the log-modulus [3], which overcomes ratios' limitations and is amenable to knowledge discovery. An effective KDD of financial reports would quicken and lighten the analysis process, freeing analysts to concentrate on specific cases thus improving their efficiency.

Section II describes the KDD challenge being tackled; Section III offers theoretical considerations supporting the use of log-modulus; Section IV presents results of controlled experiments where log-modulus are compared with ratios; Section V highlights expected benefits.

II. FINANCIAL ANALYSIS, A KDD CHALLENGE

Financial reports are standardized data-sets prepared and published by business companies on a regular basis. They contain, besides non-numerical data, a collection of monetary amounts with an attached meaning: revenues of the period, different types of expenses, asset and liability values at the end of the period and others. Such amounts are obtained via a process involving the recognition and aggregation into "accounts", of similar transactions relating to the period. The resulting "set of accounts" is made available to the public together with non-numerical information in the form of a financial report.

Amongst investors, regulators and banks, an extremely popular value-added product is the database containing current and past financial reports of companies listed in one or several regions. This database typically includes complementary information, such as an extended identification data, industrial and economic classifications, the rating of outstanding debts and the market value of shares. Financial services companies, such as Thomson-Reuters or Standard & Poor's respectively sell "Datastream" and "Compustat" databases, two examples amongst others of such product. Analysts routinely access financial reports via databases.

Attributes examined by financial analysts are hierarchically linked: the meaning of one depends on the meaning of others higher up in a hierarchy (Figure 1). The top attribute, which allows all the others to be meaningful, is whether a report is trustworthy or not. If the report is free from

 ISBN: 978-1-61208-477-0

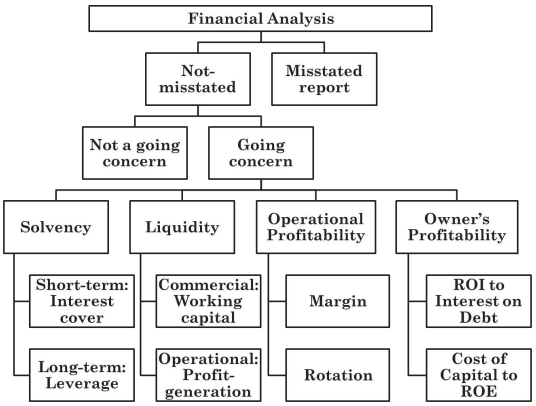

Figure 1. Uppermost dependencies in the hierarchy of financial attributes.

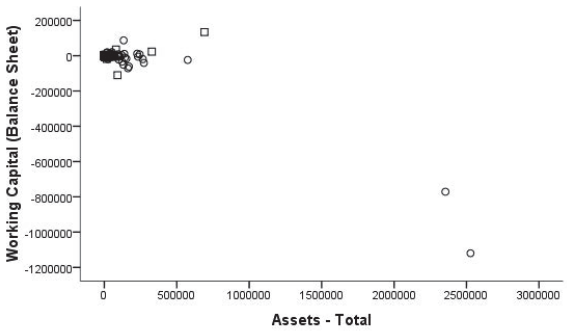

Figure 2. Influential cases in a scatter-plot of two ratio components, where some 3,000 cases are concentrated in a small region.

manipulation then it may be asked whether the company is a going concern or not. Only in going concerns it makes sense to assess attributes, such as Liquidity, which also are at the root of hierarchies.

Knowledge discovery in databases of financial reports is the process of assigning each company in the database a set of logical classes/numerical values pertaining to attributes forming taxonomies similar to those of Figure 1. The assignment process is carried out using a corresponding set of models which, in turn, are built using "supervised learning" where algorithms learn to recognise classes from instances where diagnostics are already made; but unsupervised learning is also possible [4]. When completed, such process greatly facilitates the task of analysts, allowing them to concentrate on companies and conditions where algorithms may not be able to produce accurate diagnostics. If, for most of the attributes, the modelling is unreliable then knowledge discovery is of little use. Such is the present situation, where only one of the many attributes analysts work with is predicted accurately.

Financial analysis of a company is typically based on the comparison of two monetary amounts taken from published reports. For instance, when a company's net income at the end of a given period is compared with assets required to generate such income, an indication of "Profitability" emerges. Pairs of items are often expressed in the form of a single value, their ratio. Since the size effect is similar for all items taken from the same company and period, size cancels out when a ratio is formed. Thus, ratios may be used to compare companies of different sizes [5]. Besides their size-removal ability, ratios directly measure attributes, which are implicit in reported statement numbers. Profitability, for instance, is identified as a specific ratio. Thus, the use of ratios has extended to cases where size-removal is not the major goal. Indeed, ratios are used because they embody the knowledge, which analysts possess [6].

Financial analysis is a rewarding albeit burdensome exercise. In the hands of an experienced analyst, a trustworthy financial report reveals the true condition of a company. Attempts to extract knowledge from such rich content did not succeed probably due to the very success of analysts. When trying to build automated, knowledge discovery algorithms applicable to databases, authors tend to imitate analysts namely in the use of ratios. But in spite of being the chief tool of analysts, ratios are inadequate to analytical knowledge discovery for two reasons: first, their statistical behaviour is atypical; second, they are themselves knowledge, focused pieces of knowledge, not just data.

Ratios are inadequate firstly because monetary amounts taken from financial reports, as well as ratios formed from them, obey a multiplicative law of probabilities, not an additive law. Ratio components, indeed any figure reported in a given set of accounts, are accumulations and, as such, they obey a specific generative mechanism where distributions are better described by the Lognormal and other similar functions with long tails (influential cases) and inherent heteroscedasticity [5]. Where the multiplicative character of financial statement data is ignored, any subsequent effort to model such data is fruitless, not so much because Ordinary Least Squares (OLS) or other assumptions are violated but due to the distorting effect of influential cases (Figure 2) and heteroscedasticity. And when predictive performance is the issue, the use of robust algorithms is not recommended because the cost of such robustness is lessened performance. Amongst the three basic types of measurement, Nominal, Ordinal and Scalar, the latter is the richest in content. When a scale is treated as an ordered sequence (as is the case of most robust algorithms), a great deal of content is lost.

In the second place, the use of ratios in knowledge-discovery entails a contradiction. When a ratio is chosen instead of other ratios, knowledge is involved. Each ratio embodies the analyst's knowledge that, when two monetary amounts are set against each other, a hidden attribute is evidenced. Ratios, therefore, convey previously known knowledge.

Analysts use ratios because they can assess only one piece of information at a time. They are not able to jointly deal with collections of distributions, their moments and variance-covariance matrices as algorithms do. Analysts need focus, machines don't. Predictive models can only lose by mimicking analysts' requirements of separation of knowledge in small bits in order to rearrange it in a recognisable way. As explained in the coming section, algorithms are able to choose amongst a set of monetary amounts, those leading to optimal models. In doing so, algorithms build their own representations in a way similar

 ISBN: 978-1-61208-477-0

to analysts' task of selecting, amongst innumerable combinations of monetary amounts, the pair which highlights a desired attribute.

III. THEORETICAL CONSIDERATIONS

Studies on the statistical characteristics of reported monetary amounts brought to light two facts. First, in cross-section the probability density function governing such amounts is nearly lognormal. Second, amounts taken from the same set of accounts share most of their variability as the size effect is prevalent [5]. Thus, variability of logarithm of account i from set of accounts j, $\log x_{ij}$, is explained as the size effect s_j, which is present in all accounts from j, plus some residual variability ε_i:

$$\log x_{ij} = \mu_i + s_j + \varepsilon_i \tag{1}$$

μ_i is an account-specific expectation. Formulations such as (1), as well as the underlying random mechanism, apply to accumulations only. Accounts, such as Net Income, Retained Earnings and others, which can take on both positive- and negative-signed figures, are a subtraction of two accumulations. Net Income, for instance, is the subtraction of Total Costs from Revenue, two accumulations, not the direct result of a random mechanism.

Given two accounts $i = 1$ and $i = 2$ (Revenue and Expenses for instance) and the corresponding reported amounts x_1 and x_2 from the same set, the logarithm of the ratio of x_2 to x_1 is

$$\log \frac{x_2}{x_1} = (\mu_2 - \mu_1) + (\varepsilon_2 - \varepsilon_1) \tag{2}$$

It is clear why ratios formed with two accounts from the same set are effective in conveying information to analysts: the size effect, s_j, cancels out when a ratio is formed. In (2), the log-ratio has an expected value $(\mu_2 - \mu_1)$. The median ratio $\exp(\mu_2 - \mu_1)$ is a suitable norm against which comparisons may be made while $\exp(\varepsilon_2 - \varepsilon_1)$ indicates the deviation from such norm observed in j. Ratios thus reveal how well j is doing no matter its size. For instance, if the median of Net Income to Assets ratio is 0.15, any company with one such ratio above 0.15, no matter small or large, is doing better than the industry.

In (2), upward or downward deviations from the log of the industry norm are the result of subtracting two residuals, each of them size- and account type-free. The deviation $\varepsilon_2 - \varepsilon_1$ from industry norms/benchmarks plays the crucial role of conveying to analysts the size-free, company-specific data they seek. It is clear, however, that $\varepsilon_2 - \varepsilon_1$ is only part of the size-free, company-specific information available in x_1 and x_2. When the ratio is formed, all variability common to x_1 and x_2 is removed. Residuals ε_1 and ε_2 are uncorrelated and the size-free, company-specific information contained in x_1 and x_2 but not conveyed by $\varepsilon_2 - \varepsilon_1$ is the variable orthogonal to $\varepsilon_2 - \varepsilon_1$, which is $\varepsilon_2 + \varepsilon_1$ [7]. Therefore, $\varepsilon_2 + \varepsilon_1$ is size-free information not conveyed by the ratio.

It is thus demonstrated that the exclusive use of ratios as model predictors curbs the information offered to the algorithm. Only one dimension of the size-free information, $\varepsilon_2 - \varepsilon_1$, is made available while the other dimension, $\varepsilon_2 + \varepsilon_1$, is ignored.

Given this, it is worth asking whether amounts directly taken from reports would not do a better job than ratios as predictors in statistical models. Such possibility is attractive but raises questions. It is attractive because predictors obeying (1) behave exceedingly well: distributions are nearly Normal, relationships are homoscedastic and influential cases, when present, are true outliers. Indirectly, log-transformed numbers allow the use of powerful algorithms which make the most of existing content. In the downside, one obvious concern is how to deal with accounts, which can take on both positive- and negative-signed figures. Logarithms can only deal with positive values.

An equally pressing concern is how to interpret coefficients of such models. Consider the usual linear relationship where y is explained by a set of predictors $x_1, x_2, \ldots$

$$y = a + b_1 x_1 + b_2 x_2 + \cdots \tag{3}$$

If, instead of $x_1, x_2, \ldots$ log-transformed predictors obeying (1) are included in (3), such relationship becomes

$$y = A + b_1 \varepsilon_1 + b_2 \varepsilon_2 + \cdots + (b_1 + b_2 +) s_j \tag{4}$$

where $A = a + b_1 \mu_1 + b_2 \mu_2 + \ldots$ is a constant value and residuals $\varepsilon_1, \varepsilon_2, \ldots$ now play the role of linear predictors. The term $(b_1 + b_2 + \ldots) s_j$ apportions the proportion of s_j (size) variability required by y. Coefficients $b_1, b_2, \ldots$ are under a constraint: their summation $b_1 + b_2 + \ldots$ must reflect the extent and sign of size-dependence in y; and where y is size-independent, $b_1 + b_2 + \ldots$ must be zero so as to bar information conveyed by s_j from entering the relationship.

Suppose, for instance, that y is indeed size-independent. Moreover, y is being predicted by two accounts only, x_1 and x_2. In this case $b_2 = -b_1 = b$ and (4) becomes $y = a + b(\mu_2 - \mu_1) + (\varepsilon_2 - \varepsilon_1)$ or

$$y = a + b \log \frac{x_2}{x_1} \tag{5}$$

In other words, a ratio is automatically formed so that size is removed from the relationship modelling y. Given the variety of companies' sizes found in cross-section relationships, the predictive power of s_j on y is, in most practical cases, small or non-existent. In such type of models $b_1 + b_2 + \ldots$ coefficients will indicate, not so much the strength and sign of the relationship between the ε and y but the amount of size-related variability, which is being allocated to a given predictor in order to counterbalance size-related variability from other predictors so that y is modelled by size-independent or nearly size-independent variability. When building an optimal model, the modelling algorithm assigns the role of denominator to some predictors (negative-signed coefficients) and that of numerator to others. Logarithmic representations similar to financial ratios are thus formed. In this way, financial attributes are modelled without the intervention of the analyst. This is a notable trait of the methodology.

The second concern, how to deal with accounts, which can take on both positive- and negative-signed amounts,

 ISBN: 978-1-61208-477-0

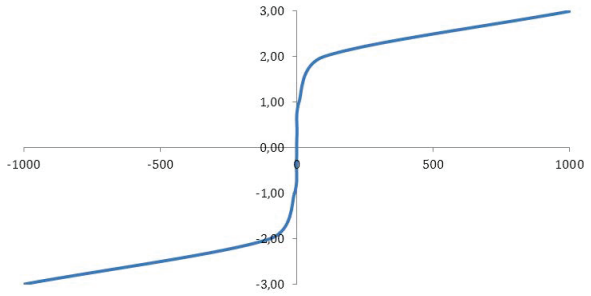

Figure 3. The x-axis represents x and the y-axis represents the log-modulus of x.

may be solved by using the "log-modulus" [3] or other similar transformation. Given variable x, the log-modulus consists of using

$$\operatorname{sgn} x \log(|x| + 1) \tag{6}$$

instead of x (Figure 3). In this way, accumulations or subtractions of accumulations, no matter their sign, become statistically well-behaved.

The coming section shows that models using the log-modulus as predictors no longer need the support of analysts (who, when selecting appropriate ratios, apportion substantial knowledge into the model) and perform better than those using ratios. Internal representations tend to assume the form of ratios in log-space because instances used in the learning of the algorithm greatly differ in size while the attribute to be predicted is indeed predictable. Models thus tend to self-organize themselves into size-independent linear combinations of predictors, efficient in predicting classes of the attribute.

Another advantage of the log-modulus transformation is that it considerably reduces the frequency of missing values in random samples. Besides reducing the power of samples, missing values in predictors are a source of bias to models because the probability that a reported number be missing often is correlated to the attribute being predicted. For instance, it is frequent to find amounts of zero in dividends and other accounts. When ratios are formed with such values in the denominator, as is the case of the ratio "Changes in Dividends relative to Previous Year", a missing case is created. Moreover, such missing case is correlated with the paying or not of dividends, an important predictor of Earnings' increases.

The log-modulus transformation solves this problem. Changes in relation to the previous period, for instance, are expressed in log-modulus as

$$\delta \log x = \log x_{t-1} - \log x_t \tag{7}$$

where t and $t-1$ express subsequent time periods and the operator $\log$ refers to (6). Since the log-modulus transformation is continuous and monotonic, changes expressed as in (7) do not generate new missing values.

Incidentally, unlikely ratios, such changes are never ambiguous: assumed values cannot have two meanings. This is not the case with ratios where negative-valued numerators and denominators lead to the same ratio sign as positive-valued numerators and denominators.

IV. CONTROLLED EXPERIMENTS

This section compares the predictive performance of ratios with that of log-modulus-transformed amounts taken from financial reports. Class proportions, period, industry, company size, the algorithm used and other characteristics, are similar for the two models being compared so as to equalise their influence on performance. The only differing characteristic is the type of predictor used. The modelling algorithm used throughout is the Binary Logistic Regression from the SPSS package.

Three experiments are performed respectively on the prediction of

1) bankruptcy, [8][1]
2) fraud [9][10][2]
3) and Earnings [11][12]

As depicted in Figure 1, bankruptcy and fraud are two basic attributes of financial analysis, directly influencing the way all other attributes are interpreted. As for Earnings, it is a good example of an attribute occupying a place further down in the hierarchy. Of the three, only bankruptcy prediction is reliable; in spite of the large research effort devoted to improving fraud detection, until today results are below the feasibility level, at 75% out-of-sample correct classification and highly unbalanced. All the previous literature, namely papers cited above, use ratios.

The first experiment replicates Altman's bankruptcy predicting model [8]. A total of 2,997 cases of US bankruptcy filings is drawn from the UCLA-LoPucki Bankruptcy Research Database [13]. Bankruptcies but the first in each company are discarded as well as cases about which detailed financial figures are not available. Two random samples of nearly 900 different cases each are drawn from the remaining (nearly 2,200) bankruptcies. The two samples contain companies listed in US exchanges and present in the Standard & Poor's "Compustat" database. They span the period 1979-2008. All sizes (Log-Total Assets deciles) and all the 24 "Global Industry Classification Standard" (GICS) groups are significantly represented in samples. Cases in the two samples are matched with an equal number of records from non-bankrupt companies. Pairing is based on the GICS group, on size decile and on year. Among financial statements fulfilling the pairing criteria, one case is randomly selected for matching and then such case is made unavailable for future matching. Although the same case is not used to match more than one bankruptcy case, cases from the same company in different years are allowed to be available for matching. The two matched samples have nearly 1,800 cases each. One of the two samples, always the same, is used as the learning-set and the other as the test-set. Due to missing observations, samples contain less than 1,800 cases:

Learning-set: non-bankrupt	845 (50.1%)
Learning-set: bankrupt	841 (49.9%)
Test-set: non-bankrupt (N)	837 (49.8%)
Test-set: bankrupt (P)	845 (50.2%)

 ISBN: 978-1-61208-477-0

TABLE I. BANKRUPTCY PREDICTION.

Bankruptcy predicting models	Ratios	Log-modulus
Non-bankrupt correct (TN)	782 (93.6%)	822 (98.2%)
Non-bankrupt incorrect (FP)	55 (6.4%)	15 (1.8%)
Bankrupt correct (TP)	819 (96.9%)	814 (96.3%)
Bankrupt incorrect (FN)	26 (3.1%)	31 (3.7%)
Precision: TP / (TP + FP)	93.71%	98.19%
Pseudo R-Square	0.595	0.693
Chi-Square	1526, 5 df	1993, 5 df

TABLE II. FRAUDULENT REPORT PREDICTION.

Fraud predicting models	Ratios	Log-modulus
Non-fraud correct (TN)	244 (69.1%)	303 (85.8%)
Non-fraud incorrect (FP)	109 (30.9%)	50 (14.2%)
Fraud correct (TP)	328 (79.8%)	371 (90.5%)
Fraud incorrect (FN)	83 (20.2%)	39 (9.5%)
Precision: TP / (TP + FP)	75.1%	88.1%
Pseudo R-Square	0.305	0.569
Chi-Square	266, 8 df	617, 8 df

Two models are then built and tested. The first model uses Altman's 5 ratios [8] as predictors while the second uses log-modulus of 5 accounts selected by the algorithm among the whole set. Test-set results for models using ratios and log-modulus are compared in Table I.

As mentioned, bankruptcy prediction is the sole case of successful modelling of financial attributes using ratios. This is due to the fact that the relationship is strong: along the last centuries, financial reports were perfected so as to highlight solvency problems. Also, Altman uses a small sample (thus limiting variability) and discarded the most notorious outliers. Even so, when the log-modulus methodology is used, performance improves and the proportion of explained variability (Cox and Snell Pseudo R-Square), as well as the overall significance of the model (Chi-Square), both increase markedly.

Log-modulus and coefficients in the model are:

Cash and Short Term Investments	+2,473
Total Liabilities	-3,532
Retained Earnings	+0,222
Tax Expense	+0,375
Cash-Flow from Operations	+0,269
Constant term	+7,129

Therefore, Total Liabilities plays the role of a denominator to the other four predictors in internally-generated linear combinations similar to ratios. Coefficients add to -0.193; such variability models the size effect.

The second experiment replicates the fraud predicting model of Beneish [9]. The methodology is similar to the above bankruptcy-prediction case. Data used for learning and testing models consists of a collection of 3,403 "Accounting and Auditing Enforcement Releases" resulting from investigations made by the US Securities and Exchange Commission. The database is from the Centre for Financial Reporting and Management of the Haas School of Business (University of California) [14]. It contains enforcement releases issued between 1976 and 2012 against 1,297 companies which had manipulated financial reports. After removing cases for which no detailed financial data is available, the database contains 1,152 releases. Manipulated reports from the same company in different years are not removed from the sample. Enron, for instance, was the object of 6 releases and all of them are included. Two random samples of nearly 550 different cases each are then drawn. The two samples contain companies listed in US exchanges and which are present in the Standard and Poor's "Compustat" database. They span the period 1976-2008. All sizes and all GICS groups are significantly represented. The two samples are matched with an equal number of reports from companies, which are neither the object of releases throughout the period nor bankrupt in the same year. Pairing is based on the GICS group, on size decile and on year. Amongst reports from companies fulfilling the pairing criteria, randomly selected cases for matching are made unavailable for future matching. Although the same case is not used to match more than one release case, cases from the same company in different years are allowed to remain as candidates to matching. Matched samples have nearly 1,100 cases each. One of the two samples, always the same, is used to build models and the other to test performance of models. Due to missing observations, the size of samples available for model-building and model-testing is less than 1,100 cases:

Learning set: non-fraud cases	335 (45.7%)
Learning set: fraud cases	398 (54.2%)
Test set: non-fraud cases (N)	353 (46.2%)
Test set: fraud cases (P)	411 (53.8%)

Two models are then built and tested. One of the models uses the 8 Beneish ratios [9] while the other uses 8 log-modulus selected by the algorithm. Since, in this case, some Beneish ratio components refer to the previous period, log-modulus are also allowed to express changes in relation to the previous period as in (6) and the algorithm has selected two such changes. Test-set results for models using ratios and log-modulus are compared in Table II.

Performance observed in the model using ratios agrees with that reported in the literature. The model using log-modulus shows a substantial increase in out-of-sample performance. Besides a clearly lower performance, the model using ratios introduces imbalance in the recognition of classes: misclassification in non-fraudulent cases is significantly higher than in fraudulent cases. It is also worth noting the proportion of explained variability (Cox and Snell Pseudo R-Square) and the Chi-Square of the model, which are less than half of that in the log-modulus model. Clearly, the latter fully uses the available variability while the former only uses a limited portion of it.

The third and last experiment involves the prediction of the sign of unexpected changes in Earnings per Share (EPS) one year ahead [11]. The characteristic features of this experiment are the large number of available cases (unexpected Earnings changes one year ahead can be estimated from the database), a weak relationship, indeed the weakest of the three relationships modelled and the absence of matching. The emphasis is placed on comparing the effect of unbalanced samples.

After estimating the classes to be predicted, a number of records is put aside, namely cases with missing values in the predicted dichotomous variable or in predictors. A total of nearly 140,000 cases remain, where some 90,000 are

 ISBN: 978-1-61208-477-0

TABLE III. INCREASE IN EPS PREDICTION.

EPS predicting models	Ratios	Log-modulus
EPS non-increases correct (TN)	42,006 (97.4%)	35,783 (85.7%)
EPS non-increases incorrect (FP)	1,101 (2.6%)	5,967 (14.3%)
EPS increases correct (TP)	4,725 (20.5%)	16,153 (70.8%)
EPS increases incorrect (FN)	18,378 (79.5%)	6,658 (29.2%)
Precision: TP / (TP + FP)	81.1%	73.0%
Pseudo R-Square	0.061	0.342
Chi-Square	4,191, 8 df	27,263, 8 df

non-increases and 50,000 are increases. Methodology has been detailed in previous experiments. The final number of cases in the learning- and test-set is:

Learning set: EPS non-increases 43,242 (64.7%)
Learning set: EPS increases 23,560 (35.3%)
Test set: EPS non-increases (N) 43,107 (65.1%)
Test set: EPS increases (P) 23,103 (34.9%)

Class proportions are significantly imbalanced; both the modelling process and the interpretation of results should reflect such imbalance [15].

From these samples, two models are built and tested. One of the models uses 8 ratios from previous authors [11] and the other uses a set of 8 log-modulus selected by the algorithm. Since, in this case too, some of the ratio components refer to the previous period, log-modulus are also allowed to express changes as in (6) and the algorithm has indeed selected two such changes. Test-set results for models using ratios and log-modulus are compared in Table III.

In this case, classification results should be interpreted in the light of the initial imbalance of classes in the training-set [15], which is 15.1%. For example, a classification accuracy of 70.6%, obtained from an initial imbalance of 15.1% means a gain, in relation to a classification made at random (without any previous information) of just $5.5\% = 70.6\% - (50\% + 15.1\%)$.

Ratios lead to an extremely small percentage of false-positives while the percentage of false-negatives is very high. The model is almost blind to unexpected increases in EPS while recognising decrease very sharply. Therefore, similarly to the previous experiment, the model based on ratios tends to amplify class imbalances.

The examination of the overall significance of the model and the proportion of explained variability shows conclusively that ratios fail to use much variability, which is clearly useful for the modelling of the relationship. This may also explain their notorious inability to produce balanced models: wherever there is neglected variability there is a bias.

V. CONCLUSION

Till the present day, effective KDD of financial reports has proved to be an elusive goal except in the case of bankruptcy prediction, just one of the many attributes involved in Financial Analysis. Log-modulus, not requiring previous knowledge while apportioning all the available variability, may overcome this stalemate.

Predictive models based on ratios incorporate knowledge from the analyst, who is required to select appropriate ratios capable of apportioning information needed to recognise specific attributes. Therefore, the modelling process is not fully automated. Log-modulus, by contrast, allow full KDD since the algorithm generates internal representations similar to ratios. It was shown that the modelling algorithm builds linear combinations of predictors able to unveil financial attributes, such as Solvency or Profitability.

It was also shown that the proposed methodology circumvents most of the difficulties associated with ratios when used as predictors in statistical models, namely the curtailing of variability apportioned by ratio components and the generation of missing cases. Finally, the use of controlled experiments has demonstrated that the log-modulus, agreeing with the statistical characteristics of data being modelled, perform better than ratios, delivering more accurate, robust and balanced models.

ACKNOWLEDGMENT

This research is sponsored by the Foundation for the Development of Science and Technology of Macau (FDCT), China, under the project number 044/2014/A1.

REFERENCES

[1] D. Bellovary and M. Akers, "A Review of Bankruptcy Prediction Studies: 1930-present," Journal of Financial Education, vol. 33, pp. 1–42, 2007.

[2] P. Dechow, C. Larson and R. Sloan, "Predicting Material Accounting Misstatements," Contemporary Accounting Research, vol. 28, no. 1, pp. 17–82, 2011.

[3] J. John and N. Draper, "An Alternative Family of Transformations," Journal of the Royal Statistical Society, Series C (Applied Statistics), vol. 29, no. 2, pp. 190–197, 1980.

[4] S. Huang, R. Tsaih and F. Yu, "Topological Pattern Discovery and Feature Extraction for Fraudulent Financial Reporting," Expert Systems with Applications, vol. 41, no. 9, pp. 4360–4372, 2014.

[5] S. McLeay and D. Trigueiros, "Proportionate Growth and the Theoretical Foundations of Financial Ratios," Abacus, vol. XXXVIII, no. 3, pp. 297–316, 2002.

[6] W. Beaver, "Financial Ratios as Predictors of Failure," Journal of Accounting Research, Suplement. Empirical Research in Accounting: Select Studies, vol. 4, pp. 71–127, 1966.

[7] D. Trigueiros, "Incorporating Complementary Ratios in the Analysis of Financial Statements," Accounting, Management and Information Technologies, vol. 4, no. 3, pp. 149–162, 1994.

[8] E. Altman, Corporate Financial Distress. Wiley (New York), 1983.

[9] M. Beneish, "The Detection of Earnings Manipulation," Financial Analysts Journal, vol. 55, no. 5, pp. 24–36, 1999.

[10] A. Sharma and P. Panigrahi, "A Review of Financial Accounting Fraud Detection based on Data Mining Techniques," International Journal of Computer Applications, vol. 39, no. 1, pp. 37–47, 2012.

[11] J. Ou and S. Penman, "Financial Statement Analysis and the Prediction of Stock Returns," Journal of Accounting and Economics, vol. 11, no. 4, pp. 295–329, 1989.

[12] J. Ou, "The Information Content of Non-Earnings Accounting Numbers as Earnings Predictors," Journal of Accounting Research, vol. 28, no. 1, pp. 144–163, 1990.

[13] url: http://lopucki.law.ucla.edu/ Retrieved: Nov. 2015

[14] url: http://groups.haas.berkeley.edu/accounting/faculty/aaerdataset/ Retrieved: Nov. 2015

[15] N. Chawla, "Data Mining for Imbalanced Datasets: an Overview," in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds. Springer US, 2005, pp. 853–867.

 ISBN: 978-1-61208-477-0

The Mining and Analysis of Data with Mixed Attribute Types

Ed Wakelam, Neil Davey, Yi Sun, Amanda Jefferies, Parimala Alva, Alex Hocking
School of Computer Science
University of Hertfordshire
Hatfield, UK
e-mail: {e.wakelam, n.davey, y.2.sun, a.l.jefferies, p.alva, a.hocking3}@herts.ac.uk

Abstract— **Mining and analysis of large data sets has become a major contributor to the exploitation of Artificial Intelligence in a wide range of real life challenges, including education, business intelligence and research. In the field of education, the mining, extraction and exploitation of useful information and patterns from student data provides lecturers, trainers and organisations with the potential to tailor learning paths and materials to maximize teaching efficiency and to predict and influence student success rates. Progress in this important area of student data analytics can provide useful techniques for exploitation in the development of adaptive learning systems. Student data often includes a combination of nominal and numeric data. A large variety of techniques are available to analyse numeric data, however there are fewer techniques applicable to nominal data. In this paper, we summarise our progress in applying a combination of what we believe to be a novel technique to analyse nominal data by making a systematic comparison of data pairs, followed by numeric data analysis, providing the opportunity to focus on promising correlations for deeper analysis.**

Keywords-Data Mining; Educational Data Mining; Data Analytics; Numeric, Nominal Data Analysis; Dimensionality reduction; Knowledge Extraction.

I. INTRODUCTION

We are initially investigating the potential to apply Artificial Intelligence (AI) techniques to improve e-learning systems in both educational and business settings [1], In particular, we are focussing upon how learning systems can be designed to adapt to individual students during the learning activity. This adaptability would enable the e-learning system to monitor and adjust the teaching based upon a wide variety of analyses of the knowledge and performance of the student. In order to achieve this, we are investigating how student attributes may be analysed and deployed.

Our first steps have been to perform a variety of analyses on open source published student data [2] in order to identify factors which correlate with student performance [3]. Significant advances in the field of data mining [4] are providing opportunities for tools to be deployed in analysing education data [5]. There have also been continued developments in Machine Learning (ML), which aims to determine how to perform important tasks by generalizing from examples [6].

These results may then be used to improve the design of adaptive learning systems [7] using contemporary AI techniques.

In section II, we discuss each of the types of student features relevant to our research: Categorical, comprising Nominal and Ordinal, and Measurement (Quantitative). Section III introduces the open source student data set which we have used to explore applicable analysis techniques. In section IV, we describe our experimental analysis of this data, summarising our results in section V. Finally, we discuss our conclusions in section VI including further work already underway and recommendations for future work.

II. EXISTING DATA ANALYSIS TECHNIQUES

A. Categorical Data

- *Nominal Features*

Nominal data is data where the feature values are labels such as male/female or yes/no. There are a number of statistical techniques available to analyse nominal data sets, notably Chi-square and Cramer's V [8]. Each has its own limitations, for example, sensitivity to sample size and a stronger than justified evidence of correlations [9].

In the case of nominal data, it is not possible to compare attributes directly in order to search for correlations. However, we can compare the correspondence between groupings of attributes and we have explored the use of what we believe to be a novel technique to do so. In this case, we have chosen to compare correlations between pairs of attributes [10]. Future work is underway to apply alternative nominal data analysis techniques to our data in order to compare our results and to identify the strengths and weaknesses of our technique.

- *Ordinal Features*

Ordinal data is a type of categorical data in which order is important. The originators of our data set do not categorise any of the student data captured in their study as ordinal.

B. Measurement (Quantitative) Data

There are a variety of statistical techniques available to analyse quantitative (numeric) data sets. In this case we have selected to use Principal Components Analysis (PCA) to reduce the dimensionality of our data and Growing Neural Gas (GNG) to identify potentially interesting clusters of data. GNG [11] has been successfully used to identify clusters in data for many applications such as the analysis of Hubble Space Telescope images [12] and automatic landmark

 ISBN: 978-1-61208-477-0

extraction in images [13]. PCA and GNG have also been successfully combined for intrusion detection [14].

III. PORTUGUESE STUDENT DATA SET

In order to investigate the predictive accuracy of student achievement data was taken from a set of students from a Portuguese study [15]. This data consists of information taken from two Portuguese secondary schools and each student has 33 attributes. The data includes three labels: first period grade, second period grade and final grade. The subjects are Mathematics (395 students) and Portuguese Language (649 students) and the data was collected during the 2005-2006 academic year. The attributes comprise 16 numeric (including the labels: first period, second period and final performance grades) and 17 nominal (Tables I and II).

TABLE I. EXAMPLES OF THE NUMERIC ATTRIBUTES

Identifier	Description
Age	Student's age (numeric: from 15 to 22)
Absences	Number of school absences (numeric: from 0 to 93)
Studytime	Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

TABLE II. EXAMPLES OF THE NOMINAL ATTRIBUTES

Identifier	Description
Gender	Student's gender (binary: "F" - female or "M" - male)
Mjob	Mother's job (nominal: "teacher", "health" care related, civil "services" (e.g., admin or police), "at_home" or "other")
Romantic	With a romantic relationship (binary: yes or no)

For consistency we have adopted the original attribute types as used in the Portuguese study, although there are a small number of the attributes defined as numeric which could be considered as ordinal.

IV. EXPERIMENTAL ANALYSIS

A. Analysis of Nominal Data

Our method is to compare the correspondence between pairs of our nominal data attributes. To illustrate, the technique, here is a worked example of a data set of 4 students, each with 2 nominal attributes (Table III).

TABLE III. EXAMPLE DATA SET

Student	Attribute 1 (a1)	Attribute 2 (a2)
s1	p	x
s2	p	y
s3	q	z
s4	p	y

After setting a counter to zero we compare every possible pairing of student attribute values in the attribute 1 column of Table III with the corresponding pair in the attribute 2 column. If the selected pair from attribute 1 have the same value and the corresponding pair from attribute 2 also have the same value then we increment the counter by 1. Similarly if they both have different values then we increment the counter by 1. Otherwise, we decrement the counter by 1 (see Table IV).

So, for example, looking at step 1 below, the values of attribute 1 are both "p" (i.e., the same), whereas the values of attribute 2 are "x" and "y" (i.e., different), so we decrement the counter by 1. However, looking at step 2, the values of attribute 1 are "p" and "q" (different), and the values of attribute 2 are "x" and "z" (different), so we increment the counter by 1.

TABLE IV. STEP BY STEP PROCESS

Step	Student pairing	a1	a2	Score	Cumulative counter
1	(s1 s2)	(p p)	(x y)	-1	-1
2	(s1 s3)	(p q)	(x z)	+1	0
3	(s1 s4)	(p p)	(x y)	-1	-1
4	(s2 s3)	(p q)	(y z)	+1	0
5	(s2 s4)	(p p)	(y y)	+1	1
6	(s3 s4)	(q p)	(z y)	+1	2

We repeat this process for all combinations of attribute values and the resultant counter totals are used to populate a correlation matrix. This is done by inserting the counter total into the correlation matrix cell which corresponds to the respective attribute. Obviously, each attribute fully correlates with itself resulting in identical values across the matrix diagonal. We normalise our resulting matrix by dividing all entries by this value to keep all correlation matrix values between -1 and +1 (see Table V).

TABLE V. NORMALISED CORRELATION MATRIX FOR ILLUSTRATIVE EXAMPLE 1

	a1	a2
a1	1	$^1/_3$
a2	$^1/_3$	1

Positive values represent positive correlations between the respective attributes, negative values represent negative correlations and the magnitude of the value represents the strength of the correlation.

For example, where there are a high proportion of student pairs where the corresponding attributes, such as Mother's job and gender are correspondingly the same or different this will result in a relatively higher correlation value (for example, 1/3 in Table V) between the two attributes.

For each attribute, we evaluate its correlation with all other attributes and find the mean value over all these correlations. As a first indicator of interesting attributes, particular attention was paid to those correlations where the magnitude of the mean value was high in comparison to the mean values of other attributes. Those correlations where the

 ISBN: 978-1-61208-477-0

magnitude was above the mean for that attribute then provided additional correlations for consideration.

We applied the technique to each of the Mathematics and Portuguese language data sets in turn. For each data set, we were then able to identify those pairs of attributes that were most strongly correlated – whether positively or negatively. This enabled us to consider the potential influences on student behaviours.

We were also able to compare the correlations in the Mathematics data set with those in the Portuguese language data set.

Using the correlation matrix generated by this technique we then produced corresponding PC1 v PC2 scatter plots for each of our Mathematics and Portuguese Language student data sets in order to visualize potential clusters for future analysis and comparison with any clusters identified in our numeric data. In order to visualize and more easily identify potential clusters we produced a PCA scatter plot for each of the four final grade intervals (using final grades 0-5, 6-10, 11-15, 16-20 as our labels) for each student data set.

B. *Analysis of Measurement Data*

After normalisation of the Mathematics and Portuguese Language student numeric data sets, respectively (by subtracting the mean and dividing by the standard deviation) we performed a linear Principal Component Analysis (PCA), plotting each of the leading three principle components, PC1 v PC2, PC2 v PC3, PC1 v PC3. In each Figure, the amount of variance accounted for by the respective principal components is reported. For example, in Figure 1 PC1 and PC2 account for 26% of the total information in the data.

In each case a visual inspection suggested possible clusters. In order to try and identify these clusters we applied GNG, with key parameters set to 50 training runs and a maximum of 200 nodes. This technique [16] identified a small number of clusters and their respective centroids as well as allowing us to identify the actual students in each cluster.

V. RESULTS

We are looking to identify interesting correlations in our student data attributes, providing the opportunity to focus on promising correlations for deeper analysis.

A. *Nominal data*

- *Mathematics students*

The top and bottom three cross-correlating attributes ranked by highest and lowest mean value are shown in Tables VI and VII respectively.

TABLE VI. HIGHEST MEAN VALUE MATHEMATICS STUDENT ATTRIBUTES

Attribute	Mean value
Higher education wish	0.23
School	0.19
Parent cohabitation	0.18

TABLE VII. LOWEST MEAN VALUE MATHEMATICS STUDENT ATTRIBUTES

Attribute	Mean value
Paid tutor	0.008
Gender	0.006
Extra-curricular activity	0.003

Our results show potential correlations may exist between the student's wish to take Higher Education and other nominal attributes - the school attended and parent cohabitation status, followed by receipt of extra educational support, Mother's job, access to the internet, the reason for choice of school and nursery school attendance.

Mother's job also shows potential correlations with other factors, including the wish for higher education, parent cohabitation, school attended, educational support and choice of school.

Paid extra tuition does not correlate strongly with other factors, even parent's jobs, which we might have expected. This is also true for students receiving educational support from within the family. However, future analyses may show that such extra tuition correlates with student performance measured by their grades.

Internet access also shows potential correlations with a number of factors, including the wish for higher education, school attended, parent cohabitation, address, the level of educational support by the school and Mother's job.

Factors which show very low correlations with others are the level of extra-curricular activities, whether the student was male or female and paid tutoring, followed by romantic relationships, Father's job, and family size.

- *Portuguese Language students*

The top and bottom three cross-correlating attributes ranked by highest and lowest mean value are shown in Tables VIII and IX respectively.

TABLE VIII. HIGHEST MEAN VALUE PORTUGUESE LANGUAGE STUDENT ATTRIBUTES

Attribute	Mean value
Paid tutor	0.20
Higher Education wish	0.18
Parent cohabitation	0.16

TABLE IX. LOWEST MEAN VALUE PORTUGUESE LANGUAGE STUDENT ATTRIBUTES

Attribute	Mean value
Family education support	0.02
Gender	0.01
Extra-curricular activity	0.003

Our results show potential correlations may exist between paid tutoring, the student's wish to take higher education and parent cohabitation followed by educational support and Mother's job.

 ISBN: 978-1-61208-477-0

Paid extra tuition shows potential correlations with a number of other factors including the level of educational support, the wish for higher education, parent cohabitation, and Mother's job. This is also true for extra educational support provided by the school, correlating with the use of paid tutors, parent cohabitation, and Mother's job.

Mother's job shows potential correlation with the use of paid tutoring, educational support, parent cohabitation and attendance at a nursery school.

Internet access only correlated modestly with other factors for Portuguese language students.

Factors which show very low correlations with others are the level of extra-curricular activities, student gender and family educational support, followed by romantic interest, guardian, Father's job and school attended.

- *Comparisons between Mathematics and Portuguese Language analysis results*

The wish to take higher education shows potential correlation with Mother's job, cohabitation status and receipt of extra educational support for both sets of students.

In both cases Mother's job correlates with other factors. In contrast, Father's job, along with romantic relationships and extra-curricular activities shows very low correlations with other factors in both sets.

Additional educational support provided by the school also shows potential correlation with a number of other factors in both sets.

In comparison with Portuguese language students, paid extra tuition in the case of Mathematics students does not correlate strongly with other factors.

Interestingly, gender, considered to be an influential factor, does not correlate well with other attributes in either set.

In the case of Mathematics students, internet access shows potential correlations with a number of factors, such as the wish to take further education, school attended, and parent cohabitation. However, in the case of Portuguese Language students, internet access shows only modest correlations.

- *Principal Component Analysis*

As described in section 1, above, a PCA projection will allow visualization of multi-dimensional data in a two dimensional representation. For each data set the initial PCA plot including all final grades proved too challenging to visualize and so we produced four plots, one for each of the four final grade intervals. We have included one example from each data set. Principle component analysis of our Mathematics and Portuguese Language student data shows no evidence of potential clustering.

For example, a PC1 v PC2 nominal data plot of Mathematics students' achieving final grades of between 11 and 15 (Figure 1).

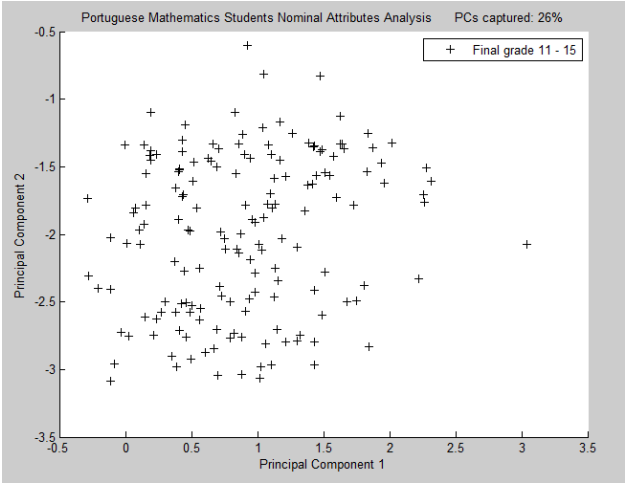

Figure 1. Mathematics nominal data PC1 v PC2 Final Grades 11-15

A further example shows a PC1 v PC2 nominal data plot of Portuguese language students' achieving grades of between 11 and 15 (Figure 2). This data plot appears to exhibit a lower boundary delineation which we believe to be a result of a predominance of very narrow variances in the attribute values in this particular data set.

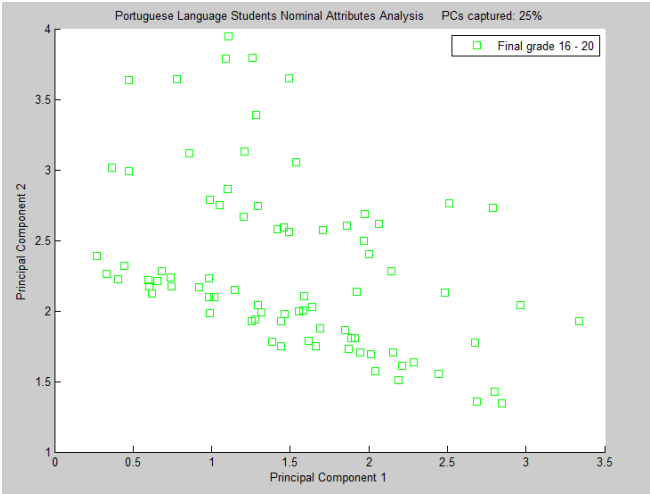

Figure 2. Portuguese Lang nominal data PC1 v PC2 Final Grades 16-20

B. *Measurement data*

- Mathematics students

GNG identified modest clustering in each of the PC1, PC2, PC3 comparisons For example, in Figure 3 we can see that three clusters have been identified. The centroids are shown in red and in each case the students in each cluster are identified in order to for look for potential correlations with the results of our nominal data analysis.

 ISBN: 978-1-61208-477-0

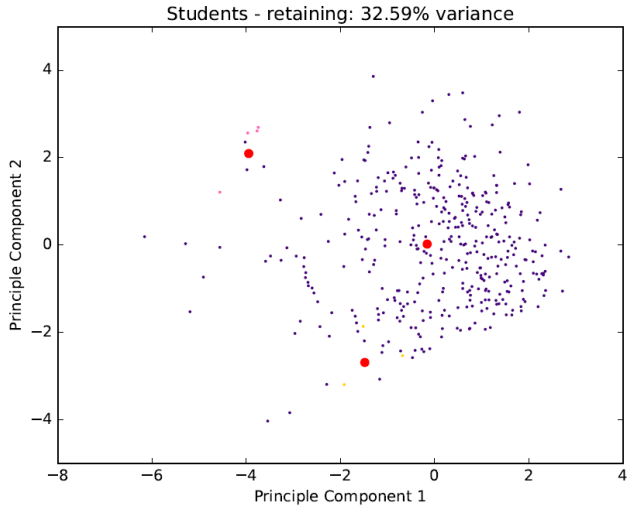

Figure 3. Mathematics students numeric data PC1 v PC2 scatter plot

- Portuguese Language students

GNG did not identify useful clustering in either of the PC1, PC2, PC3 comparisons. In all cases only one cluster was identified, for example, in Figure 4. As above, the centroids are shown in red.

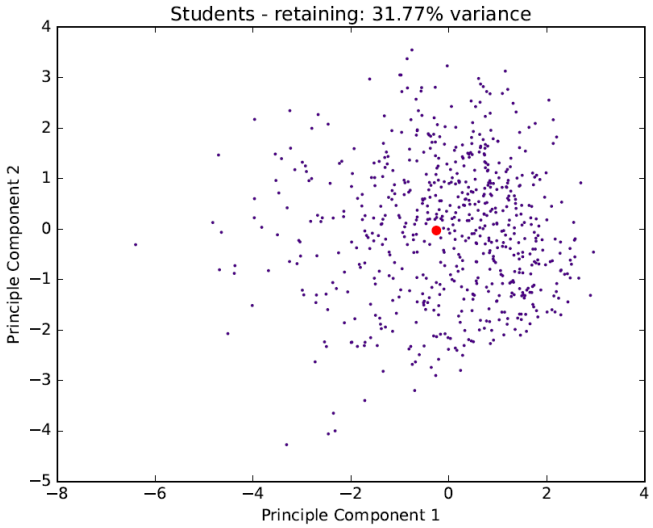

Figure 4. Portuguese Lang. students numeric data PC1 v PC2 scatter plot

We repeated the GNG analysis, adjusting the key parameters, increasing the number of training runs from 50 to 100 and maximum nodes from 200 to 600. However, this did not result in improvement. Further work is underway to identify alternative techniques to identify potential clustering in the Portuguese Language student numeric data, such as Curvilinear Component Analysis (CCA).

VI. CONCLUSION AND NEXT STEPS

In this paper, we have taken the first steps in exploring a mixed attribute type (numeric and nominal) data set provided by real student data with the objective of identifying useful potential correlations between attributes.

We have applied a novel approach to the analysis of the nominal data, comparing the correspondence between pairs of nominal attributes.

We then investigated if the analysis would identify interesting information in the data set, which to some extent it did. Our PCA plot of the Mathematics nominal data showed no evidence of clustering. Further work is underway to apply a non-linear visualization method in order to investigate potential clustering.

We then applied numeric data analysis techniques to identify clustering and potential correlations in our numeric attributes identifying some potentially interesting patterns.

In the case of our Mathematics student data using Principle Component Analysis followed by the GNG technique we were able to identify some clustering of the data, however the corresponding analysis of our Portuguese Language student data did not identify useful clusters.

Further work is underway to analyse and make comparisons between the numeric and nominal data sets to identify correlations, and subsequently to use these analyses to develop methods to predict student performance.

From the educational perspective, this would then allow us to perform follow up analyses on the extent to which different attributes can influence student achievement.

Future work includes the application of alternative nominal data analysis techniques to our nominal student data in order to compare the results and evaluate the advantages and disadvantages of these techniques in comparison with those of the technique deployed.

The novel nominal data analysis technique may provide a useful additional tool in the analysis of nominal data. We have shared the technique and corresponding MATLAB code with colleague researchers to gain further feedback on its usage and ideas on how to increase the sophistication of the method. Please contact us for a copy of the code.

REFERENCES

[1] E. Wakelam, A. Jefferies, N. Davey and Y. Sun, “The potential for using artificial intelligence techniques to improve e-Learning systems”, 2015.

[2] P. Cortez and A. Silva, “Using data mining to predict secondary school student performance”, in A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008), Porto, Portugal, April, 2008, pp. 5-12.https://archive.ics.uci.edu/ml/datasets/Student+Performance#

[3] V. Ramesh, P. Parkavi and K. Ramar, "Predicting student performance: a statistical and data mining approach". International journal of computer applications 63, no. 8, 2013, pp 0975 – 8887.

[4] P. Bhalchandra et al, "Prognostication of student’s performance: An hierarchical clustering strategy for educational dataset." In Computational Intelligence in Data Mining—Volume 1, Springer India, 2016, pp. 149-157.

[5] D. Fatima, S. Fatima, “A survey on research work in educational data mining.” IOSR Journal of Computer Engineering (IOSR-JCE), 17, 2015.

[6] T. Hastie, R. Tibshirani, J. Friedman and J. Franklin, “The elements of statistical learning: data mining, inference and prediction.” The Mathematical Intelligencer. doi:10.1007/BF02985802, 2005.

 ISBN: 978-1-61208-477-0

[7] D. Clow, "An overview of learning analytics." *Teaching in Higher Education*, 2013, pp. 683-695.

[8] A. Agresti, "Categorical data analysis." Vol. 996. New York: John Wiley & Sons, 1996.

[9] P. M. Bentler,, and D. G. Bonett, "Significance tests and goodness of fit in the analysis of covariance structures." Psychological bulletin, 88(3), 588, 1980.

[10] P. Ashrafi, "Predicting the absorption rate of chemicals through mammalian skin using Machine Learning algorithms." (Ph.D. unpublished). University of Hertfordshire, 2016.

[11] B. Fritzke, "A growing neural gas network learns topologies." Advances in neural information processing systems 7, 1995, pp. 625-632.

[12] A. Hocking, J. Geach, Y. Sun, N. Davey, N. Hine, "Unsupervised image analysis & galaxy categorisation in multi-wavelength Hubble space telescope images", Proceedings of the ECMLPKDD 2015 Doctoral Consortium (ECML 2015), 2015, pp. 105-114.

[13] E. Fatemizadeh, C. Lucas and H. Soltanian-Zadeh, "Automatic landmark extraction from image data using modified growing neural gas network." Information Technology in Biomedicine, IEEE Transactions on 7, no. 2, 77-85, 2003.

[14] G. Liu, and X. Wang, "An integrated intrusion detection system by using multiple neural networks." IEEE Conference on Cybernetics and Intelligent Systems, 2008, pp. 22-27.

[15] P. Cortez, and A. Silva, "Using data mining to predict secondary school student performance." In the Proceedings of 5th Annual Future Business Technology Conference, 2008, pp. 5–12.

[16] A. Parimala, "Using machine learning and computer simulations to analyse neuronal activity in the cerebellar nuclei during absence epilepsy." (Ph.D. unpublished). University of Hertfordshire, 2015.

 ISBN: 978-1-61208-477-0

Practical Application of the Data Preprocessing Method for Kohonen Neural Networks in Pattern Recognition Tasks

El Khatir Haimoudi[(1)], Loubna Cherrat[(2)]
[(1)] Plury-disciplinary Laboratory
FP of Larache, Abdelmalek Essaâdi University
Larache, Morocco
e-mail: helkhatir@gmail.com

Otman Abdoun[(1)], Mostafa Ezziyyani[(2)]
[(2)]Mathematics and Application Laboratory
FST of Tangier, Abdelmalek Essaâdi University
Tangier, Morocco
e-mail: cherrat@gmail.com, m.ezziyyani@fstt.ac.ma

***Abstract* — Self-Organizing Map (SOM) is a very effective solution for solving pattern recognition problems. However, some ambiguities appear during learning process with the existence of linear patterns in the learning data, in this case, the learning process lasts for a long time and the network produces irrelevant results. The work provides the resolution of the detected problem and the application of the SOM for the pattern recognition. To achieve our objective and minimize the learning time, a SOM improved model has been developed. This model uses a special block able to filter the input data and reduce the size of the learning multitude. The presented experimental test results in this work show that the improved model exceeds the standard model in terms of the recognition results accuracy and the learning time. The results obtained in this work encouraged us to think about using the improved model to develop a smart approach (SmartMaps) of Geographic Information Systems (GIS).**

Keywords- Pattern recognitions; Artificial Neural Network; self-organizing map; preliminary processing of input vectors; Data visualising; principal component analysis; power iteration algorithm.

I. INTRODUCTION

The new information technologies offer great opportunities for human activity in different areas. However, the important element of their evolution is not only the extent increase of computer technology's capacity, but also its intellectualization by the creation of new intelligent systems in the form of software or hardware models. These systems must be equipped with intellectual abilities comparable to those of humans. Their use is to solve very complex problems for classical information systems, such as the recognition, diagnosis and prediction. Recently systems based on Artificial Neural Networks (ANN) are widely used to create these systems [1]-[3]. The essential advantage of ANN is a functional similarity to biological neural networks and the universality for solving a wide range of tasks. There are a variety of architecture and learning methods for different ANN models. Currently the models based in competitive learning algorithms, like Self-Organizing Maps (SOM) and counter-propagation network are widely used in pattern recognition tasks [4]-[8]. An important and useful feature of SOM is the ability to visualize multi-parameter objects in a one-dimensional or two-dimensional space [10].

However, tests show that the use of SOM as it stands, does not give relevant results, as the learning algorithm requires normalizing input data. The consequence of this operation is the loss of some information about the initial lengths of objects, and the ratio between the absolute values of input object components. In this case and with the existence of linear patterns, the learning process takes a long time and SOM produce irrelevant results [11]. In this work, we realized a new model of the SOM which provides the introduction of a preprocessing block and data optimization. Pre-treatment process is based on a method that combines two well-known and approved algorithms: Analysis Principal Components and Iterated Power. The both map models (standard and improved) are applied to solve a task of pattern recognition; the task objective is to visualize geographical information for the African continent countries. In this work we present also the results of this practical application, and the detailed analysis of their comparison.

The article consists of five sections. In the introduction, we show the importance of intelligent systems, their application area and new means for their development. We also describe the purpose of the work, the solved problem and the future perspective. The second section comprises the description of the pattern recognition task, citing the classical and modern methods used to solve this problem. So we give details of the learning algorithm of the SOM, and the principles of its application in this domain, including the ambiguities detected in this model of ANN and possible solutions. In the third section we present the algorithm of preliminary data processing and optimization, with argument and explanation of different steps of its implementation and the benefits obtained from its application with SOM. The fourth part provides the practical application of the two

 ISBN: 978-1-61208-477-0

models for the recognition of the African continent countries. In this section we described: the approach we have followed to solve this task, the means and tools provided by the application developed for its use, as well as the results obtained and details of their analysis. As a conclusion, we mention the important moments concerning the problems encountered during the SOM applications in pattern recognition, the contribution of the proposed solution and our perceptive.

II. THE APPLICATION FEATURES OF NEURAL NETWORKS IN PATTERN RECOGNITION TASKS

The task of pattern recognition can be considered as a combination of two related subtasks: classification and clustering. The classification task is to determine the belonging of the input pattern to one of predefined classes [11]. This classification type is used for the recognition of handwritten texts, the lyrics and ECG signals. During clustering, the learning algorithm is based only on the input data without desired output. In this case, the learning process will try to identify the similarity between patterns, and similar objects will be brought to the same category (cluster); the proximity is often understood in the sense of the Euclidean metric [12] [13]. This problem occurs during the extraction of data, the study of their properties and compression. Therefore, two paradigms are identified in the problems of pattern recognition: recognition supervised based on the classification technique and unsupervised recognition where we use the clustering technique.

The classical model is based on supervised recognition methods; these are the probabilistic methods, in particular, the method based on Bayes formula, adapted for manual calculations [14]. The solution rules can be derived as probabilistic identification parameters of belonging of an object to a particular class (Bayesian method), or as a simple analytical function (discriminate analysis method). These methods have certain limitations, such as absence of reliability, because they are based only on the linear rules [15].

The modern recognition methods as neural networks cannot be used without computers. These systems are able to elaborate the classification and clustering rules, and to be used to develop intelligent systems for a wide use.

A. *The pattern recognition process with the self-organizing map of Kohonen*

Artificial neural networks are widely used for pattern recognition; these systems use specific algorithms for classification and clustering of multi-parameter objects (events, situations, processes). Currently, there are several ANN paradigms that are used in this task. However, the models which are mainly used are the ones using competitive learning methods. In particular, we can cite the SOM [6] and the counter propagation network [7].

The Kohonen network model uses the competitive learning method. This process brings together similar objects in same cluster by reserving the topological relationships in input data [16] [17]. During learning, the neurons compete, and for each group of similar objects, a single winner neuron is defined. The fixed neurons represent the centers of clusters. The metric used in this operation is the Euclidean distance between the synaptic weights vectors, and the input objects vectors.

The learning procedure begins with the normalization of input data and synaptic weights to reduce the learning time [11]. This operation is based on the following algebraic formula:

$$x_i = x_i / \sqrt{\sum_{j=0}^{n-1} x_j^2} \quad (1)$$

Where: x_i – the input object component or the vector of synaptic weights;

n – The number of variables in the vector x.

The main learning algorithm passes successively through a series of iterations, and it relies only on the input data. During the learning process, it attempts to define for each group of similar objects a specific neuron qualified as winner. At the end of this procedure the topologically adjacent neurons, respond to similar input vectors.

To fix the winner's neurons, we use the metric of the Euclidean distance [5] see formula below:

$$k : \|w_k - x\| \le \|w_o - x\| \quad \forall o \quad (2)$$

Subsequently the algorithm performs a correction of synaptic weights to gradually minimize the distance between the winning neurons and the input objects. For this correction we use the following formula [6]:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha_i(t) h(d,t) \cdot \left[y_i - w_{ij}(t) \right] \quad (3)$$

where: y_i - the value of the output neuron i;

w_{ij} (t) and w_{ij} (t +1): the synaptic weights during *t* and (*t* + *1*) iterations.

α_i(t): learning rate, this coefficient can have a value between *0* and *1*, and it is calculated using the following equation:

$$\alpha_i = \alpha_0 e_t^i \quad (4)$$

where: ***i*** is the iteration number;

t is the iteration rate.

h(d, t) : neighborhood function, it is written according to the formula below:

$$h(d,t) = \begin{cases} 0, d \ge \delta(t) \\ e^{(-\frac{d}{2\delta(t)})}, d < \delta(t) \end{cases} \quad (5)$$

$$\delta(t) = \delta_0 e^{(-\frac{t}{\mu})} \quad (6)$$

where: d is the distance between the winner neuron and an x neuron.

$$\mu = \frac{n}{\log_{10}(\delta_0)} \quad (7)$$

where: $\boldsymbol{\delta_0}$ is Constant.
$\boldsymbol{n}$ is Iteration rate.

The learning process will be continued until the stabilization of the SOM, and the results will be presented as a grid of neurons in a two dimensional space.

However, the application of the SOM can give irrelevant results due to the problem of linear dependence [12]. To avoid these constraints we offer the use of an enhanced map model that can well classify data even with the presence of linear patterns. The new model included a pretreatment method and data optimization.

III. THE DATA PRETREATMENT METHODS BASED ON A GEOMETRIC APPROACH

The idea of the proposed method is to use a specific block of data preprocessing. The processing operation uses an algorithm based on two typical methods of data analysis: Principal Component Analysis (PCA) and Iterated Power (IP) [24] [32]. This combination allows filtering data to reduce the dimension of the data table and saving the most informative parameters in each multitude vectors. The new contribution of this block is the elimination of regularity between the vectors components and disappearance of the linear dependence problem, which could prevent this type of ANN to provide accurate and relevant results.

Initially, we assume that the learning data table is composed of $\boldsymbol{n}$ rows and $\boldsymbol{p}$ columns; see Figure 1.

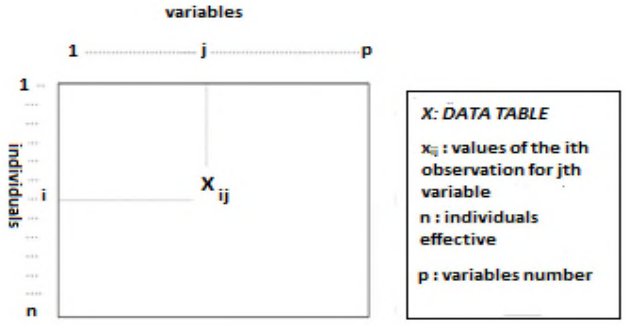

Figure 1. Learning data table

In the first step, the algorithm calculates the vector of main point $\boldsymbol{g}$. This point is the center of the points cloud in a space $\boldsymbol{F}$. See the formula below:

$$g^t = (\bar{x}^1, \ldots, \bar{x}^p) \quad (8)$$

$$\bar{x}^j = \frac{1}{n}\sum_{i=1}^{n} x_i^j \quad (9)$$

At the base of the vector $\boldsymbol{g}$ is calculated the data centered matrix, which is written in terms of $\boldsymbol{X}$ as the following way:

$$Y = X - 1g^t \quad (10)$$

where: g^t is the transposed of g, and the term centered signifies that the means of the variables $\bar{y}^j$ are zero.

The centered data matrix $\boldsymbol{Y}$ is used in this step for calculating the variance-covariance matrix $\boldsymbol{V}$, which is written as a function of Y as follows:

$$V = \frac{1}{n} Y^t Y \quad (11)$$

where: Y^t is the transposed of $\boldsymbol{Y}$.

The V matrix is presented as follows:

$$V = \begin{pmatrix} S_1 & . & . & S_{1p} \\ S_{21} & . & . & . \\ . & . & . & . \\ S_{p1} & . & . & S_p \end{pmatrix}$$

where: $\boldsymbol{S_{KL}}$ is the covariance of the variables k and l, and $\boldsymbol{S_k}$ is the variance of the variable k.

In the last step, in order to develop the correlation matrix $\boldsymbol{R}$ we must calculate the two diagonal matrices $\boldsymbol{D_{1/S^2}}$ and $\boldsymbol{D_{1/S}}$ as a function of $\boldsymbol{V}$ as follows:

$$D_{1/S^2} = \frac{1}{Diag(V)} \quad (12)$$

$$D_{1/S} = \frac{1}{Diag(D_{1/S^2})} \quad (13)$$

The matrix $\boldsymbol{R}$ is composed of linear correlation coefficients between the variables p. It summarizes and shows the structure of linear dependencies between these variables. The matrix is symmetric, and the component values of its diagonal equal $\boldsymbol{1}$. $\boldsymbol{R}$ calculates as a function of $\boldsymbol{V}$ as follows:

$$R = D_{1/S} V D_{1/S} \quad (14)$$

where: $D_{1/s}$ is a diagonal matrix, its diagonal is composed by the values $\frac{1}{s_1}, \ldots \ldots \ldots, \frac{1}{s_p}$.

Now it is the time to apply the iterative power method to search the eigenvectors [32]. These vectors are the rows of the final matrix of input objects $\boldsymbol{M}$.

 ISBN: 978-1-61208-477-0

$$M = \begin{bmatrix} x_{11} & . & . & x_{1l} \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & . & . & x_{nl} \end{bmatrix}$$

The figure below presents the proposed algorithm flowchart.

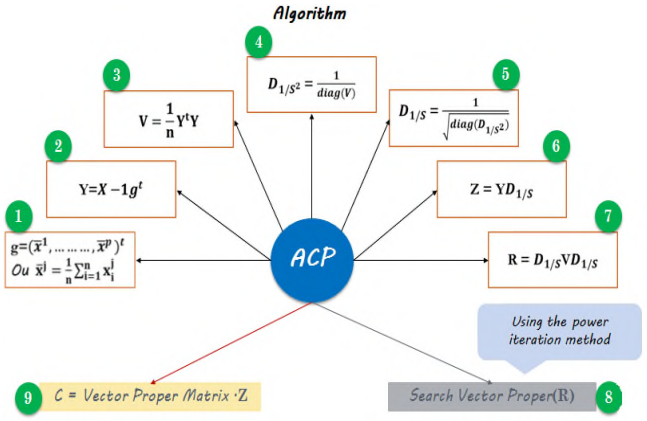

Figure 2. The proposed algorithm flowchart

In this algorithm, the first seven steps allows to calculate the correlation matrix ***R*** using the ACP method. This matrix will be used by the iterated power method to search the eigenvalues and the eigenvectors. The last two steps allows to develop the reduced final matrix based on the IP algorithm.

The new matrix calculated by using the proposed method will present the data source for learning the SOM. The results are displayed and interpreted using grids of neurons in two-dimensional space. The functional structure of the proposed model is shown below in Figure 3.

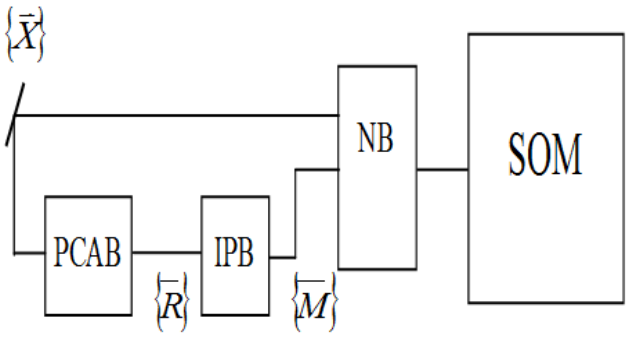

Figure 3. The functional structure of the proposed SOM model

According to the proposed algorithm diagram, and the functional structure schema of the proposed model, we can summarize the learning process of this neural network system in the following steps:

- The treatment of the initial data by the principal component analysis blocks PCAB, to obtain the correlation matrix R.
- At the base of the correlation matrix, the block of the iterated power IP seeks for eigenvectors that constitute the rows of the resulting matrix M of input objects.
- The NB blocks perform the normalization of data matrix M.
- The last step provides the phase of the network learning based on the pattern data calculated in the precedents steps.

IV. OBJECTIVE AND RESEARCH BASE

The objective of research is to establish the advantages and disadvantages of the preprocessing method of realizations based on the geometric approach in practical pattern recognition problems. The research base is the software model of the SOM developed by us using the VP.net programming platform. The application includes both learning algorithms: standard and improved, and it is able to visualize the results as neural grids in a two-dimensional space. The interpretation of learning outcomes is based on the distribution of winner neurons on the map, and the definition of the cluster which they belong.

A. Application tests and results

To study and analyze the proposed method of preliminary data processing for the SOM, we use a typical problem of pattern recognition: the Recognition of the African Continent Countries [33]. The objective of this task is to test the implementation of two SOM models (standard and modified). The recognition will be performed using the SOM instruments and tools, including the possibility of classification and clustering, as well as to view the data on the neurons grid in a two-dimensional space. The application begins with the learning step to prepare the knowledge base necessary for its operation; this database should generate relevant results. For this typical neural network, the learning results evaluation is done by using the maps which visualize the classes and clusters objects (Country). After correct learning, the SOM can be used to build an intelligent Atlas map that is able to give the necessary information about the continent countries. The learning set is composed on 52 vectors, where each one corresponds to a country. Each vector is characterized by 20 parameters (geographic location, language, area, religion, color and flags elements, etc); see Table I.

The developed software works in two modes, and supports both models: standard and improved. The learning results are interpreted by using the two maps (Class map and Clusters map) and textual data. The maps are drawn as rectangles grids, of dimension (N × N), corresponding to the number of output layer neurons. The top left rectangle presents the first neuron. For a better interpretation of the learning results, we use a coloring system where the colored rectangles represent the winner neurons. By click on every

 ISBN: 978-1-61208-477-0

rectangle the application displays the related information. The same principle is used with the map clusters; see Figure 4.

TABLE I. THE DESCRIPTIVE DATA OF THE AFRICAN CONTINENT COUNTRIES

N°	The Country	The Parameters																		
1	Algeria	1	2388	20	8	2	2	0	3	1	1	0	0	1	0	4	1	1	0	3	1
2	Angola	2	1247	7	10	5	0	2	3	1	0	0	1	0	1	4	1	0	1	4	5
3	Benin	1	113	3	3	5	0	0	2	1	1	0	0	0	0	3	1	0	0	3	3
4	Botswana	2	600	1	10	5	0	5	3	0	0	1	0	1	1	6	0	0	0	5	5
5	Burkina	4	274	7	3	5	0	2	3	1	1	0	1	0	0	4	1	0	0	4	3
6	Burundi	2	28	4	10	5	0	0	3	1	1	0	0	1	0	4	3	0	0	1	1
7	Cameroon	1	274	8	3	1	3	0	3	1	1	0	1	0	0	2	1	0	0	3	2
8	Cape Verde Islands	4	4	0	6	0	1	2	5	1	1	0	1	0	1	2	1	0	0	4	3
9	Central African Republic	1	623	2	10	5	1	0	5	1	1	1	1	1	0	2	1	0	0	6	2
10	Chad	1	1284	4	3	5	3	0	3	1	0	1	1	0	0	2	0	0	0	6	4
11	Comoro Islands	2	2	0	3	2	0	0	2	0	1	0	0	1	0	3	4	1	0	3	3
12	Congo	2	342	2	10	5	0	0	3	1	1	0	1	0	0	4	1	0	1	4	4
13	Djibouti	1	22	0	3	2	0	0	4	1	1	1	0	1	0	6	1	0	0	1	3
14	Egypt	1	1001	47	8	2	0	3	4	1	0	0	1	1	1	5	0	0	0	4	5
15	Equatorial Guinea	1	28	0	10	5	0	3	4	1	1	1	0	1	0	3	0	0	0	3	4
16	Ethiopia	1	1222	31	10	1	0	3	3	1	1	0	1	0	0	3	0	0	0	3	4
17	Gabon	2	268	1	10	5	0	3	3	0	1	1	1	0	0	3	0	0	0	3	6
18	Gambia	4	10	1	1	5	0	5	4	1	1	1	0	1	0	4	0	0	0	4	3
19	Ghana	4	239	14	1	5	0	3	4	1	1	0	1	0	1	4	1	0	0	4	3
20	Guinea	4	246	6	3	2	3	0	3	1	1	0	1	0	0	2	0	0	0	4	3
21	Guinea Bissau	4	36	1	6	5	1	2	4	1	1	0	1	0	1	2	1	0	0	4	3
22	Ivory Coast	4	323	7	3	5	3	0	3	1	1	0	0	1	0	1	0	0	0	4	3
23	Kenya	1	583	17	10	5	0	5	4	1	1	0	0	1	1	4	0	0	1	5	3
24	Lesotho	2	30	1	10	5	2	0	4	1	1	1	0	1	0	6	0	0	1	3	6
25	Liberia	4	111	1	10	5	0	11	3	1	0	1	0	1	0	4	1	0	0	6	4
26	Libya	1	1760	3	8	2	0	0	1	0	1	0	0	0	0	3	0	0	0	3	3
27	Malagasy	2	587	9	10	1	1	2	3	1	1	0	0	1	0	4	0	0	0	1	3
28	Malawi	2	118	6	10	5	0	3	3	1	1	0	0	0	1	4	1	0	0	5	3
29	Mali	4	1240	7	3	2	3	0	3	1	1	0	1	0	0	2	0	0	0	3	4
30	Mauritania	4	1031	2	8	2	0	0	2	0	1	0	1	0	0	3	1	1	0	3	3
31	Mauritius	2	2	1	1	4	0	4	4	1	1	1	1	0	0	4	0	0	0	4	3
32	Morocco	4	447	20	8	2	0	0	2	1	1	0	0	0	0	4	1	0	0	4	4

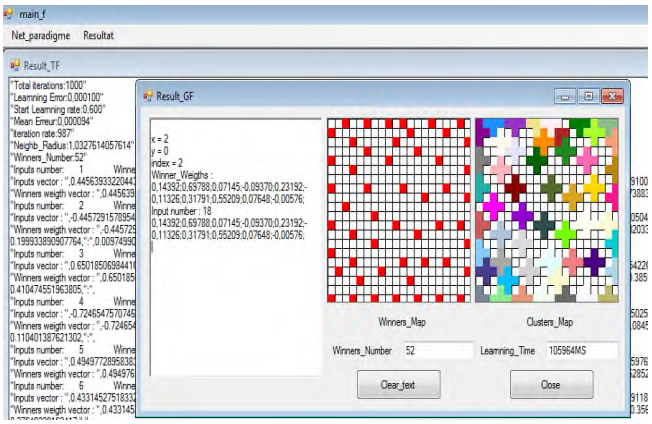

Figure 4. Graphic software interface

B. *Test results analysis*

In this section, we will try to interpret and analyze the learning results, in order to reveal the advantages and disadvantages of each model over the other. We use two specific metrics to compare the studied models: The first one is the percentage of recognition which defines a relationship between the number of winner neurons and the total number of input learning vectors. The second metric represents the learning time.

The research results presented in Figures 5 and 6 show that the standard model has defined 49 winner neurons for the 52 input objects, that present a recognition percentage equivalent to 94.23%, and a learning time that reaches 204660 MS. But the improved model has defined 52 winner neurons for the 52 individuals that present a recognition percentage reaches 100%, and a learning time not exceeding 105964 MS.

These results affirm that the new model exceeds the standard model at the level of the recognition relevance and the learning time. So we can say that with the proposed method of data pretreatment, the map possess new opportunities and able to give good results even with the existence of linear dependency in the learning data.

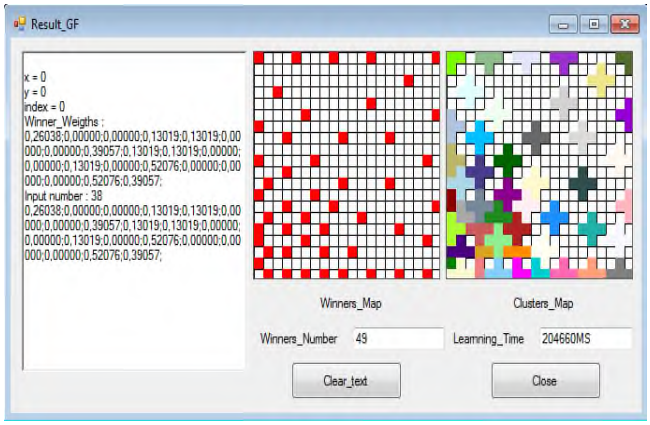

Figure 5. Learning Result recognition of African countries (modified model)

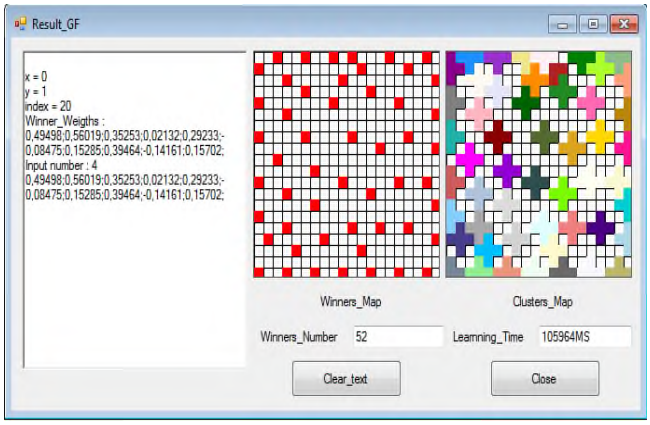

Figure 6. Learning result recognition of African countries (modified model)

The visual analysis of maps shows that for standard model the most of the winner neurons and their clusters are concentrated in the lower left half of the maps, but for the improved model these elements are well dispersed over the map surface. This improvement in the topological presentation of the results for both types of maps (winners

and Clusters) is explained by the change in the learning data structure, including the relationship between objects. This modification is performed by using data pretreatment process.

To show the impact of the input data size on the learning time, and the contribution of the method used in the improved model, the data set has been distributed to groups containing different numbers of individuals going from 5 until 52; see Table 2.

The data in Table II shows that the data pretreatment method has reduced the individual lengths, from 20 to 10 components for each individual. This decrease has allowed to the improved model reduce the learning time compared to the standard model.

TABLE II. LEARNING RESULTS FOR ALL INDIVIDUAL GROUPS

Individual numbers	Standard model				Improved model			
	Component numbers	*winner numbers*	*Iteration rate*	*Learning time in MS*	*Component numbers*	*Winner numbers*	*iteration rate*	*Learning time in MS*
5	100	5	487	8076	50	5	495	5384
10	200	10	612	21446	100	10	633	13880
15	300	15	675	37671	150	15	720	24283
20	400	20	787	52239	200	20	774	32367
25	500	25	784	65214	250	25	767	43168
30	600	30	832	66254	300	30	816	56204
35	700	35	861	106513	350	35	890	68892
40	800	40	877	112596	400	40	900	84489
45	900	45	977	148312	450	45	908	92963
52	1040	49	1000	204660	520	52	987	105964

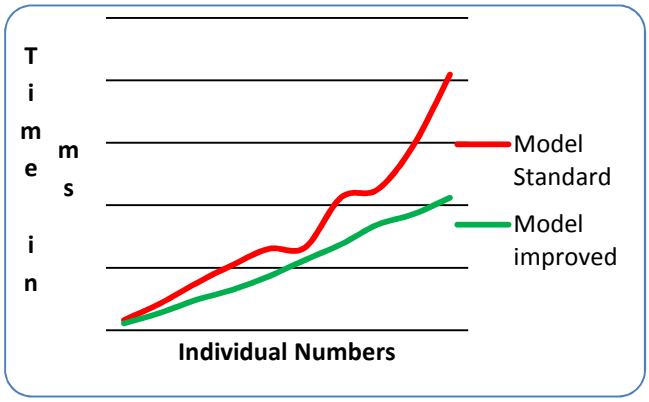

Figure 7. The graphical presentation of the learning results for both SOM models.

The graph in Figure 7 above shows the relationship between the learning time spent by both models, and the number of individuals employed. The graph curves show that the learning time for both models is increasing in parallel with the augmentation of the learning multitude size. Thus, it is observed that the learning process takes less time for the improved model than the standard model. And the difference of learning time between both models is enlarged with growth of the pattern multitudes size.

To sum up, the improved model exceeds the standard by three parameters: The relevance of the results of the recognition, learning speed and the dispersion the winner neurons on the map. The first parameter is justified by the recognition percentage, which reached 100% for the improved model, but only 94.23% for the standard. The second parameter is justified by the learning time that decreases twice using the improved model compared to the standard model. And the latter parameter is justified by equitable dispersion of winner neurons and the clusters on maps (classes and Clusters). So, the results show that the improved model has solved this task better than the standard model.

V. CONCLUSION AND FUTURE WORK

In this work we have tried to improve and implement a type of neural networks in a task of object recognition, called the Self-Organizing Map (SOM). In this work we have justified the choice of the used paradigm, and demonstrated that its direct application does not provide good results. So our objectives were determining the ambiguities and the means of their eliminations. For the first objective, via a theoretical study and experimental tests we have defined the problem that prevents the correct learning of network. To achieve the second objective, we proposed and approved a data pretreatment method, at the basis of which we have developed a new functional structure for the improved model of SOM. The results of the tests show that: The SOM is a reliable and intelligent tool for solving the recognition problems, and the method of preliminary processing of the input data enriches the SOM with new competences. Finally, the improved model exceeds the standard model in the accuracy of the results and the learning time. The obtained results encourage us to improve and apply the ANN in the various domains of human activities. In future work, firstly we will apply the new SOM model on the GIS, and then, the proposed method will be used in order to improve another ANN paradigm.

REFERENCES

[1] Deng, Geng and M.C. Ferris, “Neuro-dynamic programming for fractionated radiotherapy planning,“ Springer Optimization and Its Applications 2008. p. 47–70.

[2] M. Roman, Balabin and I. Ekaterina. Lomakina, “Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies“, J. Chem. Phys 2009. p. 131 (7).

[3] J.H. Frenster, “Neural Networks for Pattern Recognition in Medical Diagnosis“, Annual International Conference in the IEEE Engineering in Medicine and Biology Society 1990. vol. 12, N°. 3. p. 1423-1424.

[4] Shah-Hosseini, “Binary Tree Time Adaptive Self-Organizing Map“, Neurocomputing May 2011. 74 (11). p. 1823–1839.

[5] T. Kohonen, Honkela and Timo, “Kohonen network“, Scholarpedia 2011. Retrieved 2012-09-24.

[6] T Kohonen, Self-organizing maps, 2nd ed., Springer Verlag, 1997, pp. 448.

 ISBN: 978-1-61208-477-0

[7] R. Hecht-Nielsen, “Counterpropagation networks Proceedings of the IEEE First International Conference on Neural Networks“, San Diego 1987. vol.2. pp. 19-32.

[8] R. Hecht-Nielsen. Applications of counter-propagation networks. Neural Networks 1988. N.1. pp. 131-139.

[9] E.V. Gubler. Computational methods of analysis and recognition of the pathological processes Leningrad: Medicina 1978. pp. 296.

[10] Ultsch and Alfred, “U*-Matrix: A tool to visualize clusters in high dimensional data“, Department of Computer Science, University of Marburg 2003. Technical Report N.36.pp. 1-12.

[11] Michel Volle, “Analyse des données“, Economica 4e édition 1997.

[12] Yu. N. Tolstov, “Basics of multidimensional scaling“, M. KDU 2006. pp. 160 .

[13] Joe. Kim, “factorial, discriminate and cluster analysis“, Edition Ozon 2012. pp. 216.

[14] Hooper and Martyn, “Richard Price, Bayes theorem, and God“, Significance 10 (1) 2013. pp.36–39.

[15] I.P Gaydyshev. Analysis and data processing: a Special reference book. St. Petersburg: Peter 2001. pp. 752 .

[16] J. A. Freeman and D. Skapma, “Neural Networks, Algorithms, and programming technique“, Addison-Wesley publishing company 1992.

[17] B.Krose and P. van der Smagt, “ An introduction to neural networks“, The University of Amsterdam. – 1996. – pp. 135.

[18] P D. Wasserman, “Neural Computing: Theory and Practice“, ANZA Research. 1989. pp. 64.

[19] R.Christian and S.A.Yvan, “Mathématiques et technologie“, Springer Science+ Bisness 2008. pp. 431.

[20] M. Mc. Cord Nelson. W.T. Illingworth, “A practical guide to neural nets“, Addison-Wesley Publishing ompany 1991.

[21] E. Davalo and P. Naïm, “Des Réseaux de neurones“, Edition Eyrolles 1993.

[22] A. Deweze, “L’accès en ligne aux bases documentaire“, Collection MASSON 1983.

[23] J. A. Farrel and A. N. Michel, “ Associative memory via artificial neural networks“, IEEE control system magazine 1990.

[24] Jérôme Pagès. Analyse factorielle multiple avec R. EDP sciences Paris, 2013. pp. 253.

[25] C.Guinchat and Y.Skouse, “Guide pratique des techniques documentaires“, Vol 1, 2. EDICEF 1989.

[26] D. O. Hebb The organization of behavior J. Wiley and Sons NY. 1949:

[27] J Herault and C Jutten, “Réseaux de neurones et traitement de signal“,. Edition HERMES 1994.

[28] Mirkes and M. Evgeny, “Principal Component Analysis and Self-Organizing Maps“, Applet University of Leicester 2011.

[29] M. M Glybovets and A. V Olecko, “Artificial Intelligence“, K Publishing house "KM Academy" 2002. pp. 366.

[30] M. Ezziyyani, E.Haimoudi and H. Fakhori, “Toward a New Approach to Improve the Classification Accuracy of the Kohonen’s Self-Organizing Map During Learning Process“, Proc. Technical Program of International Conference on Advanced Information Technology, Services and Systems (AIT2S 15), 16-17 Dec. 2015, Settat Morocco

[31] I.P Gaydyshev, “Analysis and data processing“, A Special reference book. St. Petersburg: Peter 2001. pp. 752.

[32] Catherine Bolley, “Analyse numerique“, Ecole d'ingenieur. Nantes France 2012. pp.97.

[33] Collins Gem Guide to Flags. Collins Publishers. 1986.

 ISBN: 978-1-61208-477-0