Comparative Analysis of Supervised Machine Learning Techniques on K-12 Educational Data

Ravi Mattani
Intel Corporation
ravi.mattani@intel.com

Manjeet Rege
University of St. Thomas
Graduate Programs in Software
St. Paul, MN 55105
rege@stthomas.edu

Brandan Keaveny
Data Ethics, LLC
brandan@dataethics.net

***Abstract*— The Anonymous School District (ASD) presented in this paper is implementing a comprehensive plan to intensify the information intelligence capacity to pinpoint educational needs of every student. Their main goal is to create analytical intelligence processes specific to the research needs of the school district and deploy an infrastructure that includes implementation of state of the art data analytics tools. The initial effort towards that goal is to research several machine learning techniques. The focus of this project is to assist the ASD research team in deployment of appropriate standards and procedures for efficiently forecasting and utilizing the large amount of data collected by the district. The goal of the project is to evaluate the most efficient machine learning techniques to forecast future trends. As a result, we developed a framework to transform raw data into minable data and apply several supervised learning techniques. Experiments were conducted to analyze the best technique.**

Keywords—supervised; machine; learning; k-12; analytics

I. INTRODUCTION

One of the most important challenges educational organizations face is to make effective use of the large volumes of data collected. Educational administrators and other decision makers would want to make decisions based on facts and trends that emerge from these huge data repositories. Data analytics is the practice of applying different algorithms and statistical techniques to extract interesting, unknown trends and predict future outcomes. Specifically, it comprises of applying machine learning algorithms, and other statistical methods for data analysis. This knowledge discovery can enable educational administrators to discover trends in data and assist in decision making [1] [2].

Machine Learning [3] techniques can broadly be classified into two types: Supervised [14] and Unsupervised [15]. The latter comprise of techniques that are used to extract interpretable patterns and correlations that exist in the data. Supervised Machine Learning on the other hand includes methods to perform inference on the available data with the intention to predict future outcomes and values. Classification and regression are two types of prediction problems. Classification problems comprise of predicting categories. For example, by examining samples from past data one can generalize if a student will pass or fail a particular exam. Regression involves predicting a numerical value for a particular attribute based on historical data. Predicting graduation rate for a school based on past data would be a regression problem.

In this paper, we present our work on a project that focuses on applying supervised machine learning techniques to K-12 educational data. The Office of Accountability of the Anonymous School District (ASD) we worked on in this project is implementing a comprehensive plan to intensify the information intelligence capacity to pinpoint educational needs of every student. Their main goal is to create analytical intelligence processes specific to the research needs of the school district and deploy an infrastructure that includes implementation of data mining tools and state of the art data analytics. The initial effort is to research several supervised machine learning techniques and provide a tool which is fine tuned for knowledge discovery on data from multiple sources. The ASD collects and maintains large amounts of student data. This data is collected from different sources like student management systems, excel spreadsheets and other sources. This data varies in both size and scope. These data banks have considerable potential for information discovery and pattern analysis. The main focus of this project is to assist the ASD research team in deployment of appropriate standards and procedures for efficiently forecasting and utilizing the large amount of data collected by the district. The goal of the project is to evaluate the most efficient supervised machine learning techniques to forecast future trends. As a result, we develop a framework to transform raw data into minable data and apply several predictive data mining techniques to the same.

Rest of the paper is organized as follows. Section II is Related Work that provides a discussion of applications of data mining and machine learning techniques to educational data. In Section III we provide an overview of the overall work in terms of the machine learning techniques applied. Section IV explains the structure of the educational datasets that we have analyzed and the preprocessing done before applying the machine learning techniques. The experiments and results appear in Section V with conclusions and a discussion of future work appearing in Section VI.

 ISBN: 978-1-61208-637-8

II. Related Work

Romero et al [4] survey the applications of data mining in the education industry. They analyze how educational data mining is different from traditional systems. The data, mining-objectives and techniques used in traditional e-commerce systems are significantly different. The most important goal in e-commerce mining is monetary and measured in terms of fixed business objectives. Educational data mining is more subjective and its primary goal is to improve the learning experience of students and provide decision makers with trends and analytics. They also highlight how some traditional data mining techniques can be successfully used in education-mining but due to the special characteristics of data in educational systems traditional techniques have to be modified so as to meet the requirements. They appraise the various data-mining techniques like clustering, outlier detection, statistics, visualization and sequential pattern mining that have been applied in education systems. The need for data mining techniques specifically designed to an educational perspective is classified as one of the future research areas of data mining. In [5], Agathe Merceron et al demonstrate how data mining techniques and algorithms can assist in extracting trends and patterns from data obtained from web-based educational systems. Baker et al in [6] review the current advancements in educational data mining. They pinpoint the key areas of research particularly in the field of education mining. They provide a comparative analysis of the emerging research areas. Relationship mining, clustering and prediction have been classified as methods that are gaining popularity in education data mining. Zaiane et al [7] outline how data mining can help in information extraction and analysis of online courses. In [8], Tang et al explain how methods like clustering, association rules and collaborative filtering assist in developing optimized e-learning systems. Beck et al in [9] demonstrate how predictive data mining methods can help in predicting student behavior. Siegel in [10] provides a step by step methodology for effective implementation of predictive data mining methods. This guide provides an excellent introduction to successfully implement a predictive data mining project. It provides a thorough analysis of predictive methods and historical references with regards to advancements in predictive data mining. In [11] Samui et al provide an introduction to predictive data mining process and introduce some of the most widely used predictive techniques like neural networks ,decision trees rule induction and genetic algorithms.

Our current work builds some of the aforementioned research by applying a number of supervised machine learning algorithms to educational data. Before considering advanced models, in the current work we have focused on applying some of the commonly used machine learning models such as Decision Tress, k-Nearest Neighbor, Neural Networks, and Naïve Bayes Classifier.

III. Generalized Framework

In this section, we investigate the various predictive analytics techniques for applying to the educational datasets. The objective is to mine the data collected by the school district and provide a tool which is fine tuned for knowledge discovery of educational data. The tool used for conducting these experiments is Rapid Miner. The tool is very powerful and supports almost all machine learning procedures as well as data preprocessing, modeling, transformation and data visualization. Figure 1 shows the different steps that are performed on the raw data.

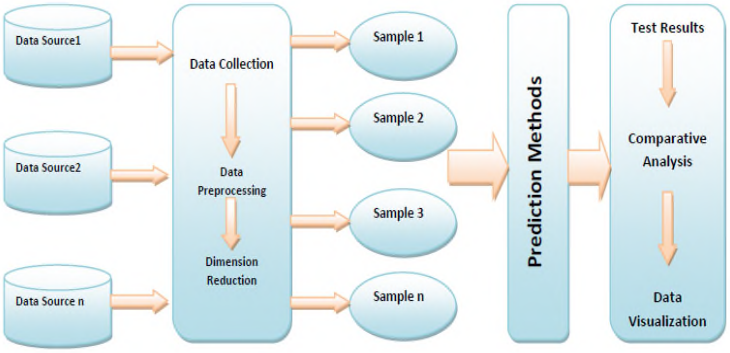

Fig. 1. Predictive data analytics and visualization process

A. Steps in Data Analytics and Visualization

Data Preprocessing: Understanding the complexity and inconsistencies that exists in the data will help us decide the most effective methods to make the data minable and achieve our objective. The activities performed in this step are:

Outlier Detection: Outliers are instances in the data that deviate from other instances and appear inconsistent. There are several algorithms and methods that can be used to detect such instances. In our case we used the statistical search based methods based on the distance measure techniques available in Rapid Miner.

Replacing Missing Values: This data cleansing process involves replacing missing instances with appropriate values to make the data complete. Several predictive mining techniques cannot handle missing values so it's important we replace these instances.

Data reduction and transformation: This step involves removing superfluous attributes from the data. If large numbers of instances in an attribute are missing, then we can eliminate that attribute. Certain techniques have limitations on the type of data that can be handled. It is necessary to transform the data into the suitable format to make it minable. For example, wherever necessary tasks such as discretizing the data, converting numerical values to binomial, converting nominal values to binomial, etc. were performed.

Perform Predictive Analytics using different techniques & evaluate performance measures: This step involves performing predictive analysis using the data mining tool and generating performance indicators that will assist in evaluating and analyzing the technique used.

B. Predictive Analytics Techniques

Decision Trees: Prediction using decision trees is one of the most popular and widely used logical methods for predictive analytics. The decision tree structure can be easily incorporated

with data that is in standard spreadsheet format. There are two fundamental data preparation and learning tasks involved with decision trees. The first task is to establish the nodes in the trees based on the data. The second task is to formulate tests for non-terminal nodes. There are several algorithms to perform these two tasks. The process involves first selecting a feature from the data which is then used to partition the data. This process is performed in recursion and applied to subdivision of data. Finally, terminal nodes are assigned with the appropriate values. One of the advantages of the decision tree approach is that it can handle high dimensional data. One can apply the various dimension reduction techniques to raw data in order to make it suitable for decision modeling. Decision trees that cover all the cases of the training data can get very complex and techniques like pruning have to be applied to reduce the complexity. Overall decision tree induction method is one of the fastest prediction techniques. Another advantage is the explanatory capability and ease of inferring the solution in decision trees. In our work, we have used the Random forest algorithm [12] to generate the decision tree. This approach is characterized by examining the data and classifying the cases on similarity measures. The solution to a new case can be found by looking for a match learnt from training data. Distance measures are used to relate a new case to an already defined case to predict the solution. This method is suitable for both classification and regression problems.

k-nearest neighbor: A distance metric is used to calculate the distance to known cases. In the k-nearest neighbor technique, k represents the number of neighbors to retrieve. Normalization of data is required to make the data suitable for this method. Feature selection plays an important role in the success of this method and it requires extensive experimentation. Overall this method produces satisfactory knowledge discovery which is based on prior experience. The k-NN approach [3] does not support missing values. Therefore, for the model to effectively predict it was important to eliminate instances with missing values.

Neural Networks: Neural Networks is a nonlinear mathematical solution for predictive analytics. This method is characterized by a network of neurons. Each neuron has a threshold value and it accepts a set of input values. A weighted sum of all the inputs to a neuron is calculated and this sum is compared to the threshold. The inputs to a neural network are features. Initially, the computation between nodes is linear. Several data preparation steps are required to make raw data suitable for neural networks. Neural networks are complex when compared to linear prediction techniques and several optimization methods have to be applied to improve the performance. Significant data reduction is required so as to limit the input. One of the drawbacks of neural network approach in predictive analytics is that the time required for training the network is high compared to other methods. For

building our neural network model, we used the Nominal to Numerical operator. This operator is used to map all non-numeric attributes to numeric values. Neural network models cannot handle non-numeric instances. This operator does not change any numeric attributes, binary attributes are converted to 0/1 and nominal attributes are converted using effect coding which is implemented in the tool. This model learns and predicts based on feed forward neural network which is trained by back propagation algorithm. We used one hidden layer along with a sigmoid activation function.

Naïve Bayes classifier: This predictive technique is based on the Bayesian Theorem [3]. It is a very popular technique as it can handle data with high dimensionality. This technique is relatively easy to understand and can generate high accuracy rates.

IV. Data Analysis and Preprocessing

It is important to analyze the raw values and attributes before we undergo data preprocessing. The design approach to perform the experiments has been based on [13].

The data sets used in this comparative analysis were provided by the ASD. All these data sets were drawn from different educational contexts. It included Student transcript GPA data, Course data (ELA) and enrollment data. In order to better understand the data distribution and the nature of the data, first an exploratory data analysis was performed. Table I shows the number of cases/instances in each dataset and the corresponding attributes. These files were extracted from the Student management system for analysis. This data was collected from 2000-2004. Each row within these files represents a student and its respective performance in each marking period. There were four marking periods that were included in our experimental analysis.

The ELA dataset has three class labels high, average and low. The number of features or predictors is 5 in the ELA dataset and the number of instances is 32936. There are two class labels pass and fail and the number of predictors is 8 in the ERSphase4GPA dataset. There are 3103 instances. In the Enrollment dataset there are 20 class labels. The data provided by ASD is the raw data dump obtained from their student management systems. We first performed data preprocessing and eliminated some of the inconsistencies in the data. After unwanted attributes were discarded the next step in the cleansing process is to remove or replace missing values in the example set with relevant instances. Many predictive techniques cannot handle missing data. We decided to replace the missing value of a numerical attribute with the mean of that attribute. If the attribute with missing value was nominal, we have used the mode of attribute for replacement. We also dropped instances that had majority of its values missing.

V. Experiments

We now describe the experiments conducted to perform comparative analysis of different predictive analytics techniques on the pre-processed ELA, ERSphase4GPA and Enrollment datasets. We compare the classification accuracy of the four methods on three different datasets and derive the best method across all. Specifically, for comparison we used the following performance measures:

- Accuracy: Percentage of correctly classified instances.
- Kappa: The kappa value presents the measure of agreement. A value of ‘0’ signifies poor agreement, i.e. the prediction was made by chance whereas ‘1’ signifies excellent prediction agreement.

TABLE I. RAW DATA ANALYSIS ON THE ASD DATASETS

Data Set	No. of cases/instances	No. of Attributes	Class/ Label	Comments
ELA 2001.csv	20359	26	**Performance** based on marking periods High, Low, Average	Dataset describes student ELA performance over four marking periods along with their final mark and course information. Raw data had large number of missing values and non significant attributes. Dataset had both numeric and nominal variables.
ELA 2002.csv	18567			
ELA 2003.csv	11585			
ELA 2004.csv	23454			
ERSphase4GPA.csv	25476	23	**Outcome** : pass or fail based on GPA	Dataset describes student GPA info over a period and student related information. Dataset had some missing values and was numeric in nature.
EOYEnrollment0102d.csv	13486	30	**Enrollment Status**: 12 different dimensions/ classes	Dataset describes student enrollment status data over years along with related student information. DOB, Area, School etc. Raw data had large number of missing values and no significant attributes Dataset had both numeric and nominal variables.
EOYEnrollment0203d.csv	10485			
EOYEnrollment0304d.csv	15785			

- Root mean square error: Measurement of error. A low root mean square error signifies that the prediction is more correct than it is wrong.

TABLE II: PERFORMANCE MEASURES FOR ELA DATASET

	Accuracy		Kappa	Root mean square error
K-NN	K=1	K=500		
	46.14 %	83.58 %	0.712	0.157
Naïve Bayes	85.84 %		0.751	0.005
Random Forest	82.80 %		0.634	0.235
Neural Net	90.11 %		0.084	0.203

Table II shows the classification accuracy, classification error, kappa and root mean square error for the ELA dataset. For K-NN, Performance measures were compared at different values of 'k'.' k' determines the flexibility of the classifier. For low 'k', we observed high bias making the classifier very flexible. Through cross validation and experimenting with various values of k we found a value that did not under fit or over fit. The accuracy increased as 'k' was increased. We achieved the highest classification accuracy at k=500. The kappa values suggest that the method has good prediction agreement. The performance of Naïve Bayes was better compared to K-nn and Random Forest method. Random forest model learns and predicts based on a set of random trees. The model created comprises of numerous random tree models. The parameters used to fine tune the model are the number of trees, the criteria used in selecting the attributes and splits. Naïve Bayes method produced the best prediction agreement. Although the best classification accuracy was obtained via the neural net method, the kappa value suggests that it had low agreement.

TABLE III: PERFORMANCE MEASURES FOR ERSPHASE4GPA DATASET

	Accuracy	Kappa	Root mean square error
K-NN	88.04%	0.743	0.275
Naïve Bayes	91.04%	0.815	0.260
Random Forest	89.81%	0.774	0.259
Neural Net	99.58%	0.014	0.074

For 'ERSphase4GPA' dataset, the K-NN performance measures remained constant at different values of 'k'. The kappa values suggest that all methods other than neural net have good prediction agreement. The performance of Naïve Bayes was better compared to K-nn and Random Forest method. This method produced the best prediction agreement. Although the best classification accuracy was obtained via the neural net method, the kappa value suggests that it had low agreement but the root mean square error was also low. Additionally, this dataset was numeric in nature, well suited for neural net method.

 ISBN: 978-1-61208-637-8

TABLE IV: PERFORMANCE MEASURES FOR ENROLLMENT DATASET

	Accuracy	Classification Error	Kappa	Root mean square error
K-NN	73.62%	26.38%	0.304	0.275
Naïve Bayes	72.08%	27.92%	0.459	0.260
Random Forest	68.71%	31.29%	0.053	0.514
Neural Net	76.65%	23.35%	0.481	0.468

For the 'Enrollment' dataset the K-NN performance measures remained constant at different values of 'k'. The performance of K-NN was better compared to Naïve Bayes and Random Forest methods. The kappa values suggest that all methods produced poor prediction agreement when compared to other datasets. The root mean square error was comparatively high for this dataset signifying that the prediction is more wrong than it is correct. The best classification accuracy was obtained via the neural net method; additionally, the kappa value suggests that it had higher agreement compared to other methods.

VI. CONCLUSIONS AND FUTURE WORK

We evaluated the performance of four supervised machine learning algorithms on three different data sets drawn from distinct educational contexts. The ERSphase4GPA dataset produced the best classification accuracy. Neural net emerged as a method producing the highest accuracy rates but this method had low agreement, i.e. prediction was made by chance for some datasets. Naive Bayes emerged as a consistent method producing reliable performance across all datasets.

There are a number of research avenues that can be followed thereby increasing the scope of this work. More experimentation with parameters within each method could effectively improve the performance. Using Deep Learning models is another interesting direction to pursue as well.

REFERENCES

[1] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods", International Journal of Database Theory and Application, 9(8), 119-136, 2016

[2] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance", In proc. of IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2015

[3] Hands-On Machine Learning with Scikit-Learn and TensorFlow, O'Reilly Media, ISBN: 1491962291

[4] C. Romero, and S. Venture, Educational Data Mining: A Survey from 1995 to 2005.Expert Systems with Applications 33, 125-146, 2007

[5] A. Merceron and K. Yacef, Educational Data Mining: a Case Study. In Proceedings of the conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology, Chee-Kit Looi, Gord McCalla, Bert Bredeweg, and Joost Breuker (Eds.). IOS Press, Amsterdam, The Netherlands, The Netherlands,467-474, 2005

[6] R. Baker and K. Yacef, the State of Educational Data Mining in 2009: A Review and Future Visions JEDM - *Journal of Educational Data Mining, Volume 1, Issue 1, October 2009 Pages 3-17, 2009*

[7] O. Zaiane, Web usage mining for a better web-based learning environment. In Proceedings of conference on advanced technology for education, 2001

[8] T. Tang and G. Mccalla, Smart recommendation for an evolving e-learning system: architecture and experiment. International Journal on E-Learning 4, 105-129, 2005

[9] J. Beck, and B. Woolf, High-level student modeling with machinelearning. Proceedings of the 5th International Conference on IntelligentTutoring Systems, 584–593, 2000

[10] E. Siegel, Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die, Wiley, ISBN: 1119145678, 2016.

[11] P. Samui, S Roy, and V. Balas, Handbook of Neural Computation, ISBN 0128113189, Academic Press, 2017.

[12] C. Smith, Decision Trees and Random Forests: A Visual Introduction For Beginners, ISBN 1549893750, 2017.

[13] L. Talavera, E. Gaudioso, Mining student data to characterize similar behavior groups in unstructured collaboration spaces, In Workshop on AI in CSCL (2004), pp. 17-23.

[14] N. Moseley, C. O. Alm and M. Rege, "Toward inferring the age of Twitter users with their use of nonstandard abbreviations and lexicon," *2014 IEEE International Conference on Information Reuse and Integration (IRI)*, 2014, pp. 219-226

[15] A. Oest and M. Rege, "Feedback-driven clustering for automated linking of web pages," *8th International Conference for Internet Technology and Secured Transactions (ICITST-2013)*, London, 2013, pp. 344-3

 ISBN: 978-1-61208-637-8