Particularized Cost Model for Data Mining Algorithms

Andrea Zanda
Universidad Politecnica Madrid
Facultad de Informatica
Madrid, Spain
andrea.zanda@alumnos.upm.es

Santiago Eibe
Universidad Politecnica Madrid
Facultad de Informatica
Madrid, Spain
seibe@fi.upm.es

Ernestina Menasalvas
Universidad Politecnica Madrid
Facultad de Informatica
Madrid, Spain
emenasalvas@fi.upm.es

***Abstract*—Ubiquitous devices demand autonomous and adaptive data mining. Despite some advances, the problem of calculating the cost associated to the execution of data mining algorithms is still a challenge. Thus, in this paper we provide a method for predicting the cost in terms of efficacy and efficiency associated to a mining algorithm, the resulting cost model as shown in our previous work can be exploited by a mechanism for predicting the best configuration of a mining algorithm according to context and resources. Recent work presents how a cost model not associated to any dataset can provide reliable estimations on efficiency and efficacy, here we present how we can improve the accuracy of such estimations by particularizing cost model to a predefined dataset. We provide the guidelines of the method and then we present a particularized cost model for C4.5 algorithm associated to a specific dataset (Parkinson's telemonitoring). Experimental results show how the particularized cost model achieves significant better estimations than the general cost model.**

***Keywords*-ubiquitous, data mining, cost model, algorithm.**

I. Introduction

The dissemination of ubiquitous devices has become a reality, nowadays such devices are able to execute almost any kind of application and collect considerable amount of data. To endow the devices with data mining services in order to exploit such amounts of data is a requirement. Applications in many domains require embedded intelligence to achieve their goals [7] [10], but their intelligence is not always personalized or adaptable. In [6] the authors provide a review of mobile care system which support the patient according to predefined mining models or to a server communication. An example on how to provide mobile devices with intelligence is in [1], an application of a neural network approach for the development of a system for knowledge classification in diabetes management. In the domain of intelligent transportation systems there are many situations in which an intelligent component is needed because internet connectivity in order to communicate with a server is not possible. In [8], the authors present a novel context-aware framework integrating intelligence for transportation systems. The system is able to: (1) learn patterns collisions by monitoring, (2) learn to recognize potential hazards in intersections and (3) warn particular threatened vehicles. Nevertheless, also in this domain the data mining framework has not been fully explored and developed. It is clear then how the integration of the mining technique directly into the devices can considerably increase the utility of ubiquitous applications, personalizing, assuring privacy and adapting to the changing world.

Data mining in ubiquitous devices has at least two requirements, to lead the process in an autonomous way and to adapt the process to the changing world. Some works in literature [3] and [2] provide approaches to adapt stream mining algorithms according to context information and available resources, but solutions for adaptable algorithms for the static case are still lacking. Further the methods applied for stream scenarios cannot be applied to batch scenarios, in fact in batch algorithms it is not possible to control the execution while the process is running, the initial algorithm configuration cannot be modified. In [4], some works providing methods for seeking the optimum neural networks algorithm configuration are presented. The main drawbacks concern the resource consumption of the methods to find the optimum and the fact they do not take into consideration external factors as context and internal resources, but only the dataset to be mined.

In [13], a mechanism able to select the best configuration for a C4.5 algorithm according to resources and context was presented. The mechanism is based on the EE-Model, which is able to estimate the efficiency and the efficacy of the C4.5 algorithm in terms of memory, CPU cycles, battery and accuracy, given the metadata of the dataset to be mined and the algorithm configuration. The model is calculated on the past behavior of the algorithm, which has been executed with different configurations and datasets. The main advantage of the model presented is the generality as it can be used to predict the efficiency and efficacy of that algorithm in any circumstances and with any dataset, but this is also its main drawback as it is not particularized. Normally in a particular device the dataset features will not vary and this is our main motivation to present a particularization of the EE-Model for a given dataset. As experiments will show, the particularization of the EE-Model for C4.5 algorithm makes it possible to get more accurate prediction on memory and CPU cycles.

The rest of the paper has been organized as follows: in

ISBN: 978-1-61208-100-7

Section 2 we focus the paper on the requirements of a cost model associated to a data mining algorithm, Section 3 sets the problems for calculating the cost model. In Section 4 we describe the guidelines in order to build a P-EE-Model and then we present a P-EE-Model for C4.5 associated to a Parkinson's tele-monitoring dataset. In Section 5 we show experimental results on the customized cost model. The Section 6 presents the conclusions and the future research.

II. Preliminaries

In [13], a mechanism to select the best algorithm configuration to execute a mining algorithm, taking into account information regarding the situation is presented. What they call situation is defined by the external factors, and it is divided into two main groups:

- Factors describing resources: memory, battery, CPU;
- Factors describing context information: information that can be sensed from sensors (location, temperature, time, etc).

Consequently, the authors divide the issue to decide the best configuration of the mining algorithm into two subproblems, on one hand how the external factors influence the requirements of the mining process in terms of efficacy, efficiency and semantics (meaning of the results), and on the other hand how the algorithm behaves when altering input data and input parameters. The main assumption under the division into two subproblem is that no matter the external factors, the algorithm inputs determine the quality of the model and performance that can be obtained. By efficiency they understand the resource consumption of the execution. On the other hand, the efficacy in a classification algorithm can be defined as accuracy (i.e., percentage of corrected classified items).

The method behind the cost model (EE-Model) presented in [13] relies on historical analysis of past execution of a particular algorithm to calculate the influence of inputs on the cost and results of the model. This is to say, information on the cost of past executions of the algorithm on different configurations and with different datasets are analyzed and knowledge discovery process is applied to extract rules that can be used to predict the bahaviour of the algorithm in new cases. As the experimental results show, the model presented there (EE-Model) provides estimations which are closer to the real efficacy and efficiency, nevertheless it presents the following drawback: it has been defined for general datasets, this is to say, historical executions analyzed consider different datasets. The features of a particular dataset can influence the behavior of the algorithm differently and this has motivated the present research in which we propose to particularize the cost-model depending on the features of a particular dataset.

Consequently it would be good to have a particularized cost model built on a determined dataset that could lead to more accurate prediction of the behavior of the algorithm both in terms of efficiency and efficacy. The underlying drawback behind is the lack of flexibility as the EE-Model customized this way would only be valid for that particular dataset. Nevertheless in real cases the dataset of a particular domain or application will change only in terms of number of records, size and distribution, all factors which our particularized cost model can adapt to. As experiments will show the particularized model can provide significantly more accurate estimations. In what follows we first present the problem and later we present the particularized EE-model.

III. Setting the problem

The same mining algorithm can lead to different resource consumption and different accuracy of the model depending on many factors, but which are the factors altering such behavior? And how do they alter such behavior? As it is depicted in Figure 1, the mining algorithm efficiency and efficacy depends on:

1) The input data;
2) The configuration.

We describe in depth these features in the next subsections to see how they affect the algorithm behavior to take advantage of these features to better predict algorithm resource consumption and performance. Note that the analysis of the semantics is out of scope in this paper.

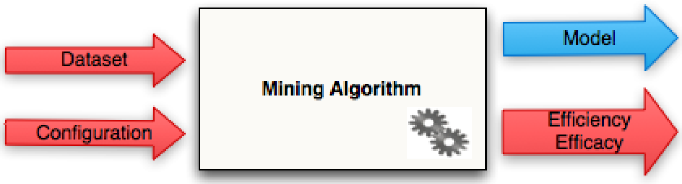

Figure 1. Mining algorithm inputs and outputs

A. Input dataset

The input dataset is the main input to the mining process. Depending on the data quality so there will be the results. Consequently we analyze in what follows how the data quality can impact the process. The quality of the data is related to:

- The number of records;
- The number of attributes;
- The type of each attribute;
- The values and distribution.

The dataset features will influence both the efficiency and the efficacy of the algorithm, in this sense for example a bad quality dataset in terms of data distribution can lead to a not precise model. Note that for example the number of columns could affect the efficiency, although it could also affect the quality of the model, increasing the number of columns leads to high dimensional problems that can be a significant obstacle to achieve high quality models. Also increasing the size of the dataset will probably result in a lower efficiency, but then the efficacy has to be explored.

B. Algorithm configuration

Setting the configuration of the algorithm means to assign values to the algorithm parameters for an execution. The configuration also determines the resource consumption and the accuracy of the results. In [5], a number of algorithms are tested with the same dataset in order to analyze their performance, the authors show the resource consumption and the accuracy achieved by testing them with different configurations. Such relations between configuration and result have to be known. Binary split option of a C4.5 algorithm for example can increase the efficiency of the algorithm because building a more branched tree, nevertheless the option can be suitable for certain types of dataset and increase the efficacy.

IV. Approach

In [13], a mechanism able to select the best configuration to execute the C4.5 algorithm according to external factors is presented. Figure 2 shows how the various phases of the process, the mechanism has a central role, it can access dataset information, configuration metadata and external factors, and it gives as result the best configuration for the mining algorithm. The EE-Model supports the mechanism by providing estimations on efficacy and efficiency of the mining algorithm execution. This solution has many advantages, first of all the system can have an estimation of the resources needed for the execution, in some cases the system can avoid executions that cannot be terminated (for battery low for example). It is also possible to avoid out of memory problems and CPU bottlenecks. In fact, for ubiquitous devices, resource aware is an important process requirement. In this paper our goal now is to: check that the EE-Model can get better results when built for a particular dataset. In Section IV-A we present how to obtain the particularized EE-Model.

A. The particularized EE-Model

The dataset features of a particular dataset can influence the behavior of the algorithm, consequently it would be good to have a particularized model for certain datasets in those domains or applications where we know the dataset features will not dramatically change. The underlying drawback behind is the lack of flexibility as the EE-Model cannot be valid for any type of dataset, but on the other hand a cost model suitable for certain purposes would improve the accuracy of the estimations. Here we will focus on the guidelines for the particularization of the EE-Model predicting C4.5 classification algorithm.

In order to build the EE-Model the steps are the following:

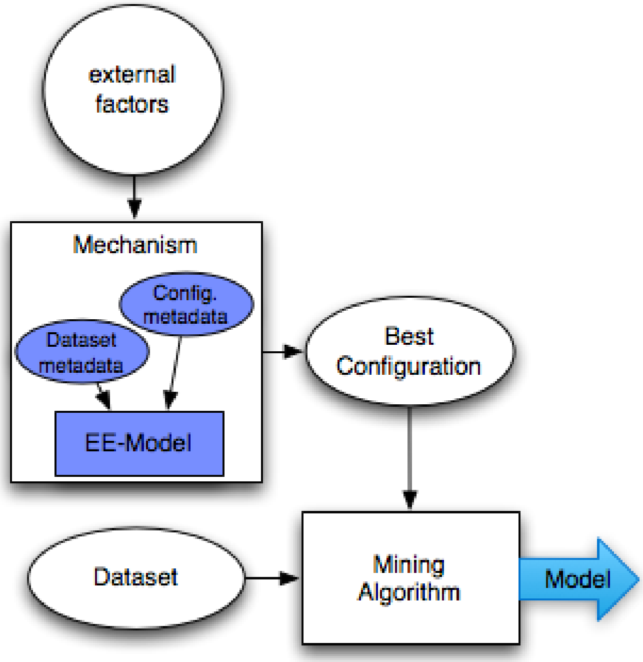

Figure 2. Approach

1) Define the set of variables to describe the executions of the algorithm:
    - Define the *condition variables* that describe the algorithm inputs as parameter settings (i.e. type of pruning) and dataset metadata. The dataset metadata in [13] is defined as number of attributes, type of attributes and so on, are not needed as they are not changing for a particular dataset, nevertheless we include richer and ad hoc information that describe that dataset. In fact the metadata has to include information such as of number of records and dataset size, class distribution and attributes distribution, in general any feature of the dataset that might change over time.
    - Define the *decision variables* that describe the set of executions information as memory and battery to name a few, or any measure is needed to predict with the model.
2) Execute a representative number of times the algorithm altering the condition variables;
3) Apply knowledge discovery process to the collected data relative to the executions in order to build one model for each decision variable. Most of times the decision variables will be numeric. According to our experience, we suggest to apply techniques as linear regression or regression tree.

In order to build the EE-Model the steps are the following:

B. *Particularized EE-Model for parkinson's tele-monitoring*

After an outline on how to build the model, here we present the EE-Model we built for a dataset relative to Parkinson's tele-monitoring [11]. The dataset is composed of a range of biomedical voice measurements with early-stage Parkinson's disease recruited for remote symptom progression monitoring, the description of the dataset attributes is given in Table I.

Table I
DATASET ATTRIBUTES DESCRIPTION

subject	Integer that uniquely identifies each subject
age	Subject age
sex	Subject gender '0' - male, '1' - female
test-ime	Time since recruitment into the trial
Jitter	Several measures of variation in fundamental frequency
Shimmer	Several measures of variation in amplitude
NHR,HNR	Measures of ratio of noise to tonal components in the voice
RPDE	A nonlinear dynamical complexity measure
DFA	Signal fractal scaling exponent
PPE	A nonlinear measure of fundamental frequency variation
total-UPDRS (CLASS)	Clinician's total UPDRS score (discretized)

Following the guidelines of Section IV-A:

1) We define the decision variables as in II, there are two measures on the efficiency and one on the efficacy of the algorithm. The accuracy is obtained with an evaluation of the model with a test set. Then we defined the condition variables as in Table III. We can notice that the dataset metadata contains on one hand information on the size of the dataset, on the attributes type and in general on the distribution of the attributes values, on the other hand the metadata associated to the algorithm parameters (in order to represent all the possible different configurations).

Table II
ALGORITHM EXECUTION INFORMATION

Memory	Memory used	Integer
CPU	Number of CPU cycles	Integer
Accuracy	Accuracy of the obtained mining model	Real

2) We obtain a dataset of historical data of execution of the algorithm in a system with 2.16 GHz Core 2 processor and 2.5GB 667 MHz DDR2 SDRAM memory. The number of execution is an important point to obtain a dataset able to represent the domain, we generated a number of 30023 covering all parameter configurations (increment of 0.10 for continuous parameters) related to the same dataset, but with different number of records and so with different class and attributes distributions.
3) This step concerns the application of data mining techniques in order to discover the relations between condition variables and decision variables. The algorithm used for memory and CPU cycles is REPTree [12], for accuracy linear regression [9]. The model for accuracy has not improved the previous results achieved, while the results for memory and CPU cycles are shown in Section V.

Table III
INPUT INFORMATION

Attribute number	Number of attributes	Integer
NInstances	Number of instances	Integer
Size	Dataset size in KB	Integer
Attribute distinct	distinct values of the column X (i.e. Class)	Integer
Attribute StdDEV	Standard deviation of the column X (i.e. Class)	Real
Attribute type	Number of columns of type Y (i.e. real)	Integer
Pruning	Whether pruning is performed. ('0' → no pruning, '1' → pruning, '2' → Reduced error pruning)	Nominal
Binary	Whether to use binary splits on nominal attributes when building the trees	Boolean
Laplace	Whether counts at leaves are smoothed based on Laplace	Boolean
CF	The confidence factor used for pruning (smaller values incur more pruning)	Real
Sub	Whether to consider the subtree raising operation when pruning	Boolean
MinNumObj	The minimum number of instances per leaf	Integer
NumFolds	Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.	Integer
Seed	The seed used for randomizing the data when reduced-error pruning is used.	Integer

V. EXPERIMENTATION

The experimentation is carried out evaluating the efficiency prediction of the presented particularized EE-Model (P-EE-Model) for Parkinson's tele-monitoring in comparison with the general EE-Model in [13]. The estimations of the two models are compared with the values of the real executions. Given a configuration and dataset metadata the EE-Model is able to estimate efficiency and efficacy, so in order to describe the experiment we first define the configurations and the dataset metadata we consider for our analysis. We define the three configurations presented in Table V, they mainly differ in the type of pruning applied in the execution. The reason is related to the fact that the pruning is the main parameter setting altering the efficiency of the algorithm. We evaluate the P-EE-Model with the Parkinson's tele-monitoring dataset we used for building the model, its description is in Table I, but here we take a sample of the original dataset having the number of records equal to 2543, the size to 424KB and different attributes and class distribution.

Then we evaluate the P-EE-Model comparing the estimations for the three configurations to the real values. Table V shows the results of the evaluation, the models provide more reliable for average memory, in fact a mean is supposed to be more stable.

Table IV
ALGORITHM CONFIGURATIONS

Config.	Prun	Bin	Lap	CF	Sub	MinOb	#Folds	S
1	1	Yes	Yes	0.25	Yes	3	–	–
2	0	Yes	No	–	0	2	–	–
3	2	No	No	–	Yes	5	5	5

Table V
EE-MODEL EVALUATION

	Average Memory	CPU cycles
Correlation Coeff.	0.95	0.99
Relative absolute error	8.0%	20.0%
Root relative squared error	9.9%	15.3%

Now we compare the performance of the P-EE-Model with the general EE-Model (G-EE-Model) in [13]. Figure 3 shows the absolute squared error obtained for the three configurations while considering first the P-EE-Model (sky blue) and then the G-EE-Model (red). The comparison is relative to the average memory and denotes a significant improvement on the accuracy of P-EE-Model estimations.

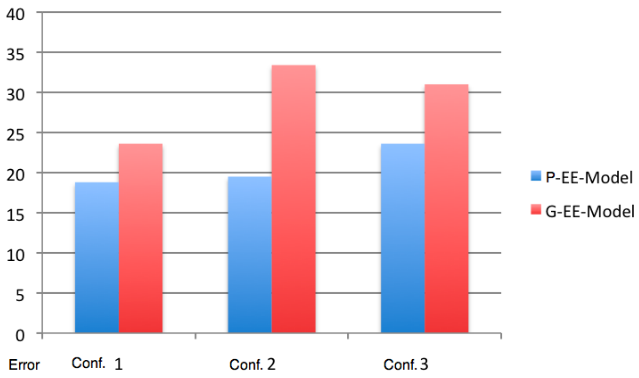

Figure 3. Comparing average memory

Figure 4 shows the comparison of the absolute squared error relative to CPU cycles, even in this case the estimations of the P-EE-Model overcome the general one. In this paper we argued the hypothesis that an EE-Model build for a particular dataset could achieve better estimations than a general one built on many datasets. According to the results above the hypothesis is verified and the estimations overcome the general model significantly.

Nevertheless, the drawback of the P-EE-Model concerns the lack of flexibility, in fact the model is only usable for the dataset on which it is built. To conclude we carried out an evaluation of the P-EE-Model which has been built on the dataset in [11], with another synthetic dataset. As we argued the estimations of such model are worse than the one provided by the general EE-Model.

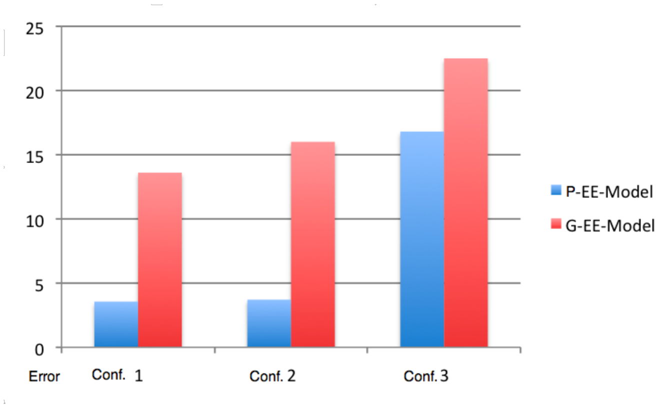

Figure 4. Comparing CPU cycles

VI. CONCLUSION

In this paper, we have presented a cost model to predict the behavior of a data mining algorithm with a specific dataset in terms of efficacy and efficiency that overcomes in accuracy the previous general cost models predicting the algorithm behavior with any kind of dataset. After describing the guidelines in order to build a particularized cost model (P-EE-Model), we present a P-EE-Model for C4.5 algorithm specific for a Parkinson's tele-monitoring dataset. According to our experimental results the E-PP-Model significantly improve the estimations of CPU cycles and average memory. Nevertheless the drawback of the P-EE-Model: less flexibility as it is associated to a specific dataset, has not to be ignored in domains of applications where the dataset can change dramatically.

REFERENCES

[1] G. Gogou, N. Maglaveras, B. V. Ambrosiadou, D. Goulis, and C. Pappas. A neural network approach in diabetes management by insulin administration. *J. Med. Syst.*, 25(2):119–131, 2001.

[2] P. D. Haghighi, M. M. Gaber, S. Krishnaswamy, and S. Loke. An architecture for context-aware adaptive data stream mining.

[3] P. D. Haghighi, A. B. Zaslavsky, S. Krishnaswamy, and M. M. Gaber. Mobile data mining for intelligent healthcare support. In *HICSS '09: Proceedings of the 42nd Hawaii International Conference on System Sciences*, pages 1–10, Washington, DC, USA, 2009. IEEE Computer Society.

[4] A. Kuzmenko and N. Zagoruyko. Structure relaxation method for self-organizing neural networks. In *ICPR*, pages IV: 589–592, 2004.

[5] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach. Learn.*, 40(3):203–228, 2000.

[6] Y.-C. Lu, Y. Xiao, A. Sears, and J. A. Jacko. A review and a framework of handheld computer adoption in healthcare. *I. J. Medical Informatics*, 74(5):409–422, 2005.

[7] C. Marx, W. Gwinner, J. Krückeberg, U. von Jan, B. Engelke, and H. K. Matthies. Mobile learning applications for education in medicine and dentistry. *Adv. Technol. Learn.*, 4(2):92–98, 2007.

[8] D. Preuveneers and Y. Berbers. Mobile phones assisting with health self-care: a diabetes case study. In *MobileHCI '08: Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*, pages 177–186, New York, NY, USA, 2008. ACM.

[9] Radhakrishna. *Linear Statistical Inference and Its Applications*. John Wiley & Sons Inc, November 1973.

[10] D. Siewiorek, A. Smailagic, J. Furukawa, A. Krause, N. Moraveji, K. Reiger, J. Shaffer, and F. L. Wong. Sensay: A context-aware mobile phone. In *ISWC '03: Proceedings of the 7th IEEE International Symposium on Wearable Computers*, page 248, Washington, DC, USA, 2003. IEEE Computer Society.

[11] A. Tsanas, M. A. Little, P. McSharry, and L. Ramig. Accurate telemonitoring of parkinsons disease progression by non-invasive speech tests. *IEEE Transactions on Biomedical Engineering*, 2009.

[12] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[13] A. Zanda, S. Eibe, and E. Menasalvas. Adapting batch learning algorithms execution in ubiquitous devices. In *MDM '10: Proceedings of the 2010 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, Kansas city, USA, 2010. IEEE Computer Society.