Proceedings of ICTAMI
International Conference on Theory and Applications in Mathematics and Informatics
17^{th}–20^{th} of September, 2015, Alba-Iulia, Romania
Computer Science Section

pp. 167-177

FEATURE SELECTION METHODS FOR MULTIDIMENSIONAL DATASETS

M. Muntean, H. Vălean, L. Căbulea

Abstract. One of the most challenging problems encountered when analysing multidimensional datasets is overabundance of features. In order to remove irrelevant, redundant, and noisy information from the data, some feature selection algorithms are used. The two-level feature selection analysed method contains Wrapper and Filter based feature search and evaluation. The experiments were performed on a database with 257 attributes and 593 instances, containing the UAB graduates answers to a questionnaire. The results showed that the two-level feature selection improved the time taken for build the classification models for multidimensional datasets, and, in some cases, improved also accuracy rates.

2010 *Mathematics Subject Classification*: 62H30, 68T99.

Keywords: Wrapper selection, Filter selection, time improvement, multidimensional dataset.

1. Introduction

The large amounts of data are collected and persistent stored in databases, increasing the need for efficient and effective analysis methods in order to use the information data. There could be a lot of patterns in a huge multidimensional database, and a lot of efficient data mining methods had been proposed to discover these models.

The main problem that appears in analysing multidimensional datasets is the decreased efficiency of mining algorithms in terms of time taken to build and evaluate models. To overcome this problem, pre-processing methods have been developed, especially feature selection methods.

2. Feature selection approach

The problem of feature selection involves finding a good set of attributes under some objective function that assigns some numeric measure of quality to the patterns discovered by the data mining algorithm [1], [2].

The objective of a data mining algorithm A is to take a training set T and discover a set of patterns P such that P optimizes some objective function F(P) that assigns some real-value measure of goodness to P. The output of A is determined by which attributes are present in the training set. We can parameterize the attributes used as a Boolean vector b, where $b_i = 0$ means attribute i is not used and $b_i = 1$ indicates that it is used [3], [4].

In general, the achievement of the optimal subset is impossible for two reasons. First, most objective functions cannot be accurate calculated, and can only be approximated. Even if it would be exact, there is the practical problem that if there are m attributes, there are 2^m possible values for b, a number of choices typically too large to search exhaustively, [3], [4]. Since we cannot always hope to find the optimal subset, we will try to find an approximating subset that will improve prediction accuracy [5]. An optimal feature subset is not necessary unique because it may be possible to achieve the same accuracy with different subsets of features (if two features are perfectly correlated, one can be replaced by the other) [2].

There are some important approaches for feature selection: Filter approach, Wrapper approach, and Embedded approach [2]. While Filter selection methods do not incorporate learning methods (Figure 1), Wrapper selection methods involves a learning algorithm to evaluate the quality of each feature subset (Figure 2). By including the learning algorithm they aim at improving accuracy. An Embedded model (Figure 3) embeds feature selection in the training process of the classifier and are usually specific to given learning machines [6].

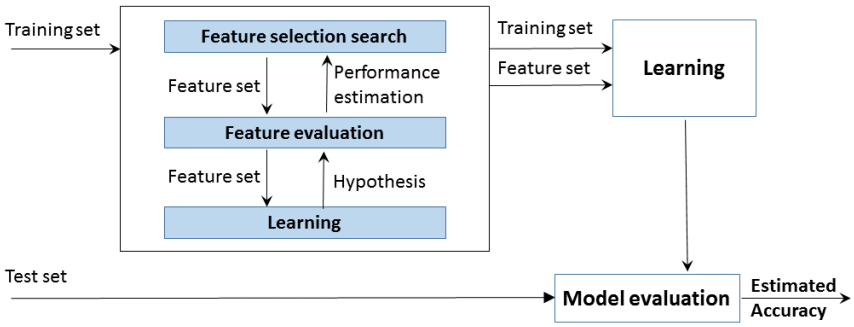

Figure 1: The Wrapper approach

The search methods that are part of Wrapper selection, most commonly include: best first search, simulated annealing, genetic algorithms, greedy stepwise forward selection, and greedy stepwise backward elimination. In this paper, best first search is used. This method performs greedy hill climbing with backtracking and it gen-

erates the successors of the best unexpanded node at each step (the node with the highest estimated accuracy). The termination condition is a number of consecutive non-improved nodes. The initial node determines the general direction of the search [7].

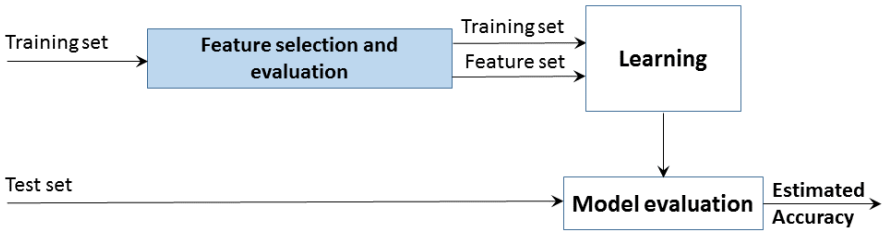

Figure 2: The Filter approach

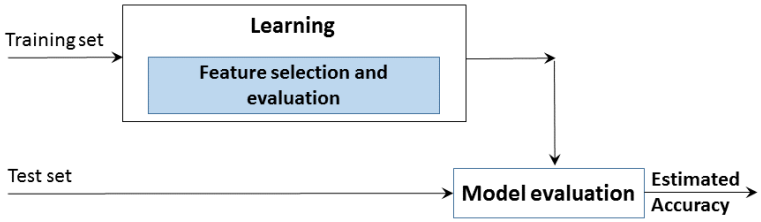

Figure 3: The Embedded approach

There are several approaches proposed in literature, that compare Filter and Wrapper feature selection results [8], [9], or use a hybrid filter-wrapper system [6], [10], in order to improve the time taken for building classification methods, and for testing these methods. In this work, we used the combined evaluation method, in order to improve the data mining process and to faster classify the proposed multidimensional dataset. In our case, for some classifiers, we have also improved the accuracy levels.

3. Feature selection methods and results

5.1. The dataset description

The UAB graduates' responses dataset contains the answers of 593 graduates of 1 Decembrie 1918 University of Alba Iulia to a questionnaire with 91 questions. The

purpose of applying this questionnaire was to evaluate if knowledge, competencies and skills obtained during the studies were sufficient to enable university graduates, promotion 2008-2009, to engage in the labour market. The dataset used for the experiments has 256 attributes (both numeric and nominal ones) and the environment that we chose to implement the experiments was the data mining tool named Weka [11]. A sample of attribute declaration in the .arff file is presented in Figure 4. The answers to this monitoring questionnaire were stored into a database and then pre-processed and saved as an .arff (Attribute Relation File Format) file (Figure 5). The "I don't know" answer was codified with the value -9, and the "I don't answer" affirmation was codified with -7. A question can have multiple choices, so the corresponding attributes have similar names (for instance STU1_1, STU1_2).

```
@relation absolventi_uab
@attribute ID numeric
@attribute HEI numeric
@attribute GY numeric
@attribute R1 numeric
@attribute R1_J numeric
@attribute R1_T numeric
@attribute R1_L numeric
@attribute R2 numeric
@attribute R3 numeric
@attribute STU1_1 numeric
@attribute STU1_2 numeric
```

Figure 4: A sample of attribute declaration in the .arff file

```
2439,130000,2009,2,-7,-7,-7,8,2005,669,713,717,725,-9,-9,-9,-9,1,1,130000,130300,130301,2,8.66,1,
2545,130000,2009,1,10000,11000,11003,3,2006,681,713,-9,-9,-9,-9,-9,-9,1,1,130000,130300,130304,1,
2562,130000,2009,1,10000,11000,11009,3,2005,-9,-9,-9,-9,669,713,-9,-9,-7,-7,-7,-7,-7,-7,-7,-7,-7,
2577,130000,2005,1,10000,11000,11003,2,2001,621,665,669,685,693,709,-9,-9,1,1,130000,130200,13020
2682,130000,2009,1,10000,11000,11002,8,1999,681,713,-9,-9,-9,-9,-9,-9,4,1,130000,130200,130205,3,
2724,130000,2009,1,10000,12000,12036,1,2005,681,713,717,9999,-9,-9,-9,-9,1,1,130000,130300,130303
2736,130000,2009,1,10000,11000,11003,4,1994,681,713,729,9999,-9,-9,-9,-9,1,1,130000,130300,130306
2778,130000,2005,1,10000,11000,11007,1,2001,620,665,693,713,-8,-9,-9,-9,1,1,130000,130100,130101,
2780,130000,2009,1,10000,11000,11003,2,1986,-9,-9,-9,-9,-9,-9,-9,-9,-7,-7,-7,-7,-7,-7,-7,-7,-7
```

Figure 5: The preprocessed answers of the graduate students

5.2. First-level feature selection

In the first-level we chose Information Gain attribute evaluation for filtering features leading to reduce dimensionality of the feature space. The Weka implementation (InfoGainAttributeEval) evaluates attributes by measuring their information gain with respect to the class. It discretizes numeric attributes first using the Minimum Descriptive Length (MDL)-based discretization method [12], being very useful

in our case, because a large number of attributes were numeric attributes. After running the evaluator (Figure 6), we decided to keep for the next level the features with average merit and average of their ranking throughout the Cross Validation higher than 0.5, meaning 75 out of 257 attributes.

```
Search Method:
        Attribute ranking.
        Threshold for discarding attributes: -Infinity

Attribute Evaluator (supervised, Class (nominal): 258 Cluster)
        Information Gain Ranking Filter

Ranked attributes:
 0.6677    171 R32_11
 0.6651    164 R32_4
 0.6651    172 R32_12
 0.6638    169 R32_9
 0.6568    175 R32_15
 0.6564    168 R32_8
 0.6536    173 R32_13
 0.6529    167 R32_7
 0.6525    179 R32_19
 0.6506    166 R32_6
 ...
 0.507     131 R29_5
 0.507     129 R29_3
 0.507     133 R29_7
 0.507     134 R29_1_M
 0.507     135 R29_2_M
 0.507     139 R29_6_M
 0.506     138 R29_5_M
 0.506     137 R29_4_M
 ...
 0.0677      7 R1_T
 0.067       8 R1_L
 0.0644      5 R1
 0           2 ID
 0           4 GY
 0           3 HEI
 0           1 Instance_number

Selected attributes: 171,164,172,169,175,168,173,167,179,166,
```

Figure 6: InfoGain feature selection results

5.3. Second-level feature selection

We used Wrapper methods only in the second level because these algorithms are too expensive for large dimensional database in terms of computational complexity and time taken since each feature set considered must be evaluated with the classifier algorithm used [13].

```
Evaluator:    weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.bayes.BayesNet
Search:       weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     absolventi_uab_clustered-weka.filters.unsupervised.attribute.Remove-V-R171,1
Instances:    593
Attributes:   76
              R32_11
              R32_4
              R32_12
              R32_9
              R32_15
              R32_8
              R32_13
              R32_7
              R32_19
              R32_6
              ...
Evaluation mode:    evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 858
        Merit of best subset found:    0.995

Attribute Subset Evaluator (supervised, Class (nominal): 76 Cluster):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.bayes.BayesNet
        Scheme options: -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 2,4,17,26,52 : 5
                     R32_4
                     R32_9
                     R31_18
                     R31_1
                     R30_13
```

Figure 7: Wrapper Subset feature selection results

A dataset containing 76 attributes (including the class attribute) was received from the previous level and was used as input for WrapperSubsetEval evaluator. On this stage we used Bayesian Network classifier and 10-fold-cross-validation and we discovered and evaluated 858 feature subsets. The optimum chosen subset was composed by 5 attributes: R32_4 (*ability to quickly gain new knowledge after graduation of study programme*), R32_9 (*ability to effectively manage work time after graduation of study programme*), R31_18 (*the possibility of participating in international internships*), R31_1 (*organization and structure of the graduated study program*) and R30_13 (*ability to exercise their authority to develop their own skill level*), with a merit (importance) of 0.995 (Figure 7).

In the final returned dataset (Figure 8) we observed that 78.6% of considered graduates succeded in their career (*cluster0*), while only 21.4% were not able to find a job, or they were not satisfied by their current job (*cluster1*).

```
@relation 'absolventi_uab_clustered-we

@attribute R32_4 numeric
@attribute R32_9 numeric
@attribute R31_18 numeric
@attribute R31_1 numeric
@attribute R30_13 numeric
@attribute Cluster {cluster0,cluster1}

@data
2,3,2,3,2,cluster0
-9,-9,-9,-9,-9,cluster1
4,4,3,4,4,cluster0
4,4,1,4,3,cluster0
3,3,3,3,4,cluster0
4,5,2,4,4,cluster0
5,4,-8,5,-9,cluster0
5,5,4,5,5,cluster0
4,4,3,2,5,cluster0
2,3,2,3,3,cluster0
4,3,2,3,3,cluster0
4,3,3,4,4,cluster0
```

Figure 8: Sample of final dataset

4. Classification results

After we found the optimum feature subset in the second level, we tested different classifiers from Weka software in order to compare the classification accuracy and

time with the ones found on the first-level of selection and on initial dataset. The obtained results are given in the table below.

Classifier	Init. acc.	Init. time	F-l acc.	F-l time	S-l acc.	S-l time
Ibk (lazy)	99.32	0	99.49	0	99.15	0
J48 (trees)	98.31	0.11	98.98	0.03	98.98	0
Jrip (rules)	98.31	0.2	97.97	0.05	98.48	0.03
PART (rules)	98.48	0.09	98.48	0.02	98.48	0
SGD (functions)	98.31	1.14	98.14	0.34	99.15	0.08
SMO PukKernel	93.25	0.83	97.3	0.2	99.32	0.05
SMO PolyKernel	99.83	0.13	99.15	0.05	98.81	0.02

Table 1. Accuracy and time results

In the classification step of first-level Filter selection, we observed that the accuracy was kept constant (and in the case of SMO Puk Kernel was even improved), while the time taken to build and evaluate models was reduced for all the classifiers used in the experiments. In the second-level Wrapper based selection, we obtained a dataset with better accuracy rates in for three classifiers (JRip, SGD and SMO Puk Kernel), and also a better classification time than the one found in the first-level for all the proposed scenarios (Figure 9 and Figure 10).

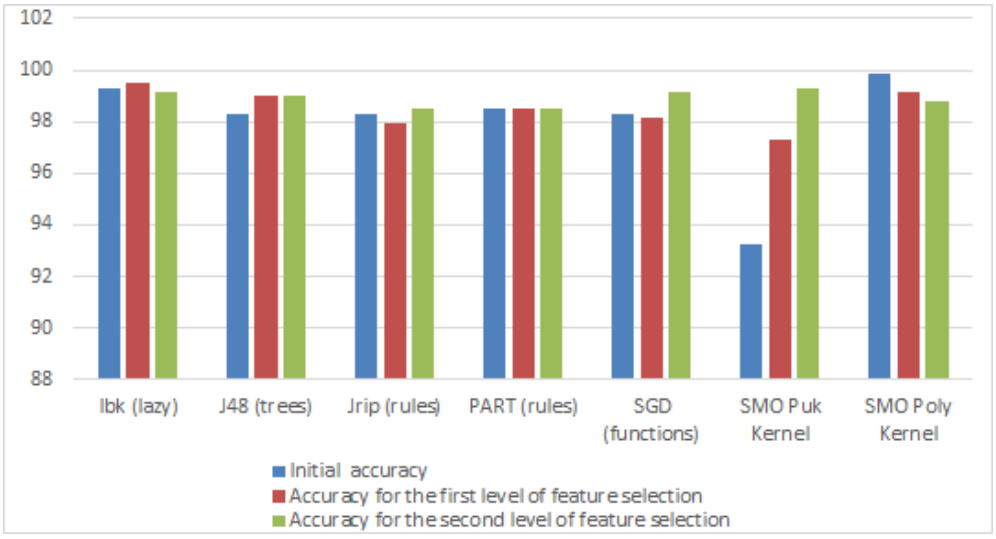

Figure 9: Classification accuracy results with respect to dataset and classifier change

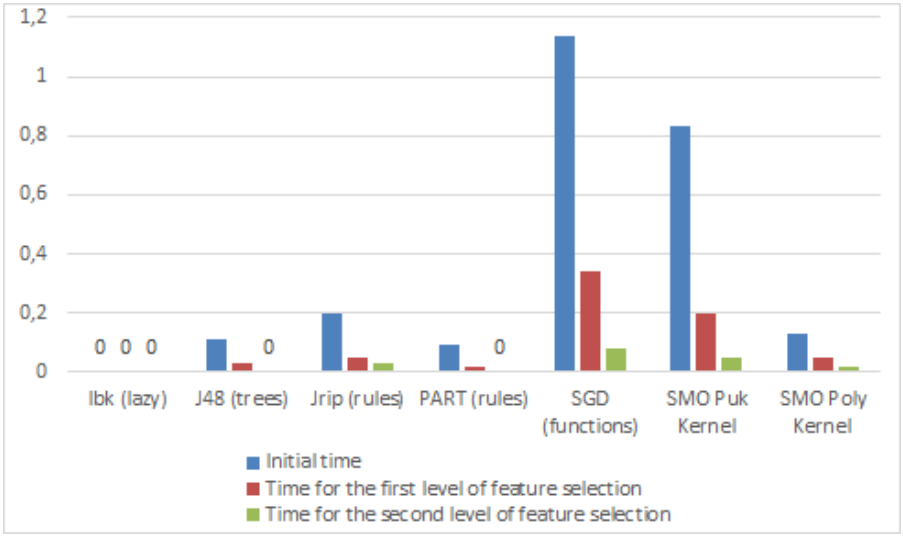

Figure 10: Classification time results with respect to dataset and classifier change

5. Conclusions

In this paper, two-level feature selection method was used in order to improve the classification time of multidimensional datasets. The data to be classified contained answers of 593 graduates of 1 Decembrie 1918 University of Alba Iulia to a questionnaire with 91 questions (the questions being pre-processed as 257 attributes). The analysed method helped us to faster classify the multidimensional proposed dataset, and, in some cases, returned higher accuracy values. A future research direction is to investigate the use of the feature selection methods on "yes/no" questionnaires, in which case special learning algorithms for pairwise data will have to be considered [14].

Acknowledgements This research was partially supported by the project 60/2.1/*S*/41750 POSDRU implemented by Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI) and the National Council for Higher Education Funding in partnership with The International Centre for Higher Education Research (INCHER) Kassel.

References

[1] I. H. Witten, and E. Frank, *Data Mining, Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, Elsevier Inc.(2005), 290.

[2] R. Kohavi, *Wrappers for Performance Enhancement and Oblivious Decision Graphs*, PhD thesis, Stanford University(1995).

[3] T. Mitchell, *Machine Learning*, The McGraw-Hill Companies, Inc.(1997), 52-78.

[4] G. H. John, *Enhancements to the Data Mining Process*, PhD Thesis, Computer Science Department, School of Engineering, Stanford University (1997).

[5] Y. Sun, S. Todorovic, and S Goodison, *Local-Learning-Based Feature Selection for High-Dimensional Data Analysis*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9(2010), 1610-1626.

[6] N. Ghayur, *A Hybrid Filter-Wrapper Approach for Feature Selection*, International Masters Thesis, Supervisor: Dr. Marco Trincavelli, Studies from the Department of Technology at Orebro University(2012).

[7] I. H. Witten, E. Frank, M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*, Elsevier (2011), 492.

[8] B. Jantawan, C.-F. Tsai, *A Comparison of Filter and Wrapper Approaches with Data Mining Techniques for Categorical Variables Selection*, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 6(2014), 4501-4508.

[9] E. Pitt, R. Nayak, *The Use of Various Data Mining and Feature Selection Methods in the Analysis of a Population Survey Dataset*, Second Workshop on Integrating AI and Data Mining (AIDM 2007), Gold Coast, Australia, Conferences in Research and Practice in Information Technology (CRPIT), Vol. 84, Kok-Leong Ong, Junbin Gao and Wenyuan Li, Ed.(2007).

[10] M. Danubianu, S.G. Pentiuc, D.M. Danubianu, *Data Dimensionality Reduction for Data Mining: A Combined Filter Wrapper Framework*, International Journal of Computers Communications and Control, Vol. 7, No. 5(2012), 824-831.

[11] http://www.cs.waikato.ac.nz/ml/weka/.

[12] I. H. Witten, E. Frank, M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*, Elsevier (2011), 491.

[13] A. G. Karegowda, A. S. Manjunath, M.A. Jayaram, *Comparative Study of Attribute Selection using Gain Ratio and Correlation based Feature Selection*, International Journal of Information Technology and Knowledge Management, Vol. 2, No. 2(2010), 271-277.

[14] A. Birlutiu *Machine learning for pairwise data: applications for preference learning and supervised network inference*, PhD thesis, Radboud Universiteit Nijmegen, 2011, ISBN: 9789088913303.

Maria Muntean
Exact Sciences and Engineering Department,
1 Decembrie 1918 University of Alba Iulia,
Alba Iulia, Romania
email: *mmuntean@uab.ro*

Honoriu Vălean
Automation Department,
Technical University of Cluj Napoca,
Cluj Napoca, Romania
email: *Honoriu.Valean@aut.utcluj.ro*

Lucia Căbulea
Exact Sciences and Engineering Department,
1 Decembrie 1918 University of Alba Iulia,
Alba Iulia, Romania
email: *cabuleal@uab.ro*