Y. Orenes et al., Int. J. of Design & Nature and Ecodynamics. Vol. 11, No. 3 (2016) 168–177

A COMPUTATIONAL EXPERIENCE FOR AUTOMATIC FEATURE SELECTION ON BIG DATA FRAMEWORKS

Y. ORENES, A. RABASA, A. PÉREZ-MARTÍN, J.J. RODRÍGUEZ-SALA & J. SÁNCHEZ-SORIANO
Miguel Hernández University of Elche, Spain.

ABSTRACT

The classification rule system is one of the predictive analytical techniques used in Big Data problems, where finding datasets with millions of rows but also with dozens of variables (attributes) is common. Classification rule systems consist of rule sets which have a so-called antecedent (variable or set of variables that can be numeric or nominal) and a consequent (target variable, provided nominal). If the antecedent variables are numerical, many generator algorithms of classification rules employ traditional methods of automatic feature selection, based on techniques already established in the scientific field, such as discriminant analysis or cluster analysis. In this paper, the authors propose the comparison of their own method of feature selection and classification, RBS (originally designed to manage only nominal variables) and classical methods of feature selection. After the formal definition of our own method, this paper presents the design of a computing experience that allows a qualitative and quantitative comparison of the adapted RBS and other methods for feature selection. Finally, optimal conditions of application of each method are discussed and future research areas in the field of automatic feature selection are identified.

Keywords: big data, classification rule systems, feature selection.

1 INTRODUCTION

Decisional systems integrate increasingly more data sources, and these are large and heterogeneous over time. Therefore, predictive models based on classification rules, e.g. ID3 type [1], need to incorporate mechanisms that allow to choose the most incidents variables in the target variable which is intended to be predicted. So, selection characteristic systems are also evolving and adapting to these changes that are mainly related to problems of high dimensionality.

A method for automatic feature selection on Big Data frameworks is proposed in this paper. The proposed method is evolved from a method of generating and managing classification rules which was initially designed to manage nominal variables only.

The problem of feature selection on Big Data problems and the general approach are exposed in Section 2 where a classical technique as discriminant analysis is introduced to compare later.

The simulation process for the synthetic dataset generation is presented in Section 3. As the feature selection method is expected to be tested on several overhead scenarios, several datasets will be generated. Additionally, a general experiment overview is shown, where it is explained in detail how datasets will be processed with different feature selection techniques and how the different generated set of attributes will be used for the corresponding rule set generation. The final rule sets (coming from the different reduced sets of attributes) will be compared under several criteria that are also presented.

That computational experiment is shown in Section 4, where an empirical comparison from the accuracy of rule systems is presented.

WIT CONFERENCES
This paper is part of the Proceedings of the International Conference on Big Data (Big Data 2016)
www.witconferences.com

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), http://www.witpress.com/journals
DOI: 10.2495/DNE-V11-N3-168-177

Finally, in Section 5, the conclusions about the potential of the method and the conditions in which seems to be more appropriate are presented and discussed. Furthermore, future research lines concerning feature selection methodologies are pointed out.

2 PROBLEM DEFINITION AND MAIN OBJECTIVE

In this section, discriminant analysis is presented with a brief formal description, including its main objectives and restrictions on variables management. This technique is commonly used in feature selection methods [2] (especially with numeric variables). This statistical technique is included in this study to establish a comparative framework for measuring the adapted attribute selection method provided in RBS algorithm.

Then, the method of generating rules, RBS, is introduced, so that in the next section it will be adapted to be able to manage with numeric variable.

2.1 Discriminant analysis

Discriminant analysis is a multivariate statistical technique to analyze whether there are differences between groups of objects (categories) over a set of (independent) variables measured on them. So that, the mechanism allows to reduce the number of independent variables and classify a future item whose variable values are known but the group which the item belongs to is unknown [3].

A linear combination of independent variables for the discriminant analysis, also called discriminant function, has the following form:

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + ... + W_pX_{pk} \quad (1)$$

Where:

Z_{jk} = discriminant score Z of the discriminant function for the object k.
a = constant (if it exists)
W_i = discriminant weight for the independent variable i.
X_{ik} = independent variable i for the object k.

Type of variables on discriminant analysis:

- Dependent variables: a qualitative variable (nominal) with as many discrete values as groups.
- Independent variables (classification or discriminant variables): variables with some type of relationship with the groups of the dependent variable. These variables can be numerical or nominal.

Objectives of the discriminant analysis:

- Determining rules for allocating individuals of the populations on a known classification.
- Finding the differences between the groups regarding the variables considered.
- Sorting individuals of unknown origin into one of the groups.
- Determining the best linear combinations (discriminant functions) of the independent variables for differentiating groups and thus classify new cases.

Stages of the discriminant analysis:

- Problem definition
- Selecting independent and dependent variables
- Selection of the sample size
- Testing the hypothesis

- Estimation Model
- Validation of discriminant functions
- Contribution of the variable to the discriminatory power
- Evaluation of the predictive function

2.2 Classification rules and feature selection using RBS

RBS is an algorithm for generating and ordering classification rules for discrete data, which incorporates a set of improvements to other algorithms of the same type, making it considerably faster. RBS [4] is an iterative algorithm that considers the rule support as the probability of a rule antecedent to occur, and the confidence as the conditional probability of a consequent to occur, given a specific antecedent. So, for a rule r=[$A \rightarrow Q$], the support and confidence can be expressed as follows:

$$Support(r=[A \rightarrow Q]) = \frac{N_A}{N}, \quad Confidence(r=[A \rightarrow Q]) = \frac{N_{A \rightarrow Q}}{N_A} \tag{2}$$

Where N is the total number of records in the dataset, N_A is the number of tuples where the rule antecedent occurs, and $N_{A \rightarrow Q}$ is the number of tuples where the rule antecedent and consequent occur. Fig. 1 plots the support and confidence of a hypothetical rule set. So, several borders are defined to divide the plane into four regions, so-named *REG-1*, *REG-2*, *REG-3* and *REG-0*. RBS finds the minimum and maximum borders of the support and confidence. Fig. 1 shows an example of borders positioning and rules distribution.

- Bs: Border of support that is calculated from the number of possible antecedents in a given rule set.
- Bs_U: Upper border of support, corresponding to the rule confidence average for rules over Bs.

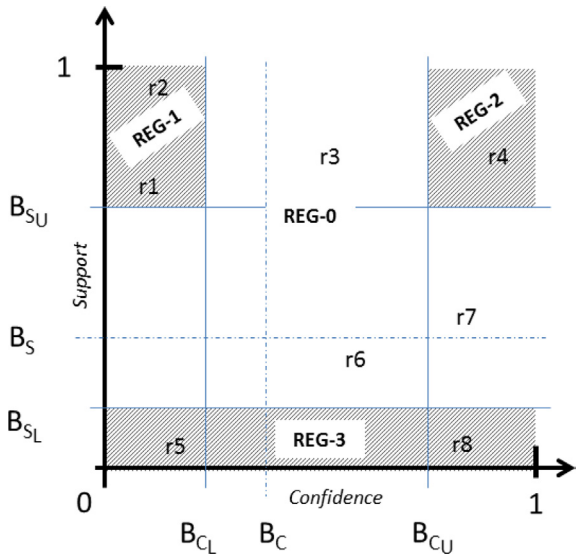

Figure 1: RBS regions.

- Bs_L: Lower border of support, corresponding to the rule confidence average for rules under Bs.
- Bc: Border of confidence that is calculated from the number of possible consequents in a given rule set.
- Bc_U: Upper border of confidence, corresponding to the rule support average for rules over Bc
- Bc_L: Lower border of confidence, corresponding to the rule support average for rules under Bc.

Moreover, RBS defines and calculates *waci* (weighted attribute correlation index) a measurement of the importance of the selected group of attributes for each rule set system Rr.

$$waci(Rr) = \frac{w_1 \mid REG-1 \mid + w_2 \mid REG-2 \mid + w_3 \mid REG-3 \mid}{\mid R \mid (w_1 a(REG-1) + w_2 a(REG-2) + w_3 a(REG-3))} \tag{3}$$

Where $|REG\text{-}i|$ represents the total of rules in region i, w_i is the assigned weight (parameter) for the region i, and $a(REG\text{-}i)$ is the surface of region i.

Thus, *waci* can be considered as a measure of variable correlation to model a given class variable. So, RBS provides both a list of the most significant combination of variables (to classify the target variable) and the classification rule sets for each one of those combinations.

Although RBS has been successfully applied to several classification problems, e.g., in Medicine [5] it has a very important constraint: it only manages nominal variables.

3 COMPUTATIONAL EXPERIMENT

3.1 Dataset definition and semi-synthetic dataset generation process

In order to compare discriminant analysis and RBS methods, a set of eight synthetic datasets have been simulated. These datasets are similar to real data of hotel booking management on East coast of Spain. Each dataset has the following eight variables:

Antecedent variables:

- Season: off season, Christmas, Easter and summer.
- Advance in reserve: number of days before booking.
- Advance in reserve (discrete): <7, 8–15, >15 (days).
- Accommodation days: number of days at the hotel.
- Accommodation days (discrete): <2, 3–7, 8–15, >15 (days).
- Country: Spain (SP), Germany (GE) and United Kingdom (UK).
- Room type: individual (ind), double (dob), triple (tri) and suite (sui).

Consequent variable:

- Daily spending (discrete): <70, 70–100, 101–140 and >140 (euros).

Some characteristics of the simulation:

- Nearly 60% of the rooms are booked in summer.
- The least booking season is Christmas.
- About 50% of the accommodations are booked more than 15 days before.
- About 55% of the rooms are booked for more than 15 days of accommodation and they are in summer.

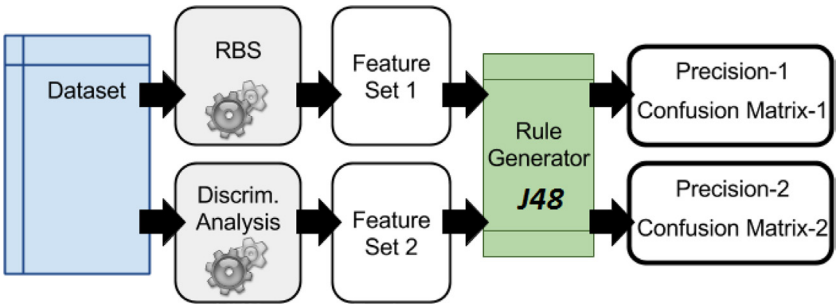

Figure 2: Research method.

- Countries are uniformly distributed.
- Individual rooms are booked less than the rest.
- Some random noises are introduced to obtain different datasets.

Different sizes of datasets are simulated to make computational tests. Sets of 1,000, 10,000, 100,000 and 1,000,000 records (tuples) have been generated with the above-mentioned characteristics.

3.2 Experiment overview

Each of the datasets is processed with discriminant analysis method and RBS method, providing their corresponding sets of selected features which are used for the generation of two different classification rule sets. Each of those rule sets reaches specific precision ratios that are compared through their corresponding confusion matrices provided by J48 Classification Tree. J48 is a Java implementation (on WEKA software [6]) of the original C4.5 Classification Tree algorithm that is an updated version of the classic ID3 [1]. C4.5 (also its corresponding J48 implementation) manages both nominal and numerical (categorical) antecedents.

The discriminant analysis has been implemented by using R statistical software version 3.2.3 [7] with the package MASS [8], while RBS has been implemented by using C language. Both of them have been developed in a virtual server with Linux Centos v6 (64 bits) Operating System, 2 CPUs to 2.8 GHz and 2 GBytes RAM.

4 EMPIRICAL COMPARISON

4.1 Quantitative comparison

Computing the datasets of 100,000 records, both R discriminant analysis and feature selection of RBS take between 1 and 2 seconds. Therefore, there are not substantial differences in computation time over 100,000 records datasets. (Nor substantial differences between the accuracies of these methods over 100,000 files tuples are appreciated) (See Table 1).

Computing the 1,000,000 records dataset, R discriminant analysis takes 8.25 seconds on average, while RBS feature selection takes 12.47 seconds on average.

However, although times of discriminant analysis method performed with R are slightly better than those obtained by Feature Selection of RBS, the former provides only the optimal combination of variables, while the second provides the 20 best combinations and also their corresponding classification rule sets.

Table 1: Classification accuracy (%) for discriminant analysis and RBS method. Empirical comparison to 100,000 tuples datasets.

Datasets	1	2	3	4	5	6	7	8
Discriminant analysis	84,09	84,21	84,05	84,37	84,24	84,13	83,97	83,81
RBS Feature selection	84,18	84,03	83,86	84,53	84,33	83,85	84,02	83,78

Table 2: Computing time (seconds) for the 8 datasets of 100,000 records.

Datasets	1	2	3	4	5	6	7	8
Discriminant analysis	0.5	0.6	0.7	0.6	0.7	0.5	0.6	0.5
RBS Feature selection	1.5	1.4	2.0	1.5	1.6	1.5	1.7	1.4

Table 3: Computing time (seconds) for the 8 datasets of 1,000,000 records.

Datasets	1	2	3	4	5	6	7	8
Discriminant analysis	7.2	7.4	7.9	8.1	7.5	7.4	10.3	10.2
RBS Feature selection	12.2	12.1	12.2	12.1	12.6	12.1	13.3	13.1

Next, Tables 2 and 3 contain the time required for both methods for computing the 8 datasets of 100,000 records and 1,000,000, respectively.

4.2 Qualitative comparison

As described above, the comparisons of both methods (discriminant analysis and RBS) in terms of time computing and accuracy of generated rule systems, lead to very similar results with datasets of 1,000; 10,000; even 100,000 tuples (Table 1). Therefore, in order to carry out a qualitative comparison, it is necessary to increase the size of the datasets up to 1,000,000 tuples (where the computation time begins to be different in each method). So, focusing on 1,000,000 records files, the most significant variables provided by both methods are essentially the same:

In the discriminant analysis experiment, all variables are included in the discriminant model. This analysis provides more or less the same conclusion about the discriminant functions. Near of 88% of the between-group variance is explained by the first discriminant function.

RBS provides 8 tables (one per each dataset) as the one shown in Table 4, where the first column contains the number of the variables combination (from 1 to 20) ordered by *waci* (see expression (3)). Also, in brackets, the quantity of variables that forms such combination is shown. The last column shows the *waci* for each combination of variables.

So, the best variable to predict the target variable is (accommod.days.D). If two variables must be considered, the best combination is (season, accommod.days.D) with a very similar correlation index. The best three variables combination is formed by (accommod.days.D, season, advan. reserve.D).

Analogously, for each dataset (from Dataset 2 to Dataset 8) the best combinations of variables are computed, providing very similar results, with the following exceptions: For Dataset 5, the best

Table 4: Best 20 variables combinations for Dataset with 100,000 records.

#(nº)	season	advan.reserve.D	accommod.days.D	country	room	waci
#1 (1)	–	–	accommod.days.D	–		0.510
#2 (1)	season	–	–	–		0.502
#3 (2)	season	–	accommod.days.D	–		0.502
#4 (1)	–	advan.reserve.D	–	–		0.388
#5 (1)	–	–	–	country		0.209
#6 (1)	–	–	–	–	–	0.183
#7 (2)	season	advan.reserve.D	–	–	–	0.175
#8 (2)	–	–	accommod.days.D	country	–	0.170
#9 (2)	season	–	–	country	–	0.167
#10 (2)	–	advan.reserve.D	accommod.days.D	–	–	0.148
#11 (2)	–	advan.reserve.D	–	country	room	0.129
#12 (2)	season	–	–	–	–	0.125
#13 (3)	season	advan.reserve.D	accommod.days.D	–	–	0.125
#14 (3)	–	advan.reserve.D	accommod.days.D	country	–	0.107
#15 (3)	season	advan.reserve.D	–	–	–	0.101
#16 (3)	season	–	accommod.days.D	country	–	0.099
#17 (2)	–	–	–	country	room	0.096
#18 (3)	season	advan.reserve.D	–	country	–	0.092
#19 (2)	–	–	accommod.days.D	–	–	0.091
#20 (3)	season	–	accommod.days.D	–	room	0.084

combination of three variables includes country, instead of advan.reserve.D (accommod.days.D, season, country). Also, for Dataset 6, whose combination of three variables includes room, instead of accommod.days.D (room, season, advan.reserve.D).

Given the great similarity between the results obtained and in order to simplify the interpretation of results, a particular file of 1,000,000 records Dataset 2 has been chosen. For this dataset, the most significant variables provided by each method are as follows.

(i) J48 classification tree

In this method, we use the full set of variables (season, advance in reserve, accommodation days, country and room type) and by using WEKA [6] for the generation of a J48 classification tree, the final model is characterized as follows:

Correctly Classified Instances: 826,127 (82.6127%)

Thus, e.g., first row in Table 5 means that 75,346 instances from the total of rules that must be classified as "<70", were correctly classified, while 5,511 of them were incorrectly classified as "70–100"; none incorrectly classified as "101–140" and 428 were incorrectly classified as ">140".

(ii) Discriminant analysis

This method does not remove any variable and provides the full set of variables (season, advance in reserve, accommodation days, country and room type) with the corresponding weight calculated for

Table 5: Confusion Matrix of J48 method using all original variables of the dataset.

"<70"	"70–100"	"101–140"	">140"	←classified as
75,346	5,511	0	428	"<70"
40,320	27,430	0	2,103	"70–100"
20,941	144	171,362	29,320	"101–140"
5,582	0	69,524	551,989	">140"

Table 6: Confusion matrix of discriminant analysis method for the dataset.

"<70"	"70–100"	"101–140"	">140"	←classified as
61,793	3,584	13,913	1,995	"<70"
27,058	17,798	17,851	7,146	"70–100"
28,162	252,000	162,546	30,807	"101–140"
6,531	279,000	20,666	599,619	">140"

Table 7: Confusion matrix of feature selection RBS method for the dataset.

"<70"	"70–100"	"101–140"	">140"	←classified as
25,699	5,939	48,754	893	"<70"
0	29,533	38,559	1,761	"70–100"
0	155	190,591	31,021	"101–140"
0	0	9,891	617,204	">140"

Table 8: Classification accuracy (%) for discriminant analysis and RBS method. Empirical comparison to 1,000,000 tuples datasets.

Datasets	1	2	3	4	5	6	7	8
Discriminant analysis	84.18	84.16	84.22	84.14	84.20	84.18	84.18	84.28
RBS Feat. Selection	86.30	85.71	86.89	85.89	86.77	85.87	86.66	86.22

this model. With the model for each dataset, this procedure can classify the instances. Following, Table 6 shows the confusion matrix after applying discriminant analysis for the feature selection over the dataset.

Correctly classified instances: 841,756 (84.1756%)

Thus, e.g., first row in Table 6 means that from the total of instances that must be classified as "<70", 61,793 instances were correctly classified, while 3,584 of them were incorrectly classified as "70–100"; 13,913 were incorrectly classified as "101–140" and 1,995 were incorrectly classified as ">140".

(iii) Adapted feature selection RBS

The best two variables combination consists of (accommod.days.D, season) because, with a very similar correlation index, considering two antecedent variables is better than considering only one. This is the selected combination to generate the classification model and measure its accuracy. By using WEKA for the generation of a J48 classification tree, the final model is characterized as follows:

Correctly Classified Instances: 863027 (86.3027%)

Analogously to Tables 5 and 6, on Table 7 the first row means that from the total of rules that must be classified as "<70", 25,699 instances were correctly classified, while 5,939 of them were incorrectly classified as "70–100"; 48,754 were incorrectly classified as "101–140" and 893 were incorrectly classified as ">140".

Next, Tables 1 and 8 summarize the accuracy reached with discriminant analysis and the adapted RBS feature selection for the 8 datasets of 100,000 and 1,000,000 records, respectively.

5 CONCLUSIONS AND FUTURE RESEARCH LINES

Adapted RBS provides good computing time (though slightly higher than those provided by discriminant analysis) for automatic selection of features. However, it also provides their corresponding classification rule sets. This is a very significant qualitative advantage over methods that only reduce the set of attributes but do not provide their corresponding classification rule sets.

The accuracies achieved in classification using the variables provided by RBS are better than the ones provided by discriminant analysis.

In future research, the authors will apply the feature selection to new problems, with many more attributes and will compare the achieved accuracy against the one reached by applying dimension reduction with classic statistical methods. Maybe, a principal components analysis based on factorial analysis can reduce the dimension of the dataset to try again with discriminant analysis. This will be especially noticeable in the presence of a higher number of variables.

RBS is shown as a very accurate method of feature selection; however, it will be necessary to incorporate a method of discretization of numeric variables that allow the treatment of this type of variables from within the algorithm itself, rather than as a step in pre-processing stage. Also it is necessary to reduce the execution time when this technique will be applied in extremely large sets of data.

ACKNOWLEDGEMENTS

This research has been partially supported by the Research Grants Ignacio H. de Larramendi, MAPFRE Foundation (2016).

REFERENCES

[1] Quinlan, J.R., Induction of decision trees. *Machine learning*, **1**, pp. 81–106, 1986. http://dx.doi.org/10.1007/BF00116251

[2] Lê Cao, K.A., Boitard, S. & Besse, P., Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, **12**(253), pp. 1–16, 2011. http://dx.doi.org/10.1186/1471-2105-12-253

[3] Fisher, R.A., Use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**(2), pp. 179–184, 1936. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x

[4] Almiñana, M., Escudero, L.F., Pérez-Martín, A., Rabasa, A. & Santamaría, L., A classification rule reduction algorithm based on significance domains, *TOP*, **22**, pp. 367–416, 2012.
[5] Rabasa, A., Compañ, A., Agulló, J.J., Rodríguez-Sala, J.J., Santamaría, L. & Noguera, L., Data management for an anaesthesiology department optimization. *WIT Transactions on Information and Communication Technologies*, eds. A. Rabasa, C.A. Brebbia & A. Bia, WIT Press, 45, pp. 175–183, 2013.
[6] WEKA, *Waikato Environment for Knowledge Analysis*. Machine Learning Group at the University of Waikato: New Zealand, available at http://www.cs.waikato.ac.nz/ml/weka/
[7] Team, R., A language and environment for statistical computing. *R Foundation for Statistical Computing, R Core Team*, Vienna, Austria, available at http://www.r-project.org/
[8] Venables, W.N. & Ripley, B.D., *Modern Applied Statistics with S, 4th edn*, Springer: New York, 2002, ISBN 0-387-95457-0
http://dx.doi.org/10.1007/978-0-387-21706-2