The importance of adequate data pre-processing in early diagnosis: classification of arrhythmias, a case study

A. Rabasa[1], A. F. Compañ[2], J. J. Rodríguez-Sala[1] & L. Noguera[1]
[1]*Operations Research Center, Universidad Miguel Hernández de Elche, Spain*
[2]*Department of Surgery and Pathology, Universidad Miguel Hernández de Elche, Spain*

Abstract

Data management can become very complex in the context of forecasting medical problems. Data collection, storage and analysis require the highest level of accuracy possible. The successful application of data mining techniques for the early diagnosis of disease or dysfunctions is increasingly more frequent among the scientific communities. However, as in any analytical method, the precision and reliability of the models provided by these techniques is absolutely dependent on the input data. If the quality of these data is not sufficient, the final accuracy can be greatly reduced to the point that the system becomes somewhat unproductive. This paper describes the main problems and how they can be properly solved at the pre-processing stage. Some of issues addressed are, for example: the detection of missing values (due to incomplete records), identification of outliers (often due to errors in measuring or recording devices), and discretization of numerical variables (where the context allows or suggests trying numeric values as nominal segments). Considering a public data base for arrhythmia from the UCI Repository, this study uses free Data Mining software to parameterize and run forecasting models and execute several computational experiments that show how the accuracy of predictions vary according to how you implement the critical pre-processing stage. The paper concludes providing a generic procedure that aims to apply the pre-processing of data in a methodical way and depending on the problems presented by the input data, and how it should be integrated into a global process of data management.
Keywords: Data Mining, pre-processing, forecasting, medicine, arrhythmia.


WIT Transactions on Information and Communication Technologies, Vol 45, © 2013 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA130211

1 Introduction

1.1 Pre-processing basic concepts

The pre-processing stage is usually considered (Chen [1], Berry and Linnoff [2]) as a part of the general Data Mining schema (Figure 1), where, after the problem definition and data collection stages and prior to the model selection, the input data must be prepared to be adequately processed, at least from a formal and theoretical approach. However, in practice, basic pre-processing is implemented independently from the Data Mining model, while other pre-processing tasks are developed depending on it.

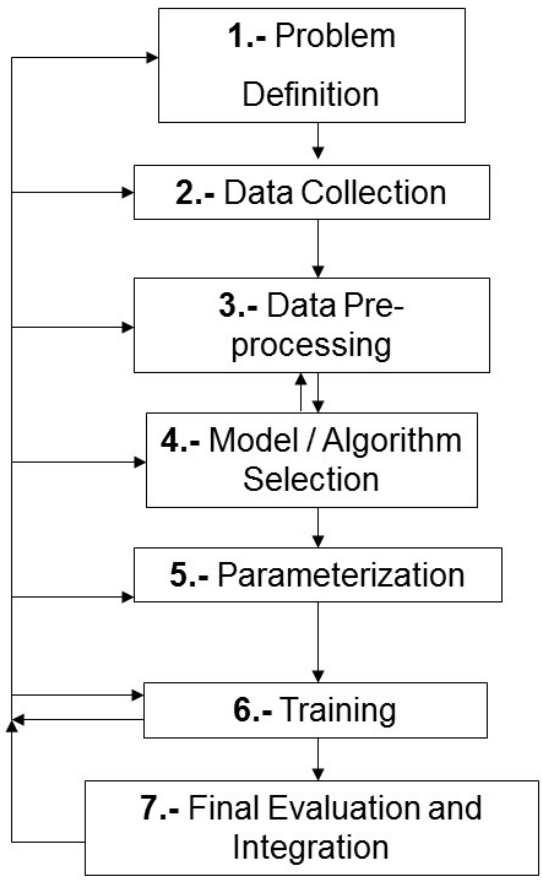

Figure 1: Pre-processing stages as a part of the Data Mining life cycle.

The most common pre-processing tasks are as follows: for example, outlier detection (and maybe the outlier's replacement), missing value detection (and maybe missing value completion), filtering attributes (statistically insignificant or high error ratio), filtering uncompleted records (with missing high ratio values), and discretization of numerical values.

Pre-processing becomes necessary in different contexts, for example, where different sources are joined, where noisy real time data must be processed, or even when final accuracy must be improved. This paper focuses on the last scenario.

1.2 Objectives

This paper presents two main objectives:

- To demonstrate how different data pre-processing procedures lead to different accuracy ratios.
- To provide a generic procedure for pre-processing.

1.3 Experiment guidelines

Different pre-processing procedures will be applied to the original Data Set (DS_0). Thus, a group of different pre-processed Data Sets will be obtained (DS_i). Each of these Data Sets will be subjected to a classification tree for a total of 10 times in order to calculate (by Cross Validation technique) the precision provided by the tree in each Data Set.

Assuming that the trees will be generated with the same criteria for construction, expansion and pruning, the differences between the accuracies achieved will only be attributable to the input Data Set in each case and they will depend entirely on the pre-processing which they have undergone.

Thus, it is possible to empirically establish how different pre-processing procedures lead to different levels of accuracy under the same predictive model. Figure 2 shows a general schema for the accuracy comparison, depending on the input Data Sets.

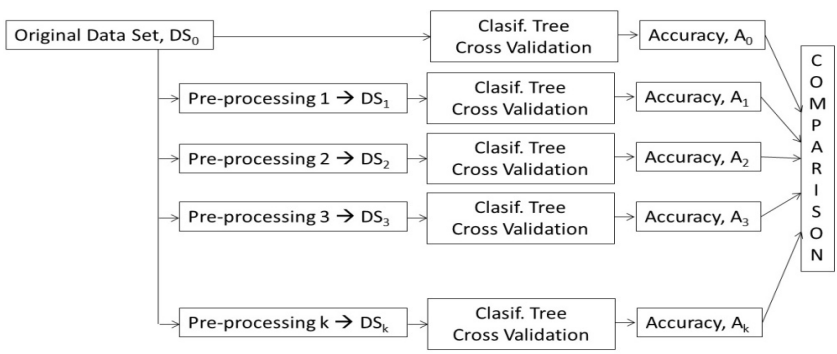

Figure 2: Accuracy comparison schema depending on different input Data Sets.

Both pre-processing and classification tree constructions are implemented using WEKA [3], the most extended Data Mining open access tool in academic frameworks.

2 Original Data Set description

Guvenir *et al.* [4] proposed an accurate method (the VF15 algorithm) for classification tasks, tested with their own cardiac data base. The original Data Set is available at UCI Machine Learning Repository [5]. The aim of their study was to classify the type of cardiac arrhythmia (there are 16 types for such a class variable). The original Data Set consists of 279 features (attributes or columns): Patients' personal data and the values registered by the electrocardiogram. Most of them (206) are linear attributes and the rest are nominal. There are several missing values. The Data Set consists of 452 patient records (rows or instances).

In this paper the same original Data Set (further DS_0) is used as input data, and is subjected to a classification tree. The achieved accuracy (A_0) will be considered as accuracy of reference.

Figure 3 shows the head of DS_0, the first 15 attributes (from 280) and the first 30 records (from 452), where missing values are denoted by '?'.

f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17	f18	f19	f20	f21	f22	f23	f24	f25	f26	f27	f28	f29	f30
75	0	190	80	91	193	371	174	121	-16	13	64	-2	?	63	0	52	44	0	0	32	0	0	0	0	0	0	44	20
56	1	165	64	81	174	401	149	39	25	37	-17	31	?	53	0	48	0	0	0	24	0	0	0	0	0	0	64	0
54	0	172	95	138	163	386	185	102	96	34	70	66	23	75	0	40	80	0	0	24	0	0	0	0	0	0	20	56	52
55	0	175	94	100	202	380	179	143	28	11	-5	20	?	71	0	72	20	0	0	48	0	0	0	0	0	0	0	64	36
75	0	190	80	88	181	360	177	103	-16	13	61	3	?	?	0	48	40	0	0	28	0	0	0	0	0	0	0	40	24
13	0	169	51	100	167	321	174	91	107	66	52	88	?	84	0	36	48	0	0	20	0	0	0	0	0	0	20	44	36
40	1	160	52	77	129	377	133	77	77	49	75	65	?	70	0	44	0	0	0	24	0	0	0	0	0	0	0	40	32
49	1	162	54	78	0	376	157	70	67	7	8	51	?	67	0	44	36	0	0	24	0	0	0	0	0	0	0	52	32
44	0	168	56	84	118	354	160	63	61	69	78	66	84	64	0	40	0	0	0	20	0	0	0	0	0	0	0	44	12
50	1	167	67	89	130	383	156	73	85	34	70	71	?	63	0	44	40	0	0	28	0	0	0	0	0	0	0	56	24
62	0	170	72	102	135	401	156	83	72	71	68	72	?	70	20	36	48	0	0	36	0	0	0	0	0	0	0	52	0
45	1	165	86	77	143	373	150	65	12	37	49	26	?	72	0	40	28	0	0	20	0	0	0	0	0	0	0	40	20
54	1	172	58	78	155	382	163	81	-24	42	41	-13	?	73	0	72	0	0	0	24	0	0	0	0	0	0	0	44	44
30	0	170	73	91	180	355	157	104	68	51	60	63	?	56	0	92	0	0	0	32	0	0	0	0	0	0	28	48	20
44	1	160	88	77	158	399	163	94	46	20	45	40	?	72	0	80	0	0	0	28	0	0	0	0	0	0	20	72	0

Figure 3: Extract of the original Data Set, DS_0.

3 Classification tree models and accuracy measurement

As the Data Sets pre-processing designs presented in section 4 are based on the accuracy achieved after being processed by the J48 classification tree, it is mandatory to present these concepts previously. Thus, in this section we introduce some classification tree concepts and the accuracy measure that will be used for each preprocessed Data Set, DS_i.

3.1 Classification trees concepts

C4.5, presented by Quinnlan [6] is probably one of the most extended for solving forecasting real problems in Medicine: Block *et al.* [7] presented a comparative

study of forecasting methods in Medicine, where C4.5 was pointed as a very accurate and suitable algorithm. C4.5 has been widely used, for example to predict the posology (Chan *et al.* [8]), or even for early cancer diagnosis (e.g. Polat *et al.* [9] and Takir and Bouridane [10] among others). C4.5 was first presented as a significant improvement to ID3, which only managed categorical attributes. C4.5 generates a Classification Tree with the input data sets by boosting at each node, based on information gain. Leafs of the tree correspond to the class variable instances. Each branch (from root to leaf) is interpreted as a classification rule.

WEKA provides an accurate Java implementation of C4.5: the J48 Classification Tree.

3.2 Accuracy measurement

In this paper, a Cross Validation methodology is used for measuring the classification accuracy reached by applying the J48 algorithm over each Data Set. Accuracy, A_i, associated to each Data Set, DS_i is assumed to be the average of ten executions of the J48 algorithm. Figure 4 shows A_0=64.3805 after applying Cross Validation over DS_0 classification.

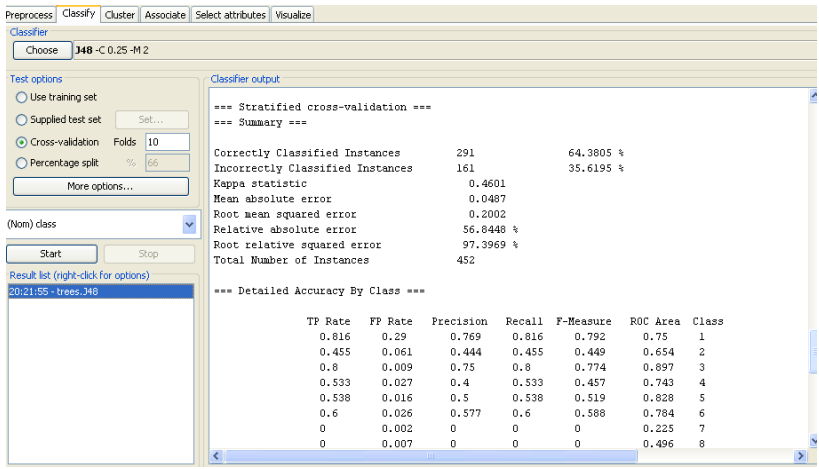

Figure 4: WEKA's cross validation over DS_0.

4 Generating different data sets from ad-hoc pre-processing

The original Data Set, DS_0, inputs into the J48 Classification Tree, and after a Cross Validation procedure, provides an accuracy ratio of reference A_0=64.38.

Next, DS_0 is subjected to different pre-processing routines under a trial and error methodology, in order to improve (or "approximately" maintain) the best accuracy ratio, A_i, achieved in previous steps.

First, the missing value detection notes that attribute f14 have an 83% missing value ratio. The rest of the columns have an acceptable missing value ratio (less than 50%). By deleting f14 attribute, we obtain DS_1. A_1=64.16 ($\approx A_0$). This is a very similar accuracy and it has one column less, so both DS_0 and DS_1 will be considered. For the next pre-processing variations, we focus on DS_1 (Figure 5).

Looking for outliers and extreme values, WEKA finds a set of attributes that could be avoided from DS_1, because there is a 10% outlier and extreme value ratio. So, by different parameterizing of outliers and extreme boundaries, we obtain DS_2 and DS_3, with A_2= 64.38 (=A_0) and A_3=61.95 (<A_0), respectively. DS_2 achieves the same accuracy level as DS_0 (with 15 attributes less), so DS_2 will be focused for the next pre-processing variations.

Sometimes the missing value replacement could improve the sample quality, so missing values on DS_2 are replaced with their corresponding attribute average value. This provides DS_4, with worse accuracy levels (A_4<A_0). In fact, there is no sense in replacing missing ECG values for other patients. Even by including f14 (with 83% missing) in DS_2 to obtain DS_5, the accuracy does not improve (A_5=64.38).

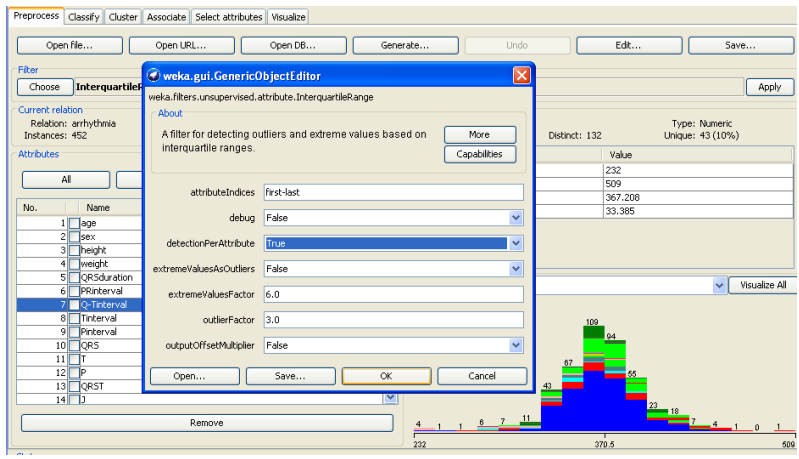

Figure 5: Outlier detection over DS_0.

This study faces a very broad problem (280 attributes per 452 records). If accuracy was maintained, a reduced set of attributes would be preferable because it would generate reduced classification rule sets. Thus, it seems to be necessary to select the most important attributes. Automatic feature selection over DS_0 and DS_2 provides DS_6 and DS_7, respectively.

Also, after applying automatic discretization DS_9 and DS_{10} were created (over DS_0 and DS_6 respectively: the Data Sets with best accuracies at the moment), both Data Sets were processed by the Classification Tree with A_9=64.38 and A_{10}=64.40. Neither improves A_6 accuracy.

Just as a test, the authors combined some of the Data Sets to increase accuracy, by considering extra attributes, removed in previous experiments. For example, defining DS_8 as the union of DS_6 and DS_7, the experiment equals the best accuracies obtained so far (A_6=68.36) but considering larger Data Sets with still too many attributes. Besides, although this case requires numerical data to be treated as numbers (and not as categorical), the authors tried to apply the WEKA's automatic discretization over the best Data Sets, DS_0 and DS_6, and neither improves accuracy.

Next, Table 1 summarizes the previously described pre-processing actions, and their respective accuracy levels.

Table 1: Pre-processing actions.

DS	*Pre-processing routine*	*Results*	*Accur.*
DS_0	No pre-processed	DS_0	64.38
DS_1	Missing values > 50% detection and attribute filtering.	$DS_0 - \{f14\}$ f14 has 83% missing	64.16
DS_2	Outliers (3) and Extreme Values (6) Detection >10%	DS_1-{f27, f53, f75, f87, f90, f123, f135, f170, f192, f210, f220, f223, f250, f260}	64.38
DS_3	Outliers (1,5) and Extreme Values (3) Detection >10%	DS_1-{f27, f51, f53, f75, f87, f88, f90, f101, f123, f135, f148, f170, f180, f192, f210, f220, f223, f250, f260}	61.95
DS_4	Missing values completed with average values	DS_2 with average missing values completed	62.61
DS_5	Recovering original missing values	DS_2 with original missing value $DS_2 + \{f14\}$	64.38
DS_6	Automatic Feature Selection over DS_0	{f5, f7, f8,f11, f15, f40, f76, f90, f93, f100, f103, f112, f114, f121, f190, f197, f211, f217, f222, f224, f228, f247, f248, f267, f277, f279}	**68.36**
DS_7	Automatic Feature Selection over DS_2	{f5, f7, f8, f11, f14, f38, f54, f66, f85, f87, f94, f97, f106, f108, f115, f181, f187, f206, f210, f215, f234, f252, f262, f264}	63.94
DS_8	*Automatic Feature Selection best union*	*$DS_6 \cup DS_7$*	*68.36*
DS_9	*Automatic discretization over DS_0*	*DS_0: all attributes are discrete*	*64.38*
DS_{10}	*Automatic discretization over DS_6*	*DS_6: all attributes are discrete*	*60.40*

Figure 6 shows the pre-processing experiment trace.

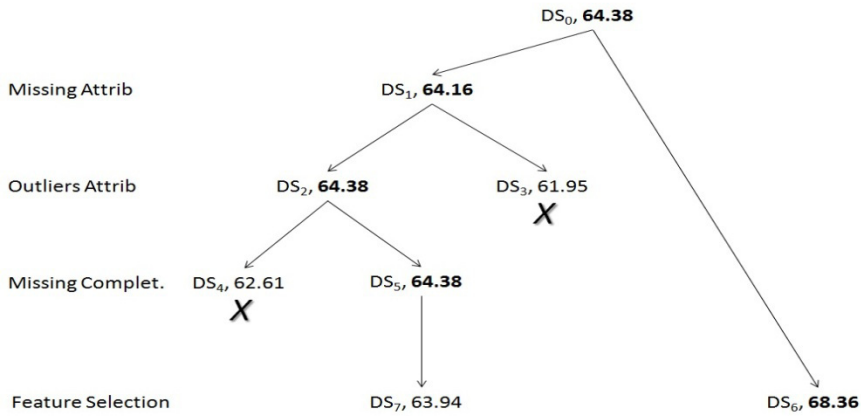

Figure 6: Pre-processing experiment trace.

5 Conclusions

5.1 An adequate pre-processing improves the final accuracy

- The accuracy initially reached (A_0=64.38) by processing the original Data Set with the J48 classification tree is significantly improved (A_6=68.36), if the same technique (equally parameterized) is applied to a well pre-processed Data Set.

- Missing values must be detected and if they appear very often at the same attribute, they must be removed. However, depending on the problem context, the missing values replacement is not suitable

- Some values, statistically considered as outliers, must be maintained in the Data Set. They may indicate certain pathologies. If wrong attributes are removed (even containing outliers), the final accuracy could worsen. So, A_3 is 2 percentage points worse than A_0.

- Where the amount of records is too small in comparison to the number of attributes, it is preferable not to delete any row and efforts must focus on the attribute selection.

- If numerical data could be discretized without loss of critical information, it could lead to much more precise forecasting, but it must depend on the expert´s criteria. In this case, the numerical attributes must be treated as numbers.

5.2 Generic procedure for pre-processing

The authors propose a fuzzy-greedy accuracy improving procedure under a trial and error methodology that progressively improves (or "approximately" maintains) the best accuracy ratio, A_i, as achieved in previous steps.

Some questions about the missing value treatment and replacement, the suitability of discretization, removing outliers and others must be considered.

Figure 7 shows a generic schema of such a procedure. The main questions are ordered and points are given about how to deal with attributes or records.

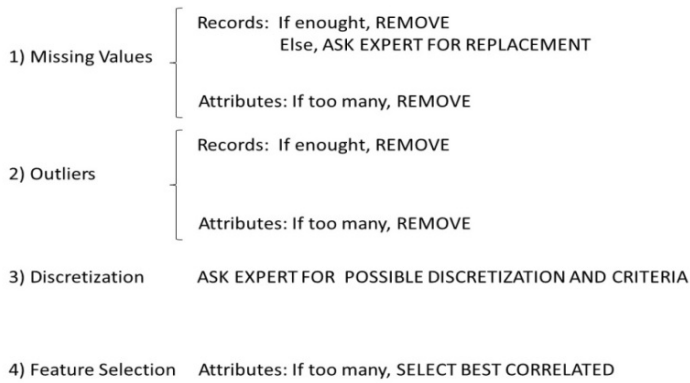

Figure 7: Generic procedure for pre-processing.

Acknowledgement

This study is part of the Bancaja-UMH project: “Minería de Datos sobre Análisis Preoperatorios para Servicios Hospitalarios Quirúrgicos y de Anestesia” (Data Mining with Presurgery Tests for Surgery and Anaesthesiology Hospital Departments).

References

[1] Chen, Z. *Data Mining and Uncertaining Reasoning*. An Integrated Approach. Wiley Interscience, 2001
[2] Berry, M.J. and Linoff, G. *Mastering Data Mining*. The Art and Science of Customer Relationship Management. Wiley, 2000
[3] Machine Learning Group at the University of Waikato, New Zeland. www.cs.waikato.ac.nz/ml/weka/
[4] Guvenir, H.A., Acar, B., Demiroz, G. and Cekin, A. A Supervised Machine Learning Algorithm for Arrhythmia Analysis. Proc. of the Computers in Cardiology Conference, Lund, Sweden, 1997
[5] Machine Learning Repository, University of California, Irvine, USA. archive.ics.uci.edu/ml/
[6] Quinlan, J. R. C4.5: *Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993
[7] Block, P., Paern, J., Hüllermeier, E., Sanschagrin, P., Sotriffer, C. and Klebe, G. Physicochemical Descriptors To Discriminate Protein–Protein Interactions In Permanent And Transient Complexes Selected By Means Of Machine Learning Algorithms. Wiley Inter Science, *Proteins: Structure, Function, and Bioinformatics* 65, pp. 607–622, 2006

[8] Chan, A.L., Chen, J.X. and Wang, H.Y. Application of Data Mining to Predict the Dosage of Vancomycin as an Outcome Variable in a Teaching Hospital Population. Dustri-Verlag. *International Journal of Clinical Pharmacology and Therapeutics* 44 (11), pp-533-538, 2006

[9] Polat, K., Sahan, S., Kodaz, H. and Gunes, S. A New Classification Method For Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System (FS-AIRS). Springer-Verlag. *Advances in Natural Computation 2, Proc. Lecture Notes in Computer Science* 3611, pp. 830-838, 2005

[10] Tahir, M.A. and Bouridane, A. Novel Round-Robin Tabu Search Algorithm For Prostate Cancer Classification And Diagnosis Using Multispectral Imagery. *IEEE-Inst. Electrical Electronics Eng. IEEE Transactions on Information Technology in Biomedicine* 10 (4), pp. 782-793, 2006