2006 IEEE International Conference on
Systems, Man, and Cybernetics
October 8-11, 2006, Taipei, Taiwan

Techniques for Missing Value Recovering in Imbalanced Databases: Application in a Marketing Database with Massive Missing Data

Luis E. Zárate, Bruno M. Nogueira, Tadeu R. A. Santos and Mark A. J. Song

Abstract— **Missing data in databases are considered to be one of the biggest problems faced on Data Mining application. This problem can be aggravated when there is massive missing data in the presence of imbalanced databases. Several techniques as imputation, classifiers and approximation of patterns have been proposed and compared, but these comparisons do not consider adverse conditions found in real databases. In this work, it is presented a comparison of techniques used to classify records from a real imbalanced database with massive missing data, where the main objective is the database pre-processing to recover and select records completely filled for a further application of the techniques. It was compared algorithms such as clustering, decision tree, artificial neural networks and Bayesian classifier. Through the results, it can be verified that the problem characterization and database understanding are essential steps for a correct techniques comparison in a real problem. It was observed that artificial neural networks are an interesting alternative for this kind of problem since it is capable to obtain satisfactory results even when dealing with real-world problems.**

I. INTRODUCTION

In the last years, KDD (*Knowledge Discovery in Database*) [1], where Data Mining is inserted, has been applied to several scientific and marketing segments. It could be mentioned areas such as industry, finance, health, telecommunications, business and many others, always aiming the non-obvious knowledge discovery and the supporting the taking of decisions.

Databases in which KDD process is applied frequently have missing data occasioned by non-controlled circumstances. Missing data are the ones which values were not added to database but for which a real value exists on the ambient that they were extracted. The presence of missing data on a database is a common fact and could be distributed in various attributes, in a same instance (record) or in a random form. Missing values could generate serious problems on knowledge extraction and on Data Mining algorithms application [2], [3], [4].

During the knowledge discovering process on a database, a very common procedure to deal with missing values consists in eliminating those attributes or instances that contain missing data, imposing restrictions to the extracted knowledge. The instances (records) or attributes elimination could result in loss of important information related to present values [4].

Other procedures suggest the missing values substitution by default values or mean values on all occurrences (imputation) [5]. However, those techniques are applied only when the number of missing values is short. Moreover, the substitution by a default value can introduce distorted information, which is not present on the event and circumstances that generate these instances [2].

Values recovery becomes, then, an extremely important point on knowledge discovery process on database, requiring careful value predictions using more advanced and elaborated techniques and procedures, together with the tacit knowledge of a problem domain expert and the preprocessing of the database [6]. All these procedures and techniques must aim not to distort information.

A possible solution for values recovering is to use classification models. To apply those techniques is necessary to select records completely filled and it can be a difficult work in the presence of databases with massive missing values in different attributes. Due to this fact, it is necessary a pre-processing of the databases to verify inconsistencies and to recover values, objectifying to obtain a set of representative records to be used by the classification algorithms. Some techniques to recover data, based on classifiers, were proposed and compared in [3]. However, most of the comparisons are made using artificial or public databases, which do not contain problems as found in real databases. This last aspect is discussed in [7].

One other alternative to recover missing value would be machine learning techniques application. Among those techniques it can be mentioned *Artificial Neural Networks* (ANN), which can learn relations between variables from instances shown to them during training process [4], [8].

One of the most common problems faced when analyzing real-world databases is an imbalanced database. They consist in those domains which contains one class represented by a minor number of examples than the others (1:100, 1:1000, 1:10000) [9], [10], [11], [12], [13], [14], [15]. When facing imbalanced datasets, some learning systems are not able to generate classification models capable to classify minority classes correctly. In order to solve this problem, some techniques were proposed to artificially balance databases. An algorithm, called *Consistent Sub-Set* (CSS) [10], will be used in this work. This algorithm balances the database removing class examples considered well defined.

The database used on the present work comes from a survey made with textile retail businessmen. Initially, each considered record is a store that answered the survey (634 stores) and each attribute is equivalent to a survey field

This work is financially supported by CNPq (National Councill for Scientific and Technologic Development- Brazil. Project CT - INFO/MCT/CNPq n 031/2004).

Authors are with Applied Computational Intelligence Laboratory, Pontifical Catholic University of Minas Gerais, Belo Horizonte - MG, Brazil zarate@pucminas.br

(totalizing 71). Due to different reasons, some survey fields were not filled, generating a database with great percentage of missing data (23,5%) and 0% of records completely filled. Among those data, it was verified that the field related to the store annual *<income>* contains a great absence percentage (33%). As a result of this field information importance to obtain the stores profile, the classification process in this work intends to estimate the income interval to which its values belong.

In this work it is presented a comparison of techniques used to recover values (annual <income>) on a real imbalanced database, with occurrence of massive missing data. This makes the process of obtaining a set of representative records, used for the recovering techniques, difficult. It was compared the efficiency of classification algorithms such as: based on pre-clustering (*K-Means*), decision tree (*C4.5*), Bayesian classifier (*Naïve Bayesian Classifier*) and on approximation of patterns through artificial neural networks (*Multi Layer Perceptron*). There were considered two situations on this comparison: one representative records set was chosen of randomly form directly of the imbalanced database and another set was chosen using the algorithm *Consistent Sub-Set* proposed in [10] to artificially balance the database.

II. UNDERSTANTING THE DATABASE

In [16] it was established that an ontological analysis through understanding and problem characterization of the domain are essential steps for a correct application of a KDD process and for an efficient techniques comparison in a real problem. In this section, the analysis of the database, assisted by a domain expert, is presented. The goal is to obtain a data structure, representative of the domain, which allows it to obtain a data set completely filled without missing values. Later, that set will be used by the data classification algorithms to estimate the income interval to which its values belong.

The database used on the present work comes from a survey made with textile retail businessmen, containing missing data in great proportions (23,5%) where 0% of the records are completely filled. Initially, each considered record is a store that answered the survey (634 stores) and each attribute is equivalent to a survey field (totalizing 71). Some main attributes of the database can be observed at Table I.

The first step is to identify the attributes related to the problem domain through data structure determination. The data structure can be supported by propositional logic that permits reasoning about variables relation into a problem domain. This will be presented on the next subsection.

A. Specifying Properties Through Propositional Logic

This section describes how propositional logic can be used to identify attributes related to ẗextile market salëdomain. Atomic propositions and their relationships are specified in order to define static structure (attributes) of the problem domain. The main idea is to find the set of attributes related to *<income>*.

TABLE I
EXAMPLE DATABASE MAIN ATTRIBUTES

Attribute	**Description**
<com_name>	Store commercial name
<address>	Store address
<owner_gen>	Store owner gender
<prop_name>	Store owner name
<region_appeal>	Store location region advantages
<cli_stratum>	Clients stratum
<income>	Annual income
<appeal>	Region appeal
<difficulty>	Region difficulties
<parking>	Parking
<evening_freq>	Evening client frequency
<charge_period>	Charge period
<client_type>	Client type

In order to write specifications that describe context properties is necessary to define a set of *atomic propositions AP*. An atomic proposition is an expression that has the form v op d where $v \in V$ - the set of all variables in the context, $d \in D$ - the domain of interpretation, and op is any relational operator. To describe sequences of transitions along time, Temporal Logic is a very useful formalism. With temporal logic it is possible to reason about the system in terms of occurrences of events. For example, it can be reason if a given event will *eventually* or *always* occur.

There are several propositions of temporal logic. These logics vary according temporal structure (*linear* or *branching-time*) and time characteristic (*continuous* or *discrete*). Temporal linear logics reason about the time as a chain of time instances. Branching-time logics reason about the time as having many possible futures at a given instance of time. Time is continuous if between two instances of time there is always another instance. Time is discrete if between two instances of time a third one cannot be determined.

For the considered database, it is used a branching-time and discrete logic known as Computation Tree Logic (CTL) to express properties of systems. CTL provides operators to be applied over computation paths. When these operators are specified in a formula they must appear in pairs: path quantifier followed by temporal operator. A *path quantifier* defines the scope of the paths over which a formula f must hold. There are two path quantifiers: **A**, meaning **all** paths; and **E**, meaning **some** paths. A temporal operator defines the appropriate temporal behavior that is supposed to happen along a path relating a formula f. For example:

- F ("in the future"or "eventually") - f holds in some point of the computation path;
- G ("globally"or "always") - f holds in all path;

A well formed CTL formula is defined as follows:

1. If $p \in AP$, then p is a CTL formula, such that AP is the set of atomic propositions;

2. If f and g are CTL formulas, then $\neg f$, $f \vee \_g$, $f \wedge g$, AFf, EFf, AGf, EGf, A[fRg], E[fRg], A[fUg], E[fUg], AXf, EXf, are CTL formulas.

The selection of a set of variables (attributes) relevant to the problem domain is entirely based on the variables relationships:

- AG((*Appeal* & ¬*Difficulty*) || (*Parking* & *Transport*)) → AF(*Income*));
- EF(*Evening Frequency* → EF(*Income*));
- A[*Charge Period* ∪ *Income*];
- AG((*Client Type* & *Frequency*) || *Client Stratum*) → AF(*Income*)).

Some attributes considered relevant to estimate the <income> interval to which its values belong, obtained through propositional logic, can be observed at Table I.

After identifying the main attributes of the data structure the following step is the database pre-processing. The aim is to obtain the set of representative records completely filled to apply the recovering algorithms.

III. PRE-PROCESSING OF THE DATABASE

In this section, the pre-processing stages as recovering missing values, attributes removal, inconsistence analysis and outliers analysis, applied to the considered database, will be discussed.

A. Missing Values Recovery by Related Attributes and Characteristics Deletion

Sometimes, it is possible to recover values that are not present on the database but are available in an indirect way. In the considered database, two actions were done:

- Values recovered by related attributes: on the database, it was possible to find some attributes representing the same information. So, it could be used the information from one of them to recover information of another attribute. For example, <owner_gen> attribute was recovered by <prop_name> attribute through gender analysis.
- Values recovered by default replacement: not filled attributes, which consider default values on the survey form were replaced by these values.

Some attributes do not have sufficient information to be considered. Thus, three criterions were established to eliminate attributes and records.

- Elimination of attributes containing short information. The entropy concept was applied for this aim [17].

$$H(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i \, log_2 \, p_i \qquad (1)$$

  Where: p_i is the s_i value occurrence probability to an attribute. Thus, for fields with nearly constant values: $p_i \approx 1$, so $H(s_i) \approx 0$. On the considered database, seven attributes were eliminated in this step.
- Attributes with large percentage of missing values do not help to characterize the problem domain. So, a threshold has been established to remove an attribute. On the present work, the threshold value was 25%, resulting a sole attribute <cli_stratum> being removed.
- As done with attributes, records containing great quantity of missing data have to be eliminated. On the database considered, a threshold value of 25% was established to eliminate records. Fortysix records were removed from database.

In databases with a great number of attributes, some of them can be eliminated after a relevance classification process.

- Facts are those attributes which importance to the problem domain is considered to be more relevant because they have essential information. On the other hand, judgments are lower importance attributes, from which it could not be extracted relevant information during data analysis and, though, can be unconsidered. The distribution of relevance degree contained at an attribute is shown in Fig. 1. After a classification of the attributes, assisted by a problem domain expert, attributes considered Highly-Fact (HF), Fact (F) and Judgment-Fact (JF) are considered necessary, while Judgment (J) and Highly-Judgment (HJ) are discarded.

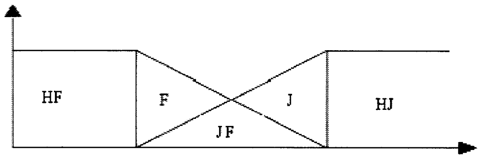

Fig. 1. Attribute importance classification

The removal is done in agreement with the following algorithm. On the example database a total of 27 attributes were removed.

```
FOR all attributes DO
    IF inf_atrib ==Judgment OR
       inf_atrib == Highly-Judgment
    THEN remove attribute
```

B. Attributes Transformation

In this section, it is presented a description of the transformations applied to the considered database, which contains dichotomic, nominal, categorical and ordinal attributes:

- The dichotomic fields were replaced by the equivalent binary number (0 or 1).
- The composite data <address> was transformed in data on the form (Longitude, Latitude) through GIS (Geographic Information System). To make possible a representative numerical classification to the addresses, the clustering technique was applied to the group the records (stores). The K-Means algorithm was chosen, defining 10 clusters a priori. Fig. 2 shows the stores grouping by localization.
- For attributes with multiple non-associated options, for example <region_appeal> - store location region advantages, each option was considered an independent attribute, being transformed to dichotomic form (marking = 1, non-marking = 0). So, the number of database attributes was expanded to 81. To other cases that consist in multiple choices of associated options it

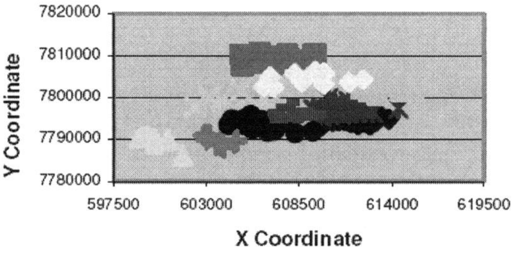

Fig. 2. Stores grouping by localization

is recommended the treatment of these data as circular data [18].

- Attributes separated by value ranges on the survey form or that consist on a simple option choice had arbitrary number associated to these values ranges or options. Four intervals were established to the attribute <income>, (2).

$$\begin{aligned} \text{Income } &\leq 61000 \in Inc_1 \\ 61000 < \text{Income} &\leq 123000 \in Inc_2 \\ 123000 < \text{Income} &\leq 377000 \in Inc_3 \\ \text{Income} &> 377000 \in Inc_4 \end{aligned} \qquad (2)$$

Due to the problem domain, on which the information falsity (annual income) is widely latent, the four income intervals were reduced to two intervals through the domain expert experience. The tacit knowledge based rule applied on this database is expressed as follows:

Considering the limit set as:

$$Limits = \{LimI_i \in \Re, LimS_i \mid LimI_i < LimS_i, \\ i = 1..N, \text{ com } LimI_{i+1} = LimS_i\} \qquad (3)$$

and the income interval ϕ taken as:

$$\Phi = \{F_i; LimI_i \leq F_i < LimS_i, i = 1..N\} \qquad (4)$$

The informed income value is defined as:

$$VInf = \{x \in F_k; p(x \in F_k) = p_k\} \qquad (5)$$

where p_k is the $x \in F_k$ probability.
If $x \in F_k$ is a false information, so:

$$p(LimI_{k+1} \leq x < LimS_n) > p(x \in F_k) > \\ p(LimI_1 \leq x < LimS_{k-1}) \qquad (6)$$

In other words:

$$p(x < LimS_N) - p(x \leq LimS_{k-1}) > p_k > \\ p(x < LimS_{k-1}) - p(x \leq LimI_1) \qquad (7)$$

Thus, it is possible to reduce the income intervals without significant information loss.The new set of income intervals $\phi*$ may be expressed as:

$$\Phi* = \{F_i; LimI_i \leq F_i < LimS_i, \, i = 1..k, N\} \qquad (8)$$

From which:

$$\begin{aligned} Income \leq 61000 \in Inc_1 * \\ 61000 < Income \in Inc_2 * \end{aligned} \qquad (9)$$

With the objective of compare the potentiality of the classification techniques, in this analysis were considered two and four intervals to the attribute <income>.

C. Inconsistent Data Identification and Outliers Analysis

Due to the marketing nature of the database, the problem domain expert alerted that false information is a common fact, generating inconsistencies mainly on the fields related to the stores income, becoming more difficult the correct information recovery process. In order to detect those records inconsistencies, the clustering technique K-Means was applied separately under the pre-classified income groups.

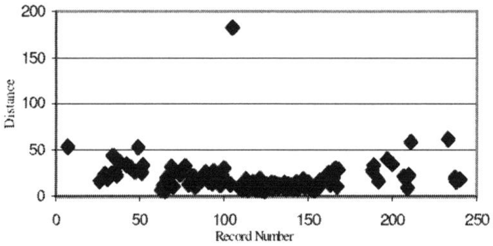

Fig. 3. Cluster 1 of income interval 1

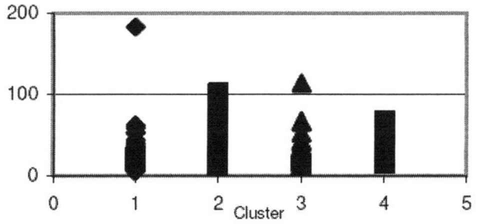

Fig. 4. Income interval 1 (4 clusters)

Fig. 3 shows the Cluster 1 for Inc_1 records distribution on the example database. Fig. 4 shows the distances between records and correspondent centroids to each interval. In both figures it is possible to observe the presence of discrepant elements (outliers), which were removed from the Inc_1 interval instances.

Objects removed from each interval obey the following rules:

$$RS_i = O_{jk} \in F_i | li_i \leq dist(O_{jk}, cent_{ki}) \leq ls_i \qquad (10)$$

Where: RS is the selected record set for the income interval "i"; F_i is the income interval number "i"; O_{jk} is the object "j"attached to cluster k; Cent_{jk} is centroid number for income interval number "i"; li_i is the inferior distance limit to the income interval number "i"; ls_i is the superior distance limit to the income interval number "i";

Each income interval Inc_i, li_i and ls_i is shown in Table II.

TABLE II
LIMIT DISTANCES TO RESPECTIVE CLUSTERS CENTROIDS

Inc_i, for i =	li_i	ls_i
1	0	60
2	10	20
3	20	50
4	20	30

IV. OBTAINING A SET OF REPRESENTATIVE RECORDS

On the considered database, after the pre-processing, 257 records (40% of the original database) and 81 attributes completely filled were obtained. As the information being recovered in that database is the one related to the stores annual income, the attribute <income> was taken as the goal of the classifiers algorithms, according to (2) and (3).

The records set was divided in two subsets: a training subset, which is presented to the classifiers in order to obtain a classification model, and a validation subset, which is used on the classifier accuracy. However, it is important to notice that the considered database contains much more records from one class than the other classes, generating an imbalanced domain that can induce some classifiers to error. Fig. 5 shows the heterogeneous distribution of the amount of records for the four income intervals.

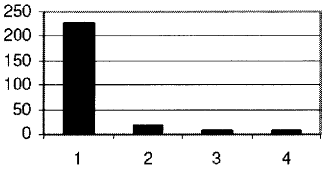

Fig. 5. Income interval x Frequency

After Consistent Sub-set Algorithm (CSS) proposed in [10], for artificial balancing of the databases, 110 records were selected as a set of representative records. The tested conditions are shown in Table III:

TABLE III
CONDITIONS TESTED

Test	Interval Number	Training Records Number	Validation Records Number	Selection Method
T1	2	110	147	Random
T2	4	110	147	Random
T3	2	110	147	CSS
T4	4	110	147	CSS

V. TECHNIQUES OF CLASSIFICATION

In this section, four representative classification techniques were analyzed: Clustering (K-Means), Decision Tree (C4.5), Artificial Neural Networks (Multi Layer Perceptron trained by backpropagation) and Bayesian Classifier (Naïve Bayesian Classifier).

A. Artificial Neural Networks Applications

To the considered database, two Multilayer Neural Networks [19] were trained. The networks were composed by perceptron neurons disposed in one hidden layer and one output layer totally connected. The used networks have 81 inputs each, with 4 and 2 outputs and 60 neurons on the hidden layer. The sigmoid function was chosen as activation function. All considered inputs represented by the E set (Table I) were mapped on 4 and 2 binary outputs mutually exclusive on the ANN (11)

$$f(E) \overset{ANN}{\rightarrow} (Inc_1, Inc_2, Inc_3, Inc_4) \tag{11}$$

For the neural networks training process, the numerical input data have to be submitted to a normalization process:

- Intending to improve the training process convergence, the data have to obey the normalization interval [0.2, 0.8];
- Data were normalized following the expressions:

$$\begin{aligned} f^a(L_0) &= L_n = (L_0 - L_{min})/(L_{max} - L_{min}) \\ f^b(L_n) &= L_0 = L_n * L_{max} + (1 - L_n) * L_{min} \end{aligned} \tag{12}$$

L_{min} and L_{max} were computed as follows

$$\begin{aligned} L_{min} &= L_{sup} - (N_s/(N_i - N_s)) * (L_{inf} - L_{sup}) \\ L_{max} &= ((L_{inf} - L_{sup})/(N_i - N_s)) + L_{min} \end{aligned} \tag{13}$$

where: L_{sup} is a variable maximum value, L_{inf} is the minimum value and N_i and N_s are the normalization limits (in this case, N_i = 0.2 and N_s = 0.8).

To begin the training process, random values between -1 and 1 were set on the connections weight.

As soon as the training process is finished, this training has to be validated. In this sense, the validation set (see Table III) was applied onto the neural network and its generated outputs were analyzed. For the tests T1 and T2, considering two and four income intervals, the accuracy rate was 89.88% and 87.82% respectively. For the tests T3 and T4, with artificial balancing of the databases, the accuracy rate was 98.05% and 91.82% respectively.

B. Decision tree application

Decision tree is one data-mining technique applied in many real-world applications as a powerful solution to classification problems. They use supervised learning methods that construct decision trees from a set of inputoutput samples. The algorithm used on the present work was the largely used C4.5 algorithm.

After the decision trees were generated, validation sets were presented to them in order to test the efficiency on the income classification. Considering the tests T1 and T2, the accuracy rate was 89.1% and 87.15%. For the tests T3 and T4, with artificial balancing of the databases, the accuracy rate was 91.05% and 91.43% respectively.

C. Cluster analysis application

The other technique used on the records classification was the classification pos - clustering. In order to group similar records, it was applied the K-Means algorithm, the simplest and most commonly used algorithm employing the Euclidean distance criterion.

Using this technique for the tests T1 and T2, the accuracy rate was 91.43% and 30.35%. For the tests T3 and T4, with artificial balancing of the databases, the accuracy rate was 91.82% and 91.05% respectively.

D. Bayesian Classifier Application

Bayesian classification is based on the Bayes Theorem, which consists in a mathematical formula used to conditional probabilities calculus. The Naïve Bayesian Classifier, or Simple Bayesian Classifier, was used in this work. It assumes the existence of conditional independence between attributes and calculates the occurrence probability of a data sample given on a class. This classifier can be formally expressed as:

$$P(X/C_i) = \prod_{t=1}^{n} P(x_T/C_i) \tag{14}$$

Where:
X is a data sample whose class label is unknown;
C_i is a class value;
x_i are values for attributes in X;
$P(X / C_i)$ is the occurrence probability of the sample X given the class C_i;
$P(x_t / C_i)$ is the occurrence probability of the value x given the class C, that can be extracted from training dataset.

Thus, a new record can be classified by calculating the occurrence probability of the given data sample to each class, and assuming that the record belongs to the class with the major occurrence probability.

For the tests T1 and T2, considering two and four income intervals, the accuracy rate was 86.38% and 78.6% respectively. For the tests T3 and T4, with artificial balancing of the databases, the accuracy rate was 72.09% and 84.04% respectively.

Fig. 6 and Fig. 7 show the accuracy rate obtained by applied techniques.

The techniques comparison results shown in Fig. 6 and Fig. 7 show that Artificial Neural Networks and Decision Trees had obtained much stable results, dealing well with the problem of imbalanced databases with two or four income intervals. The Artificial Neural Networks had obtained the expressive recovering rate of 98.05% of the missing data in the experiments with the records set obtained through balancing algorithm.

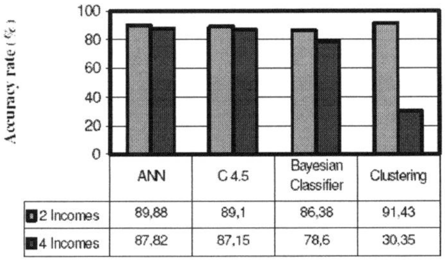

Fig. 6. Results obtained using the randomly selected training set (T1 and T2)

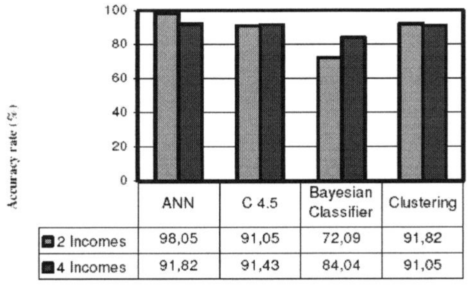

Fig. 7. Results obtained using the balanced training set (T3 and T4)

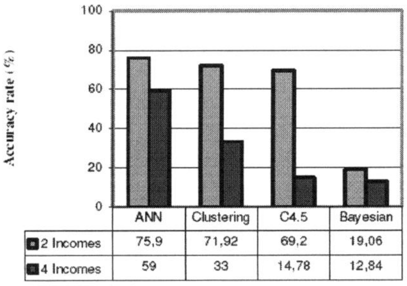

Fig. 8. Results obtained using a minor records set number

It can be observed that the clustering technique was highly sensible to the effects of imbalanced database, caused by the records of the major class.

Other important fact that can be noticed is the instability of the Bayesian classifier when it is compared with the other techniques. This occurs because the algorithm needs that the records in the training set have large variety on its field values. If it happens, the classifier cannot calculate all the possible values probability and it will not classify a certain record that have a field value that was not present in the training set.

To verify the potentiality of ANN to classify records, a new test to evaluate its generalization capacity was executed. The objective is to observe if ANN can obtain the correct class values with a minor records number.

Fig. 5 shows a heterogeneous distribution of the amount of records for the four income intervals. Intending not to polarize the neural network by the random selection of sets, 6 records of each interval were randomly selected, generating a sum of 24 training sets. The other records were reserved to the network validation process.

For the four income intervals the average learning rate was 0.4 and after 1,800,000 iterations a 0.03 global error was reached, generating a hit for all the 24 training set. To validate this network, 233 sets were chosen, obtaining a hit value of 59%. To the two income intervals, the average learning rate was 0.4 and after approximately 10,000 iterations a 0.03 global error was reached, generating a hit tax for the 20 training sets (10 of each income interval) of 100%. To validate the network, 237 sets were chosen generating a hit tax of 75.9% in 218 records that indicate Inc_1 and 42.1% in 19 records that indicate Inc_2.

Clustering, Decision tree and Bayesian Classifier were compared, too. Fig. 8 shows the results obtained. It is possible to observe that these techniques lose their efficacy when the representative records number is short.

VI. CONCLUSIONS

The presence of missing values on a database is a common fact and can generate serious problems on the knowledge extraction and on the Data Mining algorithms application. On the other hand, the instances / attributes with missing data elimination may cause the information loss and replacement by a default value may introduce distorted information on the base, which do not belong to the events and circumstances that generated it.

There are some techniques that can face missing data, but most of them fail when it exists a massive data absence. In this work the considered database has 23.5% of missing value and 0% of completely filled records. Due to this reason, it was necessary the construction of classification models from the completely filled records, to be able to recover the attribute store income interval.

Experiments show how the domain expert helps to define the data structure for an effective recovering process.

Comparing the result of the techniques applied to imbalanced databases, with the results of the same techniques applied to balanced databases, it can be observed the importance of the database balancing stage in the KDD preprocessing activity, especially when dealing with realworld databases. The preprocessing activity is considered to be one of the most important of the KDD process, where generally it is required much attention to guarantee a good final result.

The activities and methods shown in the paper were, in general, efficient in the task of records classification when facing an imbalanced database. Among these techniques, the Artificial Neural Networks and the Decision Trees, in the overall, were better capable when dealing with the problems present in a real world database. The ANN result of 98.05% correct classification when facing an imbalanced database was considered highly expressive.

Analyzing the Artificial Neural Networks results, it can be observed that, when using four income intervals, ANN had an accuracy rate much higher than the one obtained with the other techniques. This shows that ANN is better capable to face imbalanced databases.

Moreover, ANN due their capacity of generalization shown a superior performance when the number of records, considered to estimate the income, is few (20 records). This can be a fact when a real database is considered.

VII. ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of CNPq (National Councill for Scientific and Technologic Development- Brazil).

REFERENCES

[1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, MIT Press, USA, 1996.

[2] D. Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann, USA, 1999.

[3] P. Liu, L. Lei and N. Wu, "A quantitative study of the effect of missing data in classifiers", *in CIT'05: Proceedings of the 2005 Fifth international conference on Computer and Information Technology*.

[4] Y. Fujikawa, "Efficient Algorithms for Dealing with Missing values in Knowledge Discovery", *School of Knowledge Science - Japan Advanced Institute of Science and Technology*. Japan, 2001.

[5] M. Kantardzic, *Data Mining - Concepts, Models, Methods and Algorithms*, IEEE Press, USA, 2003.

[6] M. Hofmann and B. Tierney, "The involvement of human resources in large scale data mining projects", *in Proceedings of the 1st international symposium on Information and communication technologies*, Ireland, 2003, pp. 103 - 109.

[7] C. Soares, "Is the UCI Repository Useful for Data Mining?", *in Lecture Notes in Computer Science*, Volume 2902, Jan 2003, Pages 209 - 223.

[8] L.E. Zárate, B.M. Nogueira and T.R.A Santos, "Recuperação de Dados Ausentes Através de Redes Neurais artificiais - Estudo de Caso para uma Base de Dados Mercadológica", *in Congresso Brasileiro de Redes Neruais, CBRN 2005*, Natal, RN, Brazil.

[9] N. Japkowicz, "Concept-Learning in the Presence of Between- Class and Within-Class Imbalances", *in Proceedings of the Fourteenth Conference of the Canadian Society for Computational Studies of Intelligence (AI'2001)*, Canada, 2001.

[10] R.C. Prati, G. Batista and M.C. Monard, "Uma Experiência no Balanceamento Artificial de Conjuntos de Dados para Aprendizado com Classes Desbalanceadas utilizando Análise ROC", *in Proceedings of IV Workshop on Advances & Trends in AI for Problem Solving*, Chilán, 2003.

[11] G. Batista, R.C. Patri, M.C. Monard. "A Study of the Behaviour of Several Methods for Balancing Machine Learning Training Data", in *SIGKDD Explorations*, June 2004, Vol. 6 Issue 1, pp.20-29.

[12] G.M. Weiss, "Mining with Rarity: A Unifying Framework", *in SIGKDD Explorations*, June 2004, Vol. 6 Issue 1, pp.7-19.

[13] H. Guo and H.L. Viktor, "Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach", *in SIGKDD Explorations*, June 2004, Vol. 6 Issue 1, pp.30-39.

[14] T. Jo and N. Japkowicz, "Class Imbalanced versus Small Disjunct", *in SIGKDD Explorations*, June 2004, Vol. 6 Issue 1, pp.40-49.

[15] C. Phua, D. Alahakoon and V. Lee, "Minority Report in Fraud Detection: Classification of Skewed Data", *in SIGKDD Explorations*, June 2004, Vol. 6 Issue 1, pp.50-59.

[16] P. Gottgtroy, N. Kasabov and S. Mcdonell, "An Ontology engineering approach for knowledge discovery from data in evolving domains", *Knowledge Engineering and Discovery Institute*, Auckaland University of Technology - New Zealand.

[17] C.E. Shannon, "The Mathematical Theory of Communication", *in Bell System Technical Journal*, 1948.

[18] N.I. Fisher, *Statistical Analysis of Circular Data*, Cambridge Univesity Press, Australia, 1995.

[19] S. Haykin, *Redes Neurais - Princípios e Práticas*,Bookman, Brazil, 2001.