

A feature selection Bayesian approach for a clustering genetic algorithm

E. R. Hruschka Jr^{1, 3}, E. R. Hruschka² & N. F. F. Ebecken¹

¹*COPPE / Universidade Federal do Rio de Janeiro, Brasil*

²*Universidade Católica de Santos (UniSantos), Brasil*

³*Universidade Tuiuti do Paraná, Brasil*

Abstract

Feature selection is an important task in clustering problems. Some features help to find useful clusters whereas others may hinder the clustering process. In other words, some selected features can provide better clusters. Besides, the feature selection process also allows the reduction of the dataset dimensionality, improving the clustering method efficiency. This work describes a Bayesian feature selection approach for a Clustering Genetic Algorithm (CGA). The general method can be described by means of four steps: (i) apply the CGA to some selected objects (sample) of the complete dataset; (ii) consider that the obtained clusters form different classes, which can be modeled by Bayesian networks; (iii) generate a Bayesian network and employ the Markov Blanket of the class variable to the feature subset selection task; (iv) apply the CGA in the complete dataset now formed only by the selected features. Initially, we are mainly interested in evaluating if the feature selection process makes sense in the context of the CGA, which can find the *best* clustering in a dataset according to the Average Silhouette Width criterion. Thus, our first investigation supposes an ideal situation, where the CGA has actually found the right clustering in step (i). Thus, the Bayesian networks are generated not in a sample, but in the complete dataset correctly clustered/classified. In this way we can better evaluate if the proposed hybrid method is appropriate, i.e. if the features selected by means of Bayesian networks are suitable for the CGA. In this sense, we performed simulations in three datasets that are benchmarks for data mining methods - Wisconsin Breast Cancer, Mushroom and Congressional Voting Records. The results obtained in the simulations performed in the datasets formed by the selected features provided better results than those obtained in the complete datasets. Thus, we believe that the proposed method is very promising.

1 Introduction

Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [1]. Data mining is a step in this process that centers on the automated discovery of new facts and relationships in data and it consists of three basic steps: data preparation, information discovery and analysis of the mining algorithm output [2]. The data preparation step can be performed in many ways. The main techniques are data cleaning, integration, transformation, reduction, discretization and generation of concepts hierarchies [3]. Feature selection is an important method to the data reduction task and, as shown in [4], it can help in all the phases of the knowledge discovery process: it can select relevant features, remove irrelevant ones, improve the data quality, make mining algorithms work faster and enhance the comprehensibility of the mined results.

When working on a machine learning approach, the quality of the feature selection task is usually measured analyzing the classification accuracy (by means of a wrapper model for feature selection). However, it is also possible to select the most relevant features of a dataset by filter feature selection methods [4], which do not employ the classification results to select the most important features. In spite of not using any of the traditional filter feature selection measures (described in [4, 5]) we use a Bayesian network learning method to determine the causal relationship among the features. Afterward, the features having no direct causal relationship with the class feature (outside its *Markov blanket*) are considered less relevant to the model and they are removed from the dataset. Therefore, our method does not use the classification results to select the most suitable feature, so it can be seen as a Bayesian filter feature selection similar to the one used in [6]. The main difference is that in [6] a *Conditional Independence* (CI) learning method is applied instead of the Heuristic Search Learning Algorithm applied in our work.

Emerging data mining applications place special requirements on clustering techniques [7], i.e. effective treatment of high dimensionality, interpretability of results and, finally, scalability and usability. Some recent works on clustering try to handle high-dimensionality by selecting important features [8,9,10,11]. In this work we also propose an approach to select important features for clustering. The general method can be described by means of four steps: (i) apply the CGA to some selected objects (sample) of the complete dataset; (ii) consider that the obtained clusters form different classes, which can be modeled by Bayesian networks; (iii) generate a Bayesian network and employ the Markov Blanket of the class variable to the feature subset selection task; (iv) apply the CGA in the complete dataset now formed only by the selected features. Initially, we are mainly interested in evaluating if the feature selection process is appropriate to the Clustering Genetic Algorithm (CGA) [12] and our investigation supposes an ideal situation, where the CGA has actually found the right clustering in step (i). Thus, the Bayesian networks are generated not in a sample, but in the complete dataset correctly clustered/classified. In this way we can better evaluate if the

proposed hybrid method is appropriate, i.e. if the features selected by means of Bayesian networks are suitable for the CGA.

The application of the proposed method is illustrated by means of three datasets. We compare the results obtained in the original space with those obtained in the space formed by the selected features. The experimental results show that the features selected by means of the Markov Blanket are *predictive* in the CGA context. In fact, they provided better results than those obtained in the complete datasets. The next section describes the employed Bayesian feature selection method whereas section 3 briefly describes the CGA. Section 4 shows the simulation results and section 5 describes the conclusions and points out some future works.

2 Bayesian feature selection method

A Bayesian network is a directed acyclic graph (DAG) in which the nodes represent the variables and the arcs represent a causal relationship among the variables connected. A conditional probability table gives the strength of such relationship. The variables that are not connected by an arc can be considered as having no direct causal influence on them. The strength of the causal influence is given by the conditional probability $P(X_i | \Pi X_i)$ (X_i being the i -th variable, and ΠX_i the set of parents of X_i). As seen in [13], the conditional independence assumption (Markov condition) allows one to calculate the joint distribution of all variables as:

$$P(x_1, x_2, \dots, x_n | BK) = \prod P(x_i | \pi_i, BK) \quad (1)$$

Where BK represents background knowledge. Thus, a Bayesian network can be used as a knowledge representation that allows inferences. Learning a Bayesian network from data became an effervescent research topic in the last decade [14], and there are two main classes of learning algorithms to this task. The first one is based on heuristic searching methods and the second one uses the conditional independence (CI) definition to generate the network. There are also algorithms that combine these two strategies.

When using algorithms based on heuristic search, one important issue is the initial order of the dataset variables [15]. These algorithms depend on the order of the variables to determine the arcs direction such that an earlier variable is a possible parent only of the later ones. The conditional independence (CI) methods try to find the arcs direction without the need for this ordering. However, when the ordering is known the algorithms can be improved [16]. In this work we apply a heuristic search based learning algorithm [17] and the statistical χ^2 test is used to determine the initial variables order. This learning process results in a network (where each network node represents a feature) that describes the causal dependencies among the dataset features.

Let us consider a Bayesian network where Λ_A is the set of children of node A and Π_A is the set of parents of node A. The subset of nodes containing Π_A , Λ_A and the parents of Λ_A is called Markov Blanket of A (figure 1). As showed in [13], the only network nodes that have influence on the conditional distribution

of a given node A (given the state of all remaining nodes) are the nodes that form the Markov Blanket of A . Thus, after defining the network structure that models a dataset the Markov Blanket of the class variable can be used as a feature subset selection criterion.

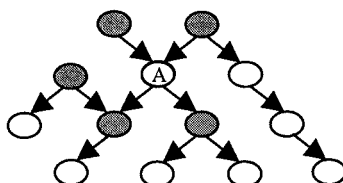


Figure 1: The shadowed nodes represent the Markov Blanket of node A .

The feature selection task is not usually employed in clustering applications [8] basically because the classes are not known a priori. However, our general method supposes that the CGA is initially applied in a dataset and, in this way, some groups of objects are initially found. After that, one can consider that these groups form classes. Therefore, one can employ a feature selection task in a similar way to the one employed in supervised learning tasks.

To verify the consistency of the generated network and provide evidence that the selected features are relevant to the model, a Bayesian classification task is performed first with all the dataset features (original dataset) and later only with the selected ones (Markov Blanket features). The Bayesian feature subset selection method applied is summarized in figure 2. Note that the classification process is not used to select the features, but only to show the consistency of the already selected ones.

- 1) Define the variable order using the statistical χ^2 test;
- 2) Learn a Bayesian network from the ordered data;
- 3) Use a Bayesian classifier to verify the consistence of the generated network;
- 4) Define the Markov Blanket of the *Class* feature;
- 5) Remove the features outside the defined Markov Blanket from the dataset;
- 6) Use a Bayesian classifier to verify the consistence of the network containing only the features inside the Markov Blanket of the class.

Figure 2: Bayesian feature subset selection process.

3 Clustering genetic algorithm

This section intends to briefly describe the Clustering Genetic Algorithm (CGA). Further details can be found in [12]. Basically, clustering is a task where one seeks to identify a finite set of categories or clusters to describe the data. This work considers that clustering involves the partitioning of a set X of objects into a collection of mutually disjoint subsets C_i of X . Formally, let us consider a set of N objects $X = \{X_1, X_2, \dots, X_N\}$ to be clustered, where each $X_i \in \mathcal{R}^p$ is an attribute vector consisting of “ p ” real measurements. The objects must be clustered into

non-overlapping groups $C = \{C_1, C_2, \dots, C_k\}$ where k is the number of clusters, such that:

$$C_1 \cup C_2 \cup \dots \cup C_k = X, C_i \neq \emptyset, \text{ and } C_i \cap C_j = \emptyset \text{ for } i \neq j. \quad (2)$$

The problem with finding an optimal solution to the partition of N data into C classes is NP-complete [18] and because the number of distinct partitions of “ n ” objects into “ m ” clusters increases approximately as $m^n/m!$, attempting to find a globally optimum solution is usually not computationally feasible [19]. Genetic algorithms are widely believed to be effective on NP-complete global optimization problems and they can provide good sub-optimal solutions in reasonable time [20]. Thus, a clustering genetic algorithm can provide a way of finding the right clustering.

3.1 Encoding scheme

Considering that there are N objects to be clustered, a phenotype is represented as a one dimensional integer array with $(N+1)$ elements. As each data unit can be numbered from 1 to N , the i th element of a genotype represents the i th data unit, whereas the last gene represents the number of clusters. Therefore, each gene of a chromosome has a value over the alphabet $\{1, 2, 3, \dots, k\}$, where k is the maximum number of clusters. For example, considering a dataset with 20 objects one can get the following clustering: 22345123453321454552 5. This means that 5 objects $\{1, 2, 7, 13, 20\}$ form the cluster whose label is 2, and the cluster whose label is 1 has 2 objects $\{6, 14\}$. The last gene corresponds to the number of clusters encoded by the solution.

The first problem with this straightforward encoding scheme is that it is redundant (there are $5!$ chromosomes encoding the same solution). Thus, the search space is much larger than the original one. In addition, there is another problem caused by this encoding scheme when it is applied with the traditional genetic operators. It is the undesirable effect of casting context-dependent information out of context under the standard crossover, i.e. the context insensitivity [18]. This problem happens when one is dealing with the classic genetic operators of crossover and mutation, which usually produce invalid solutions. To avoid these problems, we have developed genetic operators that are group-oriented. Basically, the CGA is described in Figure 3.

- 1) Initialize a population of genotypes;
- 2) Evaluate each genotype in the population;
- 3) Apply a linear normalization;
- 4) Select genotypes by proportional selection;
- 5) Apply crossover and mutation;
- 6) Replace the old genotypes by the ones formed by 5);
- 7) If the convergence is attained, stop; if not, go to step 2).

Figure 3: Clustering Genetic Algorithm (CGA).

3.2 Objective function

Determining the optimal number of clusters in a dataset is one of the most difficult aspects in the clustering process [21]. Most of the approaches found in the literature determine the right number of clusters according to criteria based on given clusterings. However, the CGA is supposed to optimize not only the clusterings for a given number of clusters but also the number of clusters. Therefore, the CGA employs an objective function based on the Average Silhouette Width [22]. Basically, let us consider an object “i” belonging to cluster A. So the average dissimilarity of “i” to all other objects of A is denoted by $a(i)$. Now consider a different cluster C and let us calculate the average dissimilarity of “i” to all objects of C, which will be here denoted by $d(i,C)$. After computing the $d(i,C)$ for all clusters $C \neq A$ we select the smallest of those, $b(i) = \min d(i,C)$ for $C \neq A$. This number represents the dissimilarity of “i” to its neighbor cluster. Now one defines the silhouette $s(i)$ like follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3)$$

When the cluster A contains only one object we consider that $s(i)=0$, which is the most neutral choice [22]. In addition, it is easy to see that: $-1 \leq s(i) \leq 1$. The objective function is the average of $s(i)$ for $i=1,2,\dots,N$. It implies that the best value of “k” happens when the objective function value is as high as possible.

3.3 Selection

The genotypes that make part of each generation are selected according to the roulette wheel selection strategy. As this strategy does not allow negative objective function values, a constant number equals to one is summed up to each objective function value before the selection process takes place. Besides, the best genotype is always copied into the succeeding generation.

3.4 Operators

Classic genetic operators are not suitable for clustering problems [18] because they usually work with a set of unrelated genes. In clustering problems the genes are related, i.e. equal allele genes mean that the objects they represent belong to the same cluster. In other words, the groups are the meaningful building blocks of a solution. Thus, it is necessary to use operators that work with groups of genes, instead of with the genes themselves.

The crossover operator combines pieces of information coming from different genotypes and it works in the following way: First, two genotypes (A and B) are selected; Second, considering that A represents k_1 clusters, the CGA chooses randomly n $[1,k_1]$ clusters to copy in B. The unchanged groups of B are maintained and the changed ones have their objects allocated to the cluster that has the nearest centroid. In this way the child C is obtained. This same process is

employed to get child D, but now considering that the changed clusters of B are copied in A.

There are two operators for mutation. The operator 1, which only works in clusterings formed by more than two groups, eliminates a randomly chosen group and places all their objects into the remaining cluster that has the nearest centroid. The operator 2 divides a randomly selected group into two new ones. The first group is formed by the objects closer to the centroid, whereas the other group is formed by those objects closer to the farthest object to the centroid. These operators are suitable for grouping problems because they just change the genotypes in the smallest possible way, i.e. dividing groups and eliminating others in the hope of finding better solutions to the problem being solved. Besides, 50% of the genotypes are crossed-over, 25% are mutated by operator 1 and 25% are mutated by operator 2.

3.5 Initial population

Fundamentally, we have employed the methodology developed in [22] to set up the initial population. The initial clustering process is based on the selection of representative objects. The first selected object is the most centrally located in the set of objects. Subsequently, other objects are selected. Basically, the chance of selecting an object increases when it is far from the previously selected ones and when there are many objects next to it. After selecting the representative objects, the initial population is formed considering that the non-selected objects must be clustered according to their proximity to the representative ones. Considering k representative objects, the first genotype represents two clusters, the second genotype represents three clusters, ..., and the last one represents k clusters. Thus, we have employed initial populations formed by $(k-1)$ genotypes, where each genotype represents a different clustering.

4 Simulations

Clustering algorithms can be evaluated by means of *classification datasets*. To do so, the clustering algorithm is applied in the classification dataset (without the class labels) in order to verify whether it finds the correct classes/clusters or not. This work intends to show that the combination of a Bayesian feature selection method with the CGA can provide a useful hybrid system for data mining. As previously mentioned, in this work we do not evaluate the whole proposed hybrid system, which is: (i) apply the CGA to some selected objects (sample) of the complete dataset; (ii) consider that the obtained clusters form different classes, which can be modeled by Bayesian networks; (iii) generate a Bayesian network and employ the Markov Blanket of the class variable to the feature subset selection task; (iv) apply the CGA in the complete dataset, now formed only by the selected features. In fact, we are mainly interested in evaluating the soundness of the Bayesian feature selection method in the context of the CGA. Thus, our first investigation supposes that the CGA has found the *right* clustering in step (i), i.e. we are supposing an ideal situation in which the CGA is capable

of finding the right clusters in the first step. In this way, we can suppose that we are providing *ideal* conditions for the Bayesian networks to find the *best* features, i.e. the feature selection task is not influenced by the efficacy of the CGA in finding the *right* clusters. Obviously it is unlikely that the CGA will always find the *right clustering* in step (i) and the evaluation of the system as a whole is necessary to validate the proposed approach. Initially, however, we believe that it is necessary to evaluate if the selected features are adequate to the CGA. Thus, we performed simulations to provide a way of comparing the efficacy of the CGA both in the complete dataset and in the dataset formed by the selected features.

Basically, the proposed method was evaluated in three datasets [23] that are usually employed as benchmarks for data mining applications: Wisconsin Breast Cancer, Mushroom and Congressional Voting Records. The CGA was applied in the complete dataset and in the dataset formed by the selected features. In each simulation performed by the CGA, we have employed populations formed by 20 genotypes that, in turn, implies in using 21 clusters at most. Besides, we simulated the clustering process 11 times (this number is convenient to perform the Wilcoxon/Mann-Whitney statistical test) for each dataset and the maximum number of generations was set to 100. Besides, in all tables presented in this work ACR stands for Average Classification Rate and μ_s and σ_s respectively stand for the mean and the standard deviation in the considered samples.

4.1 Wisconsin breast cancer database

This database contains examples formed by 9 features and an associated class label (benign or malignant). The features are: clump thickness (x_1), uniformity of cell size (x_2), uniformity of cell shape (x_3), marginal adhesion (x_4), single epithelial cell size (x_5), bare nuclei (x_6), bland chromatin (x_7), normal nucleoli (x_8) and mitoses (x_9). The two classes are known to be linearly inseparable. The total number of objects is 699 (458 benign and 241 malignant), of which 16 have a single missing feature. We have removed those 16 objects and used the remaining ones (444 benign examples and 239 malignant examples). The Euclidean distance was employed to calculate the dissimilarities among the examples. Table 1 summarizes the obtained results in the dataset formed by all features as well as in the dataset formed by the features selected by the Markov Blanket process (x_2 , x_3 , x_4 , x_6 , x_7 , x_8). In all experiments the CGA found two clusters.

Table 1: Summary of results – Wisconsin breast cancer.

Dataset / Result	ACR - Benign	ACR - Malignant	ACR - Total
All features	$\mu_s=98.24$; $\sigma_s=0.09$	$\mu_s=90.26$; $\sigma_s=0.94$	$\mu_s=95.45$; $\sigma_s=0.30$
Selected features	$\mu_s=98.20$; $\sigma_s=0.00$	$\mu_s=91.52$; $\sigma_s=0.20$	$\mu_s=95.86$; $\sigma_s=0.07$

4.2 Mushroom database

This database consists of 8,124 examples, each with 22 symbolic features with up to 12 different values. 51.8% of the cases represent edible, and the rest non-edible (mostly poisonous) mushrooms. The features will be here called x_i ($i=1,2,\dots,22$) according to the order that they appear in the database. There are 2,480 missing values (all for feature #11). These examples were removed and we applied the proposed method in the remaining ones. The simple matching approach was employed to calculate the distances among the objects and Table 2 summarizes the obtained results. The features selected by the Markov Blanket process were x_3 (cap-color), x_5 (odor) and x_9 (gill-color). The CGA found 21 clusters in the simulations performed in the dataset formed by the selected features, whereas in the complete dataset 4 clusters were found. In order to evaluate the CGA efficacy in these cases, one considers that each cluster belongs to the class that has the higher number of objects.

Table 2: Summary of results – Mushroom.

Dataset / Result	ACR – Edible	ACR – Poisonous	ACR - Total
All features	$\mu_s=99.16$; $\sigma_s=0.17$	$\mu_s=62.15$; $\sigma_s=0.00$	$\mu_s=85.03$; $\sigma_s=0.11$
Selected features	$\mu_s=94.95$; $\sigma_s=0.00$	$\mu_s=74.21$; $\sigma_s=0.00$	$\mu_s=87.03$; $\sigma_s=0.00$

4.3 Congressional voting records database

This database includes votes for each of the U.S. House of Representatives Congressmen on 16 key votes (features). There are 435 instances (267 democrats, 168 republicans) and each of the 16 features is Boolean valued. Besides, there are 203 instances with missing values. These instances were removed and the proposed method was employed in the 232 remaining ones (124 democrats, 108 republicans). The features will be here called x_i ($i=1,2,\dots,16$) according to the order that they appear in the database. The simple matching approach was employed to calculate the distances among the examples and Table 3 summarizes the obtained results. The features selected by the Markov Blanket process were x_3 (adoption-of-the-budget-resolution), x_4 (physician-fee-freeze), x_5 (El-Salvador-aid) and x_{12} (education-spending). It is important to mention that when all features were employed the CGA found two clusters, whereas in the dataset formed by the selected features 14 clusters were found. As previously mentioned, we consider that each cluster represents the class that has the higher number of objects.

Table 3: Summary of results – congressional voting records.

Dataset / Result	ACR - Democrat	ACR - Republican	ACR - Total
All features	$\mu_s=77.86$; $\sigma_s=0.42$	$\mu_s=96.30$; $\sigma_s=0.00$	$\mu_s=86.44$; $\sigma_s=0.22$
Selected features	$\mu_s=95.97$; $\sigma_s=0.00$	$\mu_s=98.15$; $\sigma_s=0.00$	$\mu_s=96.98$; $\sigma_s=0.00$

5 Conclusions and future work

This work shows how a Bayesian Feature Selection method can be employed with a Clustering Genetic Algorithm (CGA). Basically, the proposed method uses a Bayesian filter to define the most relevant features of a dataset, and these features are used to get better clusters by the CGA. The proposed methodology was illustrated by means of simulations in three databases that are benchmarks for data mining methods: Wisconsin Breast Cancer, Mushroom and Congressional Voting Records. We compared the clusters obtained in the space formed by all the features with those obtained in the space formed by the selected features. The experimental results showed that the features selected by means of the Markov Blanket provided better results in terms of the average classification rate. In fact, we applied the Wilcoxon statistical test [24] (also known as Mann-Whitney test), supposing that the “ACR-Total” values in the datasets formed by the selected features are higher than those obtained in the complete datasets and we concluded that the results are statistically significant at the 1% significance level. Thus, the presented hybrid method is promising.

Our future work will focus on four main aspects. First, we are going to evaluate the system as a whole, performing all the proposed steps. Second, we are going to employ a wrapper model within the Markov Blanket. In this sense, we believe that it will be possible to find a subset of selected features, improving the knowledge structure without decreasing the average classification rates. Besides, it is possible that another Bayesian Network learning algorithm (based on CI for instance) can provide a different Markov Blanket, and this verification will also be performed. Third, we are going to evaluate the efficacy of the proposed method when standardized datasets are employed. Finally, we are going to evaluate our methodology in other databases, and we hope to describe the results soon.

References

- [1] Fayyad, U. M., Shapiro, G. P., Smyth, P. “From Data Mining to Knowledge Discovery : An Overview”. In: *Advances in Knowledge Discovery and Data Mining*, Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Editors, MIT Press, pp. 1-37, 1996.
- [2] Bigus, J. P., *Data Mining with Neural Networks* , First edition, USA, McGraw-Hill, 1996.
- [3] Han, J. and Kamber, M., *Data Mining, Concepts and Techniques*. Morgan Kaufmann, 2001.
- [4] Liu, H. and Motoda, H., *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic, 1998.
- [5] Yang, Y. and Pederson, J., A comparative study on feature selection in text categorization. In Proc. of the *Fourteenth International Conference on Machine Learning*, 1997.

- [6] Cheng, J., Bell, D.A., and Liu, W., Learning belief networks from data: An information theory based approach, *Proceeding of the sixth ACM International Conference on Information and Knowledge Management*, 1997.
- [7] Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P., "Automatic subspace clustering of high dimensional data for data mining applications, *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1998.
- [8] M. Dash and H. Liu, Feature Selection for Clustering. *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2000.
- [9] Bradley, P.S., Fayyad, U., Reina, C., "Scaling clustering algorithms to large datasets", *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pages 9-15, 1998.
- [10] Cheng, C., Fu, W., Zhang, Y., "Entropy-based subspace clustering for mining numerical data, *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD'99)*, 1999.
- [11] Aggarwal, C.C., Procopiuc, C., Wolf, J.L., Yu, P.S., Park, J.S., "Fast Algorithms for projected clustering, *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 61-72, 1999.
- [12] Hruschka, E. R., Ebecken, N.F.F. "A genetic algorithm for cluster analysis", *Intelligent Data Analysis (IDA)*, v.7, pp. 15-25, IOS Press, 2003.
- [13] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA., 1988.
- [14] Chickering, D. M., Optimal Structure Identification with Greedy Search, *Journal of Machine Learning Research*, 2002.
- [15] Hruschka JR, E. R., Ebecken, N. F. F., Ordering attributes for missing values prediction and data classification. In: *Data Mining III - Management Information Systems Series* - vol. 6, ed. Southampton - UK : WIT Press, 2002.
- [16] Spirtes P., Glymour C., and Scheines R. *Causation, Predication, and Search*. Springer-Verlag, New York, 1993.
- [17] Hruschka JR, E. R., Ebecken, N. F. F. Missing Values prediction with K2. *Intelligent Data Analysis (IDA)*. IOS Press, Netherlands: , v.6, n.6, 2002.
- [18] Falkenauer, E., *Genetic Algorithms and Grouping Problems*, John Wiley & Sons, 1998.
- [19] Arabie, P., Hubert, L. J., An Overview of Combinatorial Data Analysis (Chapter 1). *Clustering and Classification*, ed. P. Arabie, L.J. Hubert, G. DeSoete, World Scientific, 1999.
- [20] Park, Y., Song, M., "A Genetic Algorithm for Clustering Problems", *Proceedings of the Genetic Programming Conference*, University of Wisconsin, July, 1998.
- [21] Cole, R. M., *Clustering with Genetic Algorithms*, Master of Science Thesis, Department of Computer Science, University of Western Australia, 1998.



192 Data Mining IV

- [22] Kaufman, L., Rousseeuw, P. J., *Finding Groups in Data, An Introduction to Cluster Analysis*, Wiley Series in Probability and Mathematical Statistics, 1990.
- [23] Merz, C.J., Murphy, P.M., UCI Repository of Machine Learning Databases, <http://www.ics.uci.edu>, Irvine, CA, University of California, Department of Information and Computer Science.
- [24] Triola, M. F., *Elementary Statistics*, 7th Edition, Addison Wesley Longman, 1999.