A new approach for Bayesian classifier learning structure via K2 Algorithm

Heni Bouhamed [1], Afif Masmoudi [2], Thierry Lecroq[1], Ahmed Rebaï [3]

[1] University of Rouen, LITIS EA 4108, 1 rue Thomas Becket, 76821 Mont-Saint-Aignan cedex, France
Heni.bouhamed@yahoo.fr
Thierry.lecroq@univ-rouen.fr
[2] Department of Mathematics, Faculty of Science of Sfax, Soukra B.P 802 Sfax, Tunisia
Afif.masmoudi@fss.rnu.tn
[3] Bioinformatics Unit, Centre of Biotechnologie of Sfax, 3018 Sfax, Tunisia
Ahmed.rebai@cbs.rnrt.tn

Abstract. It is a well-known fact that the Bayesian Networks' (BNs) use as classifiers in different fields of application has recently witnessed a noticeable growth. Yet, the Naïve Bayes' application, and even the augmented Naïve Bayes', to classifier-structure learning, has been vulnerable to certain limits, which explains the practitioners' resort to other more sophisticated types of algorithms. Consequently, the use of such algorithms has paved the way for raising the problem of super-exponential increase in computational complexity of the Bayesian classifier learning structure, with the increasing number of descriptive variables. In this context, the present work's major objective lies in setting up a further solution whereby a remedy can be conceived for the intricate algorithmic complexity imposed during the learning of Bayesian classifiers' structure with the use of sophisticated algorithms. Noteworthy, the present paper's framework is organized as follows. We start, in the first place, by to propose a novel approach designed to reduce the algorithmic complexity without engendering any loss of information when learning the structure of a Bayesian classifier. We, then, go on to test our approach on a car diagnosis and a Lymphography diagnosis databases. Ultimately, an exposition of our conducted work's interests will be a closing step to this work.

Keywords: Bayesian Classifier; structure learning; classification; clustering; modeling; algorithmic complexity, K2 algorithm.

1 Introduction

It is worth noting that efficient classifiers can be reached through the use of Bayesian networks [1, 2, 3]. In fact, a Bayesian Classifier relative to a problem with p variables is characterized by the distinction of having $p + 1$ nodes. Indeed, all Bayesian classifiers model the fact of belonging to a certain class by means of a discrete node dubbed "class node". This node is discrete and multinomial having k modality. The class node is distinct for not owning a parent node. Regarding the other p variables, which we call descriptive variables, they are denoted X_i (i from 1 to p).

The Bayesian classifier with the simplest structure is the Naïve Bayesian Network (RBN) [9], also called Naïve Bayes classifier. Nevertheless, no correlations between the attributes are taken into account with respect to the Naïve Bayes, where all features contribute to the classification in the same way. The classification node takes advantage of the information provided by each attribute independently of the information provided by other features-still; this may not be optimal for the classification task. Hence, various proposals have been suggested in a bid to enrich the Naive Bayesian Network structure to make it account for correlations between different attributes. In [2], for instance, the authors have proposed a Tree-Augmented Naïve Bayes (TAN) approach to enrich the network structure. According to this approach, a tree structure is applied for the classification to be achieved [20, 5]. The tree structure has the advantage of having a low degree of complexity, along with the ability to avoid over fitting problems. However, it restricts the number of parents, other than the classification node, to exactly one single parent for each node, which turns out to be a strong constraint. So, the resulting structure appears to neglect the case where a variable is correlated with several other variables. Besides, it outlooks the case where a variable is conditionally independent of all other variables within the classification node. In which case, the node representing that variable only needs the class node as a parent. The addition of another parent only adds unnecessary complexity and increases the number of network parameters. Consequently, other authors [4, 5, 6, 28, 8] have proposed the use of more sophisticated methods to overcome these shortcomings, among which are: the use of the K2 algorithms [6, 24, 25, 26, 27, 28, 10, 4], the Genetic Search [7, 4], the Greedy Search [11, 4], the Annealing Simulated [8, 4], the Greedy Hill Climber [7, 4] and the Repeated Hill Climber [7, 4]. Although these algorithms have actually managed to attain performant classifiers, their application has resulted in the frequently and commonly encountered problem of structure-learning computational complexity owing to the increase in the number of descriptive variables.

Hence, a new approach has been proposed through this research work based on a structure learning upstream clustering, which can be jointly used with the K2 algorithms pertinent to the structure learning of Bayesian Classifiers. The envisaged aim behind this framework proposal is to reduce the computational complexity and, consequently, the execution time without engendering a loss of information, in comparison to the use of the classic K2 algorithm.

2 A new clustering-based heuristic: methodology

The idea lying behind our conceived procedure lies in the rapid super-exponential surge of algorithmic complexity of learning the Bayesian Classifier structure from data [12, 13] with respect to the rise in the number of variables. To remedy this problem, our idea consists in subdividing the variables into subsets (or clusters), by learning the structure of each cluster's separately, while looking for a convenient procedure whereby the different structures could be assembled into a final structure. In this regard, it has been noticed that in the case of a Bayesian classifier learning structure, there exists one single central variable of a global interest called "class" variable. In this respect, we reckon to execute the processing of each cluster's

learning structure with the class variable, then, proceed by assembling the different various structures around this class variable as a next step.

2.1 The variables' clustering

Regarding our present work, we have chosen to use the K-means algorithm, as it is the most popular and applied in the literature, added to fact that its algorithmic complexity is linear *(O(n))* [14]. We also propose to use a hierarchical clustering algorithm along with the bootstrap technique to obtain the optimal number of clusters that will be introduced as entries in the K-means algorithm. To note, the databases that will be applied to test our approach, in the experimentation section, consist of categorical variables, and regarding the performance of clustering we will use the toolbox ClustOfVar with the software R [15]. In particular, we will use the variant K-means for categorical variables [16, 17] and the link-likelihood approach [18] (hierarchical clustering algorithm for categorical variables). To assess the stability of all possible partitions, *2* to *p-1* (where p is the total number of variables) clusters from the hierarchical clustering, we will use a feature called "Stability" (also developed in the ClustOfVar toolbox) based on the "bootstrap" technique. The result is a graph which is then a tool to help to select the number of clusters. The user can be choosing the number *K* of clusters to the heights of the first increase in the stability.

2.2 Structure learning

A structure learning has been performed for each cluster of variables including the class variable. The ultimate structure would be the assembling of the n structures obtained from each cluster around the class variable.

We will perform our tests via the K2 algorithm with, as input, the order obtained by applying the algorithm MWST (for the MWST algorithm, the initial node will be the class variable) [21]. In our study case, we would rather try to prove that the joint use of our approach together with the K2 algorithm can be beneficial in reducing the computational complexity without losing information.

Note that in our work, we will use the BNT toolbox [22] running on the Matlab software (2010 version) to apply the MWST and K2 algorithms to structure learning. We will also apply the BNT toolbox for parameters learning and inference.

3 Experimentations procedures

3.1 Data-bases

We first test our approach, on a car diagnosis database (Car Diagnosis 2). It has *18* variables, among which is a status variable called "Car starts", the Class variable. The parameters' generating file of this data base is available on the site http://www.norsys.com/downloads/netlib/. According to these parameters, we have been able to generate *10.000* examples, among which *32* have been left aside for the references' testing phase. We also apply our approach to a Lymphography diagnosis

database (Lymphography). It is made up of *19* variables, among which is a status variable called "Diagnosis", the Class variable. This lymphography domain has been obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia (available on request on the site http://archive.ics.uci.edu/ml/datasets/Lymphography). Among the *148* instances of data, *32* have been left aside for the references' testing phase.

3.2 Clustering

Regarding the clustering, we are going to use the stability function (bootstrap approach using the mean of corrected rand criterion [19]) of the toolbox ClustOfVar [16] after the application of an hirarchical ascendant algorithm, in order to estimate, approximately, the number of clusters to be entered in the algorithm K-means.

Using the stability graphics, the optimal number of clusters selected, for "Car diagnosis 2" database, has been equal to three.

Using the stability graphics, the optimal number of clusters selected, for "Lymphography" database, has been equal to two.

3.3 The classical learning structure compared to our new heuristic

For the "Car diagnosis 2" database, the execution time has been *3.45* seconds for the classical structure learning of the entire variables. The global execution time of our approach application has been *1.45* seconds (over *1.32* seconds for cluster 1; *0.05* seconds for cluster 2 and *0.09* seconds for cluster 3). The sum of these executions' time (*1.45* seconds) remains significantly inferior to the structure learning of the entire variables simultaneously.

For the "Lymphography" database, the sum of learning structure of "cluster 1" and "cluster 2" executions' time (equal to *1.65* seconds) remains significantly inferior to the structure learning of the entire variables, simultaneously, which equals *2.67* seconds.

3.4 Both attained structures' relevant inferences and result comparisons

Our approach favors the preservation of data for the class variable's sake, we will learn the parameters of the two structures found for each of the databases studied (structure found after learning all the variables simultaneously and structure found after assembling the various structures of the clusters around the class variables). For the class variable, we are going to calculate the probabilities of its different states; given the states of the networks other nodes in respect of the two obtained Bayesian classifiers structures. Thus, a *32* database will be used for experimenting the class variables of both databases. Naturally, the experimentation examples have been excluded during the structures' learning. The statistical significance of difference between the obtained probabilities, with respect to both structures, will be measured

via the “Z” test (comparing the two observed means belonging to two different samples) [23].

The two tested Class variables are “Car starts” of the “Car Diagnosis 2” database and “Diagnosis” of the “Lymphography” database. The results are presented in graphs form (See Fig. 1 and Fig. 2) showing the Z-test variation corresponding to each variable studied according to its different possible states.

3.5 Discussion

Based on the achieved experimental results, the pairs of probabilities for the variable “diagnosis” of the "Lymphography" database are identical; the preservation of information has been complete (see Fig. 2). As for the variable “Car Start” of “Car Diagnosis 2” database, the probabilities pairs are very similar but not identical; the hypothesis H_0 has always been rejected, even with very small Z values, not exceeding the value of $|0.46|$, very distant from the threshold of $|1.96|$, as set by the Z test theory (see Fig. 1). It can, therefore, be deduced that the inference results, regarding both of learning structures approaches, are very similar even at eye sight, and without applying any statistical tests to measure the difference’s significance. Through our approach, we have managed to reduce, considerably, the algorithmic complexity of the Bayesian classifier structure learning without any significant loss of information.

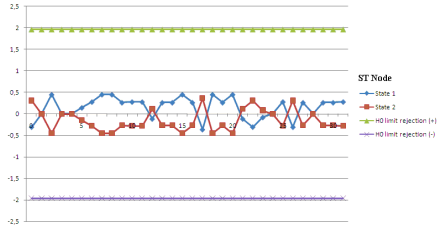

Fig. 1 *Z*-test variation for the “Car starts” variable (“Car Diagnosis 2” database).

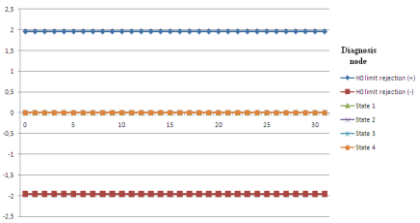

Fig. 2 *Z*-test variation for the “Diagnosis” variable (“Lymphography” database).

4 Conclusion

Within the scope of the present work, we have set up a new well-defined approach for the Bayesian Classifier structure learning from data-base, so useful that it can be jointly applied with the K2 algorithms in the aim to reduce the computational complexity of this process. we have proved that loss in data turns out to be so negligible that it does not affect the extracted Bayesian classifier stemming results during the inference stage, while saving a great deal of execution time.

References

1. P. Langley, S. Sage, Induction of Selective Bayesian Classifiers, in Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, (1994) 399-406.

2. N. Friedman, D. Geiger, M. Goldszmid, Bayesian Network classifiers. Machine Learning, (1997) 131-163.
3. F. Pernkopf, Bayesian network classifiers versus selective k-NN classifier , Pattern Recognition, (2005) 1-10.
4. M. Stuart, H. Yulan, L. Kecheng, Choosing the best Bayesian classifier : An empirical study. IAENG International Journal of Computer Science (2009) 1-10.
5. M. G. Madden, A New Bayesian Network Structure for Classification Tasks, in Proceedings of 13th Irish Conference on Artificial Intelligence & Cognitive Science, (2002) 203-208.
6. B. Lerner, R. Malka, Investigation of the K2 algorithm in learning Bayesian Network Classifiers, Applied Artificial Intelligence, (2011) 74–96.
7. H. I. Witten, F. Eibe, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, (1999).
8. S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi. Optimization by simulated annealing. Science, (1983) 671-681.
9. P. Domingos, M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss. MachineLearning, (1997) 103-130.
10. G. Cooper, E. Hersovits, A Bayesian method for the induction of probabilistic networks from data, Machine learning. 9 (1992) 309-347.
11. P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search. The MIT Press, 2nd edition, (2000).
12. P. Judea, V. Tom, A theory of inferred causation, In James Allen, Richard Fikes and Erik Sandewall, editors, KR' 91, Principles of knowledge representation and reasoning. (1991) 441-452.
13. R. W. Robinson, Counting unlabeled acyclic digraphs, Combinatorial Mathematics. 622 (1977) 28–43.
14. S. Tufféry, Data mining et statistique décisionnelle: l'intelligence des données, Editions TECHNIP. (2010).
15. A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognition Letters. 31 (2010) 651-666.
16. M. Chavent, V. Kuentz, B. Liquet, J. Saracco, ClustOfVar: an R package for the clustering of variables. The R user conference, University of Warwick Coventry UK. (2011) 44.
17. M. Chavent, V. Kuentz, J. Saracco, A partitioning method for the clustering of categorical variables. In classification as a tool for Research, Herman locarek-Junge, claus Weihs (Eds), Springer, in Proceedings of the IFCS (2009).
18. I.C.Lerman,. Likelihood linkage analysis (LLA) classification method : An example treated by hand, Biochimie. 75 (5) (1993) 379-397.
19. P. Green, A. Kreiger, A Generalized Rand-Index Method for Consensus Clustering of Separate Partitions of the Same Data Base. Journal of classification. (1999) 63-89.
20. C. Chow, C. Liu, Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory. 14 (3) (1968) 462–467.
21. O. Francois, P. Leray, Evaluation d'algorithmes d'apprentissage de structure pour les réseaux bayésiens, In Proceedings of 14ème Congrès Francophone Reconnaissance des Formes et Intelligence Artificielle. (2004) 1453-1460.
22. K. Murphy, The BayesNet Toolbox for Matlab, Computing Science and Statistics: Proceedings of Interface. 33 (2001) http ://www.ai.mit.edu/~murphyk/Software/BNT/bnt.html.
23. R.C. Sprinthall, Basic Statistical Analysis, Seventh Edition. 2003.
24. K. Ezawa, M. Singh, S. Norton, Learning goal oriented Bayesian networks for telecommunications risk management, In Proceedings of the Thirteenth International Conference on Machine Learning (1996) 139-147.
25. A. Porwal, E. Carranza, M. Hale, Bayesian network classifiers for mineral potential mapping, Computers & Geosciences 32 (2006) 1–16.
26. R. Malka, B. Lerner, Classification of fluorescence in situ hybridization images using belief networks, Pattern Recognition Letters 25 (2004) 1777–1785.
27. R. Estevam, J. Hruschka, N. Ebecken, Towards efficient variables ordering for Bayesian networks classifier, Data & Knowledge Engineering 63 (2007) 258–269.
28. J. A. Carta, S. Velázquez, J.M. Matías, Use of Bayesian networks classifiers for long-term mean wind turbine energy output estimation at a potential wind energy conversion site, Energy Conversion and Management 52 (2011) 1137–1149.