© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Ordering attributes for missing values prediction and data classification

E. R. Hruschka Jr., N. F. F. Ebecken
COPPE / Federal University of Rio de Janeiro, Brazil.

Abstract

This work shows the application of the bayesian *K2* learning algorithm as a data classifier and preprocessor having an attribute order searcher to improve the results. One of the aspects that have influence on the K2 performance is the initial order of the attributes in the data set, however, in most cases, this algorithm is applied without giving special attention to this preorder. The present work performs an empirical method to select an appropriate attribute order, before applying the learning algorithm (K2). Afterwards, it does the data preparation and classification tasks. In order to analyze the results, in a first step, the data classification is done without considering the initial order of the attributes. Thereafter it seeks for a good variable order, and having the sequence of the attributes, the classification is performed again. Once these results are obtained, the same algorithm is used to substitute missing values in the learning dataset in order to verify how the process works in this kind of task. The dataset used came from the standard classification problems databases from UCI Machine Learning Repository. The results are empirically compared taking into consideration the mean and standard deviation .

1. Introduction

The aim of the present work is to show how the definition of a good attribute preorder can have influence on a classification task (with and without missing values) results. To achieve such objective a preorder searcher is implemented, and it prepares the data to a bayesian classifier algorithm that learns from such data and classifies the objects.

A bayesian classifier uses a bayesian network as a knowledge base [1]. This network is a directed acyclic graph (DAG) in which the nodes represent the

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

variables and the arcs represent a causal relationship among the variables connected. The strength of such relationship is given by a conditional probability table. For an introduction on bayesian networks see [1 and 2].

Once one has a bayesian network (that can be obtained from a human specialist, or from the learning from data algorithm), and an inference algorithm to be applied into the network, the classification can be performed.

In our work, we use a version of the K2 algorithm [3] to learn from data. It assumes that the attributes are discrete; the data set has only independent cases; and all the variables (attributes) must be preordered. Considering these assumptions, the algorithm will look for a bayesian structure which best represents the database.

With the bayesian network already defined, we need to perform inferences to have the classification. There are many methods of inference in a bayesian network [2] and they work propagating evidences in a network in order to obtain the desired answers, that's why most of these methods are called *evidence propagation methods.* The *bayesian conditioning evidence propagation* algorithm is one of the ways used to propagate information (evidences) in a bayesian network when the network is not singly connected [2]. It consists in changing the connectivity of the original network and generating a new structure. This new structure is created by searching for the variables that break the loops in the network (cutset) and instantiating them. This cutset search is a complex task [2], but once the new structure is created, the propagation can be implemented in a simpler way. In this work the *general bayesian conditioning* (GBC) [4] is used. It considers that in a data mining prediction work most of the values of the attributes are given, so instead of looking for a good cutset, the algorithm simply instantiates all the variables (attributes) that have no missing value (except the class attribute) and performs the propagation in the network. For a more detailed view on other propagation methods and conditioning algorithms see [1, 2 and 5].

With the algorithms described above this work performs the classification with and without generating the best preorder attributes. In the next section some related work is pointed out. In section three the classification process is described and the results are shown. The conclusions are presented in the last section and some future work is settled.

2. Related work

In the last two decades the knowledge networks theory has been studied and applied in a broadening way. Learning bayesian (or knowledge) networks is a computer based process that aims to obtain an internal representation of all the constraints of a target problem. This representation is created by trying to minimize the computational effort to deal with the problem [2,6 and 7].

The bayesian learning process can be divided into two phases. The first one is the network structure learning (called structure learning), and the second is the probability distribution table definitions (called the numerical parameters learning). The first phase is used to define the most suitable network structure to

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

represent the target problem. In the second step, once the structure is already defined, the numerical parameters (probability distribution tables) have to be set.

The first results with structure learning is shown in the Chow and Liu [8] work, in this learning process, the structure can be a tree with k nodes. It assesses a joint probability distribution P (that represents the problem model) and looks for a tree structure representing the probability distribution which is closer to P. Rebane and Pearl [9] proposed an algorithm to be applied along with Chow and Liu's. It improves the method by allowing the learning of a poly-tree structure instead of a tree.

There are many other learning methodologies, and some bayesian ones can be found in [6, 10, 11,12, and 13).

The missing values problem is an important issue in data mining. Thereby there are many approaches to deal with it [14]:

- Ignore objects containing missing values;
- Fill the gaps manually;
- Substitute the missing values by a constant;
- Use the mean of the objects in the same class as a substitution value;
- Get the most probable value to fill the missing values. It can be done with the use of regression, bayesian inference or decision trees. This process can be divided into missing values in training and test cases [15].

The bayesian *bound and collapse* algorithm [13] works in two phases: *bounding* samples that have information about the missing values mechanism and encoding the other ones in a probabilistic model of non-response. Afterwards the *collapse* phase defines a single value to substitute the missing ones.

The learning from data having missing values using the K2 algorithm proposed by Hruschka Jr. and Ebecken [5] uses the same algorithm used for predicting the missing values and classifying the prepared data. That work points out other learning from data, having a missing values approach.

In this work, the method applied to substitute the missing values and learn from data is described in [5], but instead of using the original attribute order, here we search for the best order before performing the learning. In the next section the method is shown in more details.

3. Data classification

The dataset used is called IRIS and was taken from the UCI Machine Learning Repository [16]. It contains 150 objects (50 in each one of the three classes) having four attributes and a class attribute. The class has three possible nominal values (Iris Setosa, Iris Versicolour and Iris Virginica)and the other attributes are numerical ones (called 1.sepal length; 2.sepal width; 3.petal length; and 4.petal width). There is no missing value in the data.

The reason for using this small dataset, containing only 4 attributes and 150 objects, is that the ordering process presented in this work is an exhaustive search, thereby, if the dataset presented too many attributes, the process would become too slow (see more details about this ordering process in section 3.2).

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

In the next section we present a naive discretization method which is performed to suit the data to the learning and classification algorithms.

3.1. Discretization and dataset division

As we are using a bayesian method, the data must be discrete [2]. The IRIS dataset has continuous attributes, so a discretization was done.

A naive discretization was performed (for more details about discretization methods and their effects on the data analysis see [17]). The first step was to multiply all the values in the dataset by 0.1, it converted all the values into integer ones. Afterwards, the value 43 was subtracted from all values of the first attribute "sepal length", 20 from the second attribute "sepal width" values, 10 from the third attribute "petal length" values and 1 from the last attribute "petal width" values. An example of the discretization is shown below.

Table 1. Data Discretization

	Original data	Discretization	Final discrete data
1 – Sepal length	5.1	(5.1 * 0.1) – 43	8
2 – Sepal width	3.5	(3.5 * 0.1) – 20	15
3 – Petal length	1.4	(1.4 * 0.1) – 10	4
4 – Petal width	0.2	(0.2 * 0.1) – 1	1

The nominal class definition was converted into numerical values as following:

Table2. Class numerical values

Nominal class values	Numerical Class value
Iris-setosa	0
Iris-versicolor	1
Iris-virginica	2

Having the discrete data, it was divided into five datasets, each one having a training and a test subset. It was done by dividing the original sample into a *training* (80% of the data – 120 objects) and a *test* (20% of the data – 30 objects) sample five times. The division was made using a random number generator [18] to select the objects from the original sample. After the division, the objects that belong to a specific *test* sample are not present in any of the other four. Thus, if all the *test* samples are concatenated, they will result in the whole original sample (with 150 objects), therefore the tests will evaluate all the objects of the sample minimizing the bias of classifying the same objects which were used in the training process or classifying only a subset of the dataset [5]. The results with all the data sets are in table 3.

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Table 3. Results of the *tests* samples.

Class	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Mean	Standard Dev.
0	90	62,5	80	100	88,88	84,27	14,08
1	25	60	36,36	55,55	25	40,38	16,61
2	77,77	46,15	80	21,42	77,77	60,62	26,02
Total	**61.29**	**54.84**	**64.52**	**51.61**	**62.96**	**59,04**	**5,55**

The classification results shown in table 4 were obtained without considering the attributes ordering. The aim of the present work is to show improvements that can be obtained in the classification results if the attribute ordering is taken into account. Therefore, the next section presents the procedure for finding the best attribute order.

3.2. Ordering the attributes

The ordering process adopted is a simple exhaustive search for the best classification results. As there are four attributes, there are 24 different possible orderings.

For each possible ordering, the procedure for dividing the original dataset and classifying the five tests samples were applied, and the classification results were compared. The best outcome was achieved with the 19th ordering (table 4).

Table 4. Results of the 19th ordering *tests* samples.

Class	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Mean	Standard Dev.
0	100	100	100	72,72	100	94,54	12,19
1	72,72	75	80	100	60	77,54	14,55
2	88,88	81,81	69,23	61,53	100	80,29	15,32
Total	**87,1**	**83,87**	**80,65**	**74,19**	**81,48**	**81,45**	**4,77**

Comparing the results of the classification with and without the ordering process one can see that the results are promising. This better classification happens because of the K2 algorithm property [3] that considers the variables order to define the causal relationship between the problem variables. When testing all the possible ordering, there were some that brought worse results than the classification using the original order and some which brought better ones. Thus, one can see that the improvement in the classification results will depend on the quality of the original order. Anyway, searching for the best order will provide the guarantee that the classification results are not being prejudiced by the position of the variable in the dataset. Certainly, more examples should be tested, and a method that requires less computational effort must be developed (see more details in the conclusions section).

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

3.3. Missing values

To use this method having a dataset containing missing values, the substitution of missing values procedure proposed in [5] was adopted. As the IRIS database doesn't have any missing value in the original sample, and as we would like to observe the method when applied to samples with missing values attributes, we introduced some and performed the classification again to analyze the results.

Missing values were randomly [18] introduced in the attributes 1, 2, 3 and 4 (Sepal length, Sepal width, Petal length and Petal width) separately. Three new samples were generated for each one of the attributes, the first one having 10% of missing values (10% dataset), the second having 20% (20% dataset) and the third having 30% (10% dataset). Afterwards, the substitution of missing values was initiated.

Using the original sample (complete sample), four new samples were generated to be used as training samples to the substitution process (one sample to each attribute missing values substitution). Thus, a complete sample having the attribute (with missing values) positioned as "class attribute" was generated to attribute 1, 2, 3 and 4. Thereafter, the ordering process (section 3.2) was applied to each one. Therefore, a bayesian network, having the best variable order, was found to substitute missing values in each attribute, and it was used in the substitution process.

To verify the quality of the substitution process, a classification using the sample with substituted missing values was performed. The classification results using the 10% dataset are shown in table 5. In table 6 one can see the results corresponding to 20% dataset, and finally, table 7 shows the classification with the 30% dataset.

Table 5. Classification with 10% of missing values.

	Missing values only in attribute 1		Missing values only in attribute 2		Missing values only in attribute 3		Missing values only in attribute 4	
Class	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
0	96	5,47	96	5,47	93,68	5,90	90,18	7,08
1	75,81	11,34	81,32	13,57	75,55	21,71	81,36	22,81
2	69,94	9,93	78,80	16,43	88,49	11,67	77,49	15,92
Total	**79,33**	**8,73**	**83,33**	**6,23**	**83,73**	**6,19**	**81,52**	**3,96**

Table 6. Classification with 20% of missing values.

	Missing values only in attribute 1		Missing values only in attribute 2		Missing values only in attribute 3		Missing values only in attribute 4	
Class	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
0	96,51	4,78	94,34	5,56	92	13,03	93,58	8,79
1	89,62	12,56	73,74	31,29	73,34	12,95	84,01	7,75
2	83,94	12,73	76,94	23,21	61,55	32,22	80,26	14,56
Total	**87,53**	**2,48**	**78,66**	**8,80**	**73,27**	**10,11**	**83,59**	**3,70**

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Table 7. Classification with 30% of missing values.

	Missing values only in attribute 1		Missing values only in attribute 2		Missing values only in attribute 3		Missing values only in attribute 4	
Class	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
0	94,69	7,41	96,92	6,88	92,84	11,59	95,95	5,58
1	81,72	13,05	77,47	18,16	47,76	10,87	81,81	12,85
2	75,89	18,97	78,47	18,12	65,74	21,83	79,12	12,36
Total	**81,66**	**4,55**	**82,16**	**4,64**	**66,97**	**11,89**	**84,24**	**5,65**

It's worth saying that the classification results showed used all the datasets with the missing values already substituted, and the datasets containing missing values are independent from one another. Consequently, the objects containing missing values in the *10% dataset* may not be the same in the *20%* and in the *30% datasets*.

The datasets having the missing values substituted maintained the classification results very close to the classification having the complete data (except when the missing values were in attribute 3). More studies have to be done on this aspect, because the properties of the attributes may have an influence on these results, but one can see that as a first result, the numbers are promising.

More discussion and future work are presented in the next section.

4. Conclusion and future work

The results shown in the earlier section reveal that looking for an appropriate attribute order can improve the results in the classification task (at least when classifying data with the method used in this work). Hence, it's worthwhile to employ the ordering before classifying. Nevertheless, the procedure adopted to find the best order should be improved. The introduction of some pruning heuristics may be a good way to minimize the computational effort necessary for this search and permit the application in larger datasets.

When applying the attribute ordering process into the substitution of missing values with the method presented in [5], the results are not so determining, anyway they show that the classification was done without introducing great bias (even having 30% of missing values in one attribute). Except in the dataset containing missing values in the attribute 3, the classification results were consistent, revealing that the classification pattern was maintained. To assert that the substitution doesn't disturb the classification in any kind of data, more studies have to be performed.

The achieved results are encouraging and point to some interesting and promising future work.

The attribute ordering can be seen as a feature selector, and applying it to select the most relevant attributes in a dataset for a classification or clustering task, may bring about interesting results.

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

The substitution of missing values in datasets containing it in more than one attribute of the same object would reveal some interesting characteristics of the method.

The combination of this data preparation technique with other clustering or classification theories would reveal whether the method is robust or not.

5. References

[1] Jensen, F. V., *An Introduction to Bayesian Networks*. Springer-Verlag, New York, 1996.

[2] Pearl, J., *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, 1988.

[3] Cooper G. & Herskovitz, E.. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9, 309-347, 1992.

[4] Hruschka Jr., E. R. & Ebecken, N.F.F, Missing values prediction with K2. To appear in *Intelligent Data Analysis,* 2002.

[5] Castillo, E., Gutiérrez, J. M., Hadi, A. S., *Expert Systems and Probabilistic Network Models. Monographs in Computer Science*, Springer-Verlag, 1997.

[6] Heckerman, D., *A tutorial on learning bayesian networks*. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, Microsoft Corporation, 1995.

[7] Buntine, W., A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 1995

[8] Chow, C. K., & Liu, C. N., Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory* IT-14:462-67, 1968.

[9] Rebane, G. & Pearl, J., The recovery of causal poly-trees from statistical data. *Proceedings of Third Workshop on Uncertainty in Artificial Intelligence*, pp 222-228, Seattle, 1987.

[10] Buntine, W., Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, (2):159-225, 1994a.

[11] Heckerman, D., Geiger, D., Chickering, D. M., *Learning bayesian networks: The combination of knowledge and statistical data.* Technical Report MSR-TR-94-09 (Revised), Microsoft Research, Advanced Technology Division, July 1994.

[12] Bouckaert, R. R., *Bayesian belief networks: from inference to construction.* PhD thesis, Faculteit Wiskunde en Informatica, Utrech Universiteit, June 1995.

[13] Ramoni, M., Sebastiani, P., *An Introduction to Bayesian Robust Classifier.* KMI Technical Report KMI-TR-79, Knowledge Media Institute, The Open University, 1999.

[14] Han, J. & Kamber, M., *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

[15] Liu, W. Z., White, A. P., Thompson, S. G. and Bramer, M. A., Techniques for Dealing with Missing Values in Classification. *Advances in Intelligent*

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Data Analysis, Lecture Notes in Computer Science, LNCS1280, pages 527-536, 1997.

[16] Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annual Eugenics*, 7, Part II, 179-188 (1936); also in *Contributions to Mathematical Statistics*, John Wiley, NY, 1950.

[17] PYLE, D., *Data Preparation for Data Mining.* Morgan Kaufmann Publishers, 1999.

[18] Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P., *Numerical Recipes in C: The Art of Scientific Computing*. Second Edition, Cambridge University Press, 1992.