World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:5, No:3, 2011

Improving Classification in Bayesian Networks using Structural Learning

Hong Choon Ong

***Abstract*—**Naïve Bayes classifiers are simple probabilistic classifiers. Classification extracts patterns by using data file with a set of labeled training examples and is currently one of the most significant areas in data mining. However, Naïve Bayes assumes the independence among the features. Structural learning among the features thus helps in the classification problem. In this study, the use of structural learning in Bayesian Network is proposed to be applied where there are relationships between the features when using the Naïve Bayes. The improvement in the classification using structural learning is shown if there exist relationship between the features or when they are not independent.

***Keywords*—**Bayesian Network, Classification, Naïve Bayes, Structural Learning.

I. Introduction

BAYESIAN networks are factored representations of probability distributions that generalize the naive Bayesian classifier. They are a cross fertilization of ideas between the artificial intelligence, probability and statistics. They have received a lot of attention from both scientists and engineers across a number of fields [1]. Bayesian networks are different from other knowledge-based systems tools because uncertainty is handled in mathematically rigorous yet efficient and simple way. They give compact representation of joint probability distributions via conditional independence. Bayesian network is a data mining technique which not only enables efficient uncertainty reasoning with hundreds of variables, but also enables humans to understand the modelled domain better

A Bayesian network or directed acyclic graphical model is a probabilistic graphical model $G = (N, A)$ whose nodes N, and arcs A, represent random variables X, and direct correlations between the variables respectively [2]. The model selection in a Bayesian network consists of two steps. The first step entails learning the structure or causal dependencies which encodes the conditional independence relationships in the data. The second step involves fitting the parameters of the local distribution given the structure learned in the previous step. Many of the Bayesian network construction algorithms are based on the "node sequence already known" condition and the purpose is to reduce and simplify the complexity of the structure.

Hong Choon Ong, senior lecturer, School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia; (e-mail: hcong@cs.usm.my).

II. Classification in Bayesian Network

One of the most important areas in data mining is classification. There are a number of popular rule based algorithms for classification [3]. The classification algorithms are in the supervised learning group because they build a model based on supplied classes. Classification covers a wide range of data mining and many approaches have been discovered. These approaches include various rule-based classification algorithms like decision tree based algorithms [4], partial decision trees [5], support vector machines [6], neural networks [7] and Naive Bayes [8]. A classifier is a global model which gives a concise and clear description for each class by using attributes or features of data files.

One of the most important areas in data mining is classification. There are a number of popular rule based algorithms for classification [3]. The classification algorithms are in the supervised learning group because they build a model based on supplied classes. Classification covers a wide range of data mining and many approaches have been discovered. These approaches include various rule-based classification algorithms like decision tree based algorithms [4], partial decision trees [5], support vector machines [6], neural networks [7] and Naive Bayes [8]. A classifier is a global model which gives a concise and clear description for each class by using attributes or features of data files.

Naïve Bayes are effective because instead of estimating an n-dimensional distribution for $X_1, \ldots, X_n$, given the class which is too costly, they estimate n one-dimensional conditional distributions as can be seen from the equation above. They have been found in varied applications such as [10] for text classifications and have been found to outperform many other classifiers [11]. They can also be used to predict political risk level of a country to be used for investors who intend to achieve accurate information on the stability of the business environment in [12]. The structure of the Naïve Bayes network is shown in Figure 1. However, the strong independence assumption among the features raises the question of whether a classifier with less restrictive assumption can perform even better.

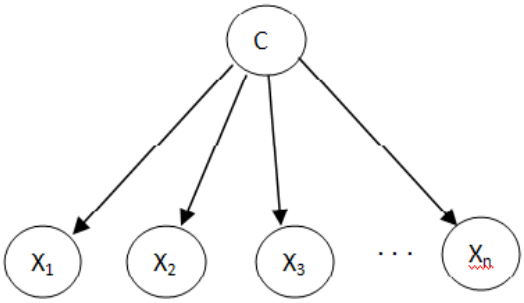

Fig.1 Illustration of the Naïve Bayes network structure

Several researchers have examined ways of achieving better performance than Naive Bayes. Friedman et al. [11] in particular consider among other structures, Tree Augmented Naïve Bayes, which allows arcs between the children of the classification node. Among the approaches which are less restrictive include Tree Augmented Naïve Bayes (TAN) and the general Bayesian networks [13]. A common feature of these networks is that the class node is treated as the parent of all the features. For TAN, each variable has as parents, the class variable and at most one other attribute. Many variants of the Naïve Bayes have been developed and some may combine more than one approach [14]. Comparison has been made on the Naïve Bayes, TAN and the general Bayesian networks but in a number of cases the general Bayesian networks are better while in others, Tan and Naïve Bayes are better [13]. Differences in experimental methodology might account for some of the disparities in conclusions drawn from work in [13] and that of [11].

III. Learning in Bayesian Network

A Bayesian network is composed of the network structure and its conditional probabilities. The main purpose of learning in a Bayesian network is twofold, which is to determine first the structure of the network (model selection) and then, the set of conditional probability tables associated with the structure learned (parameter learning) [15]. Several algorithms have been proposed for the inductive learning of general Bayesian network.

Structural learning is computationally complex because the number of possible structures is extremely huge. For classification problems, the class node is treated as the first node in the ordering and other nodes are treated as they appear in the dataset. Learning the structure implies learning the conditional independence from observations. There are two groups of approaches for Bayesian network structure learning [16]. In one approach, probabilistic relations using the Markov property with conditional independence test is used to learn and analyze the network structure and then a graph is built which satisfies the corresponding d-separation statements.

Alternatively, a score metric is used on each candidate Bayesian network and some heuristic search algorithm is used to maximize it. Greedy search algorithms, such as hill-climbing or tabu search can be used [17]. The criterion for selecting a structure is such that it maximizes the likelihood and at the same time minimizes the complexity. The log-likelihood score favour the complete graph structures but in order to avoid over fitting a penalty function is added based on the number of parameters.

The best Bayesian network is the one that best fits the data and leads to the scoring based algorithms that seek a structure that maximizes the scoring function. Log-likelihood (loglik) score is equivalent to the entropy used by Witten and Frank [18] where the maximized likelihood is decomposed by the network structure and the complexity penalty decomposes too.

Score functions commonly used in both discrete and continuous data are penalized likelihood scores such as the Akaike and Bayesian Information criteria, posterior densities such as the Bayesian Dirichlet and Gaussian equivalent scores and entropy-based measures such as the minimum Description Length [18]. The Akaike Information criteria (AIC) and Bayesian Information criteria (BIC) are decomposable scores for learning Bayesian network structures which are independent of the data but depends on the structure. The Bayesian Dirichlet Equivalent (bde) score also uses Bayesian analysis to evaluate a network given the dataset. The logarithm of the K2 score is another Dirichlet posterior density score developed by Cooper and Herskovits [19].

The job of estimating and updating the parameters of the global distribution of the network is greatly simplified by applying the Markov property when the structure of the network has been learned from the data. The network which is learned will represent an approximation to the probability distribution governing the domain. With enough samples, this approximation will be a good estimate. Therefore, we can use this network to compute the probability of class C, given the values of the attributes. The predicted class C, given a set of attributes X_1, X_2, …,X_n is simply the class that attains the maximum posterior $P_G(C \mid X)$ where $X = (X_1, X_2, ..., X_n)$ and P_G is the probability distribution representing the Bayesian network G.

IV. Experimenntal Methodology

Our main objective is to show that structural learning among the attributes in a classification dataset help to give a much better accuracy than the Naïve Bayes classifier. In this study 6 datasets meant for classification are used and taken from the UCI repository of machine learning datasets [20]. The 6 datasets with the corresponding number of attributes or features, number of instances and number of classes are given in Table 1.

Table 1
UCI Machine Learning Repository data sets for classification used.

Dataset	No. of Instances	No. of attributes	No. of classes
Hayes-Roth	160	5	3
Lenses	24	4	3
Tic-Tac-Toe Endgame	958	9	2
SPECT Heart	267	22	2
Shuttle Landing Control	15	6	2
Balance Scale	625	4	3

The Hayes-Roth data set involves human subjects study and has 5 numeric-valued attributes. It is a classification task which brings the background knowledge to bear on the attribute values and their relationship. The lenses data set is the most famous data set used in data mining and is used for fitting contact lenses. It is complete and noise free and the examples highly simplified the problem. Tic-Tac-Toe Endgame data set is a binary classification task on possible configurations of tic-tac-toe game. The database encodes the complete set of possible board configurations at the end of the tic-tac-toe games. The SPECT Heart data set describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified as either normal or abnormal. The database was processed to extract features that summarize the original SPECT images. Shuttle Landing Control data set is a small database of all nominal values. It generates comprehendible rules for determining the conditions under which an auto landing would be preferable to manual control of the spacecraft. The balance Scale data set was generated to model psychological experimental results. Each example is classified as having the balance scale tip to the right, tip to the left or be balanced.

Naïve Bayes classifier is used to do classification for all the 6 datasets and an example is shown of the Naive Bayes network structure in Figure 2 for the Shuttle Landing Control dataset which has 15 instances, 2 classes and 6 attributes or features.

It was suggested by Yingyu et al [21] that researchers use Bayesian network learning to explore the potential relationship between the variables. Structural learning using hill climbing algorithm is used in this study to learn the dependencies and causal relationship among the attributes or features. After running the learning algorithm by using the bnlearn package [22], it can be determined whether there are any dependencies among the attributes. If there are, then arcs are constructed to show the dependencies as shown in Figure 3 where there is an arc from the node 'sign' to the node 'wind' for the Shuttle Landing Control dataset and Figure 4 for the Tic-Tac-Toe Endgame dataset where there are 5 directed arcs among the attributes, both using the hill climbing algorithm. Figure 5 shows the network structure with added relationship after the structural learning using the hill climbing algorithm on the attributes from the SPECT Heart data set

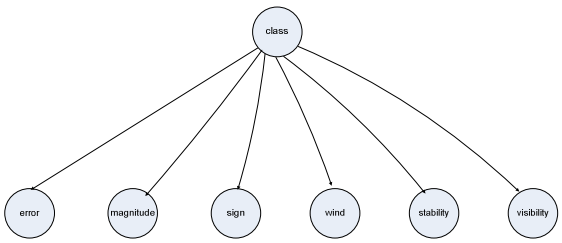

Fig. 2. The Naïve Bayes network structure from Shuttle Landing Control data set

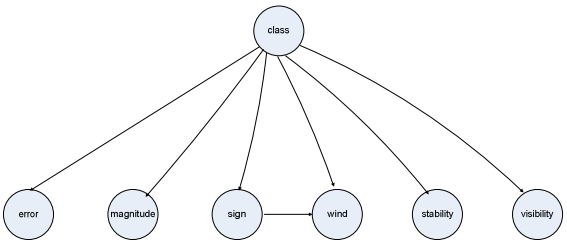

Fig. 3. The Naïve Bayes network structure with added relationship after structural learning using the hill climbing algorithm on the features/attributes from Shuttle Landing Control data set.

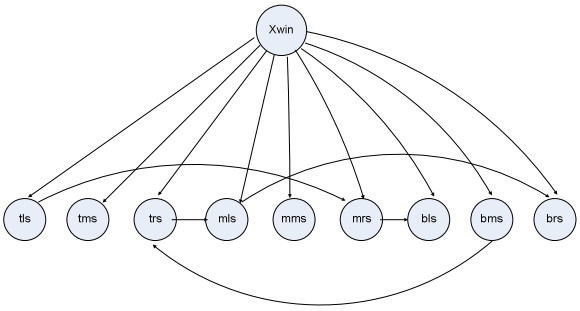

Fig. 4. The Naïve Bayes network structure with added relationship after structural learning using the hill climbing algorithm on the features/attributes from Tic-tac-toe data set

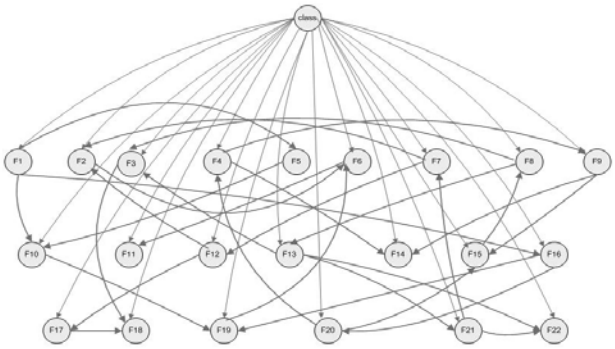

Fig. 5. The Naïve Bayes network structure with added relationship after structural learning using the hill climbing algorithm on the features/attributes from SPECT Heart data set.

V. Results

Table 2 lists the accuracy for the Naïve Bayes classifier and the classifier with added relationship in the presence of dependencies obtained among the attributes using the hill climbing algorithm. 10 fold cross validation is used in this study. Owing to the fact that Naïve Bayes classifiers assume independence of the attributes or feature, only when there are dependencies among the attributes or feature from the structural learning, do we have the added relationship.

The cross validation splits the data, D into 10 approximately equal parts $D_1, D_2, \ldots, D_{10}$ and learns on the data $D \backslash D_i$, $1 \leq i \leq 10$ with one part left out. The part D_i left out is used as the test set. The learning procedure is carried out a total of 10 times on different training. Finally, the 10 error estimates are averaged to yield an overall error estimate. The accuracy based on the percentage of successful prediction on the test sets of each dataset is given in Table 2.

All the learning algorithms from the bnlearn package show no relationship among the feature/attributes in the Lenses and Balance Scale dataset. As such, there is no added relationship in the structure among the attributes nodes. However, all the remaining 4 datasets which have relationship and have arcs constructed, showed an improvement in the accuracy over the Naive Bayes classifier as shown in Table 2 and Figure 7. Hence, it is advisable to perform structural learning on the attributes when doing a classification problem using Naïve Bayes classifier in situations where the attributes are not independent.

To further show and test our findings above, several score functions like the log-likelihood scores, logarithm of the K2 score, Bayesian Dirichlet Equivalent, Akaike Information Criterion and Bayesian Information Criterion are shown in Table 3 for estimating the fitting of the algorithms. It can be seen from Table 3 that all the scores for Tic-Tac-Toe Endgame, SPECT Heart and Shuttle Landing Control are higher for the added structure as compared to the Naive Bayes network.

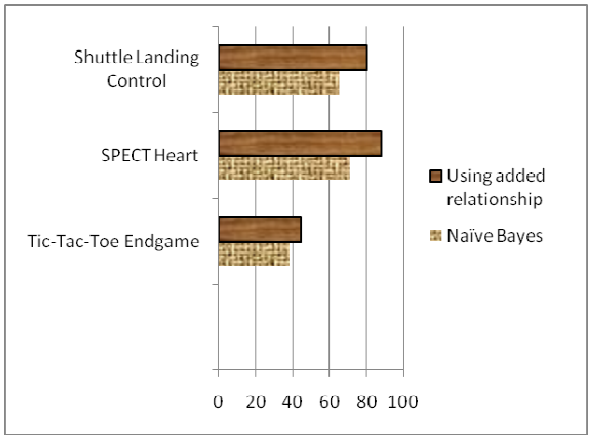

Fig. 6. Comparison of accuracies using the added structural relationship

TABLE 2
Comparison of accuracy between Naïve Bayes network and the network with added relationship through structural learning of the features using 10 fold cross validation

UCI Dataset	Attributes relation	Accuracy using Naïve Bayes	Accuracy using added relationship among attributes
Hayes-Roth	no	67.47%	-
Lenses	no	60.00%	-
Tic-Tac-Toe Endgame	yes	38.25%	44.65%
SPECT Heart	yes	70.99%	87.95%
Shuttle Landing Control	yes	65.00%	80.00%
Balance Scale	no	83.19%	-

VI. Conclusions

In this study, structural learning is proposed to be applied when the attributes in a classification dataset are not independent. The results show that learning the dependencies among the attributes or features and constructing the added relationship when doing classification helps to improve the accuracy of the Naïve Bayes classification, especially when the dependency is strong

For further work, structural learning can be applied using other score based algorithms besides the hill climbing algorithm. Alternatively, the constraint based methods can be applied in places where the dependency is weak and may not be captured by the hill climbing algorithm.

Table 3.
Comparison of various scores for the three data sets with added relationship

Dataset	Network	loglik	K2	bde	AIC	BIC
Tic-Tac-Toe Endgame	Naïve Bayes	-9697.467	-9796.478	-9779.634	-9734.467	-9824.466
	With added structure	-9123.78	-9462.339	-9546.326	-9312.78	-9772.509
SPECT Heart	Naïve Bayes	-2558.81	-2653.211	-2752.324	-2603.81	-2676.510
	With added structure	-1926.523	-2171.540	-2657.125	-2049.523	-2248.236
Shuttle Landing Control	Naïve Bayes	-81.13365	-111.3118	-114.4997	-112.1331	-123.1078
	With added structure	-69.74502	-108.4316	-115.2126	-108.7450	-122.552

ACKNOWLEDGMENT

This work was supported in part by the Universiti Sains Malaysia (USM) Research University grant no. 1001/PMATH/817037

REFERENCES

[1] A. Darwiche, *Modelling and Reasoning with Bayesian Networks.* Cambridge University Press, Los Angeles, 2009.

[2] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: networks of plausible inference.* Morgan Kaufmann, 1988.

[3] M. M. Mazid, S. B M. Shawkat Ali and K. S. Tickle, Input space reduction for Rule Based Classification. *WSEAS Transactions on Information Science and Applications.* 2010, Vol. 7, Issue 6, pp. 749-759.

[4] J. R. Quinlan, *C4. 5: programs for machine learning.* San Mateo, CA: Morgan Kaufmann, 2003

[5] I. H. Witten and E. Frank, Generating accurate rule sets without global optimization, *in Proceedings of the Fifteenth International Conference*, San Francisco, CA, 1998.

[6] V. N. Vapnik, *The nature of statistical learning theory*: Springer Verlag, Heidelberg, DE, 1995.

[7] S. L. Ang, H. C. Ong, and H. C. Low, Criterion in selecting the clustering algorithm in Radial Basis Functional Link Nets *WSEAS Transactions on Systems.* 2008, Vol. 11, Issue 7, pp. 1290-1299.

[8] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, *in The 20th Int'l Conference on Very Large Databases*, Santiago, Chile, 1994.

[9] J. Cheng and R. Greiner, Learning Bayesian Belief Network Classifiers: Algorithms and System, *Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence* 2001, pp. 141 – 151.

[10] A. Almonayyes, Multiple Explanations Driven Naïve Bayes Classifier, *Journal of Universal Computer Science* 2006, Vol. 12, Issue 2, pp. 127-139.

[11] N. Friedman, D. Geiger and M. Goldszmidt, Bayesian Network Classifiers, *Machine Learning*, Vol. 29, 1997, pp. 131-163.

[12] Siavash Asadi Ghajarloo. Mining Implicit Knowledge to Predict Political Risk by Providing Novel Framework with using Bayesian Network. *World Academy of Science, Engineering and Technology*. 2011, Vol. 74, pp. 656-663.

[13] M. G. Madden, On the classification performance of TAN and general Bayesian networks *Knowledge-Based Systems,* 2009, Vol. 22, Issue 7, pp. 489-495.

[14] K. M. Al-Aidaroos, A. A. Bakar, and Z. Othman, Naive Bayes variants in classification learning, *International Conference on Information Retrieval and Knowledge Management* 2010, pp. 276-281.

[15] P. Sebastiani, M. M. Abad, and M. F. Ramoni, *Bayesian Networks. Data Mining and Knowledge Discovery Handbook*, Eds. Maimon, O. and Rokach, L. Part 2, Chapter 10, 175-208, Springer, New York 2010.

[16] M. Scutari, Learning Bayesian Networks with the bnlearn R Package, *Journal of Statistical Software*, 2010, Vol. 35, Issue 3, pp. 1–22.

[17] M. Scutari and K. Strimmer, *Introduction to Graphical Modelling. Chapter for the upcoming Handbook of Statistical Systems Biology* Balding, D., Stumpf, M., Girolami, M. eds. 21 pages. 2010.

[18] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition Morgan Kaufmann, San Francisco, 2005.

[19] G. F. Cooper and E. A. Herskovits, Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 1992, Vol. 9, pp. 309-347.

[20] A. Frank, and A. Asuncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], (Irvine, CA: University of California, School of Information and Computer Science, 2010).

[21] G. Yingyu, L. Chunping and Qin,Y. Study on Factors of Floating Women's Income in Jiangsu Province Based on Bayesian Networks. *Advances in Neural Network Research and Applications*, *Lecture Notes in Electrical Engineering* 2010, Vol. 67, Issue 9, pp. 819-827.

[22] R Development Core Team. A Language and Enviroment for Statistical Computing. [http://www.R-project.org]. *R Foundation for Statistical Computing*, Vienna, Austria, 201.