Quality Measures for Bayesian Network Classifier Design

Marián Mach, Miroslav Mráz

Department of Cybernetics and Artificial Intelligence, Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic, Marian.Mach@tuke.sk

Abstract: The paper is dedicated to classification of documents into one of available classes. The role of a classifier is played by Bayesian network classifiers having the structure of an augmented Naive Bayes classifier. The focus is on quality measures enabling to compare different Bayesian networks and select the better one while searching a space of possible network structures. Experiments with several quality measures and several types of network structures were carried out using an English document collection.

Keywords: document classification, Bayesian networks, Bayesian classifier, augmented Naive Bayes, quality measure, recall and precision

1 Introduction

Classification task steadily attracts attention of many practitioners and researchers thanks to new applications related with information acquisition from the Web (for example see [5]). Most of used approaches ignore relationships among attributes. Since considering these relationships can be a way how to increase the performance of classification, new classification methods are called for. The use of Bayesian networks seems to be very promising from this point of view.

The problem with this approach is that many algorithms for learning Bayesian networks need a measure in order to compare different networks and select the better one. Since classification is not a typical task performed by Bayesian networks (a special network structure is required), widely used quality measures are not guaranteed to be suitable for this task.

2 Defining a Lingo

The following notation is used in this paper:

B – Bayesian network over random variables $\{X_1,\ldots,X_n\}$, $B=(S,\theta)$

S – structure of a Bayesian network

θ - parameters of a Bayesian network

X_i – discrete variable having r_i possible values

C – class variable

A_i – attribute variable, i=1,...,n-1

D – set of training cases, $||D|| = N$

Pa_i – set of parents of variable X_i, having q_i possible combination of values

N_{ijk} – number of cases in D where random variable X_i is in configuration k and its parents are in configuration j

N_{ij} – number of cases in D where parents of random variable X_i are in configuration j

3 Bayesian Network Classifier

Using Bayesian networks for the classification task [7] imposes some constraints on the network structure. In our case variables can be divided into two groups. One variable represents classification classes – the variable is multinomial having so many different possible values as the number of classification classes is. This variable is called class variable. The other variables represent the presence or absence of features which are important from the point of classification. Those variables are binary and are called attribute variables.

Similarly, arcs can be divided into two groups as well:

- Classification arcs – arcs between the class variable and one of attribute variables. This type is compulsory, there must be defined at least one classification arc in Bayesian network classifier.
- Augmenting arcs – arcs between two attribute variables. They are optional.

The presence and the number of classification and augmenting arcs depend on a particular network structure.

Learning of a Bayesian network classifier is based on a set of training data. It results into determining the exact network structure, prior probability of the class variable, and local conditional probability distributions corresponding to attribute variables. For example, it is possible to use learning algorithms aiming at approximation of joint probability distribution of all variables.

The classification itself can be carried out as calculation of maximum posterior probabilities of available classification classes resulting from the presence or absence of relevant features.

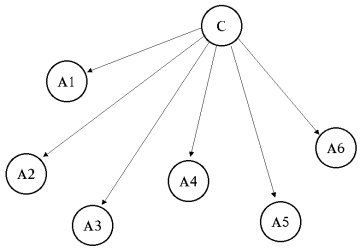

Figure 1
Naive Bayes structure

It is possible to use different types of structures which differ from each other by possible arrangements of arcs, both classification and augmenting ones. Most popular structure is the one depicted in Figure 1. This structure represents Naive Bayes classifier (NB), containing only classification arcs between the class variable and each attribute variable.

3.1 Augmented Naive Bayes

Many network structures used by Bayesian network classifiers are of 'augmented Naive Bayes classifier' type. They modify the basic NB structure in two ways – by adding augmenting arcs (representing relationships among attribute variables), and by removing classification arcs.

We have considered the following structures (Figure 2):

STAN (Selective Tree Augmented Naive Bayes). This structure imposes a tree-like relationship structure on attribute variables. But unlike [3], not each attribute variable must be related with the class variable. It is possible to omit some classification arcs (but not all – at least one should be present).

Learning is performed as searching a network candidate space (to learn candidate network structures a method for learning tree-like structures by Chow and Liu [2] can be used). The search starts with an empty set of classification arcs and a greedy procedure tries to add as many classification arcs as possible (an arc is added only if the resulting structure is better according to a quality measure).

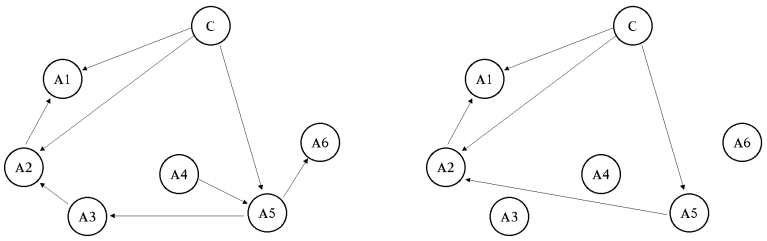

Figure 2
STAN (left) and STAND (right) network structures

STAND (Selective Tree Augmented Naive Bayes with Discarding). This structure is based on the same principle as the STAN structure. The only difference is, that those attribute variables, which are not related with the class variable directly, are not considered when generating tree-like relationship structure. As a result, these attribute variables are considered to be completely irrelevant.

4 Quality Measures

Quality measure selection is crucial for estimating the quality of learned Bayesian networks. Since the selected measure guides the whole search through the space of all network structures, the decision on its selection heavily influences the quality of the final network structure – with a direct impact on the quality of the classification process.

Basically, all quality measures can be divided into two categories – global and local measures [7]. Global measures are those widely used in Bayesian network business. While assessing a network, they take the complete network into account without preference for a particular part of the network. Therefore, each variable is important equally – the one representing classification classes as well as the one standing on behalf of a particular attribute.

On the other hand, local quality measures prefer a particular part of the network and the quality evaluation is based on this part. Not surprisingly, the variable representing classification classes is a hot candidate for the preferred part of the network when solving a classification task – the network will be evaluated at the class variable.

4.1 Heckerman-Geiger-Chickering Measure

This measure represents a global quality measure. It aims at the posterior probability of the network structure *S* given a training data set *D*.

$$p(S \mid D) = \frac{p(S,D)}{p(D)} \tag{1}$$

Since *D* is the same for each evaluated network (i.e. it plays the role of a constant), only *p(S,D)* part is considered (using chain rule):

$$Q_{HGC} = \log p(S,D) = \log p(S) + \log p(D \mid S) \tag{2}$$

Since *p(S)* is unknown, we assume a uniform probability distribution over all possible *S*. It results in a possibility to omit the structure probability element.

The second constituent of the formula can be expressed using the Bayesian Dirichlet metric [4]. It enables to transform the quality measure into the following formula

$$Q_{HGC} = \sum_{i=1}^{n}\left[\sum_{j=1}^{q_i}\left[\log\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})} + \sum_{k=1}^{r_i}\log\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}\right]\right] \tag{3}$$

where α_{ijk} and α_{ij} represent Dirichlet prior parameters, and Γ represents the gamma function

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt \tag{4}$$

4.2 Standard Bayesian Measure

This measure is also a global one. It considers the quality of approximation of the joint probability distribution. The evaluated quality can be measured as posterior probability of the network *B* given training samples *D*:

$$p(B \mid D) = p(S,\theta \mid D) = \frac{p(S,\theta,D)}{p(D)} \tag{5}$$

Similarly to the Q_{HGC} measure, *D* is the same for each evaluated network. Therefore, only *p(S,θ,D)* part is taken into account. The use of chain rule enables us to substitute it with the product of three probabilities

$$p(S,\theta,D) = p(S)p(\theta \mid S)p(D \mid \theta,S) \tag{6}$$

Naturally, the measure formulated in an above given way will prefer larger networks, since they have more parameters enabling better fit. On the other hand, the more parameters, the bigger chance to overfit the network to training data leading to poor generalisation. Therefore, the quality measure penalises the size of the network.

The composition of both parts of the measure provides the following formula

$$Q_{SB} = \log p(S) + \log p(\theta \mid S) + \log p(D \mid S, \theta) - \frac{1}{2} Dim(B) \log N \tag{7}$$

Assuming that all variables in the network have multinomial distribution leads to

$$Q_{SB} = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} (N_{ijk} + \alpha_{ijk} - 1) \log \frac{N_{ijk} + \alpha_{ijk} - 1}{N_{ij} + \alpha_{ij} - r_i} - \frac{1}{2} Dim(B) \log N \tag{8}$$

where *p(S)* was neglected for it is a constant (we assume each network structure to be equally probable).

This size of a Bayesian network can be calculated as the number of free parameters required to completely specify the joint probability distribution

$$Dim(B) = \sum_{i=0}^{n} (r_i - 1) q_i = \sum_{i=0}^{n} (r_i - 1) \prod_{X_p \in Pa_i} r_p \tag{9}$$

4.3 Local Criterion Measure

This measure is an example of a local quality measure. It is based on an idea to employ posterior probability of test data collection given a network structure and training data

$$p(x^{(l)} \mid D^{(l)}, S) \tag{10}$$

where $D^{(l)}=\{x^{(1)},x^{(2)},...,x^{(l-1)}\}$ represents training cases, while the l[st] case serves as a test case. A measure based directly on the above given posterior probability would represent a global measure since the complete network (all variables) is taken into account. Transforming this idea into the form of a local criterion (only the variable representing classification classes matters) provides the following formula

$$Q_{LC} = \sum_{i=1}^{N} \log p(c^{(l)} \mid a^{(l)}, D^{(l)}, S) \tag{11}$$

where $c^{(l)}$ is value of the class variable and $a^{(l)}$ is the configuration of attributes in the 1st case.

The measure assumes the use of a certain number of cases (all but last are used to train the network, the last one is used as a test case) while this number incrementally increases.

4.4 Leave-one-out Cross Validation

This measure is based on the same idea as the previous one, including the same transformation from a global measure to a local one. The only difference is, that the number of used cases does not increase incrementally but it is constant – all cases are used. One of them plays the role of a test case while the others are utilised to train the network. The quality measure has the form

$$Q_{LOO} = \sum_{i=1}^{N} \log p\left(c^{(l)} \mid a^{(l)}, V_l, S^h\right) \tag{12}$$

where V_l represents the data set D with the 1st case removed.

4.5 Φ-fold τ-times Cross Validation

This local measure is based on the same idea of the marginal probability of the class variable as the previous local measures. The difference is in the management of dividing the data set into test and training parts. The data cases are divided into Φ disjoint subsets which are approximately of the same size – one of them plays the role of a test set while the others represent training data. In this way there is Φ possibilities to form a training set - test set pair. Moreover, the Φ-fold validation is repeated τ times.

The formula for this quality measure has the following form

$$Q_{CV\Phi\tau} = \sum_{j=1}^{r} \sum_{i=1}^{\Phi} \sum_{l=1}^{\left\|W_i^{(j)}\right\|} \log p\left(c^{(l)} \mid a^{(l)}, V_i^{(j)}, S\ \right) \tag{13}$$

where V_i stands for training data while W_i represents test data.

In particular case $\Phi=||D||$ and $\tau=1$ this measure is the same as Leave-one-out cross validation measure.

5 Experiments and Results

In order to carry out experiments [6], we used an English document collection Reuters-21578. The collection was divided into training and test sets closely following ApteMod division of the collection.

The documents from the collection were preprocessed by removing auxiliary words (defined in a stoplist for the English language) and transforming the other words using Porter's stemmer. Those categories, containing less than one hundred documents, were removed from the collection. Terms with very low document frequency (less than five documents) were removed from the list of used terms. The preprocessing phase resulted in a collection of 8520 documents, 10 classification categories, and 4359 terms. Finally, the list of terms was reduced to 200 terms using information gain criterion.

The aim of our experiments was to compare the performance of different quality measures used in the learning step of Bayesian network classifier design. The comparison was based on comparing the performance of final classifiers designed using different quality measures (two different network structures were used – STAN and STAND). The performance was evaluated by employing precision and recall criteria.

Precision was calculated according to the following formula

$$\frac{TP}{TP + FP} \tag{14}$$

where TP (true positive) represents the number of test cases classified correctly and FP (false positive) stands for the number of cases which were classified into a class despite they do not belong to the class.

Recall was calculated according

$$\frac{TP}{TP + FN} \tag{15}$$

where FN (false negative) represents the number of test cases which were not classified into a class despite they belong to the class.

Since the preprocessed document collection contained ten classification categories, some kind of composition of the results achieved for each category should be used. We employed macro averaging – precision and recall are calculated for each classification category with subsequent averaging of these values.

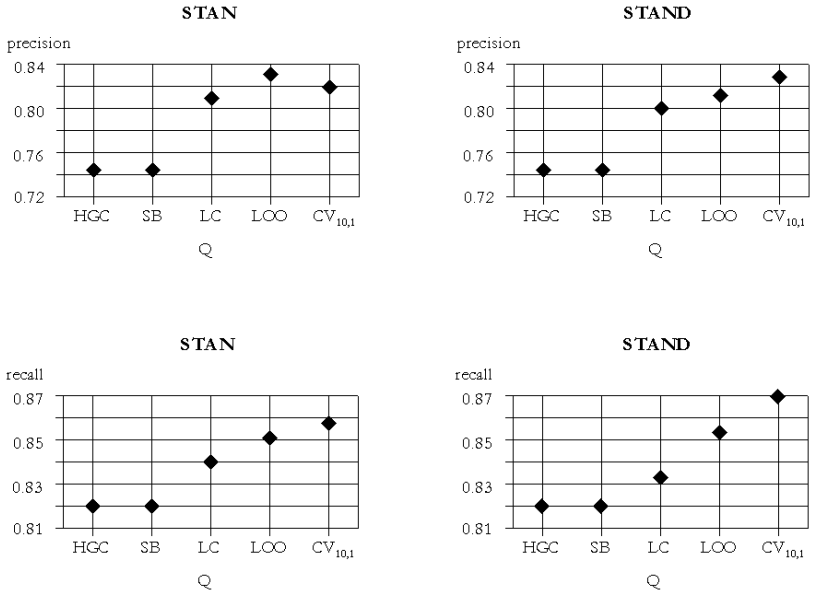

Figure 3
Experiment results

Achieved results are presented in Figure 3. A clear distinction between global and local quality measures was justified. The performance of global quality measures is not satisfactory for the document classification task. Selecting a local quality measure is a better choice.

Conclusions

The aim of the paper was to compare several quality measures for learning Bayesian networks in respect to a document classification domain. Experiments have proven the importance of selecting quality measures suitable for a particular application domain when learning Bayesian networks. A proper choice can increase the performance of the final classifier by several percentages.

Acknowledgements

The work presented in the paper was supported by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic within the 1/1060/04 project ”Document classification and annotation for the Semantic web”.

References

[1] Cooper, G. F., Herskowits, E.: A Bayesian Method for the Induction of Probabilistic Networks form Data. *Machine Learning, 9*, 1992, 309-347

[2] Chow, C. K., Liu, C. N.: Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory, 14*, 1968, 462-467

[3] Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. *Machine Learning, 29*, 1997, 131-163

[4] Heckerman, D., Geiger, D., Chickering, D. M.: Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning, 20*, 1995, 197-243

[5] Machová, K., Klimko, I.: Adaptive Web support by means of machine learning. *Proc. of the 3rd Int. conference ZNALOSTI 2004*, Brno, 2004, ISBN 80-248-0456-5, 218-225

[6] Mráz, M.: *Document classification using Bayesian networks*. MSc Theses, The University of Košice, 2004, 64 pages

[7] Sacha, J. P.: New Synthesis of Bayesian Network Classifiers and Cardiac SPEC Image Interpretation. PhD Dissertation, The University of Toledo, 1999, 172 pages