A New MDL-based Function for Feature Selection for Bayesian Network Classifiers

Mădălina M. Drugan and **Linda C. van der Gaag** [1]

Abstract. Upon constructing a Bayesian network classifier from data, the accuracy of the resulting model can often be improved upon by selecting a subset of the available features. We show that the commonly used MDL function is not suited for feature selection. We introduce a new MDL-based function that is better tailored to this task. Our experimental results demonstrate that, with the new function, classifiers are yielded that have an accuracy comparable to the ones found with the MDL function, yet include fewer features.

1 INTRODUCTION

Real-life datasets often include more features, or attributes, of the recorded instances of every-day problem solving than are strictly necessary for the classification task at hand. When constructing a Bayesian network classifier from such a dataset, these more or less redundant features may bias the output and as a consequence result in a relatively poor classification accuracy. By constructing the classifier over just a subset of the features, a less complex model is yielded that tends to have a better generalisation performance [1].

For constructing a Bayesian network classifier from data, generally a heuristic algorithm is employed that searches the space of possible models for classifiers of high quality. For comparing the qualities of alternative classifiers, often the Minimum Description Length (MDL) function is employed. This function weighs the complexity of a classifier against its ability to capture the observed probability distribution. While generally accepted as a suitable function for comparing classifiers over a fixed set of attributes, we argue that the MDL function is not quite suited for identifying and removing redundant attributes upon feature selection. We further argue that the poor feature-selection behaviour of the function can be attributed, to at least some extent, by its not using the conditional probability distribution that is of interest for the classification task.

Building upon our analysis of the feature-selection behaviour of the MDL function, we introduce the closely related MDL-FS function. The MDL-FS function differs from the MDL function only in that it encodes the conditional probability distribution over the class variable instead of the joint distribution over all variables. Since learning a conditional distribution is known to be hard, our function uses an auxiliary Bayesian network to support this task. To compare the feature-selection behaviour of the two functions in a practical setting, we conducted various experiments using different datasets. Our results indicate that the new MDL-FS function is indeed more suited for the task of feature selection than the MDL function as it yields classifiers of comparably good performance with fewer attributes.

[1] Institute of Information and Computing Sciences, Utrecht University, P.O. Box 80.089, 3508 TB Utrecht, the Netherlands. E-mail: {madalina,linda}@cs.uu.nl

2 PRELIMINARIES

We consider a set A of stochastic variables A_i, called attributes, where each A_i takes one of a finite set of values; we further consider a designated class variable C. In addition, we consider a dataset D with $N \geq 1$ labelled instances; this dataset defines the observed joint probability distribution $\hat{P}(C, A)$ over $A \cup \{C\}$. For any value assignment S^k to any subset S of variables, we write $N(S^k)$ to denote the number of instances in D for which $S = S^k$.

In general, a classifier is a function that assigns a unique class value to each unlabelled instance over the set of attributes. For reasons of space, we focus in this paper on TAN classifiers only. A TAN classifier over $A \cup \{C\}$ is a Bayesian network classifier that includes for its graphical structure an acyclic digraph in which the attributes constitute a directed tree. The classifier specifies a prior probability distribution $P(C)$ for its class variable and, for each attribute $A_i \in A$, a conditional distribution $P(A_i \mid p(A_i))$, where $p(A_i)$ are the parents of A_i in the graphical structure; these distributions with each other define a joint probability distribution $P(C, A) = P(C) \cdot \prod_{A_i \in A} P(A_i \mid p(A_i))$ over all variables involved [2].

A selective classifier includes just a subset A' of the available attributes. The attributes from A' are deemed important for the classification task at hand, whereas the attributes from $A \setminus A'$ are considered to be redundant [3]. We say that an attribute A_i is *redundant* for the class variable C given a subset of attributes $S \subseteq A \setminus \{A_i\}$, if for every value A_i^k of A_i, every value C^g of C, and every value assignment S^j to S with $N(A_i^k, S^j) > 0$, we have that

$$\frac{N(A_i^k, S^j, C^g)}{N(A_i^k, S^j)} = \frac{N(S^j, C^g)}{N(S^j)}$$

If $|S| = m$, we say that A_i is redundant for C at level m. Note that, if A_i is redundant for C given S and $N(S^j, C^g) > 0$, then C is independent of A_i given S in the observed probability distribution.

3 MDL AND FEATURE SELECTION

Building upon the MDL principle, the best classifier to explain the observed data is one that minimises the sum of the length of an encoding of the classifier itself and the length of an encoding of the data given the classifier. For a classifier $\mathcal{C}$, the MDL function is defined as

$$\mathit{MDL}(\mathcal{C} \mid D) = \frac{\log N}{2} \cdot |\mathcal{C}| - \mathit{LL}(\mathcal{C} \mid D)$$

where the term $|\mathcal{C}|$ captures the length of the encoding of the classifier; the associated term $\frac{\log N}{2} \cdot |\mathcal{C}|$ is commonly known as the penalty term of the function. The log-likelihood term $\mathit{LL}(\mathcal{C} \mid D)$ captures the length of the encoding of the observed distribution $\hat{P}(C, A)$ factorised over the graphical structure of the classifier; the term equals

$-N \cdot \sum_{A_i \in A} H_{\hat{P}}(A_i \mid p_{\mathcal{C}}(A_i)) - N \cdot H_{\hat{P}}(C)$, where $H_{\hat{P}}(\cdot)$ denotes the entropy of $\hat{P}$. The smaller the MDL value of a classifier, the better it is. The larger the value of the log-likelihood term, the better the classifier models the observed distribution. A fully connected classifier has the largest log-likelihood, yet will show a poor generalisation performance as a result of overfitting. The penalty term now counterbalances the effect of the log-likelihood term by increasing in value as a classifier becomes more densely connected.

The MDL function is generally accepted as a suitable function for comparing the qualities of alternative classifiers over a fixed set of attributes. The function is less suited, however, for feature selection. As an example, we consider, within a TAN classifier $\mathcal{C}$, an attribute A_i that is a leaf node in the tree of attributes; we assume that A_i has the attribute A_j for its parent in the tree. Upon comparing the MDL value of $\mathcal{C}$ with that of the selective TAN classifier $\mathcal{C}^-$ that is obtained by deleting A_i and its incident arcs, we find that the MDL function prefers $\mathcal{C}$ over $\mathcal{C}^-$ if and only if

$$H_{\hat{P}}(A_i \mid A_j, C) - H_{\hat{P}}(A_i) < -\frac{\log N}{2 \cdot N} \cdot (|\mathcal{C}| - |\mathcal{C}^-|)$$

Now suppose that A_i is redundant for the class variable C at level 0 and at level 1 given $\{A_j\}$. Only if the relationship of A_i with its parent A_j is very weak, will the MDL function prefer the selective classifier: informally speaking, the stronger the relationship of A_i with A_j, the closer to 0 the term $H_{\hat{P}}(A_i \mid A_j, C)$ will be. The above inequality then is likely to hold, inducing the MDL function to prefer the full classifier. In practice, the MDL function is found to eliminate upon feature selection hardly any attributes from a TAN classifier.

4 MDL-FS AND FEATURE SELECTION

Bayesian network classifiers represent a joint probability distribution $P(C, A)$ over their variables, while it is the conditional probability distribution $P(C \mid A)$ that is of interest for the classification task. In designing a new function, we build upon the assumption that the poor feature-selection behaviour of the MDL function originates, to at least some extent, from not using the conditional distribution. Our new MDL-FS function captures, like the MDL function, the joint distribution $P(C, A)$; in addition, it captures the distribution $P(A)$ over the set of attributes. Where $P(C, A)$ factorises over the structure of the classifier under study, the function uses an auxiliary network $\mathcal{S}$ to factorise $P(A)$. It now encodes the conditional distribution $P(C \mid A)$ by means of the difference between the log-likelihood of the classifier that encodes the joint distribution and the log-likelihood of the auxiliary network that encodes $P(A)$. For a classifier $\mathcal{C}$ and the auxiliary network $\mathcal{S}$, the function is defined as

$$\textit{MDL-FS}(\mathcal{C}, \mathcal{S} \mid D) = \frac{\log N}{2} \cdot |\mathcal{C}| - \textit{CLL}(\mathcal{C}, \mathcal{S} \mid D)$$

where $|\mathcal{C}|$ is as before and $\textit{CLL}(\mathcal{C}, \mathcal{S} \mid D) = \textit{LL}(\mathcal{C} \mid D) - \textit{LL}(\mathcal{S} \mid D)$. Like the MDL function, the MDL-FS function includes a penalty term and a log-likelihood term. We note that the penalty term captures the length of the encoding of just the classifier $\mathcal{C}$: we decided to exclude the auxiliary network from the term, because we are interested in the complexity of the resulting classifier only. The MDL-FS function as a consequence has no control over the complexity of the auxiliary network. Upon applying the MDL-FS function, therefore, a suitable class of auxiliary networks is set beforehand.

The MDL-FS function is better suited for feature selection than the MDL function. As an example, we consider a TAN classifier $\mathcal{C}$ and a tree-structured Bayesian network $\mathcal{S}$ of maximum log-likelihood; for ease of exposition, we assume that the graphical structure of $\mathcal{S}$ coincides with the tree of attributes of $\mathcal{C}$. We consider again a leaf attribute A_i and its parent A_j. We compare the MDL-FS value of $\mathcal{C}$ with that of the selective TAN classifier $\mathcal{C}^-$ that is obtained by deleting A_i and its incident arcs; the selective auxiliary network $\mathcal{S}^-$ is obtained accordingly. We find that the MDL-FS function prefers $\mathcal{C}$ over $\mathcal{C}^-$ if and only if

$$H_{\hat{P}}(A_i \mid C, A_j) - H_{\hat{P}}(A_i \mid A_j) < -\frac{\log N}{2 \cdot N} \cdot (|\mathcal{C}| - |\mathcal{C}^-|)$$

We suppose again that A_i is redundant for the class variable C at level 0 and at level 1 given $\{A_j\}$. We then have that $H_{\hat{P}}(A_i \mid C, A_j) = H_{\hat{P}}(A_i \mid A_j)$, regardless of the strength of the relationship between A_i and A_j. Since the above inequality does not hold, the MDL-FS function prefers the selective classifier and removes the attribute. Recall that the MDL function tends not to do so. In general, the stronger the relationship of an attribute with its neighbours in the auxiliary network and the weaker this relationship in the classifier, the more inclined the MDL-FS function will be to remove it.

Table 1. The accuracies of the constructed selective TAN classifiers.

dataset	MDL		MDL-FS	
	% sel. attr.	*accuracy*	*% sel. attr.*	*accuracy*
chess	97 ± 1	0.93 ± 0.01	46 ± 5	0.92 ± 0.01
mushrooms	93 ± 2	1.00 ± 0	55 ± 0	1.00 ± 0
splice	74 ± 8	0.95 ± 0	14 ± 1	0.95 ± 0.01
spambase	95 ± 0	0.92 ± 0.01	67 ± 3	0.92 ± 0
oesoca	97 ± 2	0.74 ± 0	60 ± 2	0.74 ± 0
ext. oesoca	95 ± 1	0.74 ± 0.01	41 ± 1	0.74 ± 0.01

5 EXPERIMENTAL RESULTS

To study the difference in behaviour of the MDL and MDL-FS functions in a practical setting, we constructed selective TAN classifiers from various datasets using both functions. For our study, we used four datasets from the UCI Irvine repository and two additional medical datasets that were generated from a real-life Bayesian network. Table 1 summarises the results of our experiments, showing the averages and standard deviations obtained over ten runs per dataset. We observe that the MDL scoring function does not substantially reduce the numbers of attributes, as expected. The MDL-FS function, with a tree-structured auxiliary network of maximum log-likelihood, on the other hand, does remove considerable numbers of attributes, without reducing the classification accuracy of the resulting classifiers.

6 CONCLUSIONS

Based upon an analysis of the feature-selection behaviour of the commonly used MDL function, we introduced a new MDL-based function that we tailored to the task of identifying and removing redundant attributes upon constructing a Bayesian network classifier from data. We argued that the MDL-FS function is more suited to the task of feature selection than the MDL function and supported our observations by experimental results obtained from various datasets.

REFERENCES

[1] A. Blum, P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, **97**, 1997, pp. 245 – 271.

[2] N. Friedman, D. Geiger, M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, **29**, 1997, pp. 131 – 163.

[3] G.H. John, R. Kohavi, K. Pfleger. Irrelevant features and the subset selection problem. *Proc. of the 11th International Conf. on Machine Learning*, 1994, pp. 121 – 129.