Learning an Optimal Naive Bayes Classifier

Miriam Martínez-Arroyo
Instituto Tecnológico de Acapulco
Acapulco, Guerrero, México
miriamma-ds@hotmail.com

L. Enrique Sucar
INAOE
Tonantzintla, Puebla, México
esucar@inaoep.mx

Abstract

The naive Bayes classifier is an efficient classification model that is easy to learn and has a high accuracy in many domains. However, it has two main drawbacks: (i) its classification accuracy decreases when the attributes are not independent, and (ii) it cannot deal with nonparametric continuous attributes. In this work we propose a method that deals with both problems, and learns an optimal naive Bayes classifier. The method includes two phases, discretization and structural improvement, which are repeated alternately until the classification accuracy cannot be improved. Discretization is based on the minimum description length principle. To deal with dependent and irrelevant attributes, we apply a structural improvement method that eliminates and/or joins attributes, based on mutual and conditional information measures. The method has been tested in two different domains with good results.

1 Introduction

A classifier, in general, produces a mapping from data (attributes) to two or more predefined classes. The naive Bayes classifier is a classification model that has several advantages [2]: it is easy to learn and understand, it is very efficient, it is in general robust and has a high accuracy. However, it has two main drawbacks: (i) its classification accuracy decreases when the attributes are not independent, (ii) it cannot deal with nonparametric continuous attributes. In this work we propose a method to learn an optimal naive Bayesian classifier that deals with both problems. The method includes two phases, discretization and structural improvement. Discretization is based on the minimum description length (MDL) principle. To deal with dependent and irrelevant attributes, we apply a structural improvement method that eliminates and/or joins attributes, based on mutual and conditional information measures. Although there is previous work in discretization, feature selection and structural improvement, we integrated all aspects in a coherent and complete methodology, which includes novel aspects in discretization and structural improvement. We have tested this method in two domains: (i) skin detection in images, and (ii) diagnosis of cervical cancer. In both domains our method produces a very simple and efficient classifier with nearly 100% accuracy.

2 Related work

2.1 Bayesian classifiers

A Bayesian classifier obtains the posterior probability of each class, \( C_i \), using Bayes rule. The naive Bayes classifier (NBC) makes the simplifying assumption that the attributes, \( A \), are independent given the class, so the likelihood can be obtained by the product of the individual conditional probabilities of each attribute given the class. Thus, the posterior probability, \( P(C_i | A_1, \ldots, A_n) \), is given by:

\[
P(C_i | A_1, \ldots, A_n) = P(C_i)P(A_1 | C_i) \cdots P(A_n | C_i)/P(A)
\]

As we mentioned before, two important problems for the NBC are how to deal with dependent and continuous attributes. We know previous work in both aspects.

3 Discretization

Discretization methods for classifiers can be divided into two main types: (i) unsupervised, and (ii) supervised. Unsupervised methods do not consider the class variable, so the continuous attributes are discretized independently. Supervised methods consider the class variable, so that the division points are selected in function of the value of the class for each data point. The problem of finding the optimal number of intervals and the corresponding limits can be seen as a search problem. That is, we can generate all possible division points over the range of each attribute (where there is a change of class), and estimate the classification error for each possible partition. Unfortunately, generating and testing all possible partitions is impractical. In the
worst case, there are in the order of \(2^{MN}\) possible partitions, where \(M\) is the number of attributes and \(N\) is the number of possible partition points per attribute. For Bayesian classifiers, [6] introduces a method that, starting from an initial partition, it makes an iterative search for a better one, by joining or splitting intervals, and testing the classification accuracy after each operation. [4] does discretization while learning the structure of a Bayesian network. For a given structure, a local search procedure finds the discretization for a variable that minimizes the description length in relation to the adjacent nodes in the graph, and this is repeated iteratively for each continuous variable. Valdes et al. [9] present a technique based on evolution strategies. The attributes are discretized to maximize class predictability. This method leads to simple models and can discover irrelevant attributes. A survey of discretization methods for naive Bayes classifiers is given in [10].

3.1 Structural improvement

The naive Bayes classifier assumes that the attributes are independent given the class. If this is not true, there are two basic alternatives. One is to transform the structure of the classifier to a Bayesian network, by introducing directed arcs between the dependent attributes [3]. The disadvantage is that the simplicity of the NBC is lost. The other alternative is to transform the structure maintaining a star or tree-structured network. For this, [8] introduces 3 basic operations: (i) eliminate an attribute, (ii) join two attributes into a new combined variable, (iii) introduce a new attribute that makes two dependent attributes independent (hidden node). Figure 1 illustrates the first two operators, which are the ones we use in this work. These operations are based on statistical tests to measure the correlation of pairs of attributes given the class variable. [7] proposes an alternative algorithm for variable elimination and merging. The algorithm is based on two search procedures: (i) forward sequential selection and joining and (ii) backward sequential elimination and joining. It starts from a full (empty) structure, and selects attributes for elimination (addition) or for combination, testing the classification accuracy after each operation. The advantage of these approaches is that they preserve the simplicity and efficiency of the NBC. Previous work has considered the problems of discretization and structure improvement separately. In this work we propose a novel algorithm that combines both aspects to learn an efficient and accurate NBC.

4 Methodology

The proposed method learns a Bayesian classifier which preserves a NBC structure and at the same time considers (i) discretization of continuous variables, (ii) elimination of irrelevant attributes, (iii) elimination or combination of dependent attributes. The basic algorithm is:

1. Initialization: (a) Define the initial structure and discretization. (b) Estimate parameters for the initial structure.
2. Discretization: (a) Improve the discretization based on MDL. (b) Evaluate the discretization.
3. Structural Improvement: (a) Improve the structure based on variable elimination and combination. (b) Evaluate the structure.
4. Evaluation: Evaluate the classifier on test data.

Steps 2 and 3 are repeated iteratively until the structure–discretization can not improved. Next we describe each stage in detail.

4.1 Initialization

This step is done once to build the initial classifier. It considers all the attributes and an initial partition for the continuous attributes with two equal size intervals.

4.2 Discretization

Given the current structure, in this stage the discretization for each continuous attribute is optimized. From the initial discretization, it generates additional partitions based on the MDL principle. Each attribute is processed independently, by splitting each interval into two parts, and testing each partition using the MDL. From these, the split with the best measure is selected; and the process is repeated with
the new partition iteratively until the MDL can not be improved. The MDL measure makes a compromise between accuracy and complexity. The measure we use is similar to the one proposed in [5], which estimates the accuracy (Net-Length) by measuring the mutual information between the attribute and the class; and the complexity (Net-Weight) by counting the number of parameters required. A constant $\alpha$, in $[0, 1]$, is used to balance the weight of each aspect, accuracy vs. complexity. Thus, the quality measure is:

$$Quality = (1 - \alpha) \times \left(1 - \frac{Net - Length}{Max - Length}\right) + \alpha \times \frac{Net - Weight}{Max - Weight}$$

Where $Net - Length$ is proportional to the number of parameters required in the model, and $Net - Weight$ corresponds to the sum of the mutual information between each attribute and the class; which gives an estimate of the model precision. The maximum length, $Max - Length$, and weight, $Max - Weight$, are estimated by considering the maximum number of intervals per attribute. An $\alpha = 0.5$ gives equal weight to accuracy and complexity, while an $\alpha$ close to 1 gives more importance to accuracy.

### 4.3 Structural improvement

Given the current discretization, in this phase the structure is improved to eliminate superfluous attributes and eliminate or combine dependent attributes. This phase considers the following stages:

1. The mutual information between each attribute and the class is obtained, and those attributes that do not provide information (below a threshold) are eliminated.

2. The remaining attributes are tested for conditional mutual information (CMI) for each pair of attributes given the class. If this value is high it is an indication that the attributes are not independent.

3. For each pair of attributes with high CMI: (a) eliminate one (the one with less mutual information with the class), or (b) combine both into a single attribute. The option with better classification rate is selected.

This process is repeated until there are no more superfluous or dependent attributes.

The final stage consists on evaluating the accuracy of the final classifier with test data.

### 5 Experimental results

#### 5.1 Skin classification

An important problem in computer vision is to discriminate skin and non skin pixels in images. For this experiment we combined 3 different color models: $RGB$, $HSV$ and $YIQ$. All the attributes are continuous and there are two classes. We used samples taken from several persons under different illumination conditions. We used approx. 7,000,000 data points (pixels), 70% for training and 30% for testing. In both, training and test data sets, approx. 50% of the cases are skin and 50% non skin. After the discretization stage, we applied the structural improvement stage. Starting from the optimal discretization, the method applies the variable elimination and combination stages until the simplest classifier with maximum accuracy is obtained. The sequence of operations and the final classifier are depicted in table 1.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Operation</th>
<th># Attr.</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>9</td>
<td>RGBHSVYIQ</td>
</tr>
<tr>
<td>1</td>
<td>Elim. B</td>
<td>8</td>
<td>RGBHSVYIQ</td>
</tr>
<tr>
<td>2</td>
<td>Elim. Q</td>
<td>7</td>
<td>RGBHSVYI</td>
</tr>
<tr>
<td>3</td>
<td>Elim. H</td>
<td>6</td>
<td>RGSVYI</td>
</tr>
<tr>
<td>4</td>
<td>Join RG</td>
<td>5</td>
<td>R-GSVYI</td>
</tr>
<tr>
<td>5</td>
<td>Elim. V</td>
<td>4</td>
<td>R-GSYI</td>
</tr>
<tr>
<td>6</td>
<td>Elimi. S</td>
<td>3</td>
<td>R-GYI</td>
</tr>
<tr>
<td>7</td>
<td>Eliminate I</td>
<td>3</td>
<td>R-GY</td>
</tr>
</tbody>
</table>

#### 5.2 Diagnosis of cervical cancer

As a second test case we consider the diagnosis of cervical cancer based on the analysis of colposcopy images. Diagnosis is based on the temporal evolution of the coloration in the interest region in the image. In previous work [1], three different mathematical models were used to model the evolution of each pixel and based on this to diagnose the probability of cancer:
Model 1: the times of the main changes in the signal: $T_s, T_b, T_c$.

Model 2: a polynomial that approximates the signal, with parameters: $P_1, P_2, P_3, P_4, P_5$.

Model 3: a parabola approximation with parameters: $P_h, P_k, P_p$.

We applied our methodology by considering the 3 models, so all the parameters are the attributes for the classifier, 11 continuous attributes. There are three classes: normal, low degree and high degree. For the experiments we used 1055 sample data labeled by an expert: 800 for training and 255 for testing. After obtaining the best discretization for each attribute, the results of the structural improvement phase are summarized in table 3. The final structure has 4 attributes: $T_s-P_k, T_b, P_2, P_4$; where one of them is the combination of two attributes of the original models. This final classifier has an accuracy of 99% accuracy. In comparison, the results with the best single model were of 91%, using the same data and testing procedure.

6 Conclusions and future work

We have presented a novel method for learning an optimal naive Bayes classifier. The method includes two phases, discretization and structural improvement. The discretization phase is based on the MDL principle. The structural improvement phase eliminates or joins attributes, based on conditional information measures. The main contribution of our work is a complete method for improving the naive Bayes classifier that integrates discretization, feature selection and structural improvement, generating a simple and accurate classifier. In contrast with other extensions of the NBC, this method maintains a simple structure making the final classifier more efficient. Experimental results in two different and challenging problems show excellent results. In the future we plan to extend the methodology for learning dynamic classifiers with hidden variables.

### Table 2. Comparison of different color models for skin classification in images.

<table>
<thead>
<tr>
<th>Color Model</th>
<th>Precision in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGB</td>
<td>94</td>
</tr>
<tr>
<td>HSV</td>
<td>93</td>
</tr>
<tr>
<td>Yc,r,c_b</td>
<td>91</td>
</tr>
<tr>
<td>R-GY</td>
<td>98</td>
</tr>
</tbody>
</table>

### Table 3. Results for cancer diagnosis.

<table>
<thead>
<tr>
<th>Operation</th>
<th>Attr.</th>
<th>attributes</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Join T_s-P_k</td>
<td>11</td>
<td>T_{s-b} T_c P_{1-5} P_{5h} P_{5p}</td>
<td>92</td>
</tr>
<tr>
<td>Elim. P_1</td>
<td>9</td>
<td>T_{s-b} T_c P_{2-5} P_{5h} P_{5p}</td>
<td>94</td>
</tr>
<tr>
<td>Elim. P_p</td>
<td>8</td>
<td>T_{s-b} T_c P_{2-5} P_{5h} P_{5p}</td>
<td>94</td>
</tr>
<tr>
<td>Elim. P_h</td>
<td>7</td>
<td>T_{s-b} T_c P_{2-5} P_{5p}</td>
<td>94</td>
</tr>
<tr>
<td>Elim. P_3</td>
<td>6</td>
<td>T_{s-b} T_c P_{2-5} P_{5p}</td>
<td>95</td>
</tr>
<tr>
<td>Elim. T_b</td>
<td>5</td>
<td>T_{s-b} P_{2-5} P_{5p}</td>
<td>98</td>
</tr>
<tr>
<td>Elim. P_5</td>
<td>4</td>
<td>T_{s-b} P_{2-5} P_{5p}</td>
<td>99</td>
</tr>
</tbody>
</table>

### References


