NAIVE BAYESIAN CLASSIFIER AND CONTINUOUS ATTRIBUTES

INFORMATICA 1/92

Keywords: machine learning,
naive Bayesian classifier, continuous atributes
Ključne besede: avtomatsko učenje,
naivni Bayesov klasifikator, zvezni atributi

Igor Kononenko
University of Ljubljana
Faculty of Electrical and Computer Engineering
Tržaška 25, 61001 Ljubljana, Slovenia

Abstract

The advantages of the naive Bayesian classifier are fast and incremental learning, robustness with respect to missing data, the inclusion of all available attributes during classification and the explanation of classification as the sum of information gains. Besides the 'naivety', the weakness is also the inability to deal with continuous attributes unless they are discretized in advance. In the paper three methods for dealing with continuous attributes are proposed. The fuzzy learning method assumes fuzzy bounds of a continuous attribute during learning, the fuzzy classification method assumes fuzzy bounds during classification and the last method tries to increase the reliability of probability approximations. The performance was tested in two medical diagnostic problems.

Naivni Bayesov klasifikator in zvezni atributi
(povzetek)

Prednosti naivnega Bayesovega klasifikatorja so hitro in inkrementalno učenje, robustnost glede manjkajočih podatkov, uporaba vseh razpoložljivih atributov za klasifikacijo in razlaga odločitev kot vsota informacijskih prispevkov. Poleg naivnosti je slabost tudi nezmožnost obravnave zveznih atributov, ki morajo biti zato vnaprej diskretizirani. V članku so predstavljene tri metode za obravnavanje zveznih atributov. Mehko učenje in mehko testiranje predpostavljata mehke meje zveznih atributov med učenjem oziroma med klasifikacijo. Tretja metoda temelji na povečanju zanesljivosti aproksimacij verjetnosti. Uspešnost je bil testirana na treh medicinskih diagnostičnih problemih.

1 Introduction

The basic Bayesian formula for calculating the probability of a class given the values of attributes, describing a given object, can be used either for generation of decision trees (Michie & AlAttar 1990) or can be simplified into 'naive' Bayesian classifier by assuming the independence of attributes (Kononenko 1989).

The advantages of the naive Bayesian classifier are fast and incremental learning, robustness with respect to missing values, the inclusion of all available attributes for classification, and the ability to explain decisions as the sum of information gains (Kononenko 1990).

It was shown by several authors that, despite 'naivety', the naive Bayesian classifier performs in many real world problems approximately the same or even better than many well known inductive learning systems (Bratko & Kononenko 1987, Cestnik 1990, Kononenko 1990).

The advantages of induction of decision trees are 'nonnaivety', *simple and powerful mechanism for on line splitting of continuous attributes*, and explicit knowledge in the form of if-then rules. This paper is concerned with the problem of dealing with continuous attributes. The problem is to split the interval of possible values of a continuous attribute into subintervals to obtain as much as possible useful information for classification from a given attribute.

In the algorithms for induction of decision trees a continuous attribute is binarized on-line during learning. In a given node of a tree a bound is selected that maximizes the information gain of the attribute (Breiman et al. 1984, Paterson & Niblett 1982, Bratko & Kononenko 1987, Quinlan 1986). This approach can be applied for the naive Bayesian classifier only at the highest level by changing all continuous attributes into binary attributes. Another approach is to define all subintervals of-line before learning, either by a human expert or with an algorithm (e.g. Cestnik 1989). However, all these approaches assume, that the optimal split can be obtained using *exact bounds* of subintervals. This is unrealistic assumption as typically the slight difference among two values, one above and one below the bound, should not have drastic effect.

In this paper three new methods for dealing with continuous attributes are proposed. Two of them are based on the idea of *fuzzy bounds* and the third one is based on the reliability of probability estimation. In next section the naive Bayesian classifier is briefly described. In section 3 the three methods for dealing with continuous attributes are defined and in section 4 the experiments in two medical diagnostic problems are described. Finally in section 5 some conclusions are given and the further research is proposed.

2 Naive Bayesian Classifier

The classification problem discussed in this paper is the following: given a set of training instances, each described with a set of n attributes and each belonging to exactly one of a certain number of possible classes, learn to classify new, unseen objects. In addition, each attribute A_i has a fixed number of NV_i possible values.

Let C represent one of the possible classes. Let V_{i,J_i} be a Boolean variable having value 1 if the current instance has value J_i of i-th attribute and 0 otherwise. The conditional probability of class C given the values of all attributes is given with the following formula, derived from the Bayesian rule (Good 1950) (for brevity conditions $V_{i,j} = 1$ will be written simply as $V_{i,j}$):

$$P(C|V_{1,J_1}, \ldots, V_{n,J_n}) = P(C) \prod_{i=1}^{n} Q_i(C, J_i) \tag{1}$$

where

$$Q_i(C, J_i) = \frac{P(V_{i,J_i}|C, V_{1,J_1}, \ldots, V_{i-1,J_{i-1}})}{P(V_{i,J_i}|V_{1,J_1}, \ldots, V_{i-1,J_{i-1}})}$$

$$= \frac{P(C|V_{1,J_1}, \ldots, V_{i,J_i})}{P(C|V_{1,J_1}, \ldots, V_{i-1,J_{i-1}})} \tag{2}$$

and $P(C)$ is the prior probability of class C. It was shown in (Kononenko 1989) that the classification with ID3 like inductive learning system can be described with (1).

From (1) the naive Bayesian classifier, as used by Bratko and Kononenko (1987), is obtained if the independence of attributes is assumed. Eq. (1) remains unchanged except that factors Q_i defined with (2) are replaced with Q'_i (we will refer to changed equation (1) with (1')):

$$Q'_i(C, J_i) = \frac{P(V_{i,J_i}|C)}{P(V_{i,J_i})} = \frac{P(C|V_{i,J_i})}{P(C)} \tag{3}$$

The probabilities necessary to calculate (3) are approximated with relative frequencies from the training set. A new object is classified by calculating the probability for each class using equation (1') and the object is classified into the class that maximizes the calculated probability.

Cestnik (1990) has shown that the kind of approximation of probabilities in (3) considerably influences the classification accuracy of the naive Bayesian classifier. Let $N(C, V_{i,J_i})$ be the number of training instances with J_i-th value of i-th attribute and belonging to class C and let $N(V_{i,J_i})$ be the number of training instances with J_i-th value of i-th attribute. Usually, the probability is approximated with relative frequency, i.e.

$$\hat{P}(C|V_{i,J_i}) = \frac{N(C, V_{i,J_i})}{N(V_{i,J_i})} \tag{4}$$

However, if the training set is relatively small, the corrections are needed with respect to the assumption of initial distribution (Good 1965). Cestnik (1990) used the following formula stemming from the assumption, that initial distribution of classes is equal to $P(C)$:

$$\hat{P}(C|V_{i,J_i}) = \frac{N(C, V_{i,J_i}) + 2\hat{P}(C)}{N(V_{i,J_i}) + 2} \tag{5}$$

where the probability of class C is calculated using the Laplace's law of succession (Good 1950,1965):

$$\hat{P}(C) = \frac{N(C) + 1}{N + 2} \tag{6}$$

Cestnik has shown some nice properties of using approximation (5) in formula (3) and has shown experimentally, that the naive Bayesian classifier using approximation (5) performs significantly better than if (4) is used. The same formula was used also by Smyth and Goodman (1990).

3 Dealing with continuous attributes

The idea of all three methods defined in this section is the folowing. The task is to calculate the probabilities of all classes of an object with a given value of a continuous attribute. These probabilities should be approximated with relative frequencies calculated from the distribution of training instances with the similar value of the attribute. It is assumed, that small variations of the value of the attribute should have small effects on the probabilities. As opposed to exact bounds, where slightly different value can have drastic effects on the calculated probabilities, the bounds of intervals are here assumed to be *fuzzy.*

The three methods described in this section differ in the way how the distribution, used in the approximation of probabilities, is obtained. For all three methods the pessimistic set of possible bounds is given in advance either by a human expert or with a simple algorithm, that returns bounds with the uniform distribution of instances over all intervals. The set of bounds is pessimistic in the sense that more bounds are given than probably needed (e.g. all attributes have in advance 20 possible intervals, which is typically too detailed split). However, exact values of these initial bounds are not important and may vary without significant changes in performance.

3.1 Fuzzy learning

The fuzzy learning is performed by calculating the probability distribution for a given interval from all training instances rather than from instances that have value of a given continuous attribute in this interval. The influence of an instance is assumed to be normally distributed with mean value equal to the value of the regarded attribute and with given σ. σ is the parameter to the learning algorithm and is used to control the 'fuzziness' of the bounds. As shown in figure 1, the influence of a given instance with value v of the given continuous attribute on the distribution of interval $(b_j..b_{j+1})$ is proportional to the following expression:

$$P(v, \sigma, j) = \int_{b_j}^{b_{j+1}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-v}{\sigma}\right)^2} dx \tag{7}$$

If $\sigma = 0$ then the usual exact bounds are assumed and the distribution over classes in the

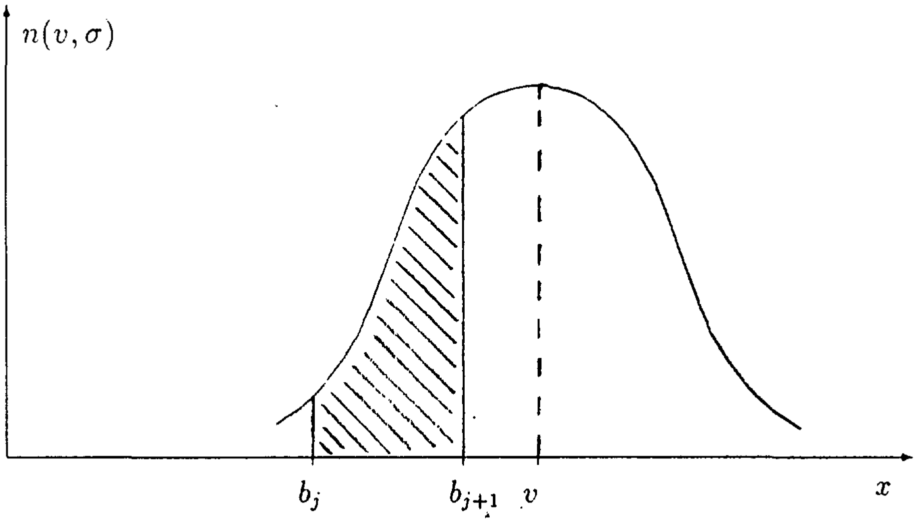

Figure 1: The normal distribution of the influence of:
- an instance over intervals of a continuous attribute for fuzzy learning,
- an interval on classification for fuzzy testing

given interval is calculated from relative frequency of training instances belonging exactly to that interval. The greater σ implies fuzzier bounds for continuous attributes. For example, for N learning instances, each with influence $P(v_k, \sigma, J_i)$, $k = 1..n$ on J_i-th interval of i-th attribute the probability that an instance belongs to that interval is calculated with:

$$\hat{P}(V_{i,J_i}) = \frac{\sum_{k=1}^{N} P(v_k, \sigma, J_i)}{N} \qquad (8)$$

3.2 Fuzzy classification

The fuzzy classification is performed by calculating the probability of all classes for a given object, given the value of the continuous attribute, from all intervals of that attribute rather than from the interval to which the object belongs. The influence of intervals is assumed to be normally distributed with mean value equal to the value of the regarded attribute of an object and with given σ.

σ is like in the previous method the parameter to the classification algorithm and is used to control the 'fuzziness' of the bounds. As shown on figure 1, the influence of a given interval on the probability of all classes is proportional to the expression (7). The expression (3) in equation (1') is here replaced with:

$$Q_i{}''(C) = \sum_{j=1}^{NV_i} P(v, \sigma, j) \times \frac{\hat{P}(C|V_{i,j})}{\hat{P}(C)} \qquad (9)$$

where v is the value of i-th attribute of a given object. Like in fuzzy learning method, for $\sigma = 0$ the usual exact bounds are assumed and the probabilities of all classes for the given object are calculated from relative frequencies of one interval only.

3.3 Reliable approximations of probabilities

The idea of the last method is to increase the reliability of approximation of the probabilities for given interval by adding to the interval the instances from neighbor intervals. For the estimation of the reliability of the probability approximation the theorem of Chebyshev (e.g. Vadnal 1979) can be used. The theorem gives the lower bound on the probability, that relative frequency f of an event after n trials differs from the factual prior probability p for less than ε:

$$P(|f - p| \leq \varepsilon) > 1 - \frac{p(1-p)}{\varepsilon^2 n} \qquad (10)$$

The lower bound is proportional to n and to

Table 1: Characteristics of two medical data sets.

domain	thyroid	rheumatology
# instances	884	355
# classes	4	6
# attributes	15	32
average # vals/att	9.1	9.1
average #missing data	2.7	0.0
majority class	56%	66%
entropy (bit)	1.59	1.70
# continuous atts	7	22
average # intervals/att	17.3	10.0

ε^2. In our case we are interested in the reliability of the approximation of probability (5). Therefore the number of trials n in (10) is equal to $N_{V_{i,J_i}}$, i.e. the number of training instances having value inside interval V_{i,J_i} of attribute A_i. As prior probability p is unknown, in our experiments for approximation of p at the right-hand side of (10) the worst case was assumed, i.e. $p = 0.5$. It remains to determine the value of ε. The influence of interval J_i of i-th attribute to class C is proportional to the difference between (5) and (6). If the influence is greater the reliability of approximation of (5) should be greater. Therefore ε should be proportional to the influence. As we regard the influence of an interval for all the classes C_j, $j = 1..n$, for ε the average difference is used:

$$\varepsilon = \sum_{j=1}^{n} \hat{P}(C_j) \times |\hat{P}(C_j|V_{i,J_i}) - \hat{P}(C_j)| \quad (11)$$

From above formulas it follows that the interval is unreliable if:

$$1 - \frac{1}{4\varepsilon^2 N_{V_{i,k}}} < P_r \quad (12)$$

where P_r is the lower bound of the probability (10) and is a parameter of the learning algorithm for controlling the reliability of approximations of probabilities. Now we can define an algorithm for postprocessing the distributions obtained by usual learning for naive Bayesian classifier. The algorithm is as follows:

```
for each continuous attribute $A_i$ do
  for each interval $V_{i,j}$ do
    while unreliable interval $V_{i,j}$ do
      add appropriate % (possibly 100%)
        of instances from neighbor
        intervals of attribute $A_i$
```

Note that here one training instance can influence more than one interval of a continuous attribute, analogously to fuzzy learning method.

4 Experiments in two medical diagnostic problems

We experimented with the naive Bayesian classifier and with three methods for dealing with continuous attributes in two medical diagnostic problems: diagnosing thyroid diseases and rheumatology. The characteristics of data sets used in our experiments are summarized in table 1. The medical data sets were collected at the University Medical Center in Ljubljana.

One run was performed by randomly selecting 70% of instances for learning and 30% for testing. Results are averages of 10 runs. For fuzzy learning and fuzzy testing methods parameter σ was determined with the following formula for each continuous attribute A_i:

$$\sigma_i = SIG \times \frac{upperbound_i - lowerbound_i}{\#intervals_i}$$

Table 2: Results of fuzzy learning and fuzzy classification.

		thyroid		rheumatology	
method	SIG	acc(%)	inf.(bit)	acc(%)	inf.(bit)
-	0.0	69.7	0.79	67.4	0.51
f.learn.	0.3	71.7	0.84	67.9	0.54
f.learn.	0.4	71.9	0.85	67.8	0.53
f.learn.	0.5	72.1	0.85	68.1	0.54
f.learn.	0.6	72.3	0.85	68.1	0.54
f.learn.	0.7	72.5	0.86	68.1	0.54
f.learn.	0.8	72.5	0.85	68.0	0.54
f.class.	0.3	71.4	0.84	59.6	0.38
f.class.	0.4	71.6	0.84	60.7	0.40
f.class.	0.5	70.8	0.82	62.6	0.45
f.class.	0.6	70.8	0.82	62.0	0.43
both	0.3	72.2	0.85	68.9	0.58
both	0.4	72.2	0.85	68.7	0.57
both	0.5	72.3	0.86	69.0	0.58
both	0.6	72.2	0.85	69.4	0.59

Table 3: Results with reliable approximations of probabilities (for $P_r = 0.0$ the result is of the original naive Bayesian classifier without reliable approximation of probabilities).

	thyroid		rheumatology	
P_r	acc(%)	inf.(bit)	acc(%)	inf.(bit)
0.0	69.7	0.79	67.4	0.51
0.1	70.6	0.80	67.0	0.51
0.2	71.0	0.80	67.5	0.52
0.3	70.9	0.80	67.9	0.53
0.4	70.9	0.80	68.3	0.53
0.5	70.7	0.79	69.2	0.55
0.6	70.5	0.79	69.3	0.53

Table 4: The comparison of performance of different classifiers in two medical domains.

classifier	thyroid		rheumatology	
	acc(%)	inf(bit)	acc(%)	inf(bit)
f.learn.	73	0.86	68	0.54
f.class.	72	0.84	61	0.40
f.learn&class	72	0.86	69	0.58
reliable app.	71	0.80	69	0.55
naive Bayes	70	0.79	67	0.51
Assistant	73	0.87	61	0.46
physicians	64	0.59	56	0.26

where SIG is parameter that was varied in our experiments. If $SIG = 1$ then σ_i is equal to the average interval length for i-th attribute. Therefore fuzziness is the function of average length of intervals. For last method the parameter P_r from (12) was varied.

Besides the average percent of correct guesses, the average information score per answer (Kononenko & Bratko 1991) was measured. Information score is the measure that eliminates the influence of prior probabilities of classes and can be applied to various kinds of incomplete and probabilistic answers. This measure is necessary as in each domain a classifier that classified each instance into the majority class (see table 1) would achieve high classification accuracy.

Results are given in tables 2 and 3. In table 2, the results are given for fuzzy learning, fuzzy classification and both methods together for various values of parameter SIG. In table 3 the results of reliable approximations of probabilities are presented for various values of parameter P_r. The combination of the latter method with other two did not give any improvements. In table 4 the results are compared with the accuracy of ID3-like inductive learning system Assistant (Cestnik et al. 1987), the naive Bayesian classifier without any method for dealing with continuous attributes and with the performance of physicians experts. The performances of physicians are the averages of four physicians experts in each domain that were tested in University Medical Center in Ljubljana.

5 Discussion

The results from table 4 show that proposed methods for dealing with continuous attributes perform better than splitting of attribute's values with exact bounds. The performance of the naive Bayesian classifier with these methods achieves and outperforms the performance of Assistant inductive learning system as well as the performance of physicians specialists. The performance of physicians is the worse, probably due to the inability to see the patient during the diagnostic process, when they were tested. Such diagnosing is, of course, unusual and unnatural for physicians. The results are presented to show that the performance of the learning systems is high enough, and not to show that the systems are better than physicians.

All three proposed methods for dealing with continuous attributes are based on the idea that a continuous attribute should not be discretized with exact bounds. It is interesting that none of the methods uses the information gain of the attribute as the measure for appropriate split. Further research should concentrate on the selection of appropriate values of parameters P_r and SIG. Obviously, optimal value of two parameters may differ among different attributes in the same problem domain. The appropriate value of the parameter may depend on the information gain of the attribute as well as on the amount of noise associated with values of the attribute.

The problem with the naive Bayesian classifier is the independence assumption. In some cases this may be too unrealistic assumption. But it seems that in the data used by human experts there are no strong dependencies between attributes because attributes are properly defined. With the independence assumption the reliability of approximating factors with relative frequencies is much greater. This is supported with experimental results. The naive Bayesian classifier despite its naiveness achieved good classification accuracy. There is a trade-off between the reliability of approximating probabilities and the errors due to the independence assumption (Kononenko 1989). An algorithm that tries to optimize this trade-off is described elsewhere (Kononenko 1991).

Acknowledgements

Physicians specialists Sergej Hojker and Vlado Pirnat from University Medical Center in Ljubljana provided the medical data and tested the physicians' performance. This research was done in the Artificial Intelligence Laboratory at the Faculty of Electrical & Computer Engineering in Ljubljana and was supported by Slovenian Research Community.

References

I.Bratko & I.Kononenko, (1987) Learning Rules from Incomplete and Noisy Data, in B. Phelps (ed.) *Interactions in Artificial Intelligence and Statistical Methods*, Hampshire: Technical Press.

Breiman, L. Friedman, J.H., Olshen, R.A., Stone C.J. (1984), *Classification and Regression Trees*, Belmont, California: Wadsworth Int. Group.

Cestnik B., Kononenko I.& Bratko I., (1987) ASSISTANT 86 : A knowledge elicitation tool for sophisticated users, in: I.Bratko, N.Lavrac (eds.): *Progress in Machine learning*, Wilmslow, England: Sigma Press.

Cestnik B. (1989) Informativity based splitting of numerical attributes into intervals, *Proc. IASTED Intern. Conf. Expert Systems Theory & Applications*, Zurich, Switzerland, June 26-28, 1989, pp.59-62.

Cestnik B., (1990) Estimating Probabilities: A Crucial Task in Machine Learning, *Proc. European Conf. on Artificial Intelligence*, Stockholm, Sweden, August 1990, pp.147-149.

Good, I.J., (1950) *Probability and the Weighing of Evidence*, London: Charles Griffin.

Good I.J., (1965) *The Estimation of Probabilities*, Cambridge: M.I.T. Press.

Kononenko I., (1989) ID3, Sequential Bayes, Naive Bayes and Bayesian Neural Networks. *Proc. 4th European Working Session on Learning*, Montpeiller, France, 4-6 December 1989, pp.91-98.

Kononenko I. (1990) Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition, in: B.Wielinga, J.Boose, B.Gaines, G.Schreiber, M. van Someren (eds.) *Current Trends in Knowledge Acquisition*, Amsterdam: IOS Press.

Kononenko I. (1991) Semi-Naive Bayesian Classifier, *Proc. 5th European Working Session on Learning*, Porto, Portugal, March 1991, pp.206-219.

Kononenko, I. & Bratko, I., (1991) Information Based Evaluation Criterion for Classifier's Performance. *Machine Learning*, Vol. 6, pp.67-80.

Michie, D., A. Al Attar (1991) Use of Sequential Bayes with Class Probability Trees. in: J.E. Hayes-Michie, D.Michie & E. Tyugu (eds.) *Machine Intelligence 12*, Oxford: Oxford University Press.

Quinlan, J.R. (1986) Induction of Decision Trees. *Machine Learning*. Vol. 1, no. 1, pp. 81-106.

Paterson, A., Niblett, T. (1982) The ACLS User Manual. Intelligent Terminals Ltd. Glasgow.

Smyth P. & Goodman R.M. (1990) Rule Induction Using Information Theory in: G.Piarersky & W.Frawley (eds.) *Knowledge Discovery in Databases*, MIT Press.

Vadnal A. (1979) *An Elementary Introduction to Probability Calculus*, (in Slovenian), Ljubljana: Državna Založba Slovenije.