SEMI-NAIVE BAYESIAN CLASSIFIER

Igor KONONENKO
University of Ljubljana, Faculty of electrical & computer engineering
Tržaška 25, 61001 Ljubljana, Yugoslavia

Abstract

In the paper the algorithm of the 'naive' Bayesian classifier (that assumes the independence of attributes) is extended to detect the dependencies between attributes. The idea is to optimize the tradeoff between the 'non-naivety' and the reliability of approximations of probabilities. Experiments in four medical diagnostic problems are described. In two domains where by the experts opinion the attributes are in fact independent the semi- naive Bayesian classifier achieved the same classification accuracy as naive Bayes. In two other domains the semi-naive Bayesian classifier slightly outperformed the naive Bayesian classifier.

Keywords: machine learning, Bayesian classifier, approximations of probabilities, (in)dependence of events

1 Introduction

Let $A_i, i = 1 \ldots n$ be a set of attributes, each having values $V_{i,j}, j = 1 \ldots NV_i$. Let C_j be one of m possible classes. If the values of attributes for a given object are obtained in sequential order from A_1 to A_k, $1 \leq k \leq n$, the probability of class C_j can be updated using the sequential Bayesian formula (Good 1950):

$$P(C_j|V_{1,J_1}, \ldots, V_{k,J_k}) = P(C_j) \prod_{i=1}^{k} \frac{P(C_j|V_{1,J_1}, \ldots, V_{i,J_i})}{P(C_j|V_{1,J_1}, \ldots, V_{i-1,J_{i-1}})} \quad (1)$$

where J_i represents the index of the value of attribute A_i for current object to be classified. The correctness of eq. (1) is obvious as the right hand side can be abbreviated to obtain the identity. If the values of all attributes are known then $k = n$. If in (1) the independence of attributes is assumed, the naive Bayesian formula is obtained:

$$\hat{P}(C_j|V_{1,J_1},\ldots,V_{n,J_n}) = P(C_j)\prod_{i=1}^{n}\frac{P(C_j|V_{i,J_i})}{P(C_j)} \tag{2}$$

Both formulas, (1) and (2) can be used to classify new objects, given the set of training examples with known classes from which the prior probabilities can be approximated. An object is classified to class with maximal probability calculated with (1) or (2). In fact, formula (1) is appropriate for induction of decision trees, if the selection of the next attribute to be tested is assumed to be independent of an object to be classified (Kononenko 1989). In the case of a decision tree, the values $V_{1,J_1},\ldots,V_{k,J_k}$ in (1) represent the path from the root to the leaf of a tree.

If a limited number of training data is available, the approximation of the probability with relative frequency:

$$\hat{P}(C_j|V_{1,J_1},\ldots,V_{k,J_k}) = \frac{N_{C_j,V_{1,J_1},\ldots,V_{k,J_k}}}{N_{V_{1,J_1},\ldots,V_{k,J_k}}} \tag{3}$$

becomes unreliable due to small number of training instances having the same values of attributes as the new object to be classified. This is also the reason of applying various pruning techniques when generating decision trees. Smaller k in (3) implies greater denominator which implies better approximation of probability.

On the other hand, in naive formula (2) the approximation of probabilities on the right hand side with relative frequencies is much more reliable. In addition, Cestnik (1990) has shown that instead of using relative frequencies it is better to use the following formula for approximation of probabilities on the right-hand side of (2) to still improve the reliability of approximations:

$$\hat{P}(C_j|V_{k,J_k}) = \frac{N_{C_j,V_{k,J_k}} + 2\times P(C_j)}{N_{V_{k,J_k}} + 2} \tag{4}$$

where $P(C_j)$ is approximated using the Laplace's law of succession (Good 1950):

$$\hat{P}(C_j) = \frac{N_{C_j}+1}{N+2} \tag{5}$$

The same formula was used also by Smyth and Goodman (1990). It was experimentally verified, that the naive Bayesian formula achieves better classification accuracy than known inductive learning algorithms (Cestnik 1990) and, surprisingly, the explanation ability of naive Bayes, at least in inexact domains such as medical diagnostics, is better than the explanation ability of a decision tree (Kononenko 1990). The kind of explanation by naive Bayes is the *sum of information gains* by each attribute for/against each class for a given object, which appeared to be preferable by human experts than single if-then rule for a classified object.

However, the naivety of formula (2) can be too drastic in certain domains with strong dependencies between attributes. There is an obvious tradeoff between the 'non-naivety' and the reliability of the approximations of probabilities. In the paper an algorithm is defined that tries to optimize this tradeoff by detecting the dependencies between attributes' values.

In next section the kinds of dependencies between events are explored. In section 3 the algorithm of the semi-naive Bayesian classifier is described. In section 4 experiments in four medical diagnostic problems are described and in section 5 the results are discussed.

2 Dependence of events

By definition events X_1 and X_2 are independent if:

$$P(X_1X_2) = P(X_1) \times P(X_2) \tag{6}$$

The dependence between X_1 and X_2 is proportional to the difference between $P(X_1) \times P(X_2)$ and $P(X_1X_2)$. In the extreme we have $X_1 = X_2$ where $P(X_1X_2) = P(X_1) = P(X_2)$ or $X_1 = \overline{X_2}$, where $P(X_1X_2) = 0$. We are interested in the conditional dependence of events X_1 and X_2 with respect to class C_j. The events X_1 and X_2 are independent with respect to C_j if:

$$P(X_1X_2|C_j) = P(X_1|C_j) \times P(X_2|C_j) \tag{7}$$

Again, the dependence between X_1 and X_2 with respect to C_j is proportional to the difference between $P(X_1|C_j) \times P(X_2|C_j)$ and $P(X_1X_2|C_j)$. In the extreme we have $C_j = X_1 \veebar X_2 = (X_1 \neq X_2)$ where $P(X_1X_2|C_j) = 0$ or $C_j = (X_1 = X_2)$ where $P(X_1X_2|C_j) = P(X_1|C_j) = P(X_2|C_j)$.

A	B	C
1	1	0
1	0	1
0	1	1
0	0	0

Table 1: XOR

XOR (exclusive 'or', see table 1) is the classical nonlinear problem, that cannot be solved by the naive Bayesian classifier. XOR can be solved in one of the following ways:

- the training examples are stored; such solution is appropriate only if XOR is known in advance to be present in the data and only in exact domains.
- the classes are split into subclasses; this solution has similar constraints like the previous one.
- attributes are joint; this solution seems to be most appropriate for the semi-naive Bayesian classifier as it naturally fits onto the formula, namely $P(C_j|X_1X_2)$ can be used instead of $P(C_j|X_1)$ and $P(C_j|X_2)$ separately. Besides, instead of joining whole attributes, only single values of different attributes can be joint, which is more flexible.

It remains to define the formula for detecting the dependencies between attributes. The following formula, that is valid for every attribute (variable) A,B and C could be used (Wan & Wong 1989):

$$H(A|C) + H(B|C) - H(AB|C) \geq 0 \tag{8}$$

where

$$H(X) = -\sum_j P(X_j) \times log_2 P(X_j)$$

and

$$H(X|Y) = \sum_i P(Y_i) \times H(X|Y_i)$$
$$H(X|Y_i) = -\sum_j P(X_j Y_i) \times log_2 P(X_j Y_i)$$

and where $X_i, i = 1..n$ are possible values of attribute X.

The equality in (8) holds if attributes A and B are independent with respect to attribute C. In that case C stands for the attribute that represents classes. The dependence of attributes A and B with respect to C is proportional to the value of the left hand-side of (8).

However, eq. (8) cannot detect dependencies between single values of attributes, which could be more useful, as joining attributes' values is more flexible than joining whole attributes. Besides, eq. (8) needs a threshold above which it is useful to join attributes without loosing the reliability of approximations of probabilities. There is no obvious way to obtain such a threshold for optimizing the tradeoff between the 'non-naivety' and the reliability. In next section the formula is designed to include this tradeoff.

3 Semi-naive Bayesian classifier

When calculating the probability of class C_j in (2) the influence of attributes A_i and A_l is defined with:

$$\frac{P(C_j|V_{i,J_i})}{P(C_j)} \times \frac{P(C_j|V_{l,J_l})}{P(C_j)} \tag{9}$$

If, instead of assuming the independence of values V_{i,J_i} and V_{l,J_l}, the values are joint, the corrected influence is given with:

$$\frac{P(C_j|V_{i,J_i}V_{l,J_l})}{P(C_j)} \tag{10}$$

For joining the two values two conditions should be satisfied: the values of (9) and (10) should be sufficiently different while the approximation of $P(C_j|V_{i,J_i}V_{l,J_l})$ with relative frequency should be sufficiently reliable. For the estimation of the reliability of the probability approximation the theorem of Chebyshev (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:

$$P(|f-p| \leq \varepsilon) > 1 - \frac{p(1-p)}{\varepsilon^2 n} \tag{11}$$

The lower bound is proportional to n and to ε^2. In our case we are interested in the reliability of the following approximation:

$$\hat{P}(C_j|V_{i,J_i}V_{l,J_l})) = \frac{N_{C_jV_{i,J_i}V_{l,J_l}}}{N_{V_{i,J_i}V_{l,J_l}}} \tag{12}$$

Therefore the number of trials n in (11) is equal to $N_{V_{i,J_i}V_{l,J_l}}$, i.e. the number of training instances having values V_{i,J_i} and V_{l,J_l} of attributes A_i and A_l, respectively. As prior probability p is unknown, in our experiments for approximation of p at the right-hand side of (11) the worst case was assumed, i.e. $p = 0.5$.

It remains to determine the value of ε. As we are interested also if the values of (9) times $P(C_j)$ and (12) are significantly different we will use ε that is proportional to the difference between the two values. The joint values will influence all classes $C_j, j = 1 \ldots m$. Therefore, ε will be the average difference between (9) times $P(C_j)$ and (12) over all classes:

$$\varepsilon = \sum_{j=1}^{m} P(C_j) \times \left| P(C_j|V_{i,J_i}V_{l,J_l}) - \frac{P(C_j|V_{i,J_i})P(C_j|V_{l,J_l})}{P(C_j)} \right| \tag{13}$$

It is necessary to determine the threshold for the probability (11) above which it is useful to join two values of two attributes. In our experiments the threshold was set to 0.5. Therefore, the rule for joining two values states: join two values if the probability is greater than 0.5 that the theoretically correct (unknown) influence of values V_{i,J_i} and V_{l,J_l} differs, in average over all classes, from the used (approximated) influence for less than the difference between used influence and the influence of the two values without joining them:

$$1 - \frac{1}{4\varepsilon^2 N_{V_{i,J_i}V_{l,J_l}}} \geq 0.5 \tag{14}$$

The values can be iteratively joint so that more than two values can be joint together. In our experiments the exhaustive search was used. The number of iterations over the whole training set is approximately equal to the number of values of all attributes. The algorithm is as follows:

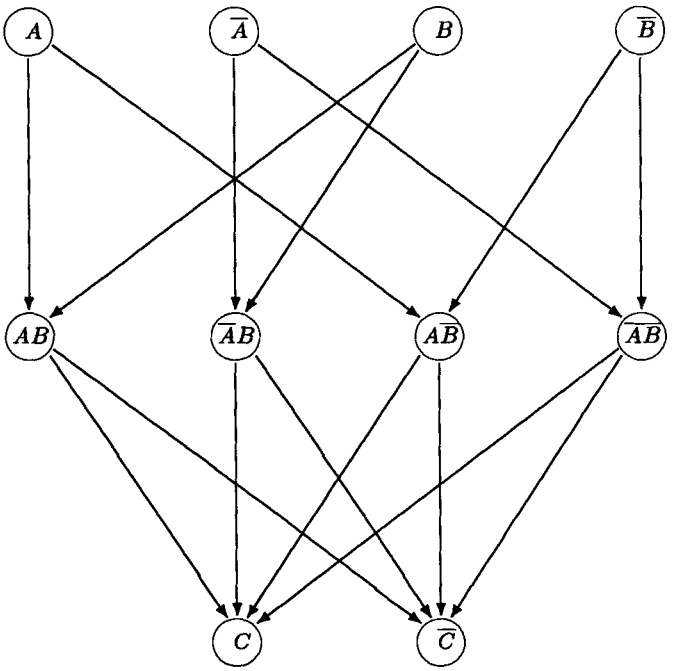

Figure 1 XOR with 8 training instances (doubled training set) solved with the semi-naive Bayesian classifier.

Determine relative frequencies needed in equation (2);
for each value $V_{i,j}$ **do**
begin
each value $V_{k,l}$ join with $V_{i,j}$;
from training set determine the relative frequencies for joint values;
discard the pairs for which $\frac{1}{4e^2 N_{V_{i,J_i} V_{l,J_l}}} > 0.5$
end

XOR as defined in table 1 is solved with the above algorithm by joining four pairs of values as shown on figure 1. However, for reliability of approximation of probabilities 8 training instances are needed obtained by doubling the training set from table 1.

domain	primary tumor	breast cancer	thyroid	rheum.
# instances	339	288	884	355
# classes	22	2	4	6
# attributes	17	10	15	32
aver. # val./attribute	2.2	2.7	9.1	9.1
aver.# missing data/inst.	0.7	0.1	2.7	0.0
majority class	25%	80%	56%	66%
entropy (bit)	3.64	0.72	1.59	1.70
# all values	59	29	141	298
accuracy of physic.	42%	64%	64%	56%

Table 2: Characteristics of data sets for four medical diagnostic problems

4 Experiments in medical diagnostics

The semi-naive Bayesian classifier, as defined in the previous section, was tested in four medical diagnostic problems: localization of primary tumor, prognostics of breast cancer recurrence, diagnostics of thyroid diseases and rheumatology. The data used in our experiments was obtained from University medical center in Ljubljana. Basic characteristics of four data sets are presented in table 2.

Diagnostic accuracy of physicians is the average of four physicians specialists in each domain from University medical center in Ljubljana tested on randomly selected subsets of patients. The diagnostic accuracy of physicians together with the number of classes and entropy roughly shows the difficulty of a classification problem. The average number of values per attribute together with the number of instances and the number of classes roughly shows the (un)reliability of relative frequencies obtained from data sets.

The percent of instances belonging to the majority class represents the classification accuracy of a simple classifier that classifies each object into the majority class. Such simple classifier

	primary tumor		breast cancer		thyroid		rheumatology	
	(%)	(bit)	(%)	(bit)	(%)	(bit)	(%)	(bit)
physicians	42	1.22	64	0.05	64	0.59	56	0.26
Assistant	44	1.38	77	0.07	73	0.87	61	0.46
naive Bayes	51	1.58	79	0.18	70	0.79	67	0.51
semi-naive Bayes	51	1.58	79	0.18	71	0.81	68	0.55
# accepted pairs	0.4		1.5		32.3		17.9	

Table 3: Results of the semi-naive Bayesian classifier in four medical diagnostic problems compared with the performance of other classifiers.

would significantly outperform physicians in rheumatology (for 10%) and in breast cancer (for 16%). This shows that the classification accuracy is not an appropriate measure for estimating the classification performance. For that reason the *average information score* of classifiers answers as defined in (Kononenko & Bratko 1991) was also measured. This measure eliminates the influence of prior probabilities of classes on the classification performance.

In our experiments the formulas (4) and (5) were used for approximating the probabilities in equation (2). In each experiment the whole data set was randomly split into 70% of instances for learning and 30% of instances for testing. The results in table 3 are averages of 10 experiments in each domain. In the table the number of accepted pairs of attributes' values is also presented. The results are compared with the performance of the naive Bayesian classifier, ID3 like inductive learning system Assistant (Cestnik et al. 1987) and physicians specialists.

It came out that joining of attributes' values in primary tumor and breast cancer is unnecessary as, by the opinion of physicians specialists, the attributes are in fact independent. The results of semi-naive and naive Bayes in these two domains are identical. However, in diagnostics

	naive Bayes	semi-naive Bayes	Assistant (non-naive Bayes)
generates	probabilities	probabilities	decision tree
knowledge	implicit	implicit	explicit
explanation	inf. gains	inf. gains	if-then rule
# atts used	all	all	few
missing data	insensitive	insensitive	sensitive
prob.approx.	reliable	reliable	unreliable
independence	assumed	not assumed	not assumed
speed	fast	slow	slow
incremental	yes	no	no
mulival. atts	sensitive	sensitive	insensitive
domains	inexact	inexact	exact

Table 4: Characteristics of three classifiers

of thyroid diseases and in rheumatology joining of attribute's values slightly improves the performance.

Assistant achieved better performance in thyroid problem due to the binarization of attributes. In thyroid data most of attributes are continuous having initially 20 subintervals of values (see table 2) which are treated by Bayesian classifier as 20 discrete values. Therefore in the Bayesian classifier the data set is split into 20 subsets when approximating probabilities, which makes the approximation unreliable. On the other hand, in Assistant attributes are binarized (the values of each attribute are joint into two subsets) which leads to increased reliability of the probability approximation.

5 Discussion

In table 4 the characteristics of naive, semi-naive Bayes and Assistant (which represents non-naive Bayes, see section 1) are sketched. The generated knowledge by Assistant is in the form of a decision tree while naive and semi-naive Bayes generate probabilities. The top part of a decision tree typically shows the structure of the problem. The decision tree can be used without a computer to classify new objects, therefore it is the kind of *explicit knowledge*. On the other hand, the probabilities generated by naive and semi-naive Bayes cannot be directly used to classify new objects. This kind of knowledge is *implicit*. The physicians found both types of knowledge as interesting and useful information.

The explanation of classification of a new object in Assistant is simply the if-then rule used for the classification while in naive and semi-naive Bayes the explanation is the sum of information gains from all attributes for/against the conclusion. Physicians preferred the sum of information gains as more natural explanation, similar to the way physicians diagnose (Kononenko 1990).

While if-then rules typically include too few attributes for reliable classification (Pirnat et al. 1989), naive and semi-naive Bayes use all available attributes. Besides, learning of decision rules and classification with decision rules is very sensitive to missing data. Missing value of an attribute in naive and semi-naive Bayes is simply ignored.

The major advantage of naive and semi-naive Bayes is reliability of approximation of probabilities. Due to small number of training instances covered by single decision rule the final decision of a rule is unreliable. Pruning of decision trees partially overcomes this problem, however due to pruning rules are shortened and more attributes are discarded from diagnostic process.

If attributes are human defined (as was the case in medical data used in our experiments) attributes are usually relatively independent, as humans tend to think linearly. However, independence assumption is often unrealistic. Semi-naive Bayes overcomes the independence assumption while preserving the reliability of probability approximations. But learning is not as fast as with naive Bayes and it is not incremental. Incremental versions of semi-naive Bayes can be developed similarly to incremental versions of ID3 (Van de Velde 1989).

Assistant with on-line binarization of attributes successfully solves the problem of continuous

and multivalued attributes (as shown in the case of thyroid diseases, see table 3). The naive Bayesian classifier assumes that all attributes are discrete. Therefore, continuous attributes must be converted to discrete by introducing a number of fixed bounds, loosing the integrity of training set and the order of values. The semi-naive Bayesian classifier is unable to join values of the same attribute in order to keep the training instances together for more reliable approximation of probabilities. There are two possibilities to overcome that problem:

- The use of XOR for joining the values of attributes besides AND as used in the algorithm described in section 3. XOR should join values $V_{i,l}$ and $V_{i,k}$ with similar $P(C_j|V_{i,l})$ and $P(C_j|V_{i,k})$. and small $N_{V_{i,l}}$ and/or $N_{V_{i,k}}$. Therefore, the aim of joining with XOR is opposite to that with AND: increasing the reliability of approximation of probabilities while unchanging the influence of attribute's values.
- The use of fuzzy bounds for continuous attributes can overcome both the loss of the information about the order of values as well as the loss of integrity of training set.

Semi-naive Bayes tries to optimize the tradeoff between the 'non-naivety' and the reliability of probability approximations. By lowering the threshold in (14) the reliability of probability approximations decreases and the 'non-naivety' increases, which can be useful for *exact* domains. For *inexact* (fuzzy) domains the threshold should be higher. Naive Bayes is due to independence assumption more appropriate for inexact domains while Assistant is appropriate for exact domains with, ideally, complete set of attributes *and* complete set of training instances.

An expert system shell based on the semi-naive Bayesian classifier can provide a useful tool for generating expert systems in domains, where training data is available. Dependencies between attributes can be determined automatically or, in case where there is not enough training instances for reliable probability approximations, the dependencies can be provided by human experts. Human expert can be consulted to determine also prior probabilities if not enough training data is available. The explanation ability of such a system seems powerful enough to assist experts or non-experts in their everyday work.

Acknowledgements

Collecting and assembling the experimental data would not be possible without the invaluable help of physicians specialists Dr. Matjaž Zwitter, Dr. Sergej Hojker and Dr. Vlado Pirnat from the University Medical Center in Ljubljana. I thank them for providing and interpreting the data, for testing the diagnostic performance of physicians specialists from the University Medical Center, for the interpretation of results and for the estimation of the explanation abilities of the naive Bayesian classifier. I am grateful to Padhraic Smyth for his comments on the manuscript. This research was supported by Slovenian Research Community. The reported work was done in the Artificial Intelligence Laboratory at the Faculty of Electrical and Computer Engineering in Ljubljana. I would like to thank Prof. Ivan Bratko for providing the environment for efficient scientific work.

References

Cestnik B. (1990) Estimating Probabilities: A Crucial Task in Machine Learning, *Proc. European Conference on Artificial Intelligence 90*, Stockholm, August 1990.

Cestnik, B., Kononenko, I., Bratko, I. (1987) ASSISTANT 86 : A Knowledge Elicitation Tool for Sophisticated Users. In: I. Bratko, N. Lavrač (eds.), *Progress in Machine learning.* Wilmslow, England: Sigma Press.

Good I.J. (1950) *Probability and the weighing of evidence.* London: Charles Griffin.

Kononenko, I. (1989) ID3, sequential Bayes, naive Bayes and Bayesian neural networks. *Proc. 4th European Working Session on Learning*, Montpellier, France, December 1989, pp.91-98.

Kononenko, I. (1990) Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. In: B. Wielinga et al. (eds.) *Current Trends in Knowledge Acquisition*, Amsterdam: IOS Press.

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

Pirnat V., Kononenko I., Janc T., Bratko I. (1989) Medical Estimation of Automatically Induced Decision Rules, *Proc. of 2nd Europ. Conf. on Artificial Intelligence in Medicine*, City University, London, August 29-31 1989, pp.24-36.

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) *Elementary Introduction to Probability Calculus* (in Slovene), Državna založba Slovenije, Ljubljana.

Van de Velde (1989) IDL, or Taming the Multiplexer, *Proc. 4th European Working Session on Learning*, Montpellier, France, December 1989, pp.211-225.

Wan S.J. & Wong S.K.M. (1989) A Measure for Concept Dissimilarity and its Applications in Machine Learning, *Proc. Int. Conf. on Computing and Information*, Toronto North, Canada, May 23-27 1989.