Proc. of ECAI 90, Stockholm, August 1990.

Estimating Probabilities: A Crucial Task in Machine Learning

Bojan Cestnik
Jožef Stefan Institute
Jamova 39, 61000 Ljubljana
Yugoslavia

Abstract

The evaluation of the Bayesian formula that is used as a basis of many machine learning systems is studied in detail. It is shown that, when used naively (i.e. assuming the independence of attributes), its classification accuracy heavily depends on the method for estimating conditional probabilities that are required by the formula. A new method for estimating conditional probabilities with a firm theoretical background, that substantially improves the classification accuracy of the formula (as shown by experimental results on four real world medical domains), is presented.

1 Introduction

Several machine learning systems, such as ID3 [Quinlan, 1986], CART [Breiman et al, 1984], ASSISTANT [Bratko & Kononenko, 1986; Cestnik et al, 1987], CN2 [Clark & Niblett, 1987], etc., were shown to be able to induce compact and accurate knowledge bases in many real world domains. However, in almost any domain the classification accuracy obtained by the naive Bayesian formula (with attributes' independence assumption) was reported to be slightly better than the one by the system itself. Still, expressing knowledge in an explicit symbolic form and explaining the decisions were the properties that made the above-mentioned systems successful in practice.

In the last two years several new systems like GINESYS [Gams, 1989] and LogArt [Cestnik & Bratko, 1988], that achieved slightly better classification accuracy in real world domains than the naive Bayesian formula, emerged. The common idea in these systems is the use of multiple knowledge. Instead of relying on a few of the most important attributes, such systems additionally take into account the remaining (less important) attributes to improve their decisions. The basic arguments explaining why such an inductive learning system performs better than the Bayesian classifier were as follows:

- the Bayesian classifier was implemented under the independence assumption that is often unrealistic,
- inductive learning systems have special mechanisms for handling noise in the learning data,
- inductive learning systems use multiple (although redundant) knowledge to improve performance.

At the same time, the Bayesian formula was further studied in [Gams & Drobnič, 1988], [Michie & Al Attar, 1989] and [Kononenko, 1989b]. One possible way of interpreting the naive Bayesian classification was shown by Kononenko [1989a]. He reported that the experts in testing domains accepted and understood the explanation. Moreover, some of them found its line of reasoning very similar to the one that they are using in practice.

In this article it is shown that the evaluation of naive Bayesian formula is very sensitive to the estimation of conditional probabilities. If the estimations are performed correctly, its classification accuracy increases substantially. The independence assumption is often unrealistic but there is a trade-off between not making this assumption and the quality of probabilities estimations; without the independence assumption conditional probabilities are estimated from smaller samples. So, especially in domains with relatively small number of examples it might be wise to adopt such an assumption. It is also shown that the mechanisms to combat noise in the learning data can as well be incorporated in the probability estimation function. And last but not least, the Bayesian classifier considers all available attributes in the classification. This is exactly what some inductive learning systems are trying to do by utilizing multiple knowledge in terms of confirmation rules [Gams, 1989] or redundant rules [Cestnik & Bratko, 1988].

2 The Bayesian formula

The learning problem addressed in this paper can be defined as follows:

Given: A set of objects for learning, described with attributes and their values. Every object belongs to one class.

Find: A classification rule that fits the learning set and can be used for classifying new objects into classes.

Let $A_i, i = 1..n$, be a set of attributes each having a certain number of possible values j_i. Let an event of attribute A_i having a value j be denoted by A_i^j or shortly V_i. Let C denote a class.

The Bayesian formula is used to compute the conditional probability of a class C given the evidence of attributes $V_1, V_2, V_3, \cdots$:

$$\Theta = p(C|V_1V_2V_3\cdots) = \frac{p(CV_1V_2V_3\cdots)}{p(V_1V_2V_3\cdots)} =$$

$$= p(C)\frac{p(C|V_1)}{p(C)}\frac{p(C|V_1V_2)}{p(C|V_1)}\frac{p(C|V_1V_2V_3)}{p(C|V_1V_2)}\cdots \qquad (1)$$

If the independence of the attributes $V_1, V_2, V_3, \cdots$ is assumed, (1) can be further simplified to obtain:

$$\Theta = p(C)\frac{p(C|V_1)}{p(C)}\frac{p(C|V_2)}{p(C)}\frac{p(C|V_3)}{p(C)}\cdots \qquad (2)$$

So, the formula (2) can be interpreted as follows: take the apriori probability of C and multiply it by a factor $h(i)$ for every attribute to obtain the final aposteriori probability, where $h(i)$ has a form:

$$h(i) = \frac{p(C|V_i)}{p(C)}$$

3 Problems with the evaluation of the Bayesian formula

Let $n(V_i)$ denote the number of examples where V_i is observed, and $n(CV_i)$ the number of examples where both V_i and C are observed. First, let us consider the approximation of probabilities with relative frequencies:

$$p(C|V_i) = \frac{n(CV_i)}{n(V_i)}$$

Problems arise when $n(V_i) = 0$ and/or $n(CV_i) = 0$. In case that $n(V_i) = 0$ (division by 0) $h(i)$ is usually set to 1 (it is assumed that V_i has no influence to the final probability Θ). If $n(V_i) > 0$ and $n(CV_i) = 0$ then $h(i)$ becomes 0, and Θ becomes 0, even if $n(V_i)$ is effectively small, meaning that the estimation is not reliable.

According to our knowledge, the naive Bayesian formula that is referred to in section 1 (when compared with other inductive learning systems) was evaluated as described above, thus approximating probabilities with relative frequencies.

As observed, problems are encountered when $n(V_i)$ and/or $n(CV_i)$ are small. However, Laplace's law of succession [Good, 1950; Good, 1965] can be used to alleviate the problem. It states that if in the sample of N trials there were n successes, the probability of the next trial being successful is $(n+1)/(N+2)$, assuming that the initial distribution of successes and failures is uniform.

Let us explore the same situations as above: if $n(V_i) = 0$ then $h(i) = 1/(2p(C))$. If $p(C)$ is decreasing then $h(i)$ is growing; the reason for this strange behavior lies in the initial assumption. If $n(V_i) > 0$ and $n(CV_i) = 0$ then $h(i) = 1/((n(V_i)+2)p(C))$. Again, $h(i)$ is proportional to $1/p(C)$. However, since $h(i)$ is a monotonically decreasing function of $n(V_i)$, the reliability of the estimation is incorporated in $h(i)$.

4 Estimation of probabilities

In [Good, 1965] it is suggested that instead of assuming a uniform initial distribution (like in Laplace's law of succession) a more flexible and convenient class of initial probability densities should be used. Let the initial probability density function be of the form:

$$f(x) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$$

where $a > 0, b > 0$ and $B(a,b)$ is Beta function.

Then, after n successes in N trials, the mathematical expectation of the probability of a success in the next trial is equal to:

$$q(n,N) = \frac{n+a}{N+a+b} \qquad (3$$

where $a > 0$ and $b > 0$.

Note that Laplace's law of succession is only a special case of (3) when a and b equal 1.

Let us prove that this approximation with an appropriate selection of parameters a and b has the following four properties:

1) $q(0,0) = \frac{a}{a+b} = p(C)$

$p(C)$, apriori probability of class C, must be greater than 0. This can be achieved by estimating it by Laplace's law of succession. According to our new method for estimating conditional probabilities the parameters a and b should then be set such as to satisfy the desired property. Let $a + b = m$. The value of m is domain dependent. It is related to the amount of noise in the domain. m can be small if little noise is expected and should grow if the

amount of noise is substantial. In the experiments in section 5 the value 2 was used for m, like in Laplace's law of succession. Once m is determined, parameters a and b can be set as follows: $a = p(C)m$ and $b = m - a$. It can then be proved that $q(0,0) = \frac{p(C)m}{m} = p(C)$.

2) $q(0,N) = \frac{a}{N+a+b} > 0$ because $a > 0$,
$q(N,N) = \frac{N+a}{N+a+b} < 1$ because $b > 0$.

3) $q(N+1,N+1) = \frac{N+1+a}{N+1+a+b} = 1 - \frac{b}{N+a+b+1} > 1 - \frac{b}{N+a+b} = \frac{N+a}{N+a+b} = q(N,N)$.

4) $q(0,N+1) = \frac{a}{N+1+a+b} < \frac{a}{N+a+b} = q(0,N)$.

5 Experimental results

The naive Bayesian formula was tested on four medical domains that are described in more detail in [Bratko & Kononenko, 1986]. Results, shown in Table 1, represent the average of 10 experiments. Each time 70% of the original examples were randomly taken for training and the remaining 30% for testing. All the methods used the same random split in one test.

Domain	#classes	Bayes1	Bayes2	Bayes3
Lymphography	4	79.0	43.6	84.8
Hepatitis	2	82.6	83.2	84.9
Breast cancer	2	77.4	77.4	78.6
Primary tumor	22	48.2	25.9	51.2

Table 1: Results of experiments in four medical domains. In all cases the Bayesian formula is evaluated naively, however, Bayes1 approximates probabilities with relative frequencies, Bayes2 with Laplace's law of succession and Bayes3 with the new method, described in section 4.

6 Discussion

The results in Table 1 show that the naive Bayesian classifier performs better with a proper probabilities' estimation method (Bayes3) than when the two other estimation methods (Bayes1 and Bayes2) are used. Since many sophisticated inductive learning systems [e.g. Clark & Niblett, 1987; Cestnik et al, 1987] are reported to achieve slightly lower classification accuracy than naive Bayesian classifier using approximation with relative frequencies, the new method also performs substantially better than the above-mentioned systems.

In Table 1 it can be observed that estimating probabilities by Laplace's law of succession in Bayes2 (assuming a uniform apriori distribution) is often unrealistic, especially in multi-class decision problems, like in lymphography and primary tumor. The cause for worse performance of Bayes2 on these two domains is explained in section 3.

The new method for estimating probabilities, presented in section 4, has a firm theoretical background [Good, 1965]. Since the estimated probabilities are treated as random variables with associated distributions, several interesting properties like, for example, variance, can also be observed. The method in general can be used by any inductive learning system that gives probabilistic answers.

The attributes' independence assumption, made by the naive Bayesian method, is often unrealistic. However, in real world domains there is a trade-off between not making this assumption and the quality of probability estimations; without the independence assumption conditional probabilities are estimated from smaller samples.

References

Bratko, I., Kononenko, I. (1986), Learning diagnostic rules from incomplete and noisy data, *AI Methods in Statistics*, UNICOM Seminar, London, December 1986.

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

Cestnik, B., Kononenko, I., Bratko, I. (1987), ASSISTANT 86: A Knowledge-Elicitation Tool for Sophisticated Users, *Progress in Machine Learning*, Eds. I.Bratko & N.Lavrač, Sigma Press.

Cestnik, B., Bratko, I. (1988), Learning redundant rules in noisy domains, *Proc. of ECAI 88*, Munchen.

Clark, P., Niblett, T. (1987), Induction in Noisy Domains, *Progress in Machine Learning*, Eds. I.Bratko & N.Lavrač, Sigma Press, Wilmslow.

Gams, M., Drobnič, M. (1988), Approaching the Limit of Classification Accuracy, *Informatica vol. 2*, Slovene Society for Informatics, Ljubljana.

Gams, M. (1989), New Measurements Highlight the Importance of Redundant Knowledge, *Proc. of the Fourth European Working Session on Learning*, Montpellier, December 1989.

Good, I.J. (1950), *Probability and the Weighting of Evidence*, Charles Griffing & Company Limited, London.

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

Kononenko, I. (1989a), Interpretation of neural networks decision, *Proc. IASTED Intern. Conf. Expert Systems Theory & Applications*, Zurich, June 1989.

Kononenko, I. (1989b), ID3, Sequential Bayes, Naive Bayes and Bayesian Neural Networks, *Proc. of the Fourth European Working Session on Learning*, Montpellier, December 1989.

Michie, D., Al Attar, A. (1989) Use of sequential Bayes with class probability trees. *Machine Intelligence 12*, Eds. J.Hayes-Michie, D.Michie, E.Tyugu, Oxford: Oxford University Press.

Quinlan, J.R. (1986), Learning from noisy data, *Machine Learning vol. 2*, Eds. R.Michalski, J.Carbonell and T.Mitchel, Palo Alto, CA: Tioga.