Robust Approach for Estimating Probabilities in Naive-Bayes Classifier

B. Chandra[1], Manish Gupta[2], and M.P. Gupta[1]

[1]Indian Institute of Technology, Delhi
Hauz Khas, New Delhi, India 110 016
bchandra104@yahoo.co.in
[2]Institute for Systems Studies and Analyses
Metcalfe House, Delhi, India 110 054

Abstract. Naive-Bayes classifier is a popular technique of classification in machine learning. Improving the accuracy of naive-Bayes classifier will be significant as it has great importance in classification using numerical attributes. For numeric attributes, the conditional probabilities are either modeled by some continuous probability distribution over the range of that attribute's values or by conversion of numeric attribute to discrete one using discretization. The limitation of the classifier using discretization is that it does not classify those instances for which conditional probabilities of any of the attribute value for every class is zero. The proposed method resolves this limitation of estimating probabilities in the naive-Bayes classifier and improve the classification accuracy for noisy data. The proposed method is efficient and robust in estimating probabilities in the naive-Bayes classifier. The proposed method has been tested over a number of databases of UCI machine learning repository and the comparative results of existing naive-Bayes classifier and proposed method has also been illustrated.

1 Introduction

Classification has wide application in pattern recognition. In classification, a vector of attribute values describes each instance. Classifier is used to predict the class of the test instance using training data, a set of instances with known classes. Decision trees [13], k- nearest neighbor [1], naive- Bayes classifier [6,7,8] etc. are the commonly used methods of classification. Naive-Bayes classifier (NBC) is a simple probabilistic classifier with strong assumption of independence. Although attributes independence assumption is generally a poor assumption and often violated for real data sets, Langley et. al. [10] found that NBC outperformed an algorithm for decision-tree induction. Domingos et.al. [4] has also found that this limitation has less impact than might be expected. It often provides better classification accuracy on real time data sets than any other classifier does. It also requires small amount of training data. It is also useful for high dimensional data as probability of each attribute is estimated independently. There is no need to scale down the dimension of the data as required in some popular classification techniques.

A. Ghosh, R.K. De, and S.K. Pal (Eds.): PReMI 2007, LNCS 4815, pp. 11–16, 2007.

NBC has a limitation in predicting the class of instances for which conditional probabilities of each class are zero i.e. conditional probability of any of the attribute value for every class is zero. To rectify this problem, the Laplace-estimate [3] is used to estimate the probability of the class and M-estimate [3] is used to estimate conditional probability of any of the attribute value. The results obtained from these estimates have more error of classification for noisy data i.e more number of classes or more number of attributes. A novel approach based on the maximum occurrence of the number for which conditional probabilities of any of the attributes are zero for a given instance, is proposed in the paper. The proposed method has been tested over a number of databases of UCI machine learning repository [12]. In proposed approach, the classification accuracy is much better even for noisy data as compared to that of basic approach of estimating probabilities in NBC and of Laplace-estimates and M-estimates.

The overview of NBC along with the estimation of probabilities in NBC is given in section 2 of the paper. The proposed approach has been described in section 3 with limitation of NBC using discretization and overcoming the limitation by proposed approach. Section 4 presents results and discussion of all approaches over several databases from UCI machine learning repository. Comparative evaluation of proposed approach with the existing basic and estimate approach are also present in this section. Concluding remarks is given in the last section of the paper.

2 Naive-Bayes Classifier (NBC)

NBC [6,7,8] is a simple probabilistic inductive algorithm with strong attribute independence assumption. NBC learns from training data and then predicting the class of the test instance with the highest posterior probability. Let C be the random variable denoting the class of an instance and let $\mathbf{X} < X_1, X_2, \ldots, X_m >$ be a vector of random variables denoting the observed attribute values. Let c represent a particular class label and let $\mathbf{x} < x_1, x_2, \ldots, x_m >$ represent a particular observed attribute value vector. To predict the class of a test instance $\mathbf{x}$, Bayes' Theorem is used to calculate the probability

$$p(C = c \mid \mathbf{X} = \mathbf{x}) = \frac{p(C = c)p(\mathbf{X} = \mathbf{x} \mid C = c)}{p(\mathbf{X} = \mathbf{x})} \tag{1}$$

Then, predict the class of test instance with highest probability. Here $\mathbf{X}= \mathbf{x}$ represents the event that $X_1 = x_1 \wedge X_2 = x_2 \wedge \ldots X_m = x_m$. $p(\mathbf{X} = \mathbf{x})$ can be ignored as it is invariant across the classes, then equation (1) becomes

$$p(C = c \mid \mathbf{X} = \mathbf{x}) \propto p(C = c)p(\mathbf{X} = \mathbf{x}|C = c) \tag{2}$$

p(C = c) and $p(\mathbf{X} = \mathbf{x}|C = c)$ are estimated from the training data. As attributes $X_1, X_2, \ldots, X_m$ are conditionally independent of each other for given class then equation (2) becomes

$$p(C = c \mid \mathbf{X} = \mathbf{x}) \propto p(C = c)\prod_{i=1}^{m} p(X_i = x_i|C = c) \tag{3}$$

which is simple to compute for test instances and to estimate from training data. Classifiers using equation (3) are called naive-Bayes classifier.

2.1 Estimation of Probabilities in Naive-Bayes Classifier

NBC can handle both categorical and numeric attributes. For each discrete attribute, $p(X_i = x_i|C = c)$ in equation (3) is modeled by a single real number between 0 and 1, and the probabilities can be estimated with reasonable well accuracy from the frequency of instances with C=c and the frequency of instances with $X_i = x_i \wedge C = c$ in the training data. We call this approach of estimating probabilities as basic approach. Laplace-estimate [3] and M-estimate [3] are also used to compute the probabilities in equation (3). In Laplace-estimate $p(C = c) : (n_c + k)/(N + n * k)$ where n_c is the number of instances satisfying C=c, N is the number of training instances, n is the number of classes and k=*1*. In M-estimate $p(X_i = x_i|C = c) : (n_{ci} + m * p)/(n_c + m)$ where n_{ci} is the number of instances satisfying $X_i = x_i \wedge C = c$, n_c is the number of instances satisfying $C = c$, p is the prior probability $p(X_i = x_i)$ (estimated by the Laplace-estimate), and m = *2*.

In contrast to discrete attribute, for each numeric attribute the probability $p(X_i = x_i|C = c)$ is either modeled by some continuous probability distribution [8] over the range of that attribute's values or by conversion of numeric attribute to discrete one using discretization [11,14,15]. Equal width discretization (EWD) [2,5,9] is a popular approach to transform numeric attributes into discrete one in NBC. A discrete attribute X_i^c is formed for each numeric attribute X_i and each value of X_i^c corresponds to an interval $(a_i, b_i]$ of X_i. If $x_i \in (a_i, b_i]$, then $p(X_i = x_i|C = c)$ in equation (3) is estimated by

$$p(X_i = x_i|C = c) \approx p(a_i < x_i \leq b_i|C = c) \tag{4}$$

It is estimated same as for discrete attribute mentioned above. Using equation (4), equation (3) becomes as

$$p(C = c \,|\mathbf{X} = \mathbf{x}) \propto p(C = c) \prod_{i=1}^{m} p(a_i < x_i \leq b_i|C = c) \tag{5}$$

Thus for a numeric attribute of a test instance $\mathbf{x} < x_1, x_2, , x_m >$, the probability is computed by equation (5).

3 Proposed Approach

The limitation of NBC using basic approach for estimating probabilities is that if the training instance with $X_i = x_i \wedge C = c$ does not present in the training data, then $p(X_i = x_i|C = c)$ is zero which ultimately leads to $p(C = c \,|\mathbf{X} = \mathbf{x})$ as zero from equation (3). It means that for any instance, $p(C = c \,|\mathbf{X} = \mathbf{x})$ will be zero for all classes if any one of $p(X_i = x_i|C = c)$ is zero for each class. Thus, NBC cannot predict the class of such test instances. Laplace and M-estimate methods are also not well enough for noisy as well as large databases and classify the noisy data with high error of classification. To reduce the error of classification for noisy data and to overcome the problems of both the estimation of probabilities in NBC, a simplistic novel approach based on the maximum occurrence of the number for which conditional probabilities of any of the attributes are zero for a given instance has been proposed in the paper. The proposed approach is given as follows.

For given test instance $\mathbf{x} < x_1, x_2, \ldots, x_m >$, if $p(C = c\ |\mathbf{X} = \mathbf{x})$ for each class is zero, then for each class, count the occurrences of attribute values (say, n_i) for which $p(a_i < x_i \leq b_i|C = c) = 0$. Here, n_i signifies the number of attributes for which training instance with $X_i = x_i \wedge C = c$ does not present in the training data. The greater is the number n_i of a class c, the lesser is the probability that test instance $\mathbf{x}$ belong to that class c. n_i also depends upon the probability of that class in the training data. Therefore, instead of taking n_i as the significant number in deciding the class of such test instance, we compute for every attribute

$$N_i = n_i / p(C = c\ ) \tag{6}$$

Now, N_i captures the dependency of $p(C = c\ )$ in the training data. Instead of taking $p(C = c\ |\mathbf{X} = \mathbf{x}) = 0$ for each class, we take $p(C = c\ |\mathbf{X} = \mathbf{x})$ is equal to $p(C = c\ )$ for a particular class for which N_i is minimum. It means, test instance $\mathbf{x}$ with $p(C = c\ |\mathbf{X} = \mathbf{x}) = 0$ for each class would be classified to the class for which N_i is minimum.Thus,N_i is computed using equation (6) for each test instances of such type.

This new approach of estimating probabilities will solve the problem in basic approach of NBC and also performs better than Laplace and M-estimate methods for noisy and large datasets. The proposed method is simple, efficient and robust for noisy data. We observe reasonably well reduction in error of classification on several datasets using the proposed approach. The comparative results of proposed approach with the existing basic approach and estimate approach are present in the subsequent section of the paper. The result shows that proposed approach outperforms the existing approaches of estimating probabilities using discretization for NBC.

4 Results and Discussion

To determine the robustness of our approach to real world data, we selected 15 databases from the UCI machine learning repository [12]. Table 1 gives mean accuracies of the ten-fold cross validation using different approaches of estimating probabilities in naive-Bayes classifier. For each datasets, Size is the number of instances, Feature is the number of attributes, Class is the number of classes, Basic, Estimate and Proposed represent probability estimation using basic approach,Laplace and M-estimate approach and proposed approach respectively and last two columns show the significance level of paired t test that proposed approach is more accurate than basic and estimate. For each datasets, we used ten-fold cross validation to evaluate the accuracy of the three different approaches i.e. basic, estimate and proposed.

Table 1 shows that the classification accuracy of proposed approach was much better than basic approach in 8 out of the 15 databases. The proposed approach was also significantly better than estimate approach in 4 of the 15 databases. The result also indicates that proposed approach improved the results significantly in case of basic approach of estimating the probabilities in NBC whereas in case of estimate approach, it outperformed for noisy databases such as sonar, pendigits, letter recognition and segmentation. It is important to note that all the databases where proposed approach outperformed estimate approach were large in size and had more number of classes than any other datasets.

Table 1. Mean Accuracies of the ten cross validation using different approaches of estimating probabilities in Naive-Bayes Classifier

Dataset	Size	Feature	Class	Basic	Estimate	Proposed	Proposed Better with Basic?	Proposed Better with Estimate?
Letter	20000	16	26	70.78	70.75	70.82	Yes(97.8%)	Yes(97.4%)
Pendigit	10992	16	10	87.65	87.43	87.65	Equal	Yes(99%)
Segmention	2310	19	7	90.61	89.96	91.08	Yes(99.9%)	Yes(95%)
Vowel	990	10	11	69.49	70.51	70.00	Yes(97.4%)	Equal
Vehicle	846	18	4	62.71	62.12	62.47	Equal	Equal
P-I-Diabetes	768	8	2	75.32	75.58	75.06	Equal	Equal
Wdbc	569	30	2	91.4	94.21	93.33	Yes(99.1%)	Equal
Ionosphere	351	34	2	85.43	90.57	88.00	Yes(99.8%)	Equal
Liver	345	6	2	64.71	64.71	65.00	Equal	Equal
New-Thyoroid	215	5	3	90.00	92.86	90.00	Equal	Equal
Glass	214	10	7	50.95	55.24	55.71	Yes(97.4%)	Equal
Sonar	208	60	2	70.00	65.24	74.29	Yes(97.9%)	Yes(100%)
Wpbc	198	30	2	73.16	70.00	73.16	Equal	Equal
Wine	178	13	3	81.11	96.67	91.67	Yes(100%)	No(99.8%)
Iris	150	4	3	94.67	95.33	95.33	Equal	Equal

Our experiments show that proposed approach is better than basic and estimate approaches and aims at reducing NBC's error of classification. It is observed from Table 1 that the proposed approach is especially useful in classifying noisy datasets.

5 Concluding Remarks

In this paper, a robust approach for estimating probabilities in naive-Bayes classifier based on the maximum occurrence of the number for which conditional probabilities of any of the attributes are zero for a given instance has been proposed in order to overcome the limitation of existing approaches of estimating probabilities in NBC. The effectiveness of the proposed approach over the existing approaches has been illustrated using different databases of UCI machine learning repository. The proposed approach performs remarkably well in terms of classification accuracy for large and noisy datasets as compared to other estimates. The approach can play an important role for wider variety of pattern recognition and machine learning problems by estimating the probabilities for naive-Bayes classifier.

References

1. Aha, D., Kibler, D.: Instance-based learning algorithms. Machine Learning 6, 37–66 (1991)
2. Catlett, J.: On changing continuous attributes into ordered discrete attributes. In: Proceedings of the European Working Session on Learning, pp. 164–178 (1991)
3. Cestnik, B.: Estimating probabilities: A crucial task in machine learning. In: Proceedings of the 9th European Conference on Artificial Intelligence, pp. 147–149 (1990)
4. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning. 29, 103–130 (1997)
5. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and unsupervised discretization of continuous features. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 194–202 (1995)
6. Duda, R.O., Hart, P.E.: Pattern classification and scene analysis. John Wiley and Sons, New York (1973)
7. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning 29, 131–163 (1997)
8. John, G.H., Langley, P.: Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, pp. 338–345 (1995)
9. Kerber, R.: Chimerge: Discretization for numeric attributes. In: National Conference on Artificial Intelligence AAAI Press, pp. 123–128 (1992)
10. Langley, P., Iba, W., Thompson, K.: An analysis of Bayesian classifiers. In: Proceedings of the Tenth National Conference on Artificial Intelligence, pp. 223–228 (1992)
11. Lu, J., Yang, Y., Webb, G.I.: Incremental Discretization for Nave-Bayes Classifier. In: Proceedings of the Second International Conference on Advanced Data Mining and Applications, pp. 223–238 (2006)
12. Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J.: UCI Repository of machine learning databases. University of California, Irvine, CA, Department of Information and Computer Science. (1998), `http://www.ics.uci.edu/mlearn/MLRepository.html`
13. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA (1993)
14. Yang, Y., Webb, G.: A Comparative Study of Discretization Methods for Naive-Bayes Classifiers. In: Proceedings of the Pacific Rim Knowledge Acquisition Workshop, Tokyo, Japan, pp. 159–173 (2002)
15. Yang, Y., Webb, G.: On why discretization works for naive-Bayes classifiers. In: Proceedings of the 16th Australian Joint Conference on Artificial Intelligence (AI) (2003)