Gaziosmanpasa University
Graduate School of
Natural and Applied Sciences
Journal of New Results in Science

ISSN: 1304-7981 **http://jnrs.gop.edu.tr**

Received: 02.10.2012
Accepted: 02.11.2012

Editors-in-Chief: Naim Çağman
Area Editor: Ahmet Fenercioğlu

Least Squares Approach to Locally Weighted Naive Bayes Method

Umut Orhan[1], Kemal Adem[2] and Onur Comert[3]

Abstract

This study proposes a new approach which calculates the weights of Locally Weighted Naive Bayes (LWNB) developed on Naive Bayes (NB) which is known with its simple structure. In this approach, a new equation is described by assigning a powered weight to each probabilistic factor in classic NB, and it is transformed to a linear form by using a simple assumption based on a logarithmic process, and then the weights are estimated by least squares technique. The success ratios are computed on two-class datasets from UCI database. The results show that LWNB with proposed approach is more successful than classic NB. In another analysis, it is determined that the class probability factor may sometimes damage the classification success. In addition, the effects of the attributes on the classification success are researched and according to the results the new approach is also suggested in the using as a feature selection technique of the pattern recognition problems.

Keywords: *Locally Weighted Naive Bayes, Least Squares, Classification, Class Probability, Feature Selectio*

1. Introduction

Naive Bayes (NB) is a simple structured statistical prediction algorithm [1, 2]. In the principle, it performs the optimum classification by saving the dependency of each attribute on only one class. The assumption of NB method about that all attributes have the same independency and importance is not reasonable [3], and for

[1] *Corresponding Author*, Electrical and Electronics Engineering Department, Faculty of Natural Sciences and Engineering, Gaziosmanpasa University, 60250 Tokat (e-mail: umutorhan@hotmail.com)

[2] Mechatronics Engineering Department, Faculty of Natural Sciences and Engineering, Gaziosmanpasa University, 60250 Tokat (e-mail: kemal.adem@gop.edu.tr)

[3] Mechatronics Engineering Department, Faculty of Natural Sciences and Engineering, Gaziosmanpasa University, 60250 Tokat (e-mail: onur.comert@gop.edu.tr)

this reason its classification success is usually less than the other classifiers. For solving this problem, there are many researches based on feature selection or weighted attributes by using NB [3-13]. One of them is Locally Weighted Naive Bayes (LWNB). In fact, LWNB is developed on Locally Weighted Learning (LWL) which is composed of memory based learning, lazy learning and regression [14]. In LWL based methods, the preferences in some details like distance measurements, weight functions, model structures, prediction, prevention the outliers, and improving the performance are very important. In LWNB method, the most important detail is the determination of the weights. First suggested LWNB method has higher classification success than classic NB, and its main advantage is easy computability [15]. The most important step in the computing process of LWNB method is the determination of the weights assigned to data attributes. There are many techniques tried to calculate the weight vector by learning data, but Hill Climbing technique is discovered as the best [16]. Similarly, the method called as Weighted Naïve Bayes can also determine the weights, and the determined weights can be successfully applied to classic NB method [8].

In LWNB method, the neighborhoods are usually evaluated by distance function. Despite the most frequently used distance function is Euclidean distance, it is not always the best [17]. Different distance functions like Interpolated Value Difference Metric [18], Minimum Risk Metric, Extended Short and Fukunaga Metric and its logarithm SF2LOG [19] compute the distance between two samples in data by subtracting the probabilities. According to the classification success, the best function is Interpolated Value Difference Metric among Euclidean and probability based these functions [17]. The weights of the attributes can also be computed by using Kullback-Leibler measurement [3]. In addition, because NB method assumes all classes as the crisp, its successes are usually in low levels. Instead of the crisp-classes assumption, by using the new concept named as weighted-classes, the success of classic NB method is improved [5]. The researches are commonly performed in the known datasets. But, there are some specific researches which are used LWNB method such as software defect prediction [20]. The assigning of the weights to the attributes frequently used in software defect improves the success of classic NB significantly. LWNB method is not only used for the supervised learning applications like the classification, but also it can be used for the semi-supervised learning processes. The newest study related to semi-supervised learning is the method named as Instance Weighted Naive Bayes which is more successful than classic NB [21].

In this study, the weights used in the main principle of LWNB method are estimated by least squares technique. For evaluating the success of this approach, nine two-class datasets from UCI database are used. In the following sections, we first describe Bayes rule and Naive Bayes classifier to better introduce and justify our approach. After we explain the proposed model developed on Locally Weighted Naïve Bayes method, we present the datasets used, the experiments, and the results of experiments. Finally, we evaluate the study in general and conclude by interpreting on the contribution of our approach.

2. Bayes and Naive Bayes

Bayes rule suggested first by Thomas Bayes describes the relation between marginal and conditional probabilities of two different sequential events (X and Y) depended on the same stochastic process [22]. This relation is given as follow.

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \tag{1}$$

where *P*(*X*), *P*(*Y*) is the marginal probabilities, and *P*(*X*|*Y*), *P*(*Y*|*X*) is the conditional probabilities. When Bayes rule is used in classification, the state had maximum probability among all possible output states is chosen as the decision class for output. The mathematical representation of this description is shown as follow.

$$y' = \underset{y_j}{\arg\max}\, P(Y = y_j \mid X) \tag{2}$$

where *y'* is the decision class, y_j is *j*. output state, and *X* is input sample. If the input sample has many attributes, the conditional probabilities of all attributes are multiplied as seen in Equation 3.

$$P(x_1, x_2, \ldots, x_m \mid y_j) = \prod_{i=1}^{m} P(x_i \mid y_j) \tag{3}$$

X input variable had *m* attributes is presented by $X=(x_1,x_2,\ldots,x_m)$. Despite Naive Bayes (NB) classifier is based on Bayes rule, it does not compute the real probabilities of the states. NB focuses on the determination of the class with the maximum probability. Because the probabilities of all target state are the same, the denominator value in Equation 1 is ignored for the classification and multivariate NB classifier make a decision by Equation 4.

$$y' = \underset{y_j}{\arg\max}\left(P(Y = y_j)\prod_{i=1}^{m} P(X = x_i \mid Y = y_j)\right) \tag{4}$$

Because of the directly effect of the number of samples and the attributes in data, classic NB has usually low successes. To prevent this, the powered weights are assigned to all factors in Equation 4.

3. Locally Weighted Naive Bayes (LWNB)

In Locally Weighted Naïve Bayes (LWNB) method, the weights are assigned to either the samples or the attributes. In first proposed LWNB method, the weights were computed according to the distance of the nearest neighbor and user defined distance threshold, and then classic NB classified a subset composed of the samples had the weight more than the threshold [15]. In the approach suggested in this study, a powered weight is assigned to each factor in the equation of classic NB, and this new equation is then transformed to linear regression form by using the logarithm. Lastly, the weights assigned to the probabilities are computed least squares techniques.

In fact, all classification methods are developed on two-class datasets. The proposed approach in this study is also described for two-class datasets, and the most important advantage of it is based on a basic assumption. According to the assumption, the class probability can be either the maximum or the minimum. In other word, the probability of the class is supposed as the maximum value for the class desired, where the samples belong. The class probability is the minimum value in otherwise. By the help of the determined weights in this way, the probabilities are increased for desired states and decreased for undesired states. In two-class data, if the class of a sample is y_1 and other class is y_2, the probabilities of both y_1 and y_2 are preferred as the multiple of 10 for easy representation of the results in next step of the study. As is known, the sum of the probabilities in classic NB method can not be 1 because NB is ignored the denominator in Bayes rule. Therefore two equations are written as follow for each sample in the dataset.

$$\begin{aligned} P^{w_0}(Y = y_1)\prod_{i=1}^{m} P^{w_i}(X = x_i \mid Y = y_1) = 10 \\ P^{w_0}(Y = y_2)\prod_{i=1}^{m} P^{w_i}(X = x_i \mid Y = y_2) = 10^{-1} \end{aligned} \tag{5}$$

When 10-based logarithms of both expressions in Equation 5 are compute, a new equation is established in linear regression form given by Equation 6. Thus the optimum values of the weights in powered position before can be computed in easier.

$$\begin{aligned} w_0 \log_{10} P(Y = y_1) + \sum_{i=1}^{m} w_i \log_{10} P(X = x_i \mid Y = y_1) = \log_{10} 10 \\ w_0 \log_{10} P(Y = y_2) + \sum_{i=1}^{m} w_i \log_{10} P(X = x_i \mid Y = y_2) = \log_{10} 10^{-1} \end{aligned} \tag{6}$$

When the equations in Equation 6 are rearranged by generalizing, a new equation given by Equation 7 is obtained.

$$w_0 \underbrace{\log_{10} P(Y)}_{E_0} + \sum_{i=1}^{m} w_i \underbrace{\log_{10} P(X \mid Y)}_{E_i} = \begin{cases} +1 & ,Y = y_1 \\ -1 & ,Y = y_2 \end{cases} \tag{7}$$

where E_0, E_i are the abbreviations of logarithmic expressions. By computing these expressions, the original dataset with n samples and m attributes is mapped into a new dataset with $2n$ samples and m+1 attributes. Thus, the new equation is transformed into linear regression form given as follow.

$$F = \sum_{i=0}^{m} w_i E_i \quad ,D = \begin{cases} +1 & ,Y = y_1 \\ -1 & ,Y = y_2 \end{cases} \tag{8}$$

where F is the prediction of LWNB method, and D is the target values in new dataset. Lastly, the weights used in Equation 8 are computed by using Least Squares (LS) technique which can simply calculate the linear effects of the data attributes on the output. According to this technique, the optimum values of the weights should minimize the sum of squared errors [22]. The error is determined by subtracting the predicted value from the target value in the dataset. The main optimization problem relation to LS technique can be presented as the objective function in Equation 9.

$$\min \sum_{j=1}^{n} \varepsilon_j^2 = \min \sum_{j=1}^{n} \left(D_j - F_j\right)^2 = \min \sum_{j=1}^{n} \left(D_j - w_1\, E_j - w_0\right)^2 \tag{9}$$

where j is the subscribe of a sample, i is the subscribe of the attributes, ε is the error, F is the prediction, and D is the target value.

The optimization problem can simply solve by equalizing first derivation of the objective function to zero. This solution can be adapted to multivariate linear regression equation by appending a subscript related to which attribute into E value. As a result, the term of w_i is computed by Equation 10. The detected weights are written in Equation 5 and the classification is performed by choosing the state with the maximum probability. LS technique has lower computational complexity like classic NB and it need not any repetition because of its simple and stable structure.

$$w_i = \frac{\sum_{j=1}^{n} E_{ij} D_j}{\sum_{j=1}^{n} E_{ij}^2} \tag{10}$$

4. Experimental Results

For the experiments in this study, nine different two-class datasets are used from machine learning database in the web site of California University [23]. All datasets are purified from the samples with missing values and the needless attributes like identity. Table 1 shows the numbers of attributes and samples in the datasets.

Table 1. The numbers of attributes and classes of datasets used in the study

Datasets	Dataset Original Name	Samples	Attributes
Diabetes	Pima Indians Diabetes	768	8
Haberman	Haberman’s Survival	306	3
Ionosphere	Ionosphere	351	34
Mushroom	Mushroom	5644	21
Parkinsons	Parkinsons	195	22
SpectHeart	SPECTF Heart	267	44
Transfusion	Blood Transfusion	748	4
Voting	Congressional Voting Records	232	16
Wisconsin	Breast Cancer Wisconsin	683	9

Since the proposed approach developed on LWNB is independent of type of data, it is implemented on both numerical and categorical datasets. In traditional, the dataset is separated into two subsets named as train and test according to chosen validation method for classification, then the variables of the classifier is adjusted by using train set, and lastly Total Classification Accuracy (TCA) ratio usually preferred as the success criterion is computed on test set. Because a comparison between the proposed approach in the study and classic NB is aimed, any validation method was not used and the approach was not compared different linear methods. Thus the datasets are used without separating for training, and TCA measurement is presented as the training success. TCA ratio is determined by dividing the numbers of correct classified samples to the numbers of all samples in data. For numerical datasets, the probabilities are calculated by gauss probability distribution, and a very small value (10^{-9}) is added all probabilities for solving the zero-probability problem. Classic NB and LWNB with the proposed approach methods are classified by nine datasets, and the obtained TCA ratios are presented in Table 2.

As seen in Table 2, LWNB method with the proposed approach is obviously superior to classic NB for Parkinsons, SpectHeart and Voting datasets. For Diabetes dataset, it can be said otherwise. In other datasets, both of the methods have almost equal successes. Because of the having high TCA ratios in both method, it can be said that Mushroom and Wisconsin datasets are the most appropriate ones among nine datasets for the statistical classification. In addition, when the methods (NB and LWNB) are compared to each other in point of the average success, it is observed that LWNB method with the proposed approach (83.81%) is more successful than classic NB method (81.77%). Class probability factor, exists in both of NB and LWNB methods and comes from Bayes rule, is related to the numbers of the samples in each class. This factor may benefit or damage to the classification success according to the preference of the person who prepares the data. The effect of class probability factor on LWNB method is analyzed on nine datasets, and the results are given by Table 3.

Table 2. The comparison of the successes of classic NB and LWNB with the proposed approach

Datasets	TCA (NB)	TCA (LWNB)
Diabetes	76.17	67.06
Haberman	74.51	73.53
Ionosphere	82.91	83.76
Mushroom	99.72	98.14
Parkinsons	69.74	83.59
SpectHeart	69.29	79.40
Transfusion	75.13	76.20
Voting	91.81	96.98
Wisconsin	96.63	95.61
Avarage	81.77	83.81

where TCA* shows TCA values without using class probability factor. Looking at Table 3, it can be said that the class probability factor supports positive effect on the classification success in Parkinsons, Mushroom and Transfusion dataset. For Spectheart and Voting datasets, it can not be mentioned about any positive or negative effect of this factor on the success. In other four datasets (Diabetes, Haberman, Ionosphere, and Wisconsin), negative effects of class probability factor are determined on the classification success.

Table 3. The effect of class probability factor on the success of LWNB method

Datasets	TCA	TCA*
Diabetes	67.06	74.48
Haberman	73.53	75.82
Ionosphere	83.76	88.89
Mushroom	98.14	97.20
Parkinsons	83.59	83.08
SpectHeart	79.40	79.40
Transfusion	76.20	60.56
Voting	96.98	96.98
Wisconsin	95.61	96.78

In another experimental analysis in the study, the effect of each attribute in dataset is researched on the classification success by using LWNB with the new approach. For this aim, all datasets are classified again by excluding a different attribute in each repetition. The experiments are repeated as much as the numbers of attributes for each dataset, then the minimum successes (Min TCA) and the maximum successes (Max TCA) are shown in Table 4. For easy comparison, the successes (TCA) obtained by all attributes are also included in the table.

Table 4. The contributions of attributes on the classification success

.

Datasets	Min TCA	Max TCA	TCA
Diabetes	65.10	67.58	67.06
Haberman	73.53	73.53	73.53
Ionosphere	75.21	86.32	83.76
Mushroom	73.49	99.88	98.14
Parkinsons	57.95	87.69	83.59
SpectHeart	79.40	79.40	79.40
Transfusion	76.20	76.20	76.20
Voting	76.72	96.98	96.98
Wisconsin	91.51	95.61	95.61

It is obviously shown in Table 4 that there are not any difference between both of the successes by using all attributes and excluding any attribute in three datasets (Haberman, SpectHeart, and Transfusion). But in Ionosphere, Mushroom and Parkinsons datasets, the success decreases by excluding an attribute. Thus, it is determined that the existence of the excluded attribute has positive effect on the classification success. Also in the same datasets, the success increases by excluding another attribute. It can be said for them that the excluded attribute affects the classification success negatively. Similarly in Diabetes, Voting and Wisconsin datasets, it is detected that an attribute has poor effect on the success, but no attribute changes the success positively. By help of this analysis, it can be determined the effects of all attribute on the classification effect. Therefore, LWNB method with the new approach would be considered as a feature selection technique in pattern recognition problems.

5. Conclusions

In this study, we suggest a new approach on the determination of powered weights assigned to the probability factors in Locally Weighted Naïve Bayes (LWNB) method by Least Squares (LS) technique. The performance of the new approach is measured by several experiments implemented on nine different datasets frequently used in testing of machine learning methods. Three researches are aimed by implementing these experiments. In the first research, LWNB with the proposed approach is compared classic Naive Bayes, and the higher success of the new approach is emphasized. In the second research, it is showed that the class probability factor may damage the classification success for some datasets. In the last research, the contribution of each attribute on the classification success is investigated by LWNB with the new approach, and it is detected that the proposed approach can find the effect of each attribute over the classification success. This research shows us that in addition to the classification problems, the new approach can be used as a feature selection technique. As a result, we believe that LWNB method with the proposed approach contributes the literature of machine learning by means of its higher success in addition to its fast and simple structure.

References

[1] D. D. Lewis, Naive bayes at forty: The independence assumption in information retrieval, In Proceedings of the tenth european conference on machine learning, Berlin, 1998, pp. 4–15.

[2] D. J. Hand, K. Yu, Idiot's bayes: Not so stupid after all?, International statistical review, vol. 69, no.3, 2001, pp. 385-398.

[3] C. H. Lee, F. Gutierrez, D. Dou, Calculating Feature Weights in Naive Bayes with Kullback-Leibler Measure, In Proceeding on Eleventh IEEE International Conference on Data Mining, 2011, pp. 1146-1151.

[4] S. Acid, L. M. De Campos, J. G. Castellano, Learning bayesian network classifiers: Searching in a space of partially directed acyclic graphs, Machine learning, vol. 59, no. 3, 2005, pp. 213-235.

[5] H. Alhammady, Weighted Naive Bayesian Classifier, In Proceeding on International Conference on Computer Systems and Applications, 2007, pp. 437-441.

[6] J. Cerquides, R. L. De Mantaras, Robust bayesian linear classifier ensembles, In Proceedings of the sixteenth european conference on machine learning, Porto, 2005, pp. 70-81.

[7] H. Langseth, T. D. Nielsen, Classification using hierarchical naive bayes models, Machine learning, vol. 63, no. 2, 2006, pp. 135-159.

[8] S. D. S. Pedro, E. R. Hruschka, N. F. F. Ebecken, WNB: A Weighted Naïve Bayesian Classifier, In Proceeding on Seventh International Conference on Intelligent Systems Design and Applications, 2007, pp. 138-142.

[9] G. I. Webb, J. Boughton, Z. Wang, Not so naive Bayes: Aggregating one-dependence estimators, Machine learning, vol. 58, no. 1, 2005, pp. 5-24.

[10] Z. Xie, W. Hsu, Z. Liu, M. L. Lee, SNNB: A selective neighborhood based naive Bayes for lazy learning, In Proceedings of the sixth pacific-asia conference on advances in knowledge discovery and data mining, Berlin, 2002, pp. 104–114.

[11] B. Zadrozny, C. Elkan, Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers, In Proceedings of the eighteenth international conference on machine learning, San Francisco, 2001, pp. 609-616.

[12] H. Zhang, L. Jiang, J. Su, Hidden naive bayes, In Proceedings of the twentieth national conference on artificial intelligence, Pittsburgh, 2005, pp. 919-924.

[13] F. Zheng, G. I. Webb, Finding the right family: Parent and child selection for averaged one-dependence estimators, In Proceedings of the eighteenth european conference on machine learning, Heidelberg, 2007, pp. 490-501.

[14] C. G. Atkeson, A. W. Moore, S. Schaal, Locally weighted learning, Artificial intelligence review, no. 11, 1997, pp. 11-73.

[15] E. Frank, M. Hall, B. Pfahringer, Locally weighted naive Bayes, In Proceedings of the nineteenth conference in uncertainty in artificial intelligence, Acapulco, 2003, pp. 249-256.

[16] H. Zhang, S. Sheng, Learning Weighted Naive Bayes with Accurate Ranking, In Proceedings of the fourth IEEE international conference on data mining, Brighton, 2004, pp. 567-570.

[17] B. Wang, H. Zhang, Probability based metrics for locally weighted naive bayes, In Proceedings of the twentieth Canadian conference on artificial intelligence, Montreal, 2007 pp. 180–191.

[18] D. R. Wilson, T. R. Martinez, Improved heterogeneous distance functions, Journal of artificial intelligence research, no. 6, 1997, pp. 1–34.

[19] E. Blanzieri, F. Ricci, Probability based metrics for nearest neighbor classification and case-based reasoning, In Proceedings of the third international conference on case-based reasoning research and development, Seeon Monastery, 1999, pp. 14–28.

[20] B. Turhan, A. Bener, Software defect prediction: Heuristics for weighted naive bayes, In Proceedings of the second international conference on software and data technologies, Barcelona, 2007, pp. 244-249.

[21] L. Jiang, Learning instance weighted naive bayes from labeled and unlabeled data, Journal of intelligent information systems, vol. 38, no. 1, 2012, pp. 257-268.

[22] C. M. Bishop, Pattern recognition and machine learning, Springer, 2006.

[23] C. Merz, P. Murphy, D. Aha, UCI repository of machine learning databases, Irvine: Department of ICS, University of California, accessed 10 July 2012, http://www.ics.uci.edu/mlearn/MLRepository.html.