Proceedings of 2012 4th International Conference on Machine Learning and Computing
IPCSIT vol. 25 (2012) © (2012) IACSIT Press, Singapore

Assigning Hybrid-Weight for Feature Attribute in Naïve Bayesian Classifier

Bao-En Guo and Hai-Tao Liu+

Department of Information Science and Technology, Xingtai University, Xingtai 054001, China

Abstract. In this paper, a novel naïve Bayesian classifier based on the hybrid-weight feature attributes (short of "NBC_{HWFA}") is proposed. NBC_{HWFA} arranges a hybrid weight for each feature attribute by merging the effectiveness of feature attribute on classification and the dependence between feature attribute and class attribute. In order to demonstrate the feasibility and effectiveness of proposed NBC_{HWFA}, we experimentally compare our method with standard naïve Bayesian classifier (NBC), NBC with gain ratio weight (NBC_{GR}), and NBC with correlation coefficient weight (NBC_{CC}) on 10 UCI datasets. And, a statistical analysis is also given. The final results show that NBC_{HWFA} can obtain the statistically best classification accuracy.

Keywords: naïve Bayesian classifier, gain ratio, correlation coefficient, classification

1. Introduction

Because its simplicity and effectiveness, naïve Bayesian classifier (short of "NBC") [1] which is based on Bayesian probability theory is applied in a wide range of practical fields, including medical diagnosis, text classification, email filtering and information retrieval [2]. Compared with decision tree and neural network, etc., which are more frequently-used learning algorithms, NBC can also obtain the better classification performance in some applications [3]. For the supervised classification problems with a large number of samples and large condition attributes, NBC can also deal with effectively.

In order to discuss our research conveniently, we list a number of notations used in this paper:

$C=\{c_1,c_2,\cdots,c_L\}$ is the class attribute set. L is the number of decision attributes.

$E=\left\{\vec{e}_1^{(1)},\cdots,\vec{e}_{N_1}^{(1)},\vec{e}_1^{(2)},\cdots,\vec{e}_{N_2}^{(2)},\cdots,\vec{e}_1^{(L)},\cdots,\vec{e}_{N_L}^{(L)}\right\}$ is the instance set, where N_j (j=1,2, $\cdots$,L) is the number instances in the j-th class; $\vec{e}_i^{(j)}=\left\{e_{i1}^{(j)},e_{i2}^{(j)},\cdots,e_{iD}^{(j)}\right\}\left(j=1,2,\cdots,L;i=1,2,\cdots,N_j\right)$ is the i-th instance in the j-th class, D is the number of feature attributes.

$\vec{e}=\{e_1,e_2,\cdots,e_D\}$ is a new instance whose class attribute is unknown.

Referring to the above notations, naïve Bayesian classifier (short of "NBC") determines the class attribute for the new instance $\vec{e}=\{e_1,e_2,\cdots,e_D\}$ with the following equation (1). Let $C(\vec{e})$ be the class attribute of new instance $\vec{e}=\{e_1,e_2,\cdots,e_D\}$:

$$C(\vec{e})=\underset{j=1,2,\cdots,L}{\arg\max}\, p\left(c_j\middle|\vec{e}\right)=\underset{j=1,2,\cdots,L}{\arg\max}\, p\left(c_j\right)p\left(\vec{e}\middle|c_j\right). \tag{1}$$

Because NBC assumes that all feature attributes are mutually independent, the definition of $C(\vec{e})$ can be redefined as follows [4]:

$$C(\vec{e})=\underset{j=1,2,\cdots,L}{\arg\max}\, p\left(c_j\right)\prod_{d=1}^{D}p\left(e_d\middle|c_j\right), \tag{2}$$

+ Corresponding author. Tel.: + +86 15230420378.
E-mail address: haitou.liu@gmail.com.

where, $p(c_j), j=1,2,\cdots,L$ is the prior probability and $p(e_d|c_j), d=1,2,\cdots,D$ is the class conditional probability. In this paper, we let $p(c_j)=1/L$ for the sake of computing complexity. $p(e_d|c_j)$ can be calculated as the method introduced by John and Langley in [5]:

$$p(e_d|c_j)=\frac{1}{N_j h_j}\sum_{k=1}^{N_j}\exp\left[-\frac{1}{2}\left(\frac{e_d-e_{kd}^{(j)}}{h_j}\right)^2\right], \tag{3}$$

where, the parameter $h_j, j=1,2,\cdots,L$ is the function of N_j which should satisfy the following conditions: $\lim_{N_j\to+\infty} h_j(N_j)=0$ and $\lim_{N_j\to+\infty} N_j\times h_j(N_j)=+\infty$. In this paper, we let $h_j=1/\sqrt{N_j}$.

The independent assumption of NBC can always be held. Thus, some improvements to NBC have been proposed. The weighted NBC is one type of these improvements. Zhang and Sheng [6] proposed that NBC can be weighted by using gain ratio (NBC_{GR}). Gain ratio is always used to measure which of the feature attributes are the most relevant with the class attribute. The more high the dependence between feature attribute and class attribute is, the larger the gain ratio weight w_{GR} of this condition attribute is. We call this weighted NBC with gain ratio weight NBC_{GR}. Zhang and Wang etc., [7] also gave a weighted NBC with correlation coefficient (simply NBC_{CC}). They used the correlation coefficient to measure the linear correlation between feature attribute and class attribute. The more linearly dependent feature attribute will obtain a larger correlation coefficient weight w_{CC}. The advantages of NBC_{GR} and NBC_{CC} had been demonstrated by experimental comparisons on standard UCI datasets [8]. However, we find that w_{GR} only considered the effectiveness of feature attribute on classification; it ignored the relationship between feature attribute and class attribute. And, for w_{CC}, the opposite is true. So, in order to give consideration to both NBC_{GR} and NBC_{CC}, a hybrid-weight NBC is proposed in this paper. We call it NBC_{HWFA} simply. w_{GR} and w_{CC} will be combined into a hybrid-weight w_{H} in NBC_{HWFA}. The hybrid-weight w_{H} takes into account the festiveness of feature attribute and its correlation with class attribute at same time.

In order to validate the efficiency and effectiveness of our proposed method, 10 standard UCI datasets are selected as the testing pool. Then, we compare the four Bayesian learning algorithms (NBC, NBC_{GR}, NBC_{CC} and NBC_{HWFA}) on the testing pool. The 10-times of 10-fold cross-validation are used to obtain the average classification accuracy of every algorithm. Finally, the correspondingly statistical analyses are given based on two-tailed t-test with a 95 percent confidence level. The experimental results show that NBC_{HWFA} can obtain the statistically best classification accuracy among all Bayesian learning algorithms.

2. The Proposed Hybrid-Weight NBC-NBC_{HWFA}

Gain ratio is always used to measure which of the feature attributes are the most relevant with the class attribute. Zhang and Sheng [6] proposed the gain ratio weight w_{GR} which can be calculated according to the following equation (4):

$$w_{\text{GR}}(d)=\frac{\text{GR}(E,A_d)\times D}{\sum_{d=1}^{D}\text{GR}(E,A_d)}, d=1,2,\cdots,D, \tag{4}$$

where, D is the number of condition attributes of dataset E. $w_{\text{GR}}(d)$ denotes the gain ratio weight of the d-th feature attribute A_d of dataset E. $\text{GR}(E,A_d)$ is the information gain ratio of condition attribute A_d with respect to dataset E [9, 10, 11].

The correlation coefficient weight [7] is given by Zhang and Wang, etc. They used the correlation coefficient to measure the linear correlation between feature attribute and class attribute. The more linearly dependent feature attribute will obtain a larger correlation coefficient weight w_{CC}. The weight w_{CC} can be calculated by the following equation:

$$w_{\text{CC}}(d)=\frac{\text{Cov}(A_d,C)}{\sqrt{\text{D}(A_d)\times\text{D}(C)}}, d=1,2,\cdots,D, \tag{5}$$

where, $\mathrm{Cov}(A_d, C) = E\{[A_d - E(A_d)][C - E(C)]\}$ is the covariance between feature attribute A_d and C, $D(A_i) = E\{[A_i - E(A_i)]^2\}$ is the variance of A_d and $D(C) = E\{[C - E(C)]^2\}$ is the variance of C.

The advantages of $\mathrm{NBC}_{\mathrm{GR}}$ and $\mathrm{NBC}_{\mathrm{CC}}$ had been demonstrated by experimental comparisons on standard UCI datasets [8]. However, we find that w_{GR} only considered the effectiveness of feature attribute on classification; it ignored the relationship between feature attribute and class attribute. And, for w_{CC}, the opposite is true. So, in order to give consideration to both $\mathrm{NBC}_{\mathrm{GR}}$ and $\mathrm{NBC}_{\mathrm{CC}}$, a hybrid-weight NBC is proposed in this paper. Based on the weigts w_{GR} and w_{CC} mentioned above, we give the calculation formulation of hybrid weight w_{H} as follows:

$$w_{\mathrm{H}}(d) = -\left[w_{\mathrm{GR}}(d) \times \log_2 w_{\mathrm{GR}}(d) + w_{\mathrm{CC}}(d) \times \log_2 w_{\mathrm{CC}}(d)\right], d = 1, 2, \cdots, D. \tag{6}$$

From the equation (6), we can observe the following two facts: (1) when $w_{\mathrm{GR}}=w_{\mathrm{CC}}=0.5$, the weight w_{H} can reach the maximum 1. That reflects such feature attribute which not only considers the effectiveness of classification but also the linear correlation with decision attribute will obtain the maximal hybrid-weight; (2) when $w_{\mathrm{GR}}=0$, $w_{\mathrm{CC}}=1$, or $w_{\mathrm{GR}}=1$, $w_{\mathrm{CC}}=0$, or $w_{\mathrm{GR}}=w_{\mathrm{CC}}=0$, the weight w_{H} reaches the minimum 0. It shows that feature attribute does not consider the effectiveness of classification and the linear correlation with class attribute will obtain minimal hybrid-weight.

The hybrid weight tries to find a balance between the effectiveness of classification and the linear correlation with class attribute. The feature attribute which reaches this balance will obtain the larger weight. Our strategy is different from the gain ratio weight and the correlation coefficient weight. These two methods only extend their researches from single aspect. The computing complexity of our proposed method is $O(N_{\mathrm{Train}}N_{\mathrm{Test}}d)$, where N_{Train} is the number of examples in training dataset E_{Train}, $N_{\mathrm{Test}}$is the number of examples in testing dataset E_{Test}, d is the number of condition attributes. And, the computing complexities of gain ratio weight and the correlation coefficient weight are all $O(N_{\mathrm{Train}}N_{\mathrm{Test}}d+d)$. From this comparison, we can find that the hybrid weight does not increase the computing complexity of NBC obviously.

3. The Experiments and Results

In our comparative experiment, 10 UCI datasets [8] are selected which represent a wide range of domains and data characteristics. The detailed descriptions of datasets are listed in Table 1. To the 10 UCI datasets, we adopted the following two pre-processing steps in our experiment:

Table. 1: The detailed description of 10 data sets used in our experiment.

Datasets	The number of attributes	The number of classes	The distribution of classes	The number of samples
Auto Mpg	5	3	245/79/68	392
Blood Transfusion	4	2	570/178	748
Credit Approval	15	2	383/307	690
Cylinder Bands	20	2	312/228	540
Ecoli	5	8	143/77/52/35/20/5/2/2	336
Glass Identification	9	7	76/70/29/17/13/9/0	214
Haberman's Survival	3	2	225/81	306
Heart Disease	13	2	150/120	270
Ionosphere	33	2	225/126	351
Iris	4	3	50/50/50	150

Note: In our study, we only consider the classification problem with continuous attribute. When computing the hybrid weights for the feature attributes, we need discretize all the continuous attributes. However, only we obtain the hybrid weights, the Bayesian learning algorithms use the continuous attributes to determine the class attribute of testing example.

(1) Delete the nominal attributes: In our work, we mainly apply Bayesian learning algorithms to deal with classification problems of continuous condition attributes. We only want to investigate the effect imposed by condition attributes on the classification performances of Bayesian learning algorithms.

(2) Discretize all the continuous attribute-values: All the continuous attribute-values in each dataset are discretized by unsupervised filter named *Discretize* in WEKA [12-22]. Its operation can be found as follows: *weka.filters.unsupervised.attribute.Discretize*. It discretizes all continuous values by binning.

In order to eliminate the effect generated by splitting dataset randomly, we use 10 times of 10-fold cross-validation procedure to implement our experiment. The experimental procedures are arranged as the following descriptions: Every dataset is randomly divided into 10 disjoint subsets, and the size of each subset is N/10, where N is the number of samples in this dataset. This procedure is run 10 times, each time using the different one of these subsets as the testing set and combining the other nine subsets for the training set. The testing accuracies are then averaged as the final classification accuracy. Every run for different Bayesian classification algorithms (NBC, NBC_{GR}, NBC_{CC} and NBC_{HWFA}) is carried out on the same training sets and evaluated on the same testing sets. In particular, the folds of cross-validation are also same for the different Bayesian classification algorithms (NBC, NBC_{GR}, NBC_{CC} and NBC_{HWFA}) on each dataset.

Now, we will compare the four different Bayesian learning algorithms on the real datasets by using 10-times of 10-folds cross-validation. The detailed experimental results are summarized in Table 2. The table records the average accuracies and standard deviation of 10-times of 10-folds cross-validation. The number in parentheses denotes the ranking of classification performance obtained by two-tailed t-test with a 95 percent confidence level. The last line in the Table 2 summarizes the average accuracies and rankings of four Bayesian algorithms on these 10 UCI dataset.

Table. 2: The detailed experimental results on accuracy and standard deviation.

Dataset	**Bayesian learning algorithms**			
	NBC	**NBC_{GR}**	**NBC_{CC}**	**NBC_{HWFA}**
Auto Mpg	0.645±0.012 (3.5)	0.645±0.016 (3.5)	0.674±0.017 (2)	0.679±0.009 (1)
Blood Transfusion	0.702±0.006 (4)	0.734±0.007 (2)	0.704±0.010 (3)	0.744±0.007 (2)
Credit Approval	0.713±0.006 (3)	0.745±0.008 (1)	0.680±0.011 (4)	0.720±0.003 (2)
Cylinder Bands	0.713±0.014 (2)	0.697±0.012 (3)	0.690±0.012 (4)	0.743±0.015 (1)
Ecoli	0.853±0.008 (3)	0.837±0.008 (4)	0.891±0.010 (1)	0.872±0.009 (2)
Glass Identification	0.591±0.034 (3)	0.614±0.032 (1.5)	0.555±0.029 (4)	0.614±0.029 (1.5)
Haberman’s Survival	0.741±0.012 (4)	0.775±0.013 (2)	0.780±0.009 (1)	0.742±0.010 (3)
Heart Disease	0.842±0.012 (4)	0.847±0.014 (2)	0.845±0.014 (3)	0.851±0.011 (1)
Ionosphere	0.906±0.008 (4)	0.911±0.007 (3)	0.915±0.008 (2)	0.940±0.004 (1)
Iris	0.959±0.010 (4)	0.960±0.011 (2.5)	0.960±0.015 (2.5)	0.967±0.011 (1)
Average	**0.767 (3.75)**	**0.776 (2.45)**	**0.769 (2.65)**	**0.787 (1.55)**

From the comparative results we can observe that the classification performance of NBC_{HWFA} is statistically best among all Bayesian classifiers. Its ranking of accuracy is 1.55. Compared with the standard NBC, NBC_{GR} and NBC_{CC} are also superior. Their rankings are 2.45 and 2.65 respectively. The standard NBC obtains the worst classification accuracy with classification ranking 3.75. The experimental results show that our hybrid weight can help NBC to improve the classification accuracy.

4. Conclusions

In this paper, a hybrid-weight naïve Bayesian classifier is introduced. Our strategy considers the condition attribute’s effectiveness of classification and correlation with decision attribute. The final experiments demonstrate that the new method is effective and efficient.

5. Acknowledgements

The authors are grateful for the anonymous reviewers. Their many valuable and constructive comments and suggestions helped us significantly improve this work. The authors also thank Dr. Yan-Wei Zhen for her help on the improvement of language quality.

6. References

[1] L. Pat, I. Wayne, T. Kevin. An analysis of Bayesian classifiers. *In Proceedings of the tenth National Conference on Artificial intelligence*, 1992, pp. 223-228.

[2] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Journal of Machine Learning*. 1997, **29** (2-3): 131-163.

[3] B. Remco. Native Bayes classifiers that perform well with continuous variables. *In Proceedings of the 17th Australian Conference on Artificial Intelligence*, Lecture Notes Artificial Intelligence, Berlin: Springer. 2004, pp. 1089-1094.

[4] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[5] J. George, L. Pat. Estimating continuous distributions in Bayesian classifiers. *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, Aug. 1995, pp. 338-345, 18-20.

[6] H. Zahng, S.L. Sheng. Learning weighted naïve Bayes with accurate ranking. *In Proceedings of the Fourth IEEE International Conference on Data Mining*, Brighton, United Kingdom, Nov. 2004, pp. 567-570.

[7] M.W. Zhang, B. Wang, B. Zhang, and Z.L. Zhu. Weighted naïve Bayes classification algorithm based on correlation coefficients. *Journal of Northeastern University (Natural Science)*. 2008, **7**: 952-955.

[8] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml, 2011.

[9] J.R. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*. 1996, **4**: 77-90.

[10] J.R. Quinlan. Induction of decision trees. *Machine Learning*. 1986, **1**(1): 81-106.

[11] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[12] I. Witten, E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. *Second edition*. Morgan Kaufmann, 2005.

[13] U.M. Fayyad, and K.B. Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*. 1992, **8**(1): 87-102.

[14] H. Singer. Maximum entropy inference for mixed continuous-discrete variables. *International Journal of Intelligent Systems*. 2010, **25**(4): 345-364.

[15] P.K. Li, B.D. Liu. Entropy of credibility distributions for fuzzy variables. *IEEE Transactions on Fuzzy Systems*, **16**(1), pp. 123-129, 2008.

[16] J. Dougherty, R. Kohavi, M. Sahami. Supervised and unsupervised discretization of continuous features. *In Proceedings of the Twelfth International Conference on Machine Learning (ICML-1995)*, 1995, pp. 194-202.

[17] C.N. Hsu, H.J. Huang, T.T. Wong. Why discretization works for naive Bayesian classifiers. *In Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000)*, 2000, pp. 399-406.

[18] C.N. Hsu, H.J. Huang, and T.T. Wong. Implications of the dirichlet assumption for discretization of continuous variables in naive Bayesian classifiers. *Machine Learning*. 2003, **53**(3): 235-263.

[19] Y. Yang, G.I. Webb. Non-disjoint discretization for naive-Bayes classifiers. *In Proceedings of the 19th International Conference on Machine Learning (ICML-2002)*, 2002, pp. 666-673.

[20] Y. Yang, G.I. Webb. Proportional k-interval discretization for naive-Bayes classifiers. I*n Proceedings of the 12th European Conference on Machine Learning (ECML-2001)*, 2001, pp. 564-575.

[21] Y. Yang, G.I. Webb. Weighted proportional k-interval discretization for naive-Bayes classifiers. *In Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-2003)*, 2003, pp. 501-512.

[22] Y. Yang, G.I. Webb. A comparative study of discretization methods for naive-Bayes classifiers. *In Proceedings of the 2002 Pacific Rim Knowledge Acquisition Workshop in PRICAI 2002 (PKAW-2002)*, 2002, pp. 159-173.