Efficient Implementation of Class Based Decomposition Schemes for Naive Bayes Classifier

Lateefa Shaik[1], **N. Narasimha Swamy**[2]

Department of Computer Science and Systems Engineering-Andhra University, Vishakhapatnam, Andhra Pradesh

Abstract: ***The accuracy of a Naive Bayes Classifier over text classification tasks can be significantly improved by taking advantage of decomposition schemes such as Error-Correcting Output Codes. ECOC is the task of document categorization and it is a method for decomposing multi-way classification problems into binary classification tasks. By The Additive nature of the classifier all binary classifiers can be trained in single pass through the data, through this approach the training complexity is reduces O(n.t.g) to O((n+t).g).***

Keywords: ECOC, BCH, Naive Bayes Classifier

1. Introduction

The decomposition of k-class problem into a set of n binary classification problem is one of the common approaches for tackling the multi class classification problem. With focus on classification problems and explores the methods of extending the binary classification techniques to multi-class classification (classifying instances into more than two classes), we are using class based decomposition schemes to Naive Bayes classifier. The decomposition schemes are (One-against-all or One-against-one), these schemes are modelled within the error correcting output codes (ECOC).

The original need for development of ECOC (Dietterich and Bakiri 1995) is to transfer the error correcting properties to the problem of multi-class Classification. Naive Bayes algorithm can benefit from the combination of ECOC for text classification. One-against-one decomposition of Naive Bayes is similar to standard Naive Bayes (Suilzmann et. al. (2007)). By the combination of ECOC with Naive Bayes we can improve the classification accuracy.

The idea behind this procedure is binary decomposition of a Naive Bayes classifier can be computed efficiently from the estimated conditional probabilities of the simple Naive Bayes. The Naive Bayes can result erroneous probability estimates due to optimal despite violations when making independent assumptions, so we can simply using ECOC-NB.

1.1 ECOC

In ECOC required code matrix, coding matrix selection is difficult when the design of ECOC classifier, this coding matrix can be generated by using the one-against-all or one-against-one decomposition schemes with BCH codes. Where rows represent the classes and columns represent the number of classifiers, if the number of classes is lower and lower the probability to generate a suitable ECOC matrix and after getting the code matrix, each row assigned as like code word for the particular class.

To find distance between the code words we can use BCH (Bose and Ray-Chaudhuri) codes, these codes are specifies hamming distance between the code words. If the distance is greater, number of errors can be detected or corrected is to be greater. BCH code words specific length and hamming distance can be computed using generator polynomial. To generate binary BCH codes we are using bchpoly routine of Gnu Octave. Sample BCH codes (15,5,7), (31,6,15), (63,7,31), (127,8,63). Detailed description and for Information on BCH codes found in Bose and Ray-Chaudhuri (1960), Macwilliams and Sloane (1983). In this paper we improved the text classification accuracy of Naive Bayes classifier with the combination of ECOC-NB.

2. Related Work

Naive Bayes classifier is the simple probabilistic classifier by applying Bayes theorem with independence assumptions between the features. Naive Bayes is having ability to learn multi-class predicators. Classification performance can be improved by the combination of ECOC. The combination seems to be promising especially for text classification (Berger 1999; Ghani 2000). Let each example X be characterized with g values $(a1, \ldots.., ag)$ for attributes $A1,..,Ag$, and C=$\{C1,..,Ck\}$ be the set of classes. Using Bayes theorem, we can compute the probability that x belongs to class Ci as:

$$\Pr(Ci|x) == \Pr(Ci) . \Pr(a1,..,ag|Ci)/\Pr(a1,..,ag)$$

Above equation can be written as:

$$Posterior=\frac{Prior *likeli\ hood}{evidence}$$

Using the class-conditional independence assumption, we can estimate the class conditional probability

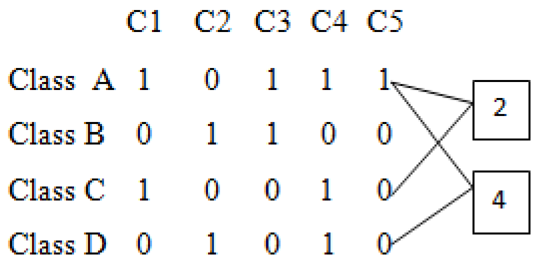
$$\Pr(a1,..,ag|Ci) = \prod_{j=1,\ldots.g} Pr(aj|Ci)$$

$\Pr(aj|Ci)$ And $\Pr(Ci)$ are estimates from the training data. The solution to multi categorization problems by identifying reducing them to multiple binary problems are then solved using a margin-based binary learning algorithm[1], The categorization of documents using ECOC (error-correcting-output codes)[2]. The error-correcting output codes provide a general-purpose method for improving the performance of inductive learning programs on multiclass problems [3].

3. Methodology

3.1 ECOC Matrix Generation Using BCH Codes

ECOC coding is a procedure for solving multi-way classification problems. The code words length hamming distance is completed using generator polynomial. Each class is initialized with randomly selected vector and it is multiplied with generator polynomial to get the code word for the class. The distance between pair of code words is should be the maximum in order to reduce misclassification of data.



	C1	C2	C3	C4	C5
Class A	1	0	1	1	1
Class B	0	1	1	0	0
Class C	1	0	0	1	0
Class D	0	1	0	1	0

If classifier C3 & C5 misclassifies an instance simultaneously then only the instance will be misclassify (A/C) by the ECOC ensemble.

When coming to B/C (or) A/B at least four classifiers need to misclassify an instance simultaneously in order to misclassify an instance between B/C or A/B. Vulnerability is between those pairs of classes-with minimum hamming distance over all pairs of classes (KC_2) where 'k' is number of classes, and it is possible to detect errors/ misclassification is less than the minimum hamming distance which is the characteristic feature of the coding scheme applied with the given data bit length and code length.

The minimum hamming distance can be obtained by the summation of two codes. If n=6 and Class A=(110011) and Class B=(10100) then A+B=(011010), w(A+B)=3 is gives the number of once in summation and d(A+B)=3=w(A+B). The hamming distance satisfies three conditions.

- d(A+B)=0 if only if A=B
- d(A,B)=d(B,A)
- d(A,B)+d(B,C)≥d(A,B)

3.2 Base Probability Reduction

General idea behind the efficient computation of arbitrary class based decompositions of Naive Bayes probability estimations can be reduced to parameter estimation of the Naive Bayes classifier is to $\Pr(aj|ci)$ and $\Pr(ci)$.

Consider a column of the coding matrix of the training set of a binary classifier. Column having set of classes as positive then it holds $Pr(C_b^+)$, based on positive samples and negative samples probabilities are calculated.

$$\Pr(C_b{}^+) = \sum_{c\in C_b^+} \Pr(c) \qquad \ldots\ldots. (1)$$

$$\Pr(a_j|C_b{}^+) = \Pr(a_j|c_1 \vee \ldots. \vee c_l) \qquad \ldots\ldots\ldots (2)$$

$$= \frac{\Pr(aj \wedge (c1 \vee \ldots.\vee cl))}{Pr(c1 \vee \ldots.\vee cl)}$$

$$= \frac{Pr(aj \wedge c1) + \ldots.+ Pr(aj \wedge cl)}{\sum_{i=1}^{l} \Pr(ci)} = \frac{\sum_{i=1}^{l} \Pr(aj \wedge ci)}{\sum_{i=1}^{l} \Pr(ci)}$$

Equation (1) and (2) shows all the necessary values $\Pr(C^+{}_b)$ and Pr $(a_{j|C_b^+})$ can be computed using $\Pr(C_i)$ and Pr $(a_j|C_i)$ as for the regular Naive Bayes. Different decompositions within ECOC framework can be applied to Naive Bayes because for reducing further probability estimation steps from training data.

3.2.1 Pre-calculation:

Discrete/nominal Attribute:

For discrete attributes Pr $(a_j|C_b^+) = \frac{(\sum_{i=1}^{l} |a_j \cap C_i|)+1}{\sum_{i=1}^{l} |C_i|+|a_j|}$

This complexity is not dependent on the number training instances. The training complexity is decreased when compared to the straight-forward application of the ECOC framework. Consider a problem with large number of attributes and classifiers lead to increase of testing complexity.

To solve the above problem we can use precalculation scheme, precalculating the probability distributions needed to the classifiers instead of aggregating the series the part-probabilities for each instance and calculating the probabilities based on attribute types.

Numeric Attribute:

For numeric attribute we can estimate the conditional probability using two procedures.

a. Normal Density Estimation:

Here conditional probability modelled as normal distribution

$$\mathbf{f_n(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\, exp^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

Where $\boldsymbol{\mu=\frac{1}{t}\sum_{m=1}^{t} x_m}$ and $\boldsymbol{\sigma\sqrt{\frac{\sum_{m=1}^{l} x_m^2 - \left(\sum_{m=1}^{t} x_m\right)^2/t}{t}}}$

Sum of observed attribute values is $\sum_{m=1}^{t} x_m$ and sum of squared values is $\sum_{m=1}^{t} x_m^2$

b. Kernel Density Estimation:

$$\mathbf{f_k(x)=\frac{1}{t.h}\sum_{m=1}^{t} K\left(\frac{x-x_m}{h}\right)}$$

Kernel density estimator also maintains all observed data values x_m and probability estimate of value x depends on its distance to these values.

3.3. Output Classification:

For 'k' classes and n bit-length of the matrix, generates 2^n possible combinations of codes, so for each class number of possible combinations are 8, i.e. n/k combinations are getting for each class.

Let $\propto_i$ is sequence of bits transmitting over having number of one in those places, when an error occur in transmitting the bits then $\in_i$ is the noise vector, the received output sequence is $\propto_i+\in_i$ and number of errors is w($\propto_i$). The Code is said to be t-error correcting, in case of t-error error correction, we use slepians decoder, by this partition the group B_n into 2^r cosets and values of r= n-log k

Step 1: Identify (2^r) cosets in B_n
Step 2: Identify coset leader
Step 3: Add coset leader to $\propto_A$to get S_A and similarly get S_B ,S_C , S_D
Classification is based on belongingness of the n-bit output vector which is formed the classification given by n classifications of the partitions S_A to S_D

a. Algorithm (ECOC-NB Training):

Requirements: *ECOC matrix M, Training set T*

(1) For each training instance(x,C_i) calculate ePrior. OBSERVE (i) is nothing but Pr (C_i)

(2) For each training instance x calculate e$Condi_{i,j}$ .OBSERVE (ValueOf (a_j))

(3) For each classifier f_i calculate Pr (C_{bj}^+) and Pr (C_{bj}^-)

(4) For each attribute a_k precalculate Pr (a_k | c_{bj}^+) and Pr (a_k | c_{bj}^-) based on attribute type and also based on estimation technique.

b. Algorithm (ECOC-NB Testing):

Requirements: matrix (m_{ij}), Pr (C_{bj}^+), Pr (C_{bj}^-), Pr(a_k | c_{bj}^+),Pr (a_k | c_{bj}^-), training instance x= (a_1 , $a_{2,}$, a_m).

(1) For each binary classifier and for each column compute bit predictions using MAKETERNARY function

(2) Return argument matrix $\sum_{j=1}^{n} m_{ij}\, b_j$

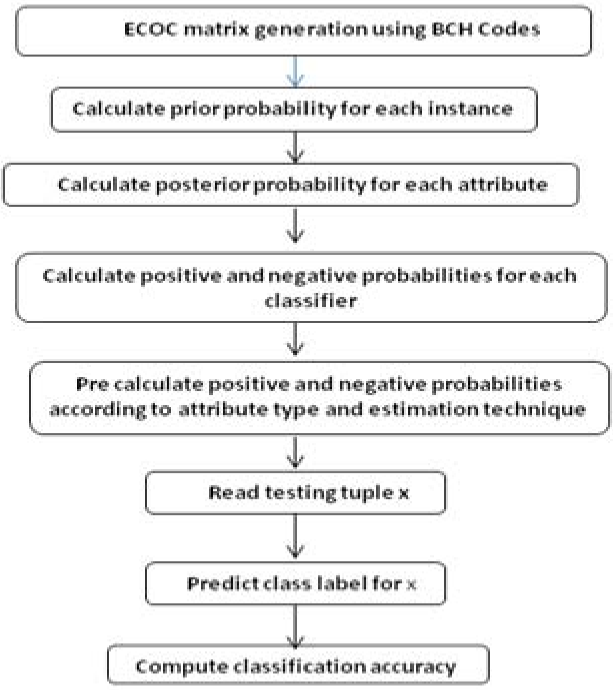
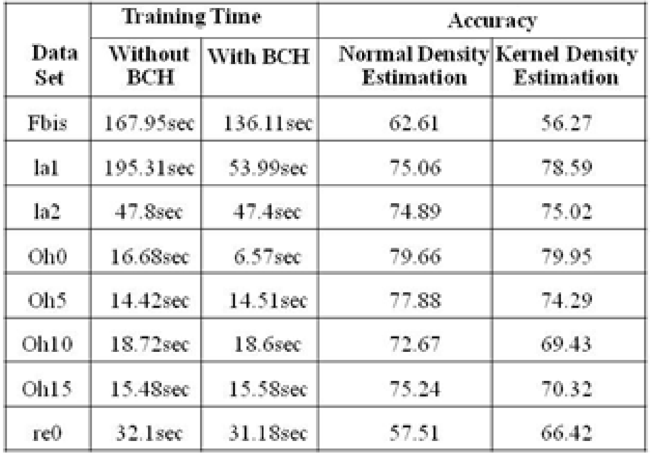

ECOC matrix generation using BCH Codes
Calculate prior probability for each instance
Calculate posterior probability for each attribute
Calculate positive and negative probabilities for each classifier
Pre calculate positive and negative probabilities according to attribute type and estimation technique
Read testing tuple x
Predict class label for x
Compute classification accuracy

Figure 1: Data Processing Diagram

4. Results

The classification accuracy of ECOC-NB is similar to standard Naive Bayes but in some cases accuracy in increased when bit-length size increased and the training time between Naive Bayes and ECOC-NB is dependent on training instances (t). The training complexity of Naive Bayes is O (n.t.g), but training time is decreased by using ECOC testing and training algorithms, while working with the large datasets and because of the additive nature of the binary classifiers can be trained in a single pass through the data. We performed classification on some data sets and through this proposed decomposition schemes training complexity is reduced to O ((n+t).g). We observed results for different data sets with increase of bit length and using two density estimation (kernel and normal) techniques.

Table 1: Training Time and Accuracy of different Data Sets



Data Set	Training Time		Accuracy	
	Without BCH	With BCH	Normal Density Estimation	Kernel Density Estimation
Fbis	167.95sec	136.11sec	62.61	56.27
la1	195.31sec	53.99sec	75.06	78.59
la2	47.8sec	47.4sec	74.89	75.02
Oh0	16.68sec	6.57sec	79.66	79.95
Oh5	14.42sec	14.51sec	77.88	74.29
Oh10	18.72sec	18.6sec	72.67	69.43
Oh15	15.48sec	15.58sec	75.24	70.32
re0	32.1sec	31.18sec	57.51	66.42

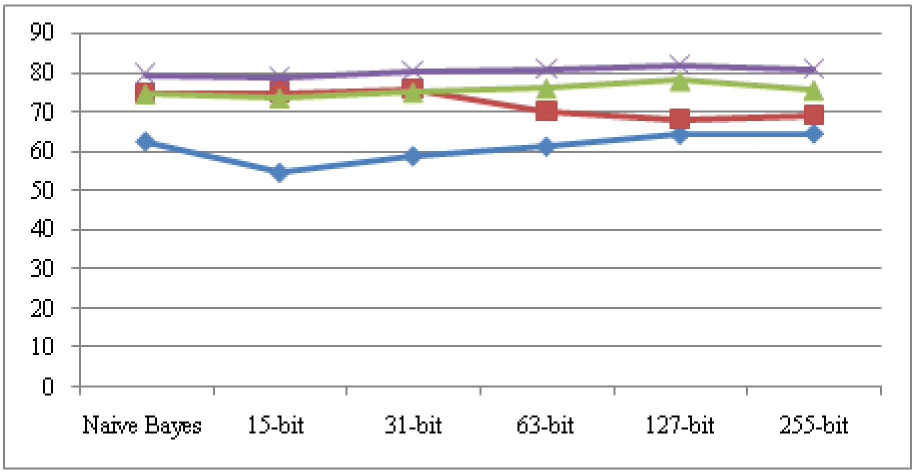


Figure 2: Accuracy comparison of Naive Bayes with different Data sets of given bit lengths

5. Conclusion

By the comparison of Naive Bayes with ECOC-NB, the straight-forward method with a training complexity of O (n.t.g) its complexity using normal and discrete density estimation methods is reduced to O ((n+t).g). By the conjunction with kernel density estimators the worst-case complexity remains the same, but in contrast it can benefit from a low number of distinct feature values.

A disadvantage of using the decomposition approach is the need for tuning parameters such as the bit-length. The cost of such a parameter tuning has become feasible. In future ECOC-NB can benefit. Naturally from specialized code types, which is an active research topic (e.g., pujol et al. 2006; Escalera et al. 2010) and the implementation of ECOC-NB is similar to regular Naive Bayes. The combination of Naïve Bayes with Error-Correcting Output Codes is almost as fast as a conventional Naïve Bayes classifier. ECOC are thus a viable technique for trying to improve the classification performance of Naïve Bayes on large-scale datasets.

References

[1] Bose and Ray-Chaudhuri(1960).
[2] Allwein, E.L.Schapire, R.E., & Singer, Y. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1, 113-141, 2000.
[3] Berger,A. Error-correcting output coding for text classification. In *Proceedings of the IJCAI-99 workshop on machine lerning for information filtering* (IJCAAI99-MLIF), Stockholm,Sweden, 1999.
[4] Ditterich, T.G., & Bakiri,G. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2,263-286.
[5] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3, 1289–1305.
[6] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. SIGKDD Explorations, 11(1), 10–18.
[7] Kittler, J., Ghaderi, R., Windeatt, T., & Matas, J. (2003). Face verification via error correcting output codes. Image and Vision Computing, 21(13–14), 1163–1169.
[8] Park, S. H., & Fürnkranz, J. (2012). Efficient prediction algorithms for binary decomposition techniques. Data Mining and Knowledge Discovery, 24(1), 40–77.
[9] Domingos, P., & Pazzani, M. J. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29(2–3), 103–130.
[10] Pujol, O., Radeva, P., & Vitrià, J. (2006). Discriminant ECOC: A heuristic method for application dependent design of error correcting output codes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(6), 1007–1012.