Improvements over Adaptive Local Hyperplane to Achieve Better Classification

Hongmin Cai[*]

School of Information Science and Technology,
The Sun Yat-sen University, P.R. China

Abstract. A new classification model called adaptive local hyperplane (ALH) has been shown to outperform many state-of-the-arts classifiers on benchmark data sets. By representing the data in a local subspace spanned by samples carefully chosen by Fisher's feature weighting scheme, ALH attempts to search for optimal pruning parameters after large number of iterations. However, the feature weight scheme is less accurate in quantifying multi-class problems and samples being rich of redundance. It results in an unreliable selection of prototypes and degrades the classification performance. In this paper, we propose improvement over standard ALH in two aspects. Firstly, we quantify and demonstrate that feature weighting after mutual information is more accurate and robust. Secondly, we propose an economical numerical algorithm to facilitate the matrix inversion, which is a key step in hyperplane construction. The proposed step could greatly low the computational cost and is promising fast applications, such as on-line data mining. Experimental results on both synthetic and real benchmarks data sets have shown that the improvements achieved better performance.

Keywords: Classification, adaptive local hyperplane, feature weighting, wrapper, mutual information, rank decomposition.

1 Introduction

Despite its age and simplicity, the Nearest Neighbor(NN) classification rule is among the most successful and robust methods for many classification problems. Many variations of this model have been reported by using various distance functions. A very interesting revision was achieved by approximating each class with a smooth locally linear manifold [20]. Recently, the authors [19] further generalized this revision by considering the feature weighting in local manifold construction, and the proposed model was called *adaptive local hyperplane* (ALH). The ALH classifier[19,22] was compared with classical classifier in many real data sets. The results were very promising.

[*] This work was supported by NSF of Guangdong Province, China under award number 9451027501002551, and the China Fundamental Research Funds for the Central Universities under award number $10ykjcll$.

P. Perner (Ed.): ICDM 2011, LNAI 6870, pp. 1–10, 2011.

Feature weighting plays an important step in ALH classifer. In general, the feature weights were obtained by assigning a continuous relevance value to each feature in hoping to enhance the classification performance of a learning algorithm by stressing on the context or domain knowledge. The feature weighting procedure is particularly useful for in instance based learning models, which usually construct the distance metrics by using all features [21]. Moreover, feature weighting could reduces the risk of over-fitting by removing noisy features thereby improving the predictive accuracy. Existing feature selection methods broadly fall into two categories, wrapper and filter methods. Wrapper methods use the predictive accuracy of predetermined classification algorithms (called base classifer), such as SVMs, as the criteria to determine the goodness of a subset of features [5,9]. Filter methods select features based on discriminant criteria that rely on the characteristics of data, independent of any classification algorithm [4,10,12]. The commonly discriminant criteria includes entropy measurement [13], Chi-squared measurement [15], correlation measurement [11], Fisher ratio measurement [6], mutual information measurement[14], and RELIEF-based measurement [18].

The key strength of the ALH classifier is in its incorporation of the feature weighting method into its nearest neighbor selection and local hyperplane construction. Thus, the data is represented in a weighted space by evaluating the feature importance in advance. However, the original feature weighting method in ALH considers the class separation criteria for individual features independently by using the *ratio of between-group to within-group sum-of-squares* (RBWSS). This criterion is known for that it omits the dependence among the features, and thus is less accurate when the tested data set being rich of redundant features. Therefore, the classification performance of ALH will be degraded.

In this paper, we proposed improvement on the standard ALH model in two aspects [19]. The first improvement is to evaluate the feature weighting scheme by mutual information, which is shown to be more accurate and robust in multi-classification problems [16,3]. The second improvement is to propose an economical numerical algorithm to low the computational cost during classification.

This paper is organized as follows. Sections 2 provides an introduction to the basics of adaptive local hyperplane (ALH) method. The previous weighting scheme was analyzed and replaced by new weighting function based on mutual information. Section 3 proposed a correction of numerical algorithm to dramatically low the computational cost during classification, thus facilitating its usage in data of large dimension. Section 4 demonstrated the performance of proposed method on benchmark data. Conclusion was presented in Section 5.

2 Adaptive Local Hyperplane and Feature Weighting Scheme

Let $\{x_i\}_{i=1}^{l}$ be a d-dimensional training data set with known class label $y_i = c$, for $i = 1, \ldots, l$ and $c = 1, \ldots, J$. In ALH algorithm, given a query sample, the first step is to find for each class the training points nearest (called *prototype*) to the query. The metric between samples was defined dependent on the feature weights. These selected prototype samples are then used to construct a local linear manifold for each class in the training set. Finally the query sample is assigned to the class associated with the closest manifold.

Adaptive Local Hyperplane. In the prototype selection stage, the feature weight is estimated by the ratio of the between-group to within-group sums of squares, called RBWSS scheme [19]:

$$r_j = \frac{\sum_i \sum_c I(y_i = c)(\bar{x}_{cj} - \bar{x}_j)^2}{\sum_i \sum_c I(y_i = c)(x_{ij} - \bar{x}_{cj})^2}, \tag{1}$$

where $I(\cdot)$ denotes the indicator function, $\bar{x}_{cj}$ denotes the jth component of class centroid of class c and $\bar{x}_j$ denotes the jth component of the grand class centroid. It is trivial to verify that the RBWSS weighting scheme ranks the feature importance by Fisher criterion, and thus is not accurate in multiple learning problems. Given the ranked feature importance, one attempts to further amplify their difference through an exponential normalization of the feature weights. This Fisher's method could rank the feature importance by a simple implementation with economic computational cost [6]. However, it tends to outweight abundant or easily separable classes if classes of the data sets are unevenly distributed [7]. To address this problem, the mutual information based criterion has been shown to be an effective measurement [10].

Mutual Information. The relevant features contain important information about the output whereas the irrelevant features contain little information regarding the output. Therefore, the task of feature weighting could be accomplished by measuring the "richness" of information concealed in data. For this purpose entropy and mutual information are introduced in Shannon's information theory to measure the information of random variables [17].

Given a discrete random variable X with its probability density function denoted as $p(x)$, the entropy of X can be defined as

$$H(X) = -\sum p(x) \log p(x) \tag{2}$$

For the case of two discrete random variables, i.e., X and Y, the joint entropy of X and Y is defined as follows:

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y) \tag{3}$$

where $p(x,y$ denotes the joint probability density function of X and Y. The common information of two random variables X and Y is defined as the mutual information between them,

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{4}$$

Quantitative measurement of feature importance for the classification task based on mutual information is one of the most effective technique for feature weighting. By resembling mutual information terms, one could obtain the quantification of the features subsets, such as *Redundancy* and *Relevance* [16]. One could use these terms to obtain features set catering to empirical needs. For instance, the scheme based on minimal-redundancy-maximal-relevance (mRMR) criterion has been developed to find a compact set of superior features at a low computational cost [16].

In this paper, we adopt the feature weighting by the mutual information criterion for the ALH classifier to overcome the limitations of the original RBWSS scheme. Synthetic examples will be given later in Section 4 to support this correction.

In the second stage, the hyperplane of class c is constructed by:

$$LH_c(q) = \{\boldsymbol{s} \mid \boldsymbol{s} = V\alpha\}, \tag{5}$$

where V is a $d \times n$ matrix composed by prototypes: $V_{\cdot i} = p_i$, with p_i being the ith nearest neighbor (called *prototype*) of class c, The parameter of $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ are solved by minimizing the distance between training samples q and the space of $LH_c(q)$ with regularization:

$$J_c(q) = \min_{\alpha}(s-q)^T W(s-q) + \lambda\alpha^T\alpha, \tag{6}$$

where $s \in LH_c(q)$, W is the diagonal matrix with $W(j,j) = w_j$ and λ is the regularization parameter.

The minimization of (7) could be achieved by solving a quadratic equation for α:

$$(V^T W V + \lambda\ I_n)\alpha = V^T W q. \tag{7}$$

At the last stage, the class label of the new comer is decided by the weighted Euclidean distance between the new comer and the local hyperplane of each class.

3 A Numerical Correction

The matrix in L.H.S of Eq. (7) is positively definite and thus classical algorithms such as QR-decomposition could be employed to find its inverse matrix [8]. In order to obtain optimal pruning parameters, such as number of the prototypes (nearest neighborhoods) and the regularizer λ, cross-validation scheme was shown to be effective and fast, thus usually serving as top choice. However, this incurs to large computation in ALH since many local hyperplane need to be constructed in Eq. 5. This problem tends to be more worse if the sample feature is in larger dimension as in many biomedical problems.

In this paper, we shall prove that the inverse matrix in L.H.S of Eq. (7) could be obtained by series of matrix multiplication instead of inversing directly, thus greatly saving the computational cost. In more details, we assume that we already derived the inverse matrix of $V_n^T W V_n + \lambda I$, consisted by n prototypes. By adding one more prototype, one is expecting to represent the local hyperplane in a less biased way, hoping to enhance its discrimination power. It implies the necessity of computing the inversion of matrix $V_{n+1}^T W V_{n+1} + \lambda I$. We will show that the inversion of $V_{n+1}^T W V_{n+1} + \lambda I$ could be updated consecutively by matrix multiplication from $V_n^T W V_n + \lambda I$. For clarity, we named this revised version as *MI-ALH.*

Theorem 1. *Suppose that $A_n = V_n^T W V_n + \lambda I$, then the matrix of $A_{n+1} = V_{n+1}^T W V_{n+1} + \lambda I$ could be formulated as:*

$$A_{n+1} = \begin{pmatrix} l_1 & \boldsymbol{l}^T \\ \boldsymbol{l} & A_n \end{pmatrix}, \tag{8}$$

and its inverse matrix is given by:

$$A_{n+1}^{-1} = G_{n+1} \begin{pmatrix} l_1^{-1} & 0 \\ 0 & A_n^{-1} - \frac{A_n^{-1}\bar{\boldsymbol{l}}\boldsymbol{l}^T A_n^{-1}}{1+\boldsymbol{l}^T A_n^{-1}\bar{\boldsymbol{l}}} \end{pmatrix} G_{n+1}^T, \tag{9}$$

where

$$\bar{\boldsymbol{l}}^T = -\frac{\boldsymbol{l}^T}{l_1} \tag{10}$$

$$G_{n+1} = \begin{pmatrix} 1 & \bar{\boldsymbol{l}}^T \\ 0 & I \end{pmatrix}. \tag{11}$$

The proof of this theorem is dependent on the following lemma and we would like to prove it at first.

Lemma 1. *Let A_{n+1} be a symmetric positively definite matrix of order $n+1$, with form of:*

$$A_{n+1} = \begin{pmatrix} l_1 & \boldsymbol{l}^T \\ \boldsymbol{l} & A_n \end{pmatrix} \tag{12}$$

where l_1 is a constant and $\boldsymbol{l}_{1\times n}^T$ is a vector. A_n is a symmetric positively definite matrix of order n. Then the inverse matrix of A_{n+1} is given by:

$$A_{n+1}^{-1} = G_{n+1} \begin{pmatrix} l_1^{-1} & 0 \\ 0 & A_n^{-1} - \frac{A_n^{-1}\bar{\boldsymbol{l}}\boldsymbol{l}^T A_n^{-1}}{1+\boldsymbol{l}^T A_n^{-1}\bar{\boldsymbol{l}}} \end{pmatrix} G_{n+1}^T \tag{13}$$

where $\boldsymbol{l}$ and G_{n+1} are defined in Eq. (10-11).

Proof. Since the matrix A_{n+1} is positively definite, it could be diagonalize by series of Gaussian elimination. The permutation matrix G could be employed to perform once Gaussian elimination. It is trivial to verify that

$$G^T A_{n+1} G = \begin{pmatrix} l_1 & 0 \\ 0 & \bar{\boldsymbol{l}}\boldsymbol{l}^T + A_n \end{pmatrix}. \tag{14}$$

According to the Sherman-Morrison-Woodbury formula [2], we know that

$$(A_n + \bar{\boldsymbol{l}}\boldsymbol{l}^T)^{-1} = A_n^{-1} - \frac{A_n^{-1}\bar{\boldsymbol{l}}\boldsymbol{l}^T A_n^{-1}}{1 + \boldsymbol{l}^T A_n^{-1}\bar{\boldsymbol{l}}}. \tag{15}$$

Therefore,

$$A_n^{-1} = G\begin{pmatrix} l_1^{-1} & 0 \\ 0 & (\bar{\boldsymbol{l}}\boldsymbol{l}^T + A_{n-1})^{-1} \end{pmatrix} G^T \tag{16}$$

$$= G\begin{pmatrix} l_1^{-1} & 0 \\ 0 & A_{n-1}^{-1} - \frac{A_{n-1}^{-1}\bar{\boldsymbol{l}}\boldsymbol{l}^T A_{n-1}^{-1}}{1+\boldsymbol{l}^T A_{n-1}^{-1}\bar{\boldsymbol{l}}} \end{pmatrix} G^T. \tag{17}$$

Now we end the proof of the **Lemma 1**.

Given the **Lemma 1**, we now continue to prove the **Theorem 1**.

Proof. Let $V_d = (P_1, P_2, \cdots, P_n)$ denote the prototype matrix. Assuming that one needs to add a new prototype $\bar{P}$ to enhance the discrimination power, it will result in a new prototype matrix $V_{n+1} = (\bar{P}, P_1, P_2, \cdots, P_n)$.

It is trivial to verify that:

$$V_{n+1}^T W V_{n+1} + \lambda I \tag{18}$$

$$= \begin{pmatrix} \bar{P}^T W \bar{P} + \lambda & \bar{P}^T W P_1 & \cdots & \bar{P}^T W P_n \\ P_1^T W \bar{P} & P_1^T W P_1 + \lambda & \cdots & P_1^T W P_n \\ \vdots & \vdots & \ddots & \vdots s \\ P_n^T W \bar{P} & P_n^T W P_1 & \cdots & P_n^T W P_n + \lambda \end{pmatrix}$$

$$= \begin{pmatrix} l_1 & \boldsymbol{l}^T \\ \boldsymbol{l} & V_n^T W V_n + \lambda I \end{pmatrix}, \tag{19}$$

where $l_1 = \bar{P}^T W \bar{P} + \lambda$ and $\boldsymbol{l}^T = (\bar{P}^T W P_1, \bar{P}^T W P_2 \cdots, \bar{P}^T W P_n)$. Therefore, the inversion of new prototype matrix could be obtained directly by Eq. (13), and we have:

$$A_{n+1}^{-1} = (V_{n+1}^T W V_{n+1} + \lambda I)^{-1} = G\begin{pmatrix} l_1^{-1} & 0 \\ 0 & A_n^{-1} - \frac{A_n^{-1}\bar{\boldsymbol{l}}\boldsymbol{l}^T A_n^{-1}}{1+\boldsymbol{l}^T A_n^{-1}\bar{\boldsymbol{l}}} \end{pmatrix} G^T. \tag{20}$$

This concludes our proof.

In summary, better classification could be obtained by adding more prototypes. However, the addition incurs to larger computational cost in solving the linear equation of Eq. (7) through matrix inversion. We have shown that the inversion of matrix could be updated directly from early inversion. This correction is fast and efficient, thus is promising for classification, even for high dimensional data.

4 Experimental Results

The mutual information based criteria has been widely applied in feature weighting and feature selection, thus facilitating its usage in machine learning [6,7,10,3]. This criteria is more accurate than RBWSS in evaluating the feature importance of multi-class data, or data being rich of redundance. We can show this by constructing a synthetic example.

In the first example, the tested data contains five feature variables in the well-known diamond shape, shown in Fig. 1, and ten noise features following standard normal distribution of $N(0, 1)$. The class label Y is completely determined by variable X_1 and X_2, both following normal distribution of $N(2, 1)$. The variable X_3 is dependent on X_1 with noise degration, and X_4 is dependent solely on X_2 with noise degration, $X_6, \cdots, X_{15}$ are noise features. The variable of X_5 satisfies $X_5 = X_3 + X4$, thus contains more information on label Y than X_3 or X_4, individually. The ideal order of the feature variable should be $X1 \approx X2 > X3 \approx X4 > X5 >>$ other noisy features. However, the top five features ranked by RBWSS is $X_8, X_{12}, X_{13}, X_{10}, X_{15}$, which are all noisy features. In comparison, the top five features ranked by mutual information (Eq. (4)) is X_1, X_2, X_3, X_4, X_5 with value of $1.0183, 0.9465, 0.7668, 0.6043, 0.4779$, respectively. The weighting scheme after mutual information demonstrate better accuracy and robustness to noises.

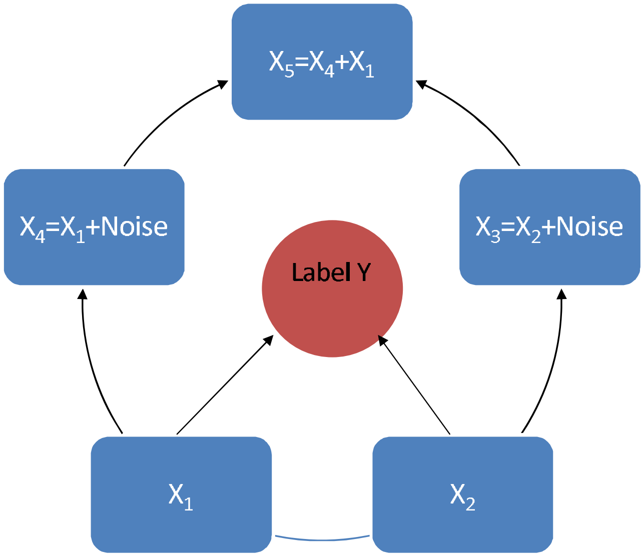

Fig. 1. A synthetic example having well-known diamond shape. The test data set is consisted by fifteen features and its label is completely determined by feature X_1 and X_2. The redundant feature X_3 and X_4 are s dependent on X_2 and X_1, respectively. The fifth feature X_5 is dependent on X_3 and X_4, degrated by noises. $X_6, \cdots, X_{15}$ are i.i.d noises following standard normal distribution.

We further demonstrate the performance of the proposed MI-ALH in classification by comparing it with ALH [19] on eleven real data sets. The tested nine benchmark data sets were downloaded from the UCI Machine Learning Repository [1], and they have been widely tested by various classification models. Three validation procedures, including the leave-one-out(LOO), 10-folds, and 20-folds cross validation, were carried out for hyperparameters estimation and accuracy testing on each dataset.

The results were summarized in Table 1. If using LOOCV, the performance of MI-ALH is slightly better than ALH. Moreover, with the decreasing of the training sample size, the performance of MI-ALH tends to better. For example, MI-ALH achieved higher classifications on 5 data sets vs lower classifications on 3 data sets under 10-fold cross validation, while 8 vs 3 in 20-fold cross validation. The outperformance obtained by MI-ALH was due to the accurate feature weighting scheme.

Table 1. Classification accuracies (%) on 11 real data sets. The better results are highlighted in bold under three different cross-validation scheme. The MI-ALH outperforms than standard one in most cases. Moreover, with the size of training sample decreasing from LOOCV to 20-fold-CV, the performance of MI-ALH was better than standard ALH, implying the accuracy and robustness of feature weighting scheme.

Validation Scheme	LOOCV		10-fold CV		20-fold CV	
Dataset	ALH	MI-ALH	ALH	MI-ALH	ALH	MI-ALH
Iris	**98.00**	97.33	96.00	96.00	96.52	95.90
Glass	75.23	**76.64**	57.40	**58.36**	61.45	**63.18**
Vote	96.98	96.98	96.56	96.56	96.52	**96.93**
Wine	98.88	**99.44**	96.63	**97.75**	98.83	**98.89**
Teach	**75.50**	74.83	**68.00**	66.71	70.23	70.00
Sonar	90.87	**91.35**	64.41	**66.33**	71.73	**72.87**
Cancer	**82.83**	82.32	**80.21**	79.71	81.56	81.06
Dermatology	97.27	**97.81**	97.00	**97.54**	96.18	**96.74**
Heart	**60.27**	59.60	57.92	57.92	57.31	**57.95**
Prokaryotic	91.68	91.68	81.153	**81.35**	88.36	**88.47**
Eukaryotic	85.08	**85.50**	75.33	**75.37**	80.31	**80.60**
Score	4 vs 5		3 vs 6		3 vs 8	

5 Conclusion

The *adaptive local hyperplane* model has been shown to be a very effective classification model on various type of data sets. However, the feature weighting scheme is less accurate in quantifying multi-class problems and samples being rich of redundance. Therefore, it leads to less accurate prototypes selection and unreliable local hyperplane construction. In this paper, we are proposing to two improvements over the standard ALH model. The first improvement is to evaluate the feature weighting scheme through mutual information. Experimental results both on synthetic and real bench mark data sets have shown the revision is more accurate and robust. The second improvement is to propose an economical numerical algorithm to facilitate the matrix inversion, which is a key step in hyperplane construction. The proposed step could greatly low the computational cost and is promising fast applications, such as on-line data mining.

References

1. Asuncion, A., Newman, D.: UCI machine learning repository (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
2. Batista, M.: A note on a generalization of sherman-morrison-woodbury formula. ArXiv e-prints (July 2008)
3. Brown, G.: A New Perspective for Information Theoretic Feature Selection. In: Twelfth International Conference on Artificial Intelligence and Statistics (2009), `http://jmlr.csail.mit.edu/proceedings/papers/v5/brown09a.html`
4. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. 3(2), 185–205 (2005), `http://view.ncbi.nlm.nih.gov/pubmed/15852500`
5. Duan, K.B.B., Rajapakse, J.C., Wang, H., Azuaje, F.: Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Transactions on Nanobioscience 4(3), 228–234 (2005), `http://view.ncbi.nlm.nih.gov/pubmed/16220686`
6. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley, Chichester (2001)
7. Forman, G.: An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3, 1289–1305 (2003)
8. Golub, G.H., Van Loan, C.H.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore (1996)
9. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422 (2002)
10. Guyon, I.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)
11. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. Ph.D. thesis, University of Waikato, Hamilton, New Zealand (1998)
12. Huang, C., Yang, D., Chuang, Y.: Application of wrapper approach and composite classifier to the stock trend prediction. Expert Systems with Applications 34(4), 2870–2878 (2008)
13. Koller, D., Sahami, M.: Toward optimal feature selection. In: Saitta, L. (ed.) Proceedings of the Thirteenth International Conference on Machine Learning (ICML), pp. 284–292. Morgan Kaufmann Publishers, San Francisco (1996)
14. Kwak, N., Choi, C.H.: Input feature selection by mutual information based on parzen window. IEEE Trans. Pattern Anal. Mach. Intell. 24, 1667–1671 (2002)
15. Liu, H., Setiono, R.: Feature selection via discretization. IEEE Transactions on Knowledge and Data Engineering 9, 642–645 (1997)
16. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 1226–1238 (2005)
17. Shannon, C.: A mathematical theory of communication. Bell Systems Technology Journal 27(3), 379–423 (1948)

18. Sun, Y.: Iterative relief for feature weighting: Algorithms, theories, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6), 1035–1051 (2007)
19. Tao, Y.: Kecman, Vojislav: Adaptive local hyperplane classification. Neurocomputing 71(13-15), 3001–3004 (2008)
20. Vincent, P., Bengio, Y.: K-local hyperplane and convex distance nearest neighbor algorithms. In: Advances in Neural Information Processing Systems, pp. 985–992. The MIT Press, Cambridge (2001)
21. Wettschereck, D., Aha, D.W., Mohri, T.: A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artificial Intelligence Review 11, 273–314 (1997)
22. Yang, T., Kecman, V., Cao, L., Zhang, C.: Testing adaptive local hyperplane for multi-class classification by double cross-validation. In: IEEE World Congress on Computational Intelligence (WCCI), pp. 1–5 (2010)