Proceedings of the Seventh International Conference on Machine Learning and Cybernetics,Kunming,12-15 July 2008

FEATURE TRANSFORMATION FOR EFFICIENTLY IMPROVING PERFORMANCE OF HSC

FU-ZHEN ZHUANG[1,2], QING HE[1], ZHONG-ZHI SHI[1]

[1]The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
[2]Graduate University of Chinese Academy of Sciences
E-MAIL: {zhuangfz, heq, shizz}@ics.ict.ac.cn

Abstract:

Hyper Surface Classification (HSC) is a novel classification method based on hyper surface which is put forward by Qing He, etc. Experiments show that HSC can efficiently and accurately classify large-size data in two dimensional space and three-dimensional space. Actually, it is difficult to deal with high dimensional data for HSC. So the dimension reduction (data rearrangement) and ensemble methods (feature subspace) are proposed for HSC. But the method based on ensemble will produce many inconsistent and repetitious data in some density dataset, which influence the classification ability of HSC. To solve the problem, a simple and effective kind of data feature transformation method for enhancing performance of HSC is proposed in this paper. The experimental results show that this method can efficiently reduce the inconsistent and repetitious data, efficiently utilize the data Information, and remarkably improve the classification performance of HSC.

Keywords:

Hyper Surface Classification; Ensemble; Classification Performance; Feature Transformation.

1. Introduction

Hyper Surface Classification (HSC) is a novel classification method based on hyper surface is put forward by He & Shi & Ren [1]-[4]. In this method, a model of hyper surface is obtained by adaptively dividing the sample space in the training process and then the hyper surface is directly used to classify large database according to whether the wind number is odd or even based on Jordan Curve Theorem in Topology. Some covering learning algorithms were proposed in the decade time. Bionic (Topological) Pattern Recognition theory is firstly proposed by Wang Shoujue as a new model for pattern recognition [5]. BPR is based on "matter cognition" instead of "matter classification", and so is thought closer to the function of human cognition than traditional statistical pattern recognition using "optimal separating" as its main principle. HSC, also a kind of cover algorithm, has the common goal of feature cognition as BPR, but the hyper surface is obtained by adaptively dividing the sample space in the training process and the classification method is based on the Jordan Curve Theorem. Experiments show that HSC can efficiently and accurately classify large-size data in two-dimensional space and three-dimensional space [1] [2]. Though HSC can classify higher dimensional data according to Jordan Curve Theorem on the theoretical plane, it is not as easy as in three-dimensional space. However, what we really need is an algorithm that can deal with data not only of massive size but also of high dimensionality. Thus in [3] a simple and effective kind of dimension reduction method without losing any essential information is proposed. Formally, the method is a dimension reduction method but in nature the method is a dimension transposition method. Moreover, we studied another different method based on ensemble in paper [4]. The most important difference between HSC ensemble and the traditional ensemble is that the sub-datasets are obtained by dividing the features rather than by dividing the sample set. Experiments show that this method has a preferable performance on high dimensional data sets, especially on those in which samples are different in each slice. For large density three-dimensional data up to 10^7, the speed of HSC is still very fast [2]. For some high dimensional density data set, the ensemble method in paper [4] will have to produce many inconsistent and repetitious data. A data feature transformation method is proposed to improve the classification performance of HSC for some high dimensional density data set.

Previous studies have shown that aggregating ensemble classifiers is often more accurate than a single classifier in the ensemble [6]-[7]. But the sub-datasets obtained by dividing the features often produces many inconsistent and repetitious samples, sometimes the number

978-1-4244-2096-4/08/$25.00 ©2008 IEEE

of available samples in the sub-datasets is much less than the size of original samples. These will lead to the poor ability of sub classifiers, and finally influence the performance of aggregating ensemble classifier. For this, feature transformation approach for improving the performance of HSC is proposed.

The rest of this paper is organized as follows: In section 2, we give an overview of Hyper Surface Classification (HSC). Then in Section 3 presents the motivation of feature transformation and transformation method. In Section 4, experimental results are presented, followed by our conclusions in Section 5.

2. Overview of the Classification Method Based on Hyper Surface

Hyper Surface Classification (HSC) is a universal classification method based on Jordan Curve Theorem in topology.

Jordan Curve Theorem. Let X be a closed set in n-dimensional space R^n. If X is homeomorphic to a sphere in $n-1$ dimensional space, then its complement $R^n \setminus X$ has two connected components, one called inside, the other called outside.

Classification Theorem. For any given point $x \in R^n \setminus X$, x is in the inside of $X \Leftrightarrow$ the wind number i.e. intersecting number between any radial from x and X is odd, and x is in the outside of $X \Leftrightarrow$ the intersecting number between any radial from x and X is even.

From the two theorems above, we conclude that X can be regarded as the classifier, which divides the space into two parts. And the classification process is very easy just by counting the intersecting number between a radial from the sample point and the classifier X. After knowing this, the very important problem is how to construct the separating hyper surface. In [1] [2], we have given the detailed training and testing steps.

Training Procedure

Step1. Input the training samples, containing k categories and d-dimensions. Let the training samples be distributed within a rectangular region.

Step2. Divide the region into $\overbrace{10\times10\times\cdots10}^{d}$ small regions called units.

*Step3***.** If there are some units containing samples form two or more different categories, then divide them into smaller units repeatedly until each unit covers at most samples from the same category.

Step4. Label each unit with $1,2,\cdots,k$ according to the category of the samples inside, and unite the adjacent units with the same labels into a bigger unit.

Step5. For each unit, save its contour as a link, and this represents a piece of hyper surface. All these pieces of hyper surface make the final separating hyper surface.

Testing Procedure

Step1. Input a testing sample and make a radial from it.

Step2. Input all the links that are obtained in the above training process.

Step3. Count the number of intersections between the radial and the first link. If the number is odd, then label the sample with the category of the link. If the number is even, go on to the next link.

Step4.If the number of intersection points between the radial and all the links is even, and then the sample becomes unrecognized.

Step5. Calculate the classification accuracy.

The classification algorithm based on hyper surface is a polynomial algorithm when the same class samples are distributed in finite connected components. Experiments show that HSC can efficiently and accurately classify large density data in two-dimensional or three-dimensional space for multi-classification. For some high dimensional density data set, the ensemble method in paper [4] will have to produce many inconsistent and repetitious data. In the following we will study a simple and efficient method to improve the performance of HSC.

3. Motivation for Feature Transformation and Transformation Technology

3.1. Motivation

In order to deal with data with high dimensionality, Hyper Surface Classifiers ensemble for high dimensional data sets is proposed in [4]. This approach based on the idea of ensemble and by attaching the same importance to each feature. Firstly they group the multiple features of the data to form some sub-datasets, then start a training process and generate a classifier for each sub-dataset, and the final decision is reached by integrating the series of classification results in the way of voting. Three features are grouped as a sub classifier in [4]. Because in this approach the sub-datasets are obtained by dividing the features rather than by dividing the sample set, so it requires that only on

the condition that there are not inconsistent samples, the size of each sub-dataset is equal to the original sample set. But actually in real data sets, the size of each sub-dataset is much smaller than the original sample set, especially for the high dimensional data sets. The smaller size of each sub-dataset, the more data information loses, and this will influence the performances of aggregating ensemble HSC classifiers. The inconsistency is that the samples which have same conditional attributes but different decisive attributes. Let us see an example following, there are ten original samples in Table.1. x_i Stands for the i-th sample, a_i stands for i-th attribute, d stands for the class or label of samples.

Table 1. Ten original samples

	a_1	a_2	a_3	a_4	a_5	a_6	$\cdots$	a_{16}	d
x_1	7	9	7	10	8	8	$\cdots$	6	*17*
x_2	2	3	3	1	1	5	$\cdots$	8	*3*
x_3	9	15	8	8	5	5	$\cdots$	9	*21*
x_4	2	3	3	2	1	6	$\cdots$	8	*12*
x_5	3	3	4	4	2	7	$\cdots$	8	*8*
x_6	2	10	3	8	1	14	$\cdots$	8	*10*
x_7	2	3	3	1	1	7	$\cdots$	10	*7*
x_8	5	6	5	4	3	5	$\cdots$	5	*25*
x_9	3	3	4	2	1	4	$\cdots$	7	*22*
x_{10}	*2*	*4*	*4*	*3*	*2*	*9*	$\cdots$	*8*	*4*

If we use the method directly in [4], there are only seven available samples in the first sub-dataset. Because x_2, x_4, x_7 are inconsistent samples and we can't use them for training, which directly leads to information loss, poor performance of sub classifiers and aggregating ensemble classifier.

3.2. Feature Transformation Method

How do we make good use of data information effectively and improve the performances of HSC learning algorithms? In order to solve these problems, a method of data feature transformation by making good use of data information effectively is proposed in this paper. In this approach, we dramatically decrease the number of inconsistent samples of each sub-dataset, which ensures the sizes of most sub-datasets close to the original sample set. We don't lose any data information among the transformation process, and the dimension of new sample is the same as the original sample.

Supposed there are n original samples. Firstly, we normalize the original samples, so that every attribute value of samples having the same range between zero and one. This step is very useful, which leverages the influence of every attribute. Secondly, a normalized sample has m attributes $x_j = (x_{j1}, x_{j2}, \cdots, x_{jm})$, then we denote the new sample after transformation as $x_j^{new} = (x_{j1}^{new}, x_{j2}^{new}, \cdots, x_{jm}^{new})$. The new attribute x_{ji}^{new} is described as,

$$x_{ji}^{new} = \frac{1}{m-2} \sum_{t=i}^{i+m-3} x_{jt} \qquad (1)$$

We can receive new attributes according to formula (1). If $t > m$, then $x_{jt} = x_{j(t \bmod m)}$, else $x_{jt} = x_{jt}$. Why use $m-2$ original attributes to construct a new attribute? One reason is that this will confirm the original data information being used fully and every original attribute being used the same frequency. After all new attributes are constructed, every original attribute is used $m-2$ times. The other reason is that we can recover the new samples back to the original ones, which also indicates no losing any information among the transformation process.

The purpose of normalization is to avoid the emergence of zero when we do sum operation. Sometimes the positive and negative attribute value may counteract each other and generate many zeros which also influence the performance of classifier. We transform the samples in Table.1, and receive the results showed in Table.2. In Table.2, we are excited to find that there have not any inconsistent samples in the first sub-dataset, as well as other sub-dataset.

Table 2. Ten samples after transformation

	a_1	a_2	a_3	a_4	a_5	a_6	$\cdots$	a_{16}	d
x_1	5.8	5.9	6.3	5.8	5.6	5.2	$\cdots$	5.5	*17*
x_2	3.2	3.7	3.9	4.4	4.4	4.4	$\cdots$	3.6	*3*
x_3	6.0	5.2	4.7	4.5	4.7	5.0	$\cdots$	5.8	*21*
x_4	1.8	2.0	2.2	2.7	2.6	2.6	$\cdots$	2.2	*12*
x_5	3.0	3.3	3.3	3.6	3.5	3.3	$\cdots$	3.2	*8*
x_6	3.9	4.2	3.7	4.2	3.6	4.1	$\cdots$	4.4	*10*
x_7	2.6	3.1	3.5	4.3	4.3	4.3	$\cdots$	3.2	*7*
x_8	3.7	4.1	4.2	3.9	4.0	3.9	$\cdots$	3.4	*25*
x_9	2.8	3.5	3.6	3.7	3.7	3.8	$\cdots$	2.9	*22*
x_{10}	*2.7*	*3.1*	*3.3*	*3.6*	*3.4*	*3.4*	$\cdots$	*2.9*	*4*

Problem. Why the method of data feature transformation can efficiently decrease the number of inconsistent samples?

Solution. An attribute is either a categorical or

numerical attribute; let's consider the categorical attribute first. Supposed attribute a_i has f_i values, before transforming the data set, the probability of two samples in sub-dataset being inconsistency is,

$$P_{before} = \frac{1}{f_i \times f_{i+1} \times f_{i+2}}, (i = 1, 2, \cdots, m-2) \quad (2)$$

But after the data set transformed, the probability of two samples in sub-dataset being inconsistency is,

$$P_{after} = \frac{1}{\prod_{i=1}^{m} f_i} \quad (3)$$

We can find that on the condition of high dimensional data, the probability of inconsistency after transformation is much smaller than the one before transformation, $P_{after} << P_{before}$. E.g provided there are ten distinct values for every conditional attribute in Table.1, then for the first sub data set $P_{after} = \frac{1}{\prod_{i=1}^{16} 10} << P_{before} = \frac{1}{10 \times 10 \times 10}$. Obviously it can dramatically decrease the number of inconsistent samples. For numerical attribute, the illustration is similar to the categorical one. The different is that we discretize the data first, and then the numerical attribute can be considered as an analogous categorical attribute.

4. Experiments and Analysis

4.1. Experimental Results

To testify the good performance of the data feature transformation, the following experiments are designed. We select the data set of Letter-recognition and spambase in UCI dataset for our experiments. Table.3 shows the description of the selected dataset.

Table3. The description of selected dataset

dataset	Size of Total dataset	Number of classes	Number of attributes	Size of selected sub-dataset
Letter-Recognition	20000	26	16	500
Spambase	4601	2	57	1000

We select two different sub-datasets from Letter-recognition, and two different sub-datasets from spambase. The size of four sub-datasets are 1000, 1000, 500 and 500 respectively. The sub-datasets are randomly selected from the original samples. HSC algorithm is used to train and test the sub-datasets before and after the sub-datasets transformed. Two sub-datasets from Letter-Recognition results are showed in Table.4. In Table.4, The second and fourth columns describe the number of available samples of each sub-dataset before and after transformation, and the third and fifth column describe the accuracy of corresponding sub-dataset before and after transformation. The row ensemble stands for the accuracy of aggregating ensemble classifiers before and after feature transformation.

Table4. Two sub-datasets from letter-recognition, each sub-dataset has 1000 samples

First sub-dataset				
	Un-Transformed		Transformed	
SubNet0	284	27.2%	1000	97.7%
SubNet1	436	41.8%	1000	98.5%
SubNet2	653	61.4%	1000	99.2%
SubNet3	631	58.2%	1000	99.8%
SubNet4	643	61%	1000	98.3%
SubNet5	557	52.5%	999	97.6%
Ensemble		*78.4%*		*99.4%*
Second sub-dataset				
	Un-Transformed		Transformed	
SubNet0	313	28.4%	999	97.4%
SubNet1	407	39.8%	1000	97.9%
SubNet2	640	60.5%	1000	97.9%
SubNet3	605	56.2%	1000	97.9%
SubNet4	618	60.9%	1000	98.2%
SubNet5	519	50.8%	999	97.7%
Ensemble		*76%*		*99.3%*

Because the size of conditional attributes of Spambase is large and page limited, the detailed results are not shown here. Instead, the brief results are shown in Table5.

Table5. Two sub-datasets from Spambase, each sub-dataset has 500 samples

First sub-dataset		
	Un-Transformed	Transformed
Ensemble	76.4%	99%
Second sub-dataset		
	Un-Transformed	Transformed
Ensemble	80%	99.6%

From the results, we can find that before the data set is transformed, there are many inconsistent and repetitious samples in the sub-datasets, leading to the size of available samples of each sub-dataset is much smaller than the size of original sample sets. In Table.4, there are only about twenty-eight percent of samples available in the worse case, compared to the one hundred percent available after transformation. The size of available samples in sub-dataset influences the performance of sub classifiers. Because of the poor performance of sub classifiers, the performance of ensemble classifier is poor too. But after the sub-dataset is

transformed, the aggregating ensemble classifier performs well. The accuracy is dramatically improved and up to ninety-nine percent. There are almost not inconsistent and repetitious samples, which is the reason for good performance and high accuracy. Though we haven't use the feature transformation method to any other classification algorithms, it is believed that the method will also work for other algorithms if the dataset is divided vertically.

4.2. Analysis

Figure.1 and Figure.2 show the visualized training Links of sub nets for the first sub-dataset from Letter-Recognition before and after feature transformation. In generally, we keep the diversity of sub classifiers to guarantee good performances of ensemble classifier, which also seen in [8] [9]. From the figures, it seems that after feature transformation, the training Links of sub nets are similar with each other and become more consistent, which violates the general regulation. Why the similar sub classifiers can improve the accuracy?

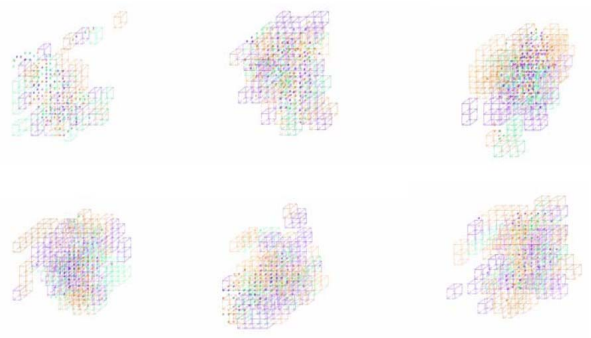

Figure 1. Links of subnets before transformation

Figure 2. Links of subnets after transformation

The intuitional interpretation is that the sub-datasets are obtained by dividing the features in the algorithm HSC rather than by dividing the sample set which usually used in other researched work. Many previous researchers have many significant contributions on the topic of relationship between diversity and ensemble accuracy, such as [9]-[11]. [9] Detailedly described the measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. In that paper, ten measures of diversity were studied, and experiments were designed to examine the relationship between the accuracy and measures of diversity. Experiment of feature subspace method had been done in that work; the accustomed diversity measures were also failure for improving the accuracy of ensemble classifier as our work.

Table 6. Two averaged pairwise measures [9] of diversity for the selected datasets

Letter-Recognition				
	First sub-dataset		Second sub-dataset	
	Un-Transformed	Transformed	Un-Transformed	Transformed
Dis	0.434	0.008	0.435	0.009
DF	0.261	0.005	0.277	0.006
Spambase				
	First sub-dataset		Second sub-dataset	
	Un-Transformed	Transformed	Un-Transformed	Transformed
Dis	0.121	0.048	0.108	0.042
DF	0.141	0.009	0.122	0.005

Conversely, it seems the more similar the sub-classifiers, the high accuracy received. We adopt two pairwise diversity measures (the disagreement *Dis* and the double fault *DF*) to measure the changes of diversity between the transformations. In Table.6, we're lucky to find that *Dis* and *DF* become smaller after transformed, which comply with the results mentioned in [9]. So the more consistent of sub classifiers the high accuracy received according to the measures *Dis* and *DF* . But we also can't obtain an explicit conclusion that how the diversity influence the ensemble accuracy. The use of diversity measures for enhancing the design of classifier ensembles is still an open question and interesting topic.

5. Conclusions

To improve the classification performance of HSC for some high dimensional density data set, a data feature transformation method is proposed in this paper. This method can efficiently decrease the inconsistent and repetitious data, and fully utilize the data information. Experimental results show that this simple method is able to

improve the performance of sub classifiers on the sub-datasets remarkably, which leads to the good performance of the aggregating ensemble classifier. Moreover, a interesting aspect of research is indicated from the results; how to measure the diversity of classifiers in feature subspace method and how to make good use of the diversity for enhancing ensemble accuracy are the future work.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 60435010, 60675010), 863 National High-Tech Program (No.2006AA01Z128), National Basic Research Priorities Programme (No.2007CB311004) and the Nature Science Foundation of Beijing (No. 4052025).

References

[1] Qing. He, Zhongzhi. Shi, Li'an. Ren. "The classification method based on hyper surface", Proceedings of the 2002 International Joint Conference on Neural Networks, pp.1499-1503, 2002.

[2] Qing. He, Zhongzhi. Shi, Li'an. Ren, E. S. Lee. "A Novel Classification Method Based on Hyper Surface", International Journal of Mathematical and Computer Modeling, pp. 395-407, 2003.

[3] Qing. He, Xiurong. Zhao, Zhongzhi. Shi. "Classification based on dimension transposition for high dimension data", SOFT COMPUTING 11(4) pp.329-334, 2007.

[4] Xiurong. Zhao, Qing. He, Zhongzhi. Shi. "Hyper Surface Classifiers Ensemble for High Dimensional Data Sets", In Wang et al. (Eds.): 3rd International Symposium on Neural Networks (ISNN 2006), LNCS 3971, pp. 1299 - 1304, 2006.

[5] Shoujie. Wang. "Bionic (Topological) Pattern Recognition-A New Model of Pattern Recognition Theory and Its Applications", ACTA ELECTRONICA SINICA 30(10), pp.1417-1420L, 2002.

[6] Larskai. Hansen, Peter. Salamon. "Neural Network Ensembles", EEE Transaction on Pattern Analysis and Machine Intelligence, 12(10), pp. 993-1001, 1990.

[7] Zhihua. Zhou, J. Wu, Y.Jiang, S.F. Chen. "Genetic Algorithm based Selective Neural Network Ensemble", International Joint Conference on Artificial Intelligence, pp. 797-802, 2001.

[8] Rushing, J., S. J. Graves, E. Criswell, and A. Lin. "A Coverage Based Ensemble Algorithm (CBEA) for Streaming Data," IEEE International Conference on Tools with Artificial Intelligence, Boca Raton, FL, 2004.

[9] Ludmila.I. Kuncheva, Christopher.J. Whitaker. "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy", Machine Learning, Springer Netherlands, 0885-6125 (Print) 1573-0565 (Online), 2003.

[10] Terry Windeatt. "Diversity measures for multiple classifier system analysis and design", Inf Fusion, 6(1), 2005, Special issue on Diversity in Multiple Classifier System.

[11] Shipp C.A., Ludmila.I. Kuncheva. "Relationships between combination methods and measures of diversity in combining classifiers", Inf Fusion, 3(2), 2002, 135-148