The Data Selection Criteria for HSC and SVM Algorithms

Qing He[1], Fuzhen Zhuang[1,2], Zhongzhi 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
{heq,zhuangfz,shizz}@ics.ict.ac.cn

Abstract

This paper makes a discussion of Consistent Subsets (CS) selection criteria for Hyper Surface Classification (HSC) and SVM algorithms. The consistent subsets play an important role in the data selection. Firstly, the paper proposes that Minimal Consistent Subset for a disjoint Cover set (MCSC) plays an important role in the data selection for HSC. The MCSC can be applied to select a representative subset from the original sample set for HSC. MCSC has the same classification model with the entire sample set and can totally reflect its classification ability. Secondly, the number of MCSC is calculated. Thirdly, by comparing the performance of HSC and SVM on corresponding CS, we argue that it is not reasonable that using the same train data set to train different classifiers and then testing the classifiers by the same test data set for different algorithms. The experiments show that algorithms can respectively select the proper data set for training, which ensures good performance and generalization ability. MCSC is the best selection for HSC, and support vector set is the effective selection for SVM.

1. Introduction

Hyper Surface Classification (HSC) is put forward by Qing He etc [12]. In this method, a model of hyper surface is obtained by adaptively dividing the samples 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. To tackle the problem of high accuracy computation demand of HSC for sparse data, it is necessary to find the approach for selecting a representative subset of the original training data, or generating a new prototype reference set from available samples. To tackle the same problem of Nearest Neighbor (NN) classification, Minimal Consistent Subset (MCS) is defined by Hart in 1968 [2]. A consistent subset of a sample set is a subset which correctly classifies all of the remaining points in the sample set. The Minimal Consistent Subset (MCS) is defined as consistent subset with a minimum number of samples. Hart pointed out that every set has a consistent subset, since every set is trivially a consistent subset of itself. Although every finite set has a minimal consistent subset, it is difficult to find the Minimal Consistent Subset, so he studied the "condensed nearest neighbor rule" (CNN). Hart's method indeed ensures consistency, but the condensed subset is not minimum size, and is sensitive to the randomly picked initial selection and the order of consideration of the input samples. There are many studies on MCS for NN such as "reduced nearest neighbor rule" of Gates [3], "iterative condensation algorithm" of Swonger [4] etc. The other related works can be found in [5]-[11]. These works aim to finding the sample subset which is the backstage manipulator of generalization ability of NN. To find the sample subset which plays an important role in the data selection for HSC, a judgmental sampling method called Minimal Consistent Subset for a disjoint Cover set (MCSC) is studied in this paper.

To emphasize the importance of MCSC, Bagging and Boosting are also used to enhance the performance of HSC [20], However, its behavior is subjected to the characteristics of MCSC. It indicates that it is very important to select data for training to learning a classifier. In this paper, we study how to find the MCSC for HSC, and the property of MCSC. For different classification algorithms, they all have their criterion to select data set for training, which comply to the consistent subset. In this paper, we also discuss the CS of algorithm SVM, and make a comparison between HSC and SVM algorithms.

This paper is organized as follows. In Section 2, we give an outline of Hyper Surface Classification (HSC). Then in Section 3 the concept and the construction of Minimal Consistent Subset for a disjoint Cover set (MCSC) of HSC are described. In section4, we give some experiments to show that the different algorithms having different data choice criteria for given data sets, followed by our conclusions in

978-0-7695-3304-9/08 $25.00 © 2008 IEEE
DOI 10.1109/ICNC.2008.334

Section 5.

2. Overview of Hyper Surface Classification Method

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 $R^n \backslash X$ has two connected components, one called inside, the other called outside.

Classification Theorem. For any given point $x \in R^n \backslash X$, 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 is in the outside of $X \Leftrightarrow$ the intersecting number between any radial from and 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 [12], 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 $\overbrace{10 \times 10 \times \cdots \times 10}(10^d)$ small regions called units.

Step3. If there are some units containing samples from two or more different categories, then divide them into smaller units recursive 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.

HSC tries to solve nonlinear multi-classification problems in the original space without having to map into higher dimensional spaces, using multiple pieces of hyper surface. It is polynomial in time complexity if samples with the same class distribute in finite connected components. Experiments show that HSC can efficiently and accurately classify large datasets in two-dimensional and three-dimensional space for multi-classification. For large three dimensional data up to 10^7, the speed of HSC is still very fast [17].

3. Minimal Consistent Subset for Disjoint Cover Set

To tackle the problem of high accuracy computation demand of HSC for sparse data, it is necessary to find the approach for selecting a representative subset of the original training data, or generate a new prototype reference set from available samples. Minimal Consistent Subset for Disjoint Cover Set (MCSC) is proposed to solve the problem.

Suppose C is the collection of all subsets for a finite sample set S. And C' is a disjoint cover set for S, i.e., a subset $C' \subseteq C$ such that each element in S belongs to one and only one member of C'. Minimal Consistent Subset for a disjoint Cover set (MCSC) C' is a sample subset combined by choosing one and only one sample from each element in the disjoint cover set C'.

For HSC method, we call sample a and b are equivalent if they belong to a same class and fall into a same unit. And the points falling into the same unit construct an equivalent class. The cover set C' is the union set of all equivalent classes in the hyper surface H. More specifically, let $\overline{H}$ be the interior of H and u is a unit in $\overline{H}$. Minimal Consistent Subset of HSC denoted by $S_{min}|_H$ is a sample subset combined by selecting one and only one representative sample from each unit included in the hyper surface, i.e.

$$S_{min|H} = \bigcup_{u \subseteq \overline{H}} \{\text{choosing one and only one } s \in u\} \quad (1)$$

Note: Selecting one and only one sample is to maintain the minimalist.

For a given sample set, we propose the following computation methods for its MCSC.

Step1. Input the samples, containing k classes and d-dimensions. Let the samples be distributed within a rectangle region.

Step2. Divide the region into $\overbrace{10 \times 10 \times \cdots \times 10}(10^d)$ small regions, called units.

Step3. If there're some units containing samples that belong to two or more different classes, then go to Step2 divide them into smaller units recursive until each unit covers at most one class of samples.

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

Step5. For each sample in the set, locate its position in the model, i.e. figure out which unit it locates in.

Step6. Combine samples that locate in the same unit into one equivalent class, then we get a number of equivalent classes.

Step7. Pick up one and only one sample from each unit or each equivalent class. Then the MCSC of HSC is obtained.

By the algorithm above, we justify Hart's statement that every set has a consistent subset, since every set is trivially a consistent subset of itself, and every finite set has a minimal consistent subset, although the minimum size is not, in general, achieved uniquely in [2]. For our method, the number of samples in each MCSC equals to the number of equivalent classes. And the number of MCSC equals to the size of Cartesian product of these equivalent classes. The method indeed ensures consistency and minimal for a given cover set. Moreover, it is not sensitive to the randomly picked initial selection and the order of consideration of the input samples.

We point out that some samples in the MCSC are replaceable, while others are not. As we can see from the process of dividing large regions into small units in the algorithm, some close samples in the same class may fall into the same unit. In that case, these samples are equivalent to each other in the building of the classifier, and we can randomly pick up one of them into the MCSC. However, sometimes there can be only one sample in a unit, and this sample plays a unique role in the forming of the hyper surface. So it is irreplaceable in the MCSC. To make the concept of MCSC based on HSC more clearly; the following two figures are listed.

We use the dataset of Breast-Cancer-Wisconsin from UCI repository, which contains 699 samples from 2 different classes. The dataset is firstly transformed into 3 dimensions by using the method in [19], and then trained by HSC. The trained model of hyper surface, composing of units in two layers, is shown in Figure1. Each unit may contain multiple samples that belong to the same class. Then we adopt the MCSC computation method to obtain the MCSC of this data set. The MCSC is also used for training, whose hyper surface structure is shown in Figure 2. The two figures are totally the same except different number of samples contained in some units. So it indicates that the selected data set—MCSC, is the best choice for HSC. Following we will study some properties of MCSC.

***Some properties of* MCSC**

LEMMA1. Given a data set $\mathcal{D}$, $\mathcal{S}_1 \subseteq \mathcal{D}$ is a MCSC, and if $\exists$ data set $\mathcal{S}_2 \subseteq \mathcal{D}$ and $\mathcal{S}_1 \subseteq \mathcal{S}_2$, then $\mathcal{S}_2$ is a consistent subset (CS).

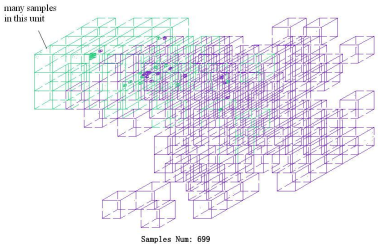

Figure 1. The Hyper Surface Structure of Breast- Cancer-Wisconsin

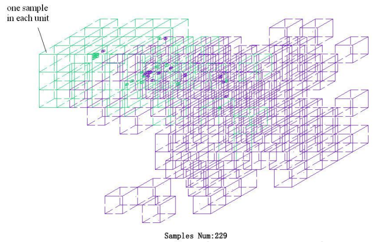

Figure 2. The Hyper Surface Structure of MCSC for Breast-Cancer-Wisconsin

PROOF. Provider $\mathcal{S}_2$ is not a CS, so for

$$\forall \mathcal{Q} \subseteq \mathcal{S}_2, \mathcal{Q} \text{ is not a CS},$$
$$\mathcal{S}_1 \subseteq \mathcal{S}_2, \mathcal{S}_1 \text{ is not a CS}.$$

But, $\mathcal{S}_1 \subseteq \mathcal{D}$ is a MCSC $\Rightarrow \mathcal{S}_1$ is CS, conflict.
Therefor, $\mathcal{S}_2$ is a CS.

LEMMA2. Given a data set $\mathcal{D}$, $\mathcal{S}_1 \subseteq \mathcal{D}$ is a MCSC, and if $\forall$ data set $\mathcal{S}_2 \subseteq \mathcal{D}$ and $\mathcal{S}_2 \subset \mathcal{S}_1$, then $\mathcal{S}_2$ is not a CS.
PROOF. Provider $\mathcal{S}_2$ is a CS, so

$$\exists \mathcal{Q}, \mathcal{M} \subset \mathcal{D}, \mathcal{Q} \cup \mathcal{S}_2 = \mathcal{S}_1, |\mathcal{S}_2| < |\mathcal{S}_1|,$$
$$\mathcal{M} \subset \mathcal{S}_2, \mathcal{M} \text{ is a MCSC}, |\mathcal{M}| \leq |\mathcal{S}_2|.$$

So, $|\mathcal{M}| < |\mathcal{S}_1|$, $\mathcal{S}_1$ is not a MCSC, conflict.
Therefor, $\mathcal{S}_2$ is not a CS.

LEMMA3. Given a data set $\mathcal{D}$, $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{D}$ ($\mathcal{S}_1 \neq \mathcal{S}_2$) are MCSC, then

$$\mathcal{Q} = \mathcal{S}_1 \cup \mathcal{S}_2, \mathcal{Q} \text{ is CS},$$
$$\mathcal{M} = \mathcal{S}_1 \cap \mathcal{S}_2, \mathcal{M} \text{ is not CS}.$$

The lemma can be proved similarly above, so it is omitted here.

For a given sample set, MCSC totally reflects its classification ability. In other words, any sample addition into the MCSC will not improve the classification ability (LEMMA1). This can be seen in Table 1, where we give the classification ability of MCSC on the data set of Breast-cancer-Wisconsin, Wine, Iris, Sonar and Wdbc. In Test I,

Table 1. The classification ability of Minimal Consistent Set

Data Set	No. Of Samples	No. Of Samples in MCSC	No. Of MCSC	Test I	Test II
Breast-Cancer-Wisconsin	699	299	28623793345289 20895091978160 1425489920000	100%	100%
Wine	178	129	2087354105856	100%	100%
Iris	150	81	369853055631360	100%	100%
Sonar	208	186	663552	100%	100%
Wdbc	569	268	390816363039941 009139983759816 964736000000000	100%	100%

Table 2. Single deletion from the MCSC of breast-cancer-wisconsin

No. of Samples in the same unit with the one deleted	ID of deleted sample	HSC Accuracy (%)	Bagging Accuracy (%)	AdaBoost Accuracy (%)
1	4	99.79	99.79	99.79
3	10	99.36	99.15	99.36
5	35	98.94	98.73	98.73
7	20	98.51	98.51	98.30
8	30	98.30	98.30	98.09
10	6	97.88	97.66	97.66
11	178	97.66	97.45	97.45
17	37	96.39	96.39	96.18
34	17	92.78	100	99.79
39	9	91.72	91.51	91.72
48	1	89.81	89.81	89.81
71	3	84.93	84.93	84.93
117	7	75.16	75.16	74.95

Table 3. Results for Waveform Data Set[†]

data set		HSC(%)	SVM(%)
MCSC for Training	Training	99.34	47.35
	Testing	99.37	62.34
SVs for Training	Training	99.42	47.6
	Testing	11.99	100
$\overline{\text{MCSC}}$ for Training	Training	100	100
	Testing	24.84	42.59
$\overline{\text{SVs}}$ for Training	Training	99.34	47.35
	Testing	99.37	62.34

[†] SVs stands for support vectors, $\overline{\text{MCSC}}$ stands for the complemental set of MCSC in the data set, $\overline{\text{SVs}}$ stands for the complemental set of SVs in the data set.

for a given data set, its MCSC is used for training and the other for testing. In Test II, ten samples are deleted from the testing set and added to the training set. We can see that after training, MCSC has the same hyper surface with the original data set, but contains much fewer samples. For any given data set, there are many different MCSC. Table 1 lists the numbers of MCSC, i.e. the size of Cartesian product of all equivalent classes.

The experimental results in Table 1 point out that any data set containing MCSC is consistent subset, on the other hand, if any samples included in the MCSC are deleted, the data set will not be consistent subset any more (LEMMA 1). The experimental results in Table 2 show that the algorithm performance is significantly influenced by the sample deleted from the MCSC. From Table 2, we also reemphasize the important of MCSC, which limits the behavior of Bagging and Boosting algorithms for enhancing HSC performance.

4. Different Data Choice Based on Different Algorithms

MCS is a universal concept. In this section, we will discuss the topic that for a given dataset, different algorithms having different choice of training data and testing data. We argue that it is not reasonable that comparing different classification algorithms by using the same train data set and the same test data. The reason is that different algorithms have different data selection criterion for the same given data set. So we can't simply test and evaluate performance another algorithms on the same data set. It is obviously that a given algorithm is not able to performance well on all different data sets, but they have the ability to select proper data set for training themselves. The following experiments are designed to confirm the opinions. We compare the performance of SVM^{lib}[18] and HSC algorithms on the data set selected deliberately. Default parameters for SVM are used except setting the gamma in kernel function as 1 and the parameter C of C-SVC, epsilon-SVR, and nu-SVR as 100. The waveform data set, which has 5000 samples, from UCI library is selected for the experiments.

In the first experiment, 4522 samples (MCSC of HSC) are selected for training and the rest $478(5000 - 4522)$ for testing. The data set is firstly preprocessed from high dimensional data to three dimensions by the method in [19]. The results of accuracy including training and testing are shown in Table 3. Conversely, we use the support vector set, which is a good choice for SVM, for training data. 4508 samples (support vectors), are selected for training and the rest $492(5000 - 4508)$ for testing. The results of accuracy also shown in Table 3.

From the results of above experiments, we can see that given a dataset, different algorithms have different data choice for the best accuracy. HSC can perform well on the training data set—MCSC deliberately selected for HSC, but SVM performs no so well. Contrarily, for the training data set—support vector set, deliberately selected for SVM, SVM performs much better than HSC. It is concluded that different algorithms have different data choice criterion. Moreover, we can find that the MCSC and support vector set are all consistent subsets, for they all can correctly classify the remaining samples by the corresponding algorithms. The separation hyper plane for SVM is determined by support vector set, does the support vector set form MCS? We can't be sure that, for there may exist some support vectors which are redundant to form a MCS.

But you can find MCS for SVM through support vector set. You can delete samples one bye one from support vector set and check the accuracy on remaining samples, until any one deleted will influence the accuracy on remaining samples.

To confirm the conclusion and make a comparison, another experiment is designed as follows. We exchange the training data set and testing data set. Results are shown in Table 3 too. From the table, we can find that the generalization ability of both HSC and SVM algorithms are all poor. The reason is that the training set selected is not representatives for the total samples space, rather than CS which can cover the samples space. The above experiments show that different algorithms have different data choice criterion, and they can perform well on the data set selected by themselves, otherwise they will have poor performance. Moreover, above experiments indicate how to choose data as training set for HSC and SVM.

5. Conclusions

Consistent Subsets (CS) selection criteria for Hyper Surface Classification (HSC) and SVM algorithms is discussed in this paper. We study the MCSC for HSC, which is a subset of the training samples space. MCSC can cover the training samples space, so it is a representative sample set for HSC. The MCSC can perform as well as the total training data set. It can also conclude that any data set containing MCSC is also consistent subset, and the samples deleted from MCSC will influence the performance. For a given data set, different algorithms always have different CS, and they can do well on the corresponding CS, but poor on other CS for other algorithms. It also indicates that algorithms have self-selection criterion for data set. MCSC is the best selection for HSC, and support vector set, though no MCS, is the effective selection for SVM and then applied to the remaining samples.

6. Acknowledgements

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

References

[1] B. V. Dasarathy. Nearest neighbor (nn) norms: Nn pattern classification techniques. *os Alamitos, CA: IEEE Computer Society Press*, 1991.

[2] P. E. Hart. The condensed nearest neighbor rule. *IEEE Trans. In formation Theory*, pages 515–516, 1968.

[3] G. W. Gates. The reduced nearest neighbor rule. *IEEE Trans. In formation Theory*, pages 431–433, 1972.

[4] C. W. Swonger. Sample set condensation for a condensed nearest neighbor decision rule for pattern recognition. *in: S. Watanade (Ed.), Frontiers of Pattern Recognition, Academic Press, New York*, pages 511–519, 1972.

[5] C. L. Chang. Finding prototypes for nearest neighbor classifiers. *IEEE Trans. Computers*, pages 1179–1184, 1974.

[6] P. A. Deliver and J. Kittler. On the edited nearest neighbor rule. In *Proc. 5th ICPR, Miami, Florida*, pages 72–80, 1980.

[7] B. V. Dasarathy. Minimal consistent set (mcs) identification for optimal nearest neighbor decision systems design. *IEEE Trans Syst .man Cybern*, pages 511–517, 1994.

[8] L. I. Kuncheva. Fitness functions in editing knn reference set by genetic algorithms. *Pattern Recognition*, pages 1041–1049, 1997.

[9] L. I. kuncheva and L. C. Bezdek. Nearest prototype classification: clustering, genetic algorithms, or random search. *IEEE Trans. Syst. Man and Cybern*, pages 160–164, 1998.

[10] V. Cerveron and A. Fuertes. Parallel random search and tabu search for the minimal consistent subset selection problem. *Lecture Notes in Computer Science*, pages 248–259, 1998.

[11] H. B. Zhang and G. Y. Sun. Optimal reference subset selection for nearest neighbor classification by tabu search. *Pattern Recognition*, pages 1481–1490, 2002.

[12] Q. He, Z. Z. Shi, and L. A. Ren. The classification method based on hyper surface. In *Proceedings of the 2002 International Joint Conference on Neural Networks*, pages 1499–1503, 2002.

[13] K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, pages 993–1001, 1990.

[14] Z. H. Zhou, J. Wu, Y. Jiang, and S. F. Chen. Genetic algorithm based selective neural network ensemble. In *International Joint Conference on Artificial Intelligence*, pages 797–802, 2001.

[15] L. Breiman. Bagging perdictiors. In *Mach. Learn.*, pages 123–140, 1996.

[16] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proc. 13th Int. Conf.: Mach. Learn., Bari, Italy*, pages 148–156, 1996.

[17] Q. He, Z. Z. Shi, L. A. Ren, and E. S. Lee. A novel classification method based on hyper surface. In *International Journal of Mathematical and Computer Modeling*, pages 395–407, 2003.

[18] Available at Chih-Jen Lin's Home Page, http://www.csie.ntu.edu.tw/Taiwan.

[19] Q. He, X. R. Zhao, and Z. Z. Shi. Classification based on dimension transposition for high dimension data. *SOFT COMPUTING*, pages 329–334, 2007.

[20] Q. He, F. Z. Zhuang, X. R. Zhao, and Z. Z. Shi. Enhanced algorithm performance for classification based on hyper surface using bagging and adaboost. In *International Conference on Machine Learning and Cybernetics*, pages 3624–3629, 2007.