MINIMAL CONSISTENT SUBSET FOR HYPER SURFACE CLASSIFICATION METHOD

QING HE*, XIU-RONG ZHAO† and ZHONG-ZHI SHI‡

The Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, 100080, China

Graduate University of Chinese Academy of Sciences
Beijing, 100039, China

*heq@ics.ict.ac.cn
†zhaoxr@ics.ict.ac.cn
‡shizz@ics.ict.ac.cn

Hyper Surface Classification (HSC), which is based on Jordan Curve Theorem in Topology, has proven to be a simple and effective method for classifying a larger database in our previous work. To select a representative subset from the original sample set, the Minimal Consistent Subset (MCS) of HSC is studied in this paper. For HSC method, one of the most important features of MCS is that it has the same classification model as the entire sample dataset, and can totally reflect its classification ability. From this point of view, MCS is the best way of sampling from the original dataset for HSC. Furthermore, because of the minimum property of MCS, every single deletion or multiple deletions from it will lead to a reduction in generalization ability, which can be exactly predicted by the proposed formula in this paper.

Keywords: Hyper Surface Classification (HSC); Minimal Consistent Subset (MCS); sampling; generalization ability.

1. Introduction

Classification has always been an important problem in machine learning and data mining. And in recent years, covering classification algorithms has been developed considerably. In Ref. 15, the covering method is applied to perform the partition of data in the transformed space. And Bionic Pattern Recognition (BPR), first proposed by Wang as a new model for pattern recognition,13 is also actually a kind of covering algorithm. BPR is based on “matter cognition” instead of “matter classification”, and is thought to be closer to the function of human cognition than traditional statistical pattern recognition using “optimal separating” as its main principle. Based on Jordan Curve Theorem, a new covering algorithm of hyper surface classification is put forward in Refs. 7 and 8. In this method, a model of hyper surface is obtained in the training process and then the hyper surface is
directly used to classify a large database according to whether the wind number is odd or even based on Jordan Curve Theorem. Experiments show that HSC can efficiently and accurately classify large-sized data in two-dimensional space and three-dimensional space. Though HSC can classify any higher dimensional data according to Jordan Curve Theorem in theory, it is not as easy to realize HSC in higher dimensional space 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 Ref. 9, a simple and effective kind of dimension reduction method without losing any essential information is proposed, which is a dimension changing method in nature. Based on the idea of ensemble, another solution to the problem of HSC on high dimensional data is proposed in Ref. 16. By attaching the same importance to each feature, firstly we 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 determination is reached by integrating the series of classification results by way of voting. Experimental results show these two solutions are both suitable for HSC on high dimensional datasets, with both achieving good results.

On the other hand, sampling plays a very important role in all classification methods. Sampling methods are classified to either probability or nonprobability. In probability sampling, each member of the population has a known nonzero probability of being selected. In nonprobability sampling, members are selected from the population in some nonrandom manner. The advantage of probability sampling is that sampling error can be calculated. Sampling error is the degree to which a sample might differ from the population. When referring to the population, results are reported plus or minus the sampling error. In nonprobability sampling, the degree to which the sample differs from the population remains unknown. Judgment sampling is a common nonprobability method. Researchers select the sample based on judgment. This is usually an extension of convenience sampling. This method intends to figure out the entire sample from some “representative” sample, even though the population includes all samples.

For example, to tackle the problem of high computational demands of nearest neighbor (NN), many efforts have been made for selecting a representative subset of the original training data. The very early study of this kind was probably the “condensed nearest neighbor rule” (CNN) presented by Hart in 1968. A consistent subset of a sample set is a subset which, when used as a stored reference set for the NN rule, correctly classifies all of the remaining points in the sample set. And the Minimal Consistent Subset is defined as consistent subset with a minimum number of elements. He also pointed out that every set has a consistent subset, since every set is trivially a consistent subset of itself. Also, every finite set has a minimal consistent subset, although the minimum size is not, in general, achieved uniquely. Hart’s method indeed ensures consistency, but the condensed subset is not minimal, and is sensitive to the randomly picked initial selection and to the order of consideration of the input samples. After that, a lot of
work has been done to reduce the size of the condensed subset, as is shown in
Refs. 1–5, 10, 11 and 14.

In this paper, the notion of Minimal Consistent Subset is applied to the Hyper
Surface Classification Method in order to enhance HSC performance and analyze
its generalization ability.

The rest of the paper is organized as follows. In Sec. 2, the outline of our
previous work is described, including the HSC method and its two solutions to high
dimensional datasets. In Sec. 3, the notion and computation method of Minimal
Consistent Subset of HSC is given, followed by its important characteristics, which
are justified by the experimental results in Sec. 4. Finally, our conclusions are stated
in Sec. 5.

2. Overview of Previous Work

In this section, we will describe the outline of previous work, on which the Minimal
Consistent Subset is based.

2.1. Kernel of HSC 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 \( \mathbb{R}^n \). If
\( X \) is homeomorphic to a sphere in \( n - 1 \) dimensional space, then its complement
\( \mathbb{R}^n \setminus X \) has two connected components, one called inside, the other called outside.

**Classification Theorem.** For any given point \( x \in \mathbb{R}^n \setminus X \), \( x \) 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 Ref. 8, we have given the detailed
training and testing steps.

**Training Procedure**

Step 1. Input the training samples, containing \( k \) categories and \( d \)-dimensions. Let
the training samples be distributed within a rectangular region.

Step 2. Divide the region into \( 10 \times 10 \times \cdots \times 10 \) small regions called units.

Step 3. If there are some units containing samples from two or more different cate-
gories, then divide them into smaller units repeatedly until each unit covers
at most samples from the same category.
Step 4. Label each unit with $1, 2, \ldots, k$, according to the category of the samples inside, and unite the adjacent units with the same labels into a bigger unit.

Step 5. 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**

Step 1. Input a testing sample and make a radial from it.

Step 2. Input all the links that are obtained in the above training process.

Step 3. 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.

Step 4. If the number of intersection points between the radial and all the links is even, then the sample becomes unrecognized.

In a word, HSC tries to solve the nonlinear multiclassification problem in the original space without having to map into higher dimensional spaces, using multiple pieces of hyper surface.

### 2.2. Properties of HSC method

#### 2.2.1. High efficiency and accuracy

HSC method is a polynomial algorithm if samples with the same categories are distributed in finite connected components. Experiments show that HSC can efficiently and accurately classify large dataset in two- and three-dimensional space for multiclassification. For three-dimensional data up to the size of $10^7$, it still runs with high speed and accuracy.\(^8\) The reason is that the time of saving and extracting hyper surface are both very short, and also the decision process is very easy by using Jordan Curve Theorem.

#### 2.2.2. Ability of generalization

The experiment of training on small scale of samples and testing on dense large scale shows that HSC has a strong ability of generalization.\(^8\) Moreover, we can see that the continuity of the hyper surface improves as the number of samples increases. In the region where samples are sparsely distributed, a big unit is required, while in the region where samples are densely distributed, a small unit is required, the local elaborate division is an important strategy, which improves the generalization ability and accuracy for dense data. However, HSC is not so good for sparse data, and this is one of the motivations to study MCS in this paper.

#### 2.2.3. Robustness

Though the data noise cannot be completely clear, it can be controlled in a local region. If a noised sample is located inside a hyper surface, the hyper surface will
change into a complex hyper surface. In this case the classification theorem is still
effective. The noise may make mistakes in classification, but the influence is con-
trolled in a local small unit.

2.2.4. Independent of sample distribution

In fact, HSC can solve the nonlinear classification problem regardless of sample
distribution in a finite region. The samples can be distributed in any way. It does
not matter whether they are distributed in the shape of interlock or crisscross. On
the contrary, some other classification methods require that the samples should
reflect the feature of data distribution.

2.3. Problems with HSC on high dimensional datasets

From the view point of theory, HSC can deal with datasets of any high dimensions
according to Jordan Curve Theorem. But in practice, it is not as easy to realize
HSC in higher dimensional space as in three-dimensional space, and there exist some
problems in both time and space in doing this directly. First of all, the number of
training samples needed to design a classifier grows as the dimension increases.
Second, it is obvious that the data structure in higher dimensional space will be
much more complex than in lower dimensional space. Moreover, it takes much more
time to unite adjacent regions with the same category to obtain a piece of separating
hyper surface.

However, what we really need is an algorithm that can deal with data not only of
massive size but also of high dimensionality. To solve this problem, we have proposed
two different methods in Refs. 9 and 16, both having achieved good results.

2.4. Solution I: dimensionality reduction

The basic idea of this method proposed in Ref. 9 is dimension reduction, i.e. trans-
forming high dimensional data into three. This simple and effective method rear-
ranges all of the numerals from higher dimensional data to lower ones, without
changing their values, but only changing their positions according to some order.
So it is a dimension reduction method formally, but it is naturally a dimension
transposition process, without losing any essential information.

2.5. Solution II: classifiers ensemble

Based on the idea of Ensemble, another approach for HSC on high dimensional
data is presented in Ref. 16. By attaching the same importance to each feature,
firstly we 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
determination is reached by integrating the series of classification results by way of
voting. 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, so in the case of no inconsistency, the size of each sub-dataset is equal to the original sample set, totally occupying a little more storage space than the original sample set. Experiments show that this method has a preferable performance on high dimensional data sets, especially when samples are different in each slice.

3. Minimal Consistent Subset of HSC

3.1. Concept and computation method of MCS

To select a representative subset of the original training data, or generating a new prototype reference set from available instances for NN, the notion of Minimal Consistent Subset (MCS) was first given by Hart. A consistent subset of a sample set is a subset which, when used as a stored reference set for the NN rule, correctly classifies all of the remaining points in the sample set. And the Minimal Consistent Subset is defined as consistent subset with a minimum number of elements. The concept can be extended to HSC, and we define the Minimal Consistent Subset of HSC as follows.

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 every element in $S$ belongs to one and only one member of $C'$. Minimal Consistent Subset (MCS) for a disjoint cover set $C'$ is a sample subset combined by choosing one sample and only one sample from each element in the disjoint cover set $C'$. For HSC method, we call sample $a$ and $b$ equivalent if they are with the same category and fall into the same unit. And the points falling into the same unit make an equivalent class. The cover set $C'$ is the union set of all equivalent classes in the hyper surface $H$. More specifically, let $H$ be the interior of $H$ and $u$ is a unit in $H$. 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.

$$S_{\min | H} = \bigcup_{u \subseteq H} \{\text{choosing one and only one } s \in u\}.$$

For a given sample set, we propose the following computation methods for its Minimal Consistent Subset.

Step 1. Input the samples, containing $k$ categories and $d$-dimensions. Let the samples be distributed within a rectangular region.

Step 2. Divide the region into $10 \times 10 \times \cdots 10$ small regions called units.

Step 3. If there are some units containing samples from two or more different categories, then divide them into smaller units repeatedly until each unit covers at most samples from the same category.

Step 4. Label each unit with $1, 2, \ldots, k$, according to the category of the samples inside, and unite the adjacent units with the same labels into a bigger unit.
Step 5. For each sample in the set, locate its position in the model, which means to figure out which unit it is located in.

Step 6. Combine samples that are located in the same unit into one equivalent class, then we get a number of equivalent classes in different layers.

Step 7. Pick up one sample and only one sample from each equivalent class to form the Minimal Consistent Subset of HSC.

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 Ref. 6. For our method, the number of samples in each Minimal Consistent Subset equals to the number of equivalent classes. And the number of Minimal Consistent Subsets equals to the size of Cartesian product of these equivalent classes. The method indeed ensures consistency and is minimal for a given sample set and HSC classifier. Moreover, it is not sensitive to the randomly picked initial selection and to the order of consideration of the input samples.

We point out that some samples in the MCS 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 within the same category 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 one of them for the MCS. However, sometimes there can be only one sample in a unit, and this sample plays a unique role in forming the hyper surface. So it is irreplaceable in the MCS.

3.2. Important features of MCS in HSC

(i) For a specific sample set, the Minimal Consistent Subset totally reflects its classification ability

From the definition of MCS and its computation steps, we can see that the model trained from MCS can correctly classify all the remaining points in the sample set. And as we know from previous work, the recall rate of HSC is 100%; therefore the model trained from MCS can also correctly classify itself. As a result, if we use the MCS for training, the model can classify the entire sample set correctly, so does the entire sample set. Actually, as can be seen from experiments in Sec. 4, the MCS has the same hyper surface with the entire sample set. Moreover, even if we add some instances into the MCS, the classification ability remains the same. So we say that the MCS totally reflects the classification ability of the original sample set.

(ii) Every single deletion from MCS will lead to failure in testing accuracy, which can be exactly predicted

Generally speaking, because of the minimum property of MCS, when we delete some samples from it, the leftover cannot correctly classify the sample set, and the testing accuracy will fall down. It is interesting to determine how much loss in the
consistency property results from an incomplete set. What is more important about HSC, we can predict the accuracy exactly.

Suppose there are \( N \) samples in a data set, and its MCS contains \( n \) samples. If the MCS is used for training and the other samples for testing, the accuracy will be 100%. When one sample is deleted from the training set and added to the testing set, the accuracy will drop to \( 1 - \frac{m}{N - n + 1} \), where \( m \) represents the number of samples that fall into the same unit with one deleted. In general, if \( K (1 \leq K \leq n) \) samples are deleted from the Minimal Consistent Subset, the accuracy will reduce to \( 1 - \frac{m_1 + m_2 + \cdots + m_K}{N - n + K} \).

(iii) *MCS is the best way of sampling for HSC*

As we know, sampling plays a very important role in all classification methods. Different ways of sampling can lead to different generalization ability. As a common kind of nonprobability sampling method, judgment sampling is one in which the researcher selects the sample based on judgment. For example, a researcher may decide to draw the entire sample from some “representative” house, even though the population includes all houses. When using this method, the researcher must be confident that these chosen samples are truly representative of the entire population.

Because MCS totally reflects the classification ability of the original sample set, it is very encouraging to select it as the representative subset. In that sense, the computation process for MCS can be regarded as a judgment sampling for the most representative examples.

(iv) *MCS is an extension of PAC learning theory in HSC*

PAC (Probably Approximately Correct) learning is a framework of learning that was proposed by Valiant in his paper.\textsuperscript{12} He gave a nice formalism for deciding how much data you need to collect in order for a given classifier to achieve a given probability of correct predictions on a given fraction of future test data.

As MCS totally reflects the classification ability of the original sample set, when we wish to learn a concept from the sample set, its MCS will be competent for this job. While satisfying the PAC learning theory, MCS provides us a tangible subset for learning from the original space.

4. Experiments and Discussions

First of all, to make the concept of Minimal Consistent Subset base on HSC more clear and vivid, the following two figures are listed.

We use the dataset of breast-cancer-Wisconsin from UCI repository, which contains 699 samples from two different categories. The dataset is firstly transformed into three dimensions by using the method in Ref. 9, and then trained by HSC. The trained model of hyper surface, composing of units in two layers, is shown in Fig. 1. Each unit may contain multiple samples from the same category. Then we adopt the MCS computation method mentioned in Sec. 3.1 to obtain the MCS of this
Fig. 1. The hyper surface structure of breast-cancer-Wisconsin.

Fig. 2. The hyper surface structure of minimal consistent subset for breast-cancer-Wisconsin.

data set. The MCS is also used for training, whose hyper surface structure is shown in Fig. 2.

From the two figures above, we can see that the hyper surface structures between the original sample set and its Minimal Consistent Subset are totally the same. The only difference between these two figures is the different number of samples contained in some units. No matter which we choose for training, either the original data set or its MCS, we get the same hyper surface.

For a specific sample set, the Minimal Consistent Subset totally reflects its classification ability. Any addition into the MCS will not improve the classification
Table 1. The classification ability of MCS.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Sample No.</th>
<th>MCS Sample No.</th>
<th>Test I</th>
<th>Test II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast-cancer-Wisconsin</td>
<td>699</td>
<td>229</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Wine</td>
<td>178</td>
<td>129</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Ten Spiral</td>
<td>33750</td>
<td>7285</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

ability, either. This can be seen from Table 1, in which MCS 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.

Note that the dataset of Ten Spiral, containing 33,750 samples from ten categories in three-dimensional space, and the hyper surface obtained by its MCS is shown in Fig. 3.

Furthermore, because of the minimum property of MCS, when we delete some samples from it, the leftover cannot correctly classify the sample set, and the testing accuracy will reduce. The relationship between loss of consistency and deleted samples has been given as a formula in Sec. 3.2.

In the following Table 2, the dataset of breast-cancer-Wisconsin is used as an example to test the formula in the case of single deletion. Applied to this dataset, the formula becomes $1 - m/(699 - 229 + 1)$, where $m$ represents the number of samples that fall into the same unit with one deleted. In Table 3, we test the formula in the case of multiple deletions. And it is $1 - (m_1 + m_2 + \cdots + m_K)/(699 - 229 + K)$.

From this table, we can see the accuracy obtained from experiments is totally consistent with that obtained by the formula, which means that when a sample is deleted from MCS, we can predict the testing accuracy exactly by the formula proposed in Sec. 3.2. This is the same with multiple deletions, which is proven by Table 3.

For single deletion from MCS, the more representative the deleted sample, the more loss there will be in accuracy. This can be concluded from both theory and
Table 2. Single deletion from MCS of breast-cancer-Wisconsin.

<table>
<thead>
<tr>
<th>ID of Deleted Sample</th>
<th>Samples in the Same ID of Deleted Unit with the One Deleted</th>
<th>Accuracy by Experiment</th>
<th>Accuracy by Formula</th>
<th>No. of the Same Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1</td>
<td>99.79%</td>
<td>99.79%</td>
<td>155</td>
</tr>
<tr>
<td>26</td>
<td>2</td>
<td>99.58%</td>
<td>99.58%</td>
<td>39</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>99.36%</td>
<td>99.36%</td>
<td>11</td>
</tr>
<tr>
<td>27</td>
<td>4</td>
<td>99.15%</td>
<td>99.15%</td>
<td>6</td>
</tr>
<tr>
<td>35</td>
<td>5</td>
<td>98.94%</td>
<td>98.94%</td>
<td>3</td>
</tr>
<tr>
<td>43</td>
<td>6</td>
<td>98.73%</td>
<td>98.73%</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>98.51%</td>
<td>98.51%</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>8</td>
<td>98.30%</td>
<td>98.30%</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>97.88%</td>
<td>97.88%</td>
<td>1</td>
</tr>
<tr>
<td>178</td>
<td>11</td>
<td>97.66%</td>
<td>97.66%</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>17</td>
<td>96.30%</td>
<td>96.30%</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>34</td>
<td>92.78%</td>
<td>92.78%</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>39</td>
<td>91.72%</td>
<td>91.72%</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>48</td>
<td>89.81%</td>
<td>89.81%</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>71</td>
<td>84.93%</td>
<td>84.93%</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>117</td>
<td>75.16%</td>
<td>75.16%</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3. Multiple deletions from MCS of breast-cancer-Wisconsin.

<table>
<thead>
<tr>
<th>Description</th>
<th>Accuracy by Prediction</th>
<th>Accuracy by Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 2, m = {1, 2}$</td>
<td>99.36%</td>
<td>99.36%</td>
</tr>
<tr>
<td>$k = 5, m = {1, 2, 3, 4, 5}$</td>
<td>96.84%</td>
<td>96.84%</td>
</tr>
<tr>
<td>$k = 10, m = {1, 2, 3, 4, 5, 6, 7, 8, 10, 11}$</td>
<td>88.13%</td>
<td>88.13%</td>
</tr>
</tbody>
</table>

experiments. For example, in Table 2, there are 117 samples in the same unit with the seventh sample, and only 10 with the sixth sample. So the seventh sample has more representative ability than the sixth. As we can be seen, when the seventh sample is deleted, the accuracy drops a lot more than when deleting the sixth one. This trend can be clearly seen in Fig. 4.

![Fig. 4. Representative ability versus accuracy.](image)
Another important feature of Minimal Consistent Subset is that for HSC method, it is the best way to sample from the original dataset. However, it is very difficult to obtain a MCS by using probability sampling method, because the probability of sampling for MCS is very small.

Take the dataset of breast-cancer-Wisconsin in Table 2, for example, the number of Minimal Consistent Subsets equals to the size of Cartesian product of all equivalent classes, i.e.

\[
1^{115} \times 2^{39} \times 3^{11} \times 4^{6} \times 5^{3} \times 6^{4} \times 7^{1} \times 8^{2} \times 10^{1} \times 11^{1} \\
\times 17^{1} \times 34^{1} \times 39^{1} \times 48^{1} \times 71^{1} \times 117^{1} \\
= 2862379334528950919781601425489920000 \\
\approx 2.86 \times 10^{40}.
\]

By using probability sampling method, the probability of sampling for MCS is

\[
2.86 \times 10^{40} / C_{699}^{229} = 2.86 \times 10^{40} / 1.822 \times 10^{270} \\
= 1.57 \times 10^{-230}.
\]

And that is almost impossible. That explains why the nonprobability judgment sampling method is used to obtain the MCS for HSC.

5. Conclusions

To select a representative subset from the original sample set, the Minimal Consistent Subset (MCS) of HSC is studied in this paper. The concept of MCS for HSC and its computation method is given. What is more, several important features of MCS for HSC are discussed, which are justified by the experimental results. The Minimal Consistent Subset totally reflects the classification ability of the entire original sample set, and every single deletion from MCS will lead to failure in testing accuracy, which can be exactly predicted by our formula. MCS is the best way of sampling for HSC method and it is an extension of PAC learning into HSC. Finally, we point out that Minimal Consistent Subset is a universal concept, but the features may be very different with different classification methods.

Acknowledgments

This work is supported by the National Science Foundation of China (60675010, 60435010 and 90604017), 863 National High-Tech Program (2006AA01Z128), National Basic Research Priorities Programme (2007CB311004) and the Nature Science Foundation of Beijing (4052025).

References

Qing He is an Associate Professor at the Institute of Computing Technology, Chinese Academy of Sciences (CAS), and he is a Professor at the Graduate University of Chinese Academy of Sciences (GUCAS). He received the B.S. degree from Hebei Normal University, Shijiazhuang, P. R. C., in 1985, and the M.S. degree from Zhengzhou University, Zhengzhou, P. R. C., in 1987, both in mathematics. He received the Ph.D. degree in 2000 from Beijing Normal University in fuzzy mathematics and artificial intelligence, Beijing, P. R. C. Since 1987 to 1997, he has been with Hebei University of Science and Technology.

He is currently a doctoral tutor at the Institute of Computing and Technology, CAS.

His interests include data mining, machine learning, classification, fuzzy clustering.

Zhong-Zhi Shi is a Professor at the Institute of Computing Technology, the Chinese Academy of Sciences, leading the Research Group of Intelligent Science.

He is a senior member of IEEE, member of AAAI and ACM, Chair for the WG 12.2 of IFIP. He serves as Vice President for Chinese Association of Artificial Intelligence, Executive President of Chinese Neural Network Council.

His research interests include intelligence science, multi-agent systems, semantic Web, machine learning and neural computing.

Xiu-Rong Zhao received the B.S. degree from the Department of Computer Science and Technology at Shandong University in 2004, and the M.S. degree from Institute of Computing Technology, Chinese Academy of Sciences in 2007.

Her research interests include machine learning and data mining.