分类超曲面算法复杂度研究

何 清[1),2)] 赵卫中[1),2)] 史忠植[1)]

[1)](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

[2)](中国科学院研究生院 北京 100039)

摘 要 分类超曲面算法是一种简单的基于覆盖的分类算法.实验证明该算法具有分类正确率高、速度快的优点.但是,关于该算法的相关理论问题需要深入研究.文中对该算法的几个相关理论问题进行了研究.首先给出并证明了在分割的最大层数给定时算法假设空间的 VC 维,在此基础上结合可能近似正确(Probably Approximately Correct,PAC)学习框架,得出了对算法样本复杂度的估计,使得分类超曲面算法保证可 PAC 学习到任意目标概念.其次,分析了算法的时间复杂度和空间复杂度.最后,给出了无矛盾样本集的概念,并证明当输入样本集是有限无矛盾样本集的条件下,算法一定是收敛的.

关键词 分类超曲面算法;VC 维;PAC 可学习性;样本复杂度

中图法分类号 TP18 **DOI 号**: 10.3724/SP.J.1016.2010.00666

Study on Complexity in Hyper Surface Classification

HE Qing[1),2)] ZHAO Wei-Zhong[1),2)] SHI Zhong-Zhi[1)]

[1)](*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[2)](*Graduate University of Chinese Academy of Sciences, Beijing* 100039)

Abstract Hyper Surface Classification (HSC) is a simple covering based classification algorithm. Experiments show that HSC can efficiently and accurately classify large-sized data in two-dimensional space and three-dimensional space. However, little research has been done on theoretical problems in HSC. This paper studies several theoretical problems in HSC. First, the paper shows that given the biggest dividing level l, the VC dimension of the hypothesis space is d^{2l}. Under the PAC theory, it arrives at the conclusion on sample complexity, the algorithm probably learn a hypothesis that is approximately correct. Then, the time complexity and space complexity are analyzed. Finally, sample set without contradiction is defined, and show that if the inputting sample set is a finite sample set without contradiction, the algorithm must be convergent.

Keywords hyper surface classification; VC dimension; PAC learnability; sample complexity

1 引 言

分类算法研究是机器学习的核心研究内容,分类能力是人类智能的最显著特征之一.基于拓扑学中的 Jordan 曲线定理,何清等[1-3]提出了一种通用的覆盖型分类方法——分类超曲面算法.该方法直接在原空间解决非线性分类问题,通过区域细化与区域合并获得由多个超平面组成的双侧闭曲面作为分类超曲面对空间进行划分,通过计算样本点关于

收稿日期:2008-05-07;最终修改稿收到日期:2009-04-09.本课题得到国家自然科学基金(60675010,60933004,60975039)、"八六三"高技术研究发展计划项目基金(2007AA01Z132)和国家"九七三"重点基础研究发展规划项目基金(2007CB311004)和国家科技支撑计划(2006BAC08B06)资助.**何 清**,男,1965 年生,研究员,博士生导师,主要研究领域为模糊数学、机器学习、人工智能.**赵卫中**,男,1981 年生,博士研究生,主要研究方向为机器学习、数据挖掘.E-mail: zhaowz@ ics.ict.ac.cn.**史忠植**,男,1941 年生,研究员,博士生导师,主要研究领域为人工智能、机器学习、多主体系统.

分类超曲面的围绕数的奇偶性即可简单方便地判定其类别. 实验表明，在二维、三维及高维空间中，分类超曲面算法分类正确率高、速度快；而存储量和计算量都很小，从而能够处理海量数据. 各种分类算法共有的局限性是：每种算法的推广性都依赖于测试数据分布与训练数据分布是否一致. 但如何获得数据分布，如何划分可以使测试数据分布与训练数据分布一致是一个难以解决的重要问题. 在分类超曲面算法方面，文献[4]通过对极小样本集的研究找到了答案，解决了为什么极小样本集会影响分类准确率的问题，还指出了极小样本集有多少种表达方式. 但是，对于该算法的相关理论问题还需要深入研究. 在本文中，我们研究了分类超曲面算法的几个相关理论问题，包括 VC 维、样本复杂度、时间复杂度、空间复杂度以及收敛性等，使分类超曲面算法的理论更加完整.

本文第 2 节介绍与本文相关的工作，包括 VC 维、PAC 框架的相关概念以及分类超曲面算法的简略描述；第 3 节分析分类超曲面算法的相关理论问题，其中首先证明算法假设空间的 VC 维，在此基础上结合 PAC 框架求得算法的样本复杂度；接着分析算法的时间复杂度和空间复杂度；最后给出算法收敛的条件；第 4 节对我们的工作进行总结.

2　相关工作

2.1　VC 维的概念

VC 维是假设空间复杂度的度量标准. 为了描述这一概念，首先引入对一个实例集合打散(Shattering)的概念[5]. 对于假设空间 H、样本集 X 以及 X 的某个子集 S，H 中的每个假设 h 导致 S 的某个划分，即 h 将 S 划分为两个子集 $\{x\in S|h(x)=1\}$ 以及 $\{x\in S|h(x)=0\}$. 当 S 的每个可能的划分都可由 H 中的某个假设来表达时，我们就称 H 打散 S.

如果一个实例集合没有被假设空间打散，那么必然存在某个概念(划分)可被定义在实例集之上，但不能由假设空间表示. 因此，H 的这种打散实例集合的能力是其表示这些实例上定义的目标概念的能力.

定义在实例集 X 上的假设空间 H 的 VC 维是可被 H 打散的 X 的最大有限子集的大小，记作 $VC(H)$[5]. 如果 X 的任意大的子集可被 H 打散，则 $VC(H)\equiv\infty$.

2.2　PAC 学习框架及样本复杂度

为了描述学习算法 L 输出的假设 h 对真实目标概念 c 的逼近程度，我们引入假设 h 对于目标概念 c 和实例分布 D 的真实错误率. 假设样本集 X 中样本按照概率分布 D 随机产生. 对于 X 中任意样本 x 及其目标值 $c(x)$ 提供给学习算法 L，学习算法 L 输出的假设 h 的真实错误率为把 h 应用到将来按分布 D 抽取的实例时期望的错误率，用 $error_D(h)$ 表示. 即 $error_D(h)\equiv\Pr_{x\in D}[c(x)\neq h(x)]$，其中符号 $\Pr_{x\in D}$ 表示在实例分布 D 的概率.

如果一个假设空间是 H 的机器学习算法 L 满足如下条件，我们就称 L 是可 PAC 学习的：给定任意的实数 ε,δ 满足 $0<\varepsilon<1$ 以及 $0<\delta<1$，如果存在一个正整数 $m_0=m_0(\delta,\varepsilon)$，对于任意的目标概念 $c\in H$ 和样本分布 D，当样本数 $m\geq m_0$ 时，学习算法 L 将以至少 $1-\delta$ 的概率输出一个假设 $h\in H$，使 $error_D(h)\leq\varepsilon$[6].

在上面的定义中，m_0 是一个学习算法以较高的概率收敛到目标概念所需样本数目的边界，称为学习算法的样本复杂度，即样本复杂度是一个学习算法所需样本数目以确保在误差 ε 范围内以至少 $1-\delta$ 的概率学习到目标概念.

2.3　分类超曲面算法概述

分类超曲面算法的理论基础是拓扑学中的 Jordan 曲线定理.

Jordan 曲线定理. 设 $X\subset R^3$ 是闭子集，X 同胚于球面 S^2，那么它的余集 $R^3\backslash X$ 有两个连通分支，一个是有界的，另一个是无界的，X 中任何一点的任何邻域与这两个连通分支均相交.

Jordan 曲线定理可以推广到高维空间.

高维空间中的 Jordan 曲线定理. 设 $X\subset S^n$ 同胚于球面 $X\subset S^n$，那么 $m\leq n$，否则 $X=S^n$. 若 $m<n$，余集的同调群为

$$H_k(S^n\backslash X)\cong\begin{cases}Z\oplus Z, & m=n-1\text{ 且 }k=0\\ Z, & m<n-1\text{ 且 }k=0\\ 0, & \text{其它}\end{cases}.$$

特别地，当 $m=n-1$ 时，$S^n\backslash X$ 由两个连通分支组成；当 $m<n-1$ 时，只有一个连通分支.

Jordan 曲线定理表明：任何由 $n-1$ 维球面经连续变形得到的双侧闭曲面都把 n 维空间分成两个区域——一个外部和一个内部，这种曲面可用于分类，我们称之为分类超曲面. 则有下面的分类定理.

分类定理. 任取 $x\in R^n\backslash X$，则 $x\in X$ 的内部$\Leftrightarrow$自 x 引出的射线与 X 的相交数(即 X 关于 x 的围

绕数)为奇数;$x\in X$ 的外部⇔自 x 引出的射线与 X 的相交数为偶数.

基于以上定理,我们可以把与球面同胚的双侧闭曲面作为分类超曲面对空间进行划分.分类超曲面可以由多个超平面构成复合超曲面,而点属于超曲面内部还是外部取决于该点引出的射线与超曲面相交数为奇数还是偶数.文献[2]中给出了详细的训练和预测步骤.

训练步骤.

1. 输入训练样本集,并使样本集分布在超立方体样本空间中.

2. 把该超立方体均匀地分割为大小相等的更小的超立方体区域,每个区域称为单元格.

3. 对于每个单元格,如果其中包含的样本属于同一类别,则转步 4;如果单元格中包含属于不同类别的样本,则把该单元格作为一个超立方体,跳转到步 2(进行局部细化操作).

4. 对于其中只包含一种类别样本的单元格,标记该单元格的类别为其中样本的类别,并将其边界存储在一个链表中.

5. 合并相邻且类别相同的单元格,标记合并后区域的类别为原单元格的类别,存储合并后区域的边界作为一个曲面片,所有的曲面片以链表形式存储,形成最终的分类超曲面.

预测步骤

1. 输入待测试的样本,并从它向外引一条射线.

2. 输入训练步骤中得到的分类超曲面.

3. 计算射线与曲面片交点的个数.如果射线与某种类别的曲面片交点的个数是奇数,则标记该样本类别为该类曲面片的类别;如果射线与所有类别的曲面片交点的个数都是偶数,则该样本类别不能确定.

4. 跳转步 1 预测下一个样本的类别.

分类超曲面算法不需要考虑选择使用何种核函数,不需要作升维变换,不需要找出支持向量,它通过单元格合并计算获得多个超平面组成的双侧闭曲面作为分类超曲面对空间划分,使得基于非凸的超曲面的分类判别变得直接、简便、易行.

3 分类超曲面算法相关理论分析

在本节中,我们对分类超曲面算法的几个相关理论问题进行了研究.首先,我们给出分类超曲面算法假设空间的 VC 维,并在 PAC 学习框架下得出样本复杂度的结论;接着我们分析了分类超曲面算法的时间复杂度和空间复杂度,最后我们讨论了该算法的收敛性和分割最大层数的问题.其中每一部分都给出了严格的证明过程.

为了叙述方便,我们以二维样本集为例进行分类超曲面算法的理论分析,所得结论可以直接推广到多维样本集的情况.

3.1 VC 维及样本复杂度

在构造分类超曲面时,假设把样本空间或单元格分割为 $d\times d$(把每一维分成 d 等份)个区域.这样分割的结果是一棵正则 d^2 叉树,树中每个中间结点都有 d^2 棵子树.我们定义层的概念:原样本空间对应正则 d^2 叉树的根结点,记作第 0 层;第 $i(i\geq 0)$ 层的区域经过局部细化步骤后所分成的区域所在层为第 $(i+1)$ 层.

每次分割后所得分类超曲面可抽象为假设 h,由分类定理,对于任意的样本 x,如果 x 在闭合的正类别曲面片内,则 $h(x)=1$;反之,如果 x 在闭合的负类别曲面片内,则 $h(x)=0$.

分类超曲面算法的假设空间记作 H.为了控制算法的泛化能力,我们限定分割的最大层数为 l,即把样本空间最多分割为 l 层.如果 l 层结点中的样本不属于同一类别将不再进行局部细化操作,按照多数原则,把该区域类别标记为其中所包含的多数样本的类别.则有下面的定理.

定理 1. 分类超曲面算法假设空间 H 的 VC 维 $VC(H)=d^{2l}$.

证明. 要证明结论,需要两步:(1)证明 H 可以打散某个元素个数为 d^{2l} 的集合;(2)证明任意元素个数为 $(d^{2l}+1)$ 的集合都不能被 H 打散.

第 1 步. 把样本空间均匀地分割为 $d^l\times d^l$ 个区域,等价于把样本空间分割为一棵 l 层的满 d^2 叉树,每个区域对应该树的第 l 层的某个叶结点.假设在每个这样的区域中各有一个样本,则共有 d^{2l} 个样本,将这 d^{2l} 个样本组成一个集合,记作 S.设 $\{S_1,S_2\}$ 是 S 的任意一个可能的划分,下面只需证明分类超曲面算法能够输出一个假设(即分类超曲面)与划分 $\{S_1,S_2\}$ 相对应,即可证明假设空间 H 可以打散样本集合 S.假设 S_1 中的样本都赋予正类别,S_2 中的样本都赋予负类别,将样本集 S 作为输入样本集,则分类超曲面算法至多将样本空间分割为 l 层(满足最大层数的限制)即可使得每个单元格中的样本属于同一类别.经过合并单元格操作后,最终所得的分类超曲面记为 h,则由分类定理,h 将 S_1 中的样本都判定为正样本,将 S_2 中的样本都判定为

负样本. 所以 h 与 S 的划分$\{S_1,S_2\}$相对应. 由以上分析可得，H 可以打散 S，所以 $VC(H)\geq|S|=d^{2l}$.

第 2 步. 设 T 是任意的含有$(d^{2l}+1)$个样本的集合. 把样本空间均匀地分割为 $d^l\times d^l$ 个区域，根据鸽巢原理[7]，则至少有一个区域，其中含有至少两个样本，不妨记作 x_1,x_2. 假设$\{T_1,T_2\}$是 T 的一个划分，满足 $x_1\in T_1,x_2\in T_2$，对于其它样本不加限制，只需要满足属于且仅属于 T_1 和 T_2 中的一个集合. 将 T 作为输入样本集，由于限制分割的最大层数为 l，则无论为 T 中样本赋予何种类别组合(每个样本可以任意赋予正类别或负类别)，应用分类超曲面算法后，x_1 和 x_2 必定在同一个单元格中，进行合并单元格操作后，二者必然在同一个区域中. 根据分类定理，最终输出的假设 h 必定将 x_1 和 x_2 同时判定为正样本或同时判定为负样本，即与假设 h 对应的划分中 x_1 和 x_2 必定属于同一个子集，所以假设 h 与划分$\{T_1,T_2\}$不一致. 由样本集 T 中样本类别组合的任意性可得，假设空间 H 不能打散 T. 又因为 T 是任意的含有$(d^{2l}+1)$个样本的集合，所以 H 不能打散任意一个含有$(d^{2l}+1)$个样本的集合，所以 $VC(H)<d^{2l}+1$.

综上所述，分类超曲面算法假设空间 H 的 VC 维 $VC(H)=d^{2l}$.　　证毕.

Blumer 等[8]给出了一个保证可能近似学习到任意目标概念所需的样本数目 m 的边界为

$$m\geq\frac{1}{\varepsilon}\left(4\log_2\left(\frac{2}{\delta}\right)+8VC(H)\log_2\left(\frac{13}{\varepsilon}\right)\right)\quad(1)$$

在 PAC 学习框架下，应用定理 1 中的结论和式(1)，我们可以得到分类超曲面算法的样本复杂度.

推论 1.　在分类超曲面算法中，每次把单元格分割为 $d\times d$ 个区域，并且限制分割的最大层数为 l，则当样本数 m 满足如下条件时，分类超曲面算法能够以概率$(1-\delta)$输出一个真实错误率小于 ε 的假设 h(即分类超曲面)：

$$m\geq\frac{1}{\varepsilon}\left(4\log_2\left(\frac{2}{\delta}\right)+8d^{2l}\log_2\left(\frac{13}{\varepsilon}\right)\right).$$

证明.　将定理 1 中的结果代入式(1)即可证明结论.

从推论 1 的结论中可以看到，算法的样本复杂度以 $1/\delta$ 的对数增长，以 $1/\varepsilon$ 对数乘以线性增长. 并且样本复杂度以参数 d 的幂次增长，以参数 l 的指数次增长. 所以可以选取合理的分割数 d 和尽可能较小的层数 l 值以使用较小数目的样本集能够准确高效地学习到目标概念.

3.2　时间复杂度和空间复杂度分析

假设输入样本集含有 n 个样本，每次进行分割和局部细化操作时，把单元格分割为 $d\times d$ 个区域，同时限制分割的最大层数为 l.

因为算法的训练步骤是构造分类超曲面的过程，预测步骤是计算从待预测样本出发的射线与训练步骤中所求的分类超曲面的相交数，所以预测步骤的时间复杂度不会超过训练步骤的时间复杂度. 故我们只需要分析训练步骤的时间复杂度.

在第 1 步中，输入训练样本集并令它们分布在样本空间内，时间复杂度为 $O(n)$.

在接下来的步骤中进行分割和局部细化的操作，由于限制最大层数为 l，所以在最坏情况下，样本空间被分割成一棵 l 层的满 d^2 叉树，树中全部结点的个数为 $1+d^2+d^{2\times2}+\cdots+d^{2\times l}=\frac{d^{2l+2}-1}{d^2-1}$，时间复杂度为 $O\left(\frac{d^{2l+2}-1}{d^2-1}\right)$，即为 $O(d^{2l})$.

构造分类超曲面的操作只是在叶结点中进行，所以时间复杂度不会超过 $O(d^{2l})$.

综上所述，基于超曲面的分类算法的时间复杂度为 $O(n+d^{2l})$. 从分析可以看出，分类超曲面算法的时间复杂度与样本点规模 n 和两个预先设定的参数 d、l 有关. 所以对于大规模样本集，我们可以设定合理的参数 d 和 l，从而能够保证较高的时间效率和分类性能.

关于空间复杂度，在分类超曲面的算法中，所需要的辅助存储空间包括划分后的正则 d^2 叉树中的结点和分类超曲面. 由时间复杂度中的分析，在最坏情况下，树中结点个数的复杂度为 $O(d^{2l})$，并且每个结点需要常数个存储单元；存储曲面片的边界只在叶结点中进行，并且每个超曲面片的边界也只需要常数个存储单元，所以总的空间复杂度仍然为 $O(d^{2l})$.

3.3　收敛性及分割的最大层数

为了证明分类超曲面算法的收敛性，我们先给出无矛盾样本集的定义.

定义 1.　对于样本集 $S=\{(x_1,y_1),\cdots,(x_n,y_n)\}$，其中 $x_i\in R^2,y_i\in\{0,1\},i=1,2,\cdots,n$，如果对于样本集 S 中任意的两个样本$(x_i,y_i),(x_j,y_j)$，$1\leq i<j\leq n$，都不出现 $x_i=x_j,y_i\neq y_j$ 即两个样本的

属性值分别对应相等,但是类别不同,则我们就称样本集 S 是无矛盾样本集.

因为分类超曲面算法是一个构造性的算法,对样本空间和单元格分割和局部细化操作进行完毕,相应的分类超曲面也随之构造完成,所以我们有如下结论.

定理 2. 对于有限无矛盾样本集,分类超曲面算法一定是收敛的.

证明. 由无矛盾样本集的定义,对于有限无矛盾样本集,不同类别的任意两个样本之间的距离大于零.又因为有限样本集的样本空间是有限的,所以分类超曲面算法经过有限步对样本空间和单元格分割和局部细化操作后,可使每个区域中的样本都属于同一类别,经过合并单元格的操作后,可以得到相应的分类超曲面,即算法收敛,结论得证. 证毕.

接下来,我们讨论根据样本分布确定分割的最大层数.所谓分割的最大层数,就是对样本空间分割后区域的最大层号.为了得到分类准确率高的分类超曲面,在划分后每个区域中的样本都属于同一类别.我们要确定满足这一要求的最小层号.

定理 3. 设全部样本点分布在 $L\times L$ 的二维样本空间内,不同类别的任意两个样本之间的最小距离为 $D_{\min}$,分割的最大层号为 l,则 $l=\left\lfloor \log_d \frac{L}{D_{\min}} \right\rfloor+1$.

证明. 由已知条件,要使划分层数不超过 l 层,并且使每个区域中的样本都属于同一类别,则必须使第 l 层区域的大小即第 l 层单元格边长小于 $D_{\min}$,这样能够保证不同类别的样本经过分割分布在不同的区域中.

经过等距均匀的分割,第 l 层区域的边长为 $\frac{L}{d^l}$,要使满足要求,则有 $\frac{L}{d^l}<D_{\min}$,经过等价变换可得 $l>\log\frac{L}{D_{\min}}$,即 $l=\left\lfloor \log_d \frac{L}{D_{\min}} \right\rfloor+1$,结论得证. 证毕.

从定理 3 的结论中可以看出,分割的最大层数 l 的取值是受参数 d 的值限制的.对于相同的样本集,当参数 d 的值减小时,l 的值增大;当参数 d 值增大时,l 的值减小.又由 3.1 节和 3.2 节的分析,算法的样本复杂度、时间复杂度、空间复杂度都以 d 的幂次增长,以 l 的指数次增长,所以在应用分类超曲面算法时,可以适当地选用较大的 d 值,从而能够选用小的 l 值,使得在保证分类准确率的同时,算法的时空效率也能得到提高.

4 结 论

在本文中,我们对分类超曲面算法的几个相关理论问题进行了研究.首先我们证明了在分割的最大层数为 l 的情况下,算法假设空间的 VC 维 $VC(H)=d^{2l}$.由得到的 VC 维并结合 Blumer 等的结论,我们给出了该算法的样本复杂度:当样本数目 $m\geq\frac{1}{\varepsilon}\left(4\log_2\left(\frac{2}{\delta}\right)+8d^{2l}\log_2\left(\frac{13}{\varepsilon}\right)\right)$ 时,分类超曲面算法保证可能近似学习到任意的目标概念.接着,我们分析了时间复杂度和空间复杂度,并得到结论:时间复杂度为 $O(n+d^{2l})$;空间复杂度为 $O(d^{2l})$.最后我们给出了无矛盾样本集的概念,并证明了当样本集是有限无矛盾样本集时,分类超曲面算法一定是收敛的.同时我们还分析了应用算法时参数 d 和 l 值的选取问题,得到结论:在保证分类准确率的同时,可以适当地选取较大的 d 值,从而能够选用较小的 l 值,使得算法时空效率能够得到提高.

参 考 文 献

[1] He Qing, Ren Li-An, Shi Zhong-Zhi. The large data direct classifying method based on hyper surface. Chinese Journal of Computers, 2003, 26(2): 206-211(in Chinese)
(何清,任力安,史忠植. 基于超曲面的海量数据直接分类法. 计算机学报, 2003, 26(2): 206-211)

[2] He Qing, Shi Zhong-Zhi, Ren Li-An, Lee E S. A novel classification method based on HyperSurface. International Journal of Mathematical and Computer Modeling, 2003, 38(3-4): 395-407

[3] He Qing, Shi Zhong-Zhi, Ren Li-An. The classification method based on hyper surface//Proceedings of the IEEE International Joint Conference on Neural Networks. Hawaii, 2002: 1499-1503

[4] He Qing, Zhao Xiu-Rong, Shi Zhong-Zhi. Minimal consistent subset for hyper surface classification method. International Journal of Pattern Recognition and Artificial Intelligence, 2008, 22(1): 95-108

[5] Mitchell Tom M. Machine Learning. New York: McGraw Hill, 1997

[6] Anthony Martin, Biggs Norman. Computational Learning Theory: An Introduction. Cambridge, England: Cambridge University Press, 1992

[7] Brualdi Richard A. Introductory Combinatorics. Upper Saddle River, NJ: Prentice Hall, 1999

[8] Blumer A, Ehrenfeucht A, Haussler D, Warmuth M. Learnability and the Vapnik-Chervonenkis dimension. Journal of the ACM, 1989, 36(4): 929-965

HE Qing, born in 1965, professor, Ph. D. supervisor. His research interests include fuzzy mathematics, machine learning and artificial intelligence.

ZHAO Wei-Zhong, born in 1981, Ph. D. candidate. His research interests include machine learning and data mining.

SHI Zhong-Zhi, born in 1941, professor, Ph. D. supervisor. His main research interests include artificial intelligence, machine learning and multi-agent system.

Background

This paper studies classification problem that belongs to the machine learning category. Hyper Surface Classification (HSC) is a simple covering based classification algorithm. The paper studies several theoretical problems in HSC. First, it shows that given the biggest dividing level l, the VC dimension of the hypothesis space is d^{2l}. Under the PAC theory, it arrives at the conclusion on sample complexity, the algorithm probably learn a hypothesis that is approximately correct. Then, the time complexity and space complexity are analyzed. Finally, sample set without contradiction is defined, and show that if the inputting sample set is finite sample set without contradiction, the algorithm must be convergent.

Hyper Surface Classification (HSC) is a novel classification method based on hyper surface is put forward by He, Shi & Ren (2002). However, what we really need is an algorithm that can deal with data not only of massive size but also of high dimensionality. Thus He, Zhao & Shi proposed a simple and effective kind of dimension reduction method without losing any essential information in 2006. A judgment sampling method based on Minimal Consistent Subset (MCS) is proposed to select of a representative subset of the original training data.

This work is supported by the National Natural Science Foundation of China (Nos. 60675010, 60933004, 60975039), the National High Technology Research and Development Program (863 Program) of China (No. 2007AA01Z132), National Basic Research Program (973 Program) of China (No. 2007CB311004) and National Science and Technology Support Plan (No. 2006BAC08B06).