中国科学院
CHINESE ACADEMY OF SCIENCES

基于超曲面的认知学习方法

何 清
heqing@ict.ac.cn
中科院计算技术研究所

中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences

内容提要

- 问题提出
- 基于超曲面的分类方法
- 极小样本集与抽样分布发现算法
- 基于超曲面的并行分类算法
- 基于超曲面的聚类算法
- 基于超曲面的离群点检测方法

中国科学院

Chinese Academy of Sciences

基于超曲面的高维数据分类

- 算法实现复杂度随维数增加而急剧增长
  需要一种高维数据的无损处理方法
- 思路：降维、换维、无损变换
  - 理论上，根据Jordan曲线定理，HSC算法可以对任意维数的数据进行分类。但在具体实现时，高维的HSC算法实现起来遇到困难。
  - 现有的两种解决方案：换维降维方法、分维集成(HSC Ensemble)方法

高维降维方法

- 目标： 高维转化为三维
- 方法
  - 把数据归一化到区间[0,1]
  - 采用主成分分析方法把各个维度重新排列
  - 3维向量的第一个分量的小数部段由原向量的第1、4、7…分量组成，依次提取这些分量小数部分的第1位、第2位、……；第二个分量由原向量的第2、5、8…分量组成；第三个分量由原向量的第3、6、9…分量组成。

中国科学院

Chinese Academy of Sciences

高维降维方法

- 维排序
- 归一化
- 紧致化
- 标准化

$$
\begin{aligned}
&(0.x_{1,1}^{(i)}x_{1,2}^{(i)}x_{1,3}^{(i)}\cdots x_{1,k}^{(i)}\cdots x_{1,l}^{(i)},\\
&0.x_{2,1}^{(i)}x_{2,2}^{(i)}x_{2,3}^{(i)}\cdots x_{2,k}^{(i)}\cdots x_{2,l}^{(i)},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&0.x_{j,1}^{(i)}x_{j,2}^{(i)}x_{j,3}^{(i)}\cdots x_{j,k}^{(i)}\cdots x_{j,l}^{(i)},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&0.x_{d,1}^{(i)}x_{d,2}^{(i)}x_{d,3}^{(i)}\cdots x_{d,j}^{(i)}\cdots x_{d,l}^{(i)})
\end{aligned}
$$

Example. Let $\mathbf{X}^{(i)}$ is a 9 dimensional vector (0.126,0.197,0.820,0.956,0,376,0.269,0.867,0.921,0.882) then $\tilde{\mathbf{X}}^{(i)}$ is (0.198256667,0.139972761,0.828268092).

Chinese Academy of Sciences

Institute of Computing Technology,Chinese Academy of Sciences

高维降维方法

- 特点
  - 降维后的3维向量小数点后的每一个数字都来自于原始高维向量。
  - 反过来，原始高维向量中的每一个数字都出现在降维后的3维向量中
  - 一种无损降维方法

中国科学院

Chinese Academy of Sciences

高维降维方法

- 实验结果

表 3.1 高维降维方法性能测试

数据集	维度	样本数	训练样本数	测试样本数	召回率	准确率
Iris	4	150	100	50	100.00%	96.00%
Wine	13	178	128	50	100.00%	90.00%
Wdbc	30	569	369	200	100.00%	95.50%
Sonar	60	208	158	50	100.00%	90.00%

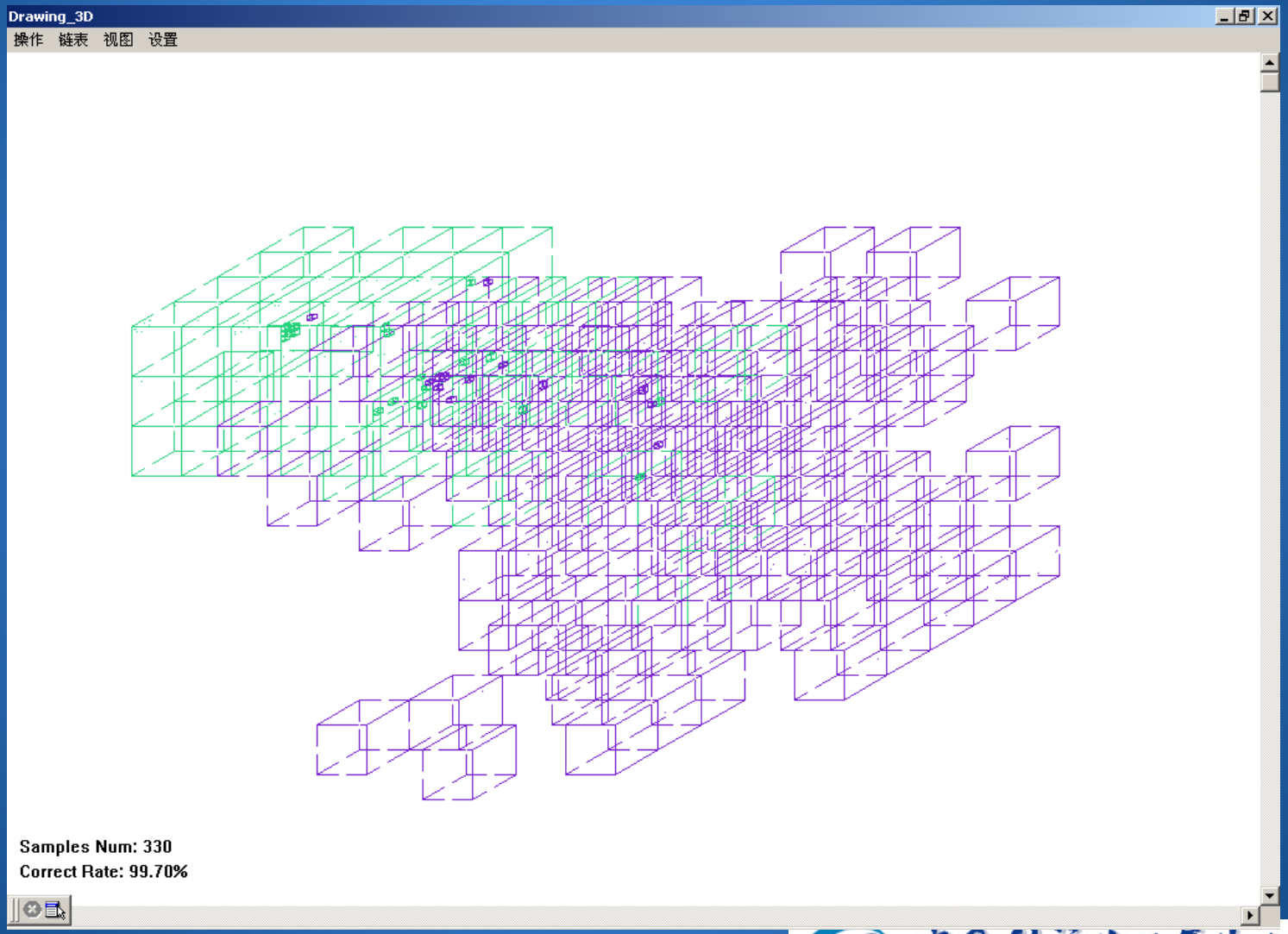
高维降维方法

- 实验结果对比

表 3.2 数据集 Breast-Cancer-Wisconsin 分类精度比较

文章	方法	分类精度
[Nauck 99]	NEFCLASS	96.50%
[Setiono 00]	NN	98.01%
[Setiono 00]	NeuroRule	98.21%
[Pena-Reyes 01]	Fuzzy CoCo	98.25%
[Abonyi 03]	DTFC	96.82%
[Borgulya 04]	FCE	97.80%
本文	HSC-DTR	99.70%

对高维数据换维分类



高维划分集成方法

- HSC处理高维数据的第二种方案
- 思路：划分－集成(Ensemble)
- 基本步骤
  - 划分－按属性划分为多个三维数据集，删除矛盾数据
  - 训练－调用HSC三维训练算法，得到多个子分类器
  - 集成－各子分类器按相对多数的原则投票

中国科学院

Chinese Academy of Sciences

高维划分集成方法

- 训练过程
  - 读入训练样本集，并记录其条件属性个数d；
  - 把训练样本集按属性划分为$\lceil d/3 \rceil$个子集，每个子集包含三个条件属性和决策属性。其中第i个子集包含属性$3i-2$，$3i-1$和$3i$，$i=1,2,3,\cdots,\lceil d/3 \rceil$。若$d$不能被3整除，则最后一个子集缺少的属性由前一个子集补足；
  - 经划分后，在每个子集中若出现矛盾数据，则将其删除。这里的矛盾数据是指两个或多个样本，其条件属性完全相同但决策属性不同；
  - 对每一个子集，启动一个HSC三维训练过程，从而得到$\lceil d/3 \rceil$个分类模型。

中国科学院

Chinese Academy of Sciences

高维划分集成方法

- 测试过程
  - (1)~(3)步同训练过程；
  - (4) 对每个子集，调用与其相对应的分类模型进行分类。因此，每一个样本都预测类别；
  - (5) 对每一个样本，采用投票的方式确定其最终类别。我们采用相对多数的投票机制，即样本的最终类别是这预测类别中得票最多的一个。

高维划分集成方法

- 实验(矛盾数据情况)

表 4.1 数据集 Pima 划分集成结果

	实际参与训练的样本数	召回率 (%)	准确率 (%)
子分类器 1	566	100	90.00
子分类器 2	542	100	85.05
子分类器 3	542	100	94.85
分类器集成	568	88.73	87.50

高维划分集成方法

- 实验

表 4.2 高维划分集成方法性能测试

数据集	维度	训练样本数	测试样本数	子分类器数	集成分类器召回率 (%)	集成分类器准确率 (%)
Iris	4	100	50	2	100	98.00
Wine	13	128	50	5	100	100
Wdbc	30	369	200	10	100	100
Sonar	60	158	50	20	100	100

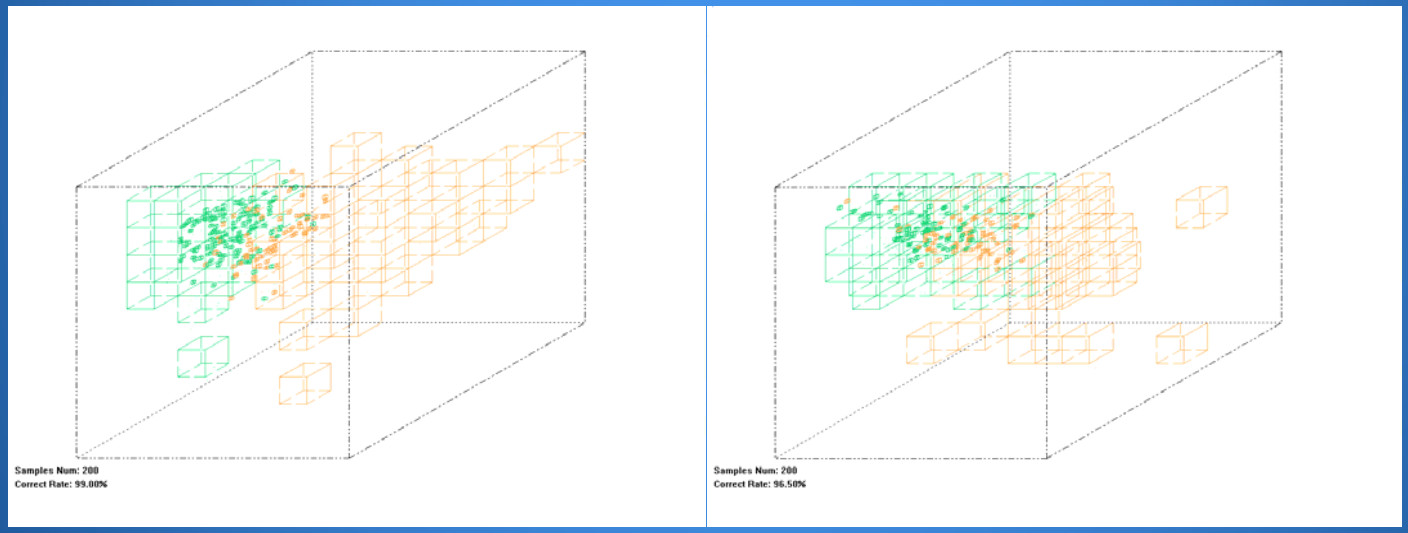
高维划分集成方法

Table 2. Classification results of HSC ensemble on the data set of Wdbc

Classifier No.	Recall	Accuracy
Classifier1	100%	99.00%
Classifier2	100%	98.00%
Classifier3	100%	97.00%
Classifier4	100%	95.00%
Classifier5	100%	94.00%
Classifier6	100%	86.50%
Classifier7	100%	97.50%
Classifier8	100%	99.50%
Classifier9	100%	89.00%
Classifier10	100%	96.50%
Classifiers Ensemble	100%	100%

高维划分集成方法

- 示例
  - Wdbc第一个和第十个子分类器结构



基于超曲面的分类方法特点

- 通过特征区域细化直接解决非线性分类问题
  - 不需要考虑使用何种核函数，不需要升维变换
- 通用可操作的分类超曲面构造法
  - 基于分类超曲面的方法通过区域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进行划分
- 独特、简便、易行的分类判别方法
  - 基于分类超曲面的方法是根据样本点关于分类曲面的围绕数的奇偶性进行分类的一种全新分类判断算法，使得基于非凸的超曲面的分类判别变得简便、易行

基于超曲面的分类法特点

- 适合多类分类，ＳＶＭ则不然
- 低维情况下算法复杂性低、分类效率高
- 推广能力较好、准确率高
  - 适合分布复杂的样本分类，对同类样本在有限连通区域连续分布的分类问题有效
  - 与SVM适用小样本不同，HSC更适合大样本
- 抗噪性
- 占用计算资源少
- 对高维数据算法复杂度高

三、极小一致集与样本抽样

- 从样本集中选择出具有代表性的样本点是覆盖型分类算法的一个重要研究问题
- “代表性”的含义在近邻法和覆盖型分类算法中有所区别
- 直接把近邻法中极小一致子集(MCS)的概念运用到覆盖型分类算法中不可行
- 提出“极小覆盖子集”的概念

极小覆盖子集

- 相关工作
  - 为减少近邻法的计算量和存储量，Hart 提出极小一致子集(Minimal Consistent Subset: MCS)的概念，用于从样本集中选择出具有代表性的样本点
  - 一致子集是指原始样本集的一个子集，用该子集进行近邻法分类时可保证原样本集完全分类正确，而极小一致子集是包含样本数最少的一致子集
  - 任何样本集都存在一致子集
  - 任何有限样本集都存在极小一致子集，但可能不唯一

中国科学院 Chinese Academy of Sciences

极小覆盖子集

- 定义(覆盖型分类算法的极小覆盖子集)
  - 对特定的训练样本集，若其子样本集训练后得到的分类模型与与原样本集训练后得到的分类模型相同，则称子样本集是原样本集的一个覆盖
  - 在一个样本集的所有覆盖中，包含样本个数最少的覆盖称为样本集的极小覆盖子集

中国科学院

Chinese Academy of Sciences

极小覆盖子集

- 基本特征
  - 覆盖性
  - 极小性
  - 存在性
- HSC极小覆盖子集的特殊性
  - 尺度依赖性
  - 极小一致性

中国科学院
Chinese Academy of Sciences

极小覆盖子集

- HSC极小覆盖子集的计算
  - 等价
    - 若样本点a与样本点b属于同一类别，并且落在分类超曲面模型的同一个单元格内，则称a与b在构造超曲面的过程中是等价的
  - 等价类
    - 落在同一单元格内的所有同类样本就构成一个等价类
  - 具有等价关系的样本点在超曲面的构造过程中起着同样的、重要的作用

中国科学院

Chinese Academy of Sciences

极小覆盖子集

- HSC极小覆盖子集的计算
  - 设训练样本集 S 经HSC算法训练后得到的分类超曲面模型为 H，u 为该模型中的一个单元格，那么从每个单元格中取出且仅取出一个样本构成的集合就称为HSC极小覆盖子集 $S_{\min}|_H$，即

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

极小覆盖子集

- 计算极小覆盖子集的基本步骤
  - 用一个方形区域覆盖所有样本点；
  - 将该区域划分成一系列小区域（单元格），直到每个小区域内包含的样本点都属于同一类别；
  - 根据小区域内样本点的类别，对所有小区域进行标记，并对每个样本点记下其所在的小区域；
  - 若两个相邻小区域的类别相同，则将其合并，重复该操作将得到分类超曲面；
  - 将落在同一小区域内的样本点合并成一个等价类；
  - 从每个等价类中选择且仅选择一个样本构成极小覆盖子集。

中国科学院

Chinese Academy of Sciences

极小覆盖子集

- 极小覆盖子集中包含的样本数目等于等价类（单元格）的个数
- 极小覆盖子集的个数等于这些等价类的笛卡尔积的大小

极小覆盖子集

- 极小覆盖子集的特征一

  对给定样本集，极小覆盖子集能够完全、充分地反映其分类能力

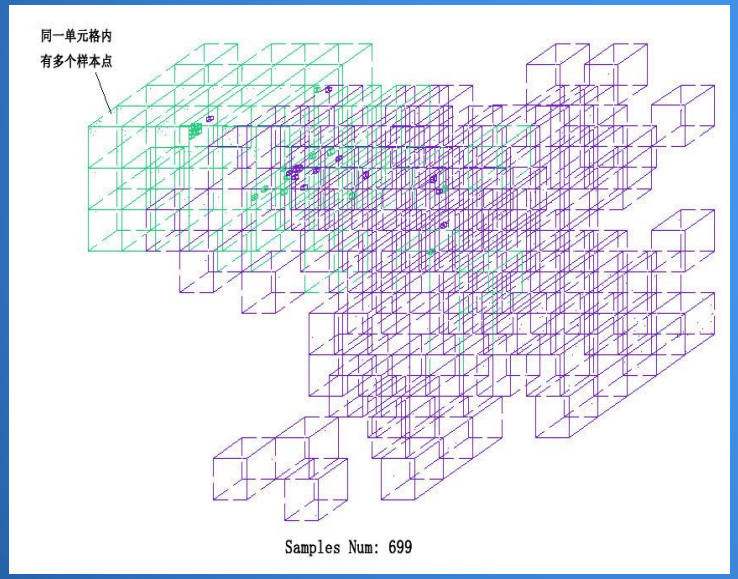
  - 极小覆盖子集的训练模型与原始样本集的训练模型完全相同
  - 极小覆盖子集与包含它的样本集都具有相同的分类能力

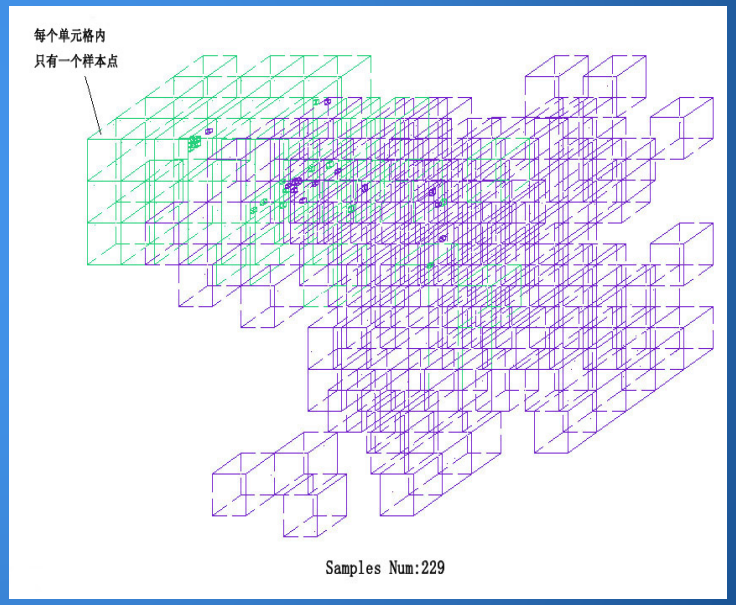
中国科学院

Chinese Academy of Sciences

极小覆盖子集

- 极小覆盖子集的特征一





极小覆盖子集

- 极小覆盖子集的特征一

表 5.1 极小覆盖子集及其超集的分类能力测试

数据集	样本数	召回率	MCS 样本数	实验 I 准确率	实验 II 准确率
Breast-cancer-Wisconsin	699	100%	229	100%	100%
Wine	178	100%	129	100%	100%
Ten Spirals	33750	100%	7285	100%	100%

极小覆盖子集

- 极小覆盖子集的特征二

假设给定样本集中有 N 个样本点，其极小覆盖子集中包含 n 个样本点；用极小覆盖子集作训练集，其余样本点作测试集，分类精度记为 $A(A=1)$。若从极小覆盖子集中删除且仅删除一个样本点并将其加入到测试集中，那么分类精度将降低为

$$A-m/(N-n+1) \qquad \text{(公式 5.4.1)}$$

其中 m 是与这个被删除样本点落在同一单元格（等价类）中的样本总数。

极小覆盖子集

表 5.2 单样本点的删除与分类精度的关系

被删除样本点ID	与被删除样本点落在同一单元格内的样本数	通过实验得到的精度	通过公式计算的精度	相同的情况数
4	1	99.79%	99.79%	155
26	2	99.58%	99.58%	39
10	3	99.36%	99.36%	11
27	4	99.15%	99.15%	6
35	5	98.94%	98.94%	3
43	6	98.73%	98.73%	4
20	7	98.51%	98.51%	1
30	8	98.30%	98.30%	2
6	10	97.88%	97.88%	1
178	11	97.66%	97.66%	1
37	17	96.39%	96.39%	1
17	34	92.78%	92.78%	1
9	39	91.72%	91.72%	1
1	48	89.81%	89.81%	1
3	71	84.93%	84.93%	1
7	117	75.16%	75.16%	1

极小覆盖子集

- 极小覆盖子集的特征二
  - 极小覆盖子集中每一个样本点的删除都会导致泛化能力的下降
  - 二者之间存在定量关系
  - 极小覆盖子集是决定分类超曲面泛化能力的幕后操纵者

中国科学院

Chinese Academy of Sciences

极小覆盖子集

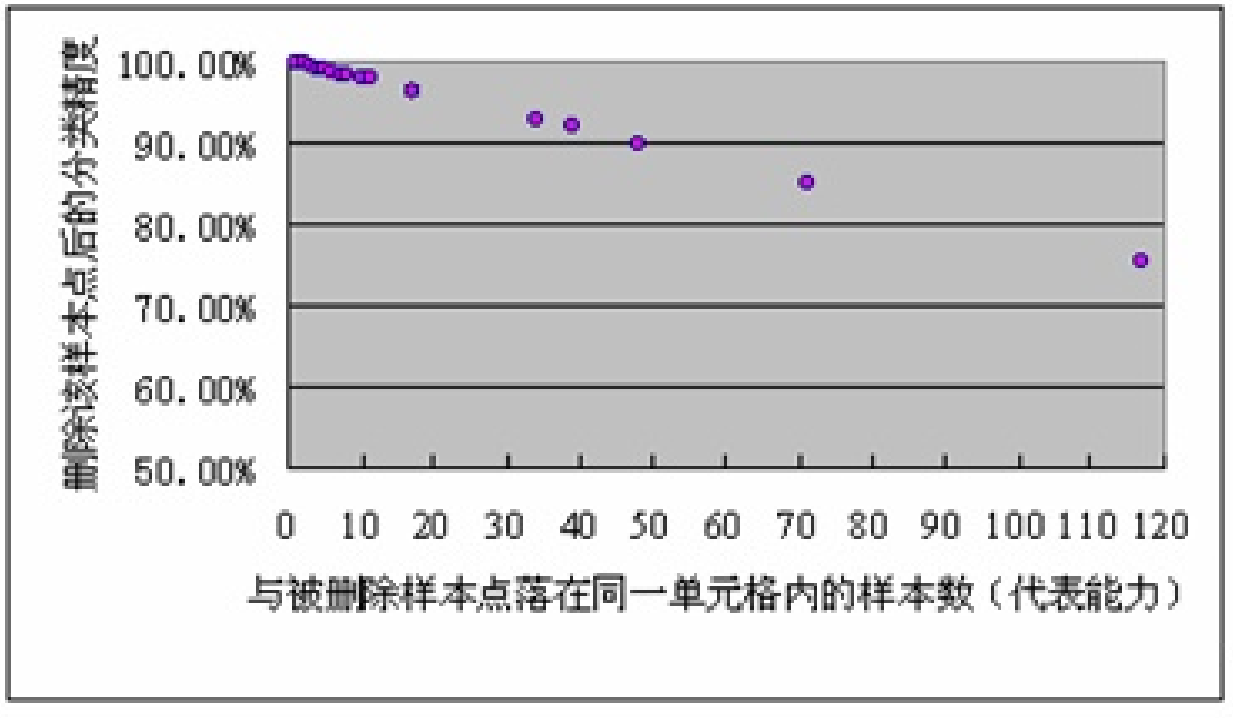


图 5.4 极小覆盖子集中被删除样本点代表能力与分类精度损失的关系

极小覆盖子集

- 极小覆盖子集的特征二

推广到一般情况，若从极小覆盖子集中同时删除 $K(1 \le K \le n)$ 个样本，并将其加入到测试集中，那么分类精度将降低为

$$A-(m_1+m_2+\ldots+m_K)/(N-n+K) \qquad \text{(公式 5.4.2)}$$

其中 m_i 是与第 $i(1 \le i \le K)$ 个被删除样本点落在同一单元格（等价类）中的样本数。

极小覆盖子集

- 极小覆盖子集的特征二

表 5.3 多样本点的删除与分类精度的关系

	通过实验得到的精度	通过公式计算的精度
$k=2$, $m=\{1,2\}$	99.36%	99.36%
$k=5$, $m=\{1,2,3,4,5\}$	96.84%	96.84%
$k=10$ $m=\{1,2,3,4,5,6,7,8,10,11\}$	88.13%	88.13%

极小覆盖子集

- 极小覆盖子集的特征三
  - 针对分类超曲面算法，极小覆盖子集是对给定数据集的最佳采样方式
  - 计算极小覆盖子集的过程也就是采样的过程
  - 极小覆盖子集具有完全代表性，并能够将原始数据集全部分类正确，故为最佳采样方式
  - 是一种非概率采样方法

中国科学院

Chinese Academy of Sciences

极小覆盖子集

例如，数据集 Breast-Cancer-Wisconsin 中含有 699 个样本，其极小覆盖子集中包含 229 个样本。根据表 5.2，不同极小覆盖子集的个数等于所有等价类的笛卡尔积的大小，即

$$1^{115} \times 2^{39} \times 3^{11} \times 4^{6} \times 5^{3} \times 6^{4} \times 7^{1} \times 8^{2} \times 10^{1} \times 11^{1} \times 17^{1} \times 34^{1} \times 39^{1} \times 48^{1} \times 71^{1} \times 117^{1}$$

$$= 2862379334528920895091978160142548992000$$

$$\approx 2.86 \times 10^{40}$$

那么，如果使用概率采样方法，能得到极小覆盖子集的概率为：

$$2.86 \times 10^{40} \Big/ C_{699}^{229} = 2.86 \times 10^{40} \Big/ 1.82 \times 10^{270} = 1.57 \times 10^{-230}$$

中国科学院

Chinese Academy of Sciences

极小覆盖子集

- 极小覆盖子集的特征四
  - 极小覆盖子集是PAC学习理论在分类超曲面算法中的推广
  - PAC学习模型给出了成功地学习目标概念所需的训练样本数目的边界
  - 针对分类超曲面算法来说，由于极小覆盖子集具有完全的代表性，从原训练集中能学习到的目标概念在其极小覆盖子集中同样能学得到
  - 极小覆盖子集在满足PAC学习模型提出的训练样本数目的边界的前提下，具体给出了训练样本集的这样一个子集

中国科学院

Chinese Academy of Sciences

Bagging和AdaBoost提升性能受限

- 提高超曲面分类器HSC对小样本的分类泛化能力
- Bagging和AdaBoost集成学习算法是否能用于提升HSC算法的分类泛化能力
- 如果能提升这种提升是否不受限制？

中国科学院

Chinese Academy of Sciences

Bagging算法

第一步：对于给定的原始训练数据集 D，包含 N 个样本，设置循环的次数为 K，和每次采样的样本数为 $n(n < N)$；

第二步：从原始训练数据集 D 随机采样，构成一个新的数据集 D_k，对于给定的算法 HSC，可以训练得到一个分类器 C_k；

第三步：重复做第二步 K 次，可以得到 K 个分类器 $C = \{C_1, \cdots, C_K\}$。

第四步：当有新的样本的时候，就可以由这 K 个分类器投票决定新样本属于哪一类，在 Bagging 训练中，每个分类器赋予相同的权重。

Adaboost算法

- Boosting 算法能提高学习算法的性能
- AdaBoost 算法生成一组分类器，然后利用它们进行投票。
- 两个算法有本质的区别
  - AdaBoost 生成分类器是顺序的，而不能是平行的：它基于前一次构造的分类器改变训练集中的权重，其目的是使不同输入分布上归纳器的期望误差最小化。最终的分类器采用加权投票，每个分类器的权重由它的性能决定

Adaboost算法

第一步：输入 N 个样本 $\{(x_1, y_1), \cdots, (x_N, y_N)\}$，然后赋予每个样本具有相同的权重，这个权重是决定该样本被采样的概率(权重越大，概率就越大)，也叫做初始的样本分布 $D^{(1)}$；

第二步：根据分布 $D^{(t)}$ 采样，用基础算法 HSC 训练分类器，得到分类器 C_t；

第三步：用分类器 C_t 对样本进行分类，用以下表达式计算分类错误率，

$$\varepsilon_t = \sum_{n=1}^{N} D_n^{(t)} I(y_n \neq C_t(x_n)) \tag{2.2}$$

第四步：计算对应的分类器的权重，如下表达式，

$$\alpha_t = \frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t} \tag{2.3}$$

第五步：用以下表达式更新数据的分布，

$$D_n^{(t+1)} = D_n^{(t)} \exp(-\alpha_t y_n C_t(x_n)) / Z_t \tag{2.4}$$

其中，Z_t 是一个归一化参数，

$$Z_t = \sum_{n=1}^{N} D_n^{(t)} \exp(-\alpha_t y_n C_t(x_n)) \tag{2.5}$$

第六步：重复第二步到第五步 T 次，可以得到 T 个分类器，最后的集成分类器为，

$$f_{Ens}(x) = \sum_{t=1}^{T} \alpha_t C_t(x) \tag{2.6}$$

实验结果

Table.1 *n* samples selected from samples set of breast-cancer-wisconsin (699 samples)

Samples Selected	200	180	160	120	100	80
HSC Accuracy	67.13%	66.28%	64.38%	63.73%	61.43%	58.32%
Bagging Accuracy	98.20%	98.27%	98.14%	97.93%	97.66%	95.96%
Samples Selected	60	50	40	30	20	10
HSC Accuracy	55.71%	54.08%	51.29%	54.71%	45.21%	36.28%
Bagging Accuracy	96.71%	95.07%	91.50%	91.18%	85.27%	77.36%

实验结果

Table.2 n samples are selected from 699, and the rest are tested

	29	79	129	229	329	429
1	88.06%	96.94%	97.72%	98.09%	98.65%	98.15%
2	88.66%	97.42%	97.72%	98.51%	97.57%	98.15%
3	89.40%	97.26%	97.54%	98.94%	98.92%	97.07%
4	89.85%	96.45%	97.37%	98.30%	97.84%	99.26%
5	89.70%	97.26%	97.54%	98.09%	97.84%	99.26%
6	91.04%	97.26%	97.19%	97.23%	98.65%	98.89%
7	89.40%	96.77%	97.19%	98.51%	98.38%	98.52%
8	90.15%	97.90%	97.37%	97.87%	97.84%	99.63%
9	88.96%	97.58%	97.89%	97.87%	98.65%	97.78%
10	91.19%	97.58%	97.37%	98.72%	98.92%	98.52%
Average Accuracy	89.64%	97.24%	97.49%	98.21%	98.32%	98.52%

实验结果

Table.3. Single deletion from the minimum sample set of breast-cancer-wisconsin

Samples in the same unit with the one deleted	ID of deleted sample	HSC Accuracy by Experiment	Bagging Accuracy by Experiment	AdaBoost Accuracy by Experiment
1	4	99.79%	99.79%	99.79%
2	26	99.58%	99.36%	99.58%
3	10	99.36%	99.15%	99.36%
4	27	99.15%	98.94%	99.15%
5	35	98.94%	98.73%	98.73%
6	43	98.73%	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%

实验结果说明

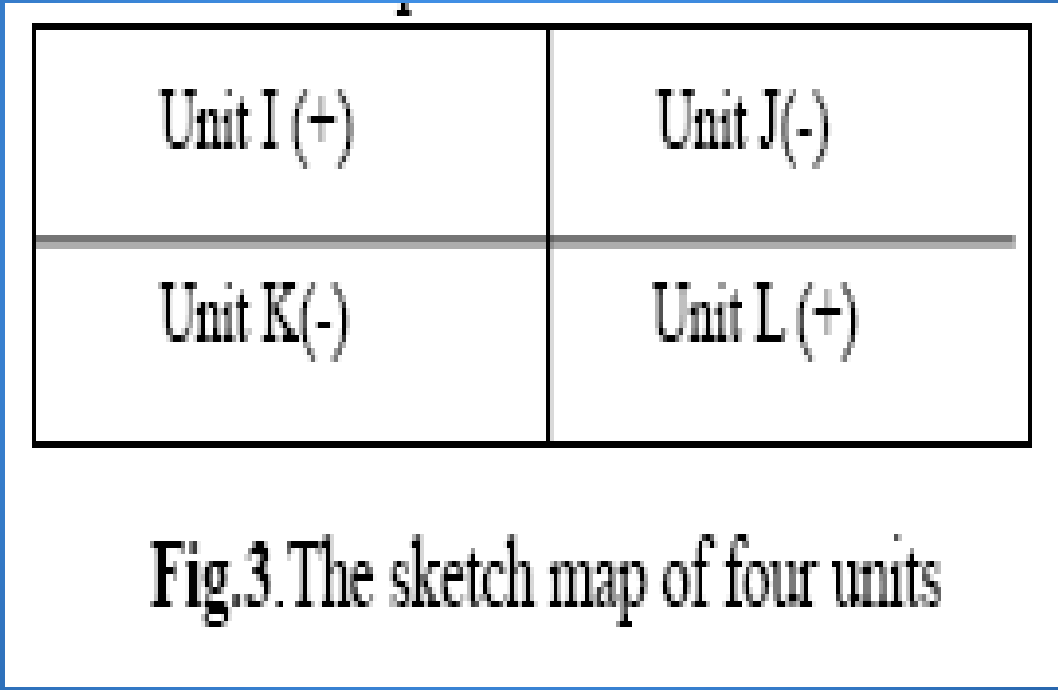


Fig.3.The sketch map of four units

实验结果

Table.4 229 samples randomly selected from 699 (no MCSC) for training, and the rest samples 699-229 are tested

	HSC by Experiment	Bagging Accuracy by Experiment	AdaBoost Accuracy by Experiment
1	64.89%	64.89%	64.89%
2	64.89%	65.11%	64.68%
3	64.89%	65.32%	65.74%
4	64.89%	65.11%	64.47%
5	64.89%	65.32%	65.53%

中国科学院

Chinese Academy of Sciences

Bagging和Adaboost的局限性

- 用Bagging和Adaboost算法得到的准确率很难超过由MCSC中样本作为训练集训练的HSC算法
- 实验结果说明了极小覆盖一致子集MCSC对Bagging和Adaboost提升行为的限制能力
- 这说明了MCSC是HSC算法泛化能力的决定因素

中国科学院

Chinese Academy of Sciences

四、超曲面分类方法的规则抽取

- 提出了基于极小覆盖子集（Minimal Cover Subset，MCS）和合并相邻单元的规则抽取与约简方法
  - 通过规则有效表示分类知识
  - 一定程度上将规则集约简至最小基数

中国科学院

Chinese Academy of Sciences

超曲面分类方法的规则抽取与约简

- 假设一个极小一致子集中的样本(*x*,*y*,*z*)，其所在层为第*i*层，X_1、Y_1和Z_1为其下限约束，X_2、Y_2和Z_2为其上限约束，则这六个约束的计算方法如下：

$$X_1 = \lfloor x \times 10^i \rfloor / 10^i \qquad X_2 = X_1 + 1/10^{-i}$$
$$Y_1 = \lfloor y \times 10^i \rfloor / 10^i \qquad Y_2 = Y_1 + 1/10^{-i}$$
$$Z_1 = \lfloor z \times 10^i \rfloor / 10^i \qquad Z_2 = Z_1 + 1/10^{-i}$$

$(0.86,\ 0.03,\ 0.23),\ layer = 1,\ \text{label}=Y$

$0.8 \le x \le 0.9,\ 0.0 \le y \le 0.1,\ 0.2 \le z \le 0.3,\ Catagory = Y$

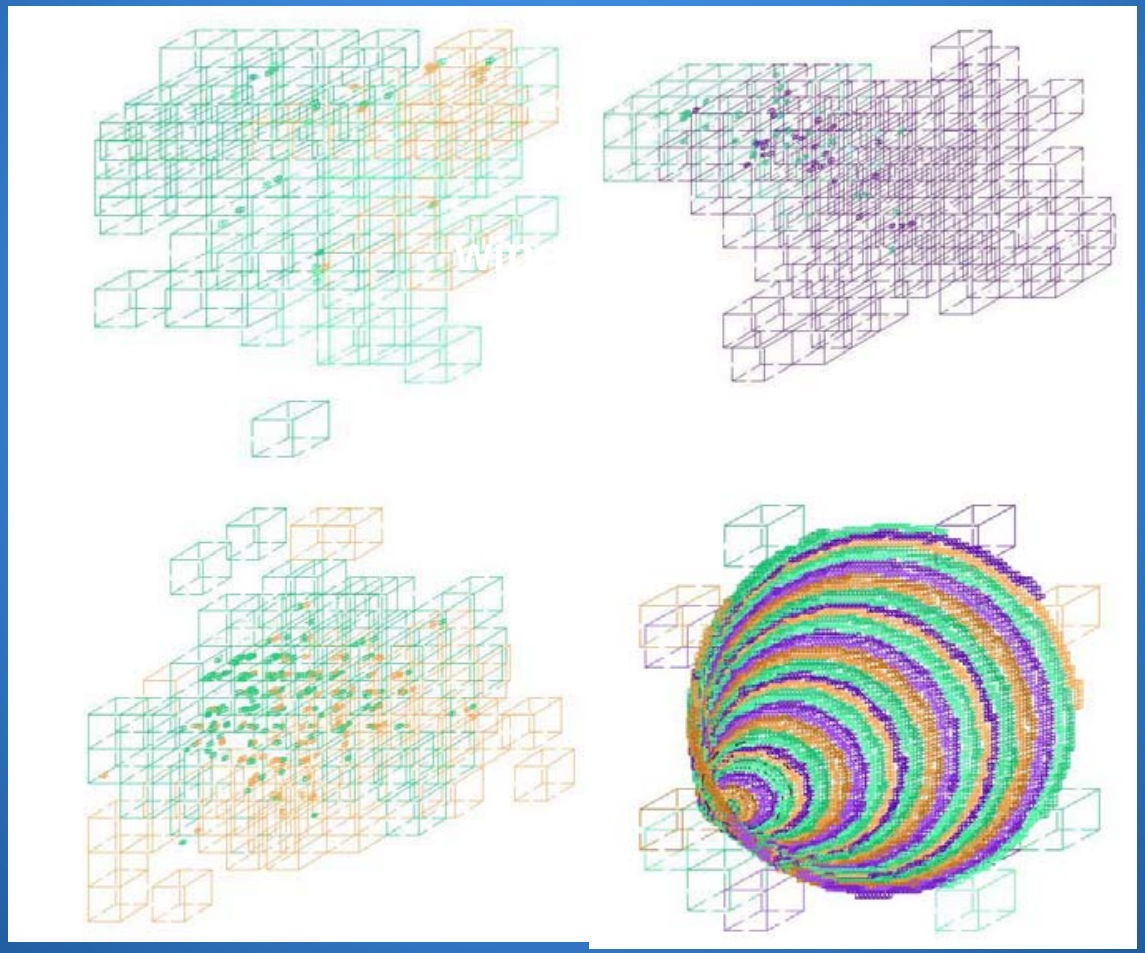
超曲面分类方法的规则抽取与约简

- 近邻规则
  - 必须是抽取于两个相邻的且分类相同的方体
  - 仅有一维，其中一条规则的前件与另一条规则该维的后件相等，而其它各维前件与后件均相等
- 关键步骤
  - 第一步，将规则导入链表，执行一遍规则融合
  - 第二步，逆置链表，执行一遍规则融合
  - 第三步，如果在经过第一步和第二步两个步骤后，规则条数保持不变，则算法终止；否则转向第一步

$$(5, 0.4, 0.5, 0.8, 0.9, 0.5, 0.6, 1)\ (84, 0.4, 0.5, 0.7, 0.8, 0.5, 0.6, 1)$$

$$\Rightarrow\ (5, 0.4, 0.5, 0.7, 0.9, 0.5, 0.6, 1)$$

超曲面分类方法的规则抽取与约简

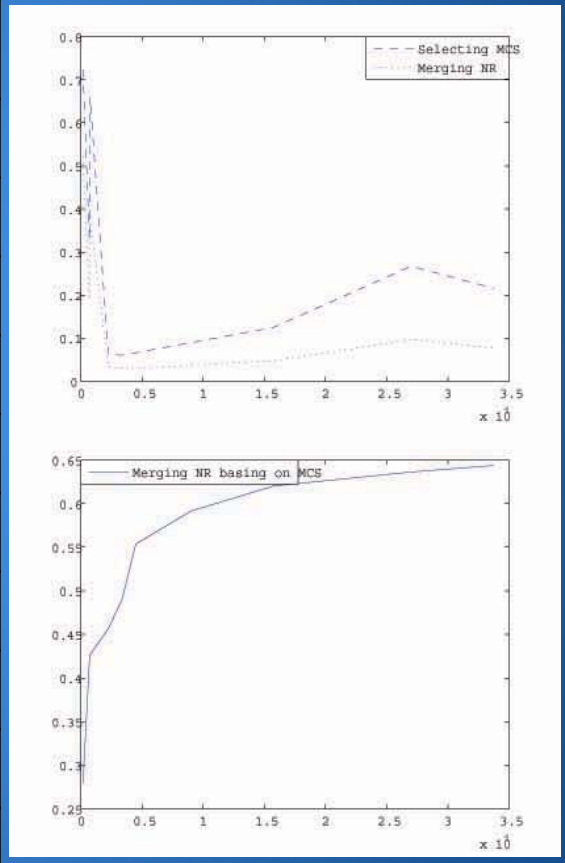


Breast cancer

spiral

规则抽取与约简实验

ataset	Size of dataset	MCS	Merging NR	Reduction Ratio
Wine	178	129	93	27.9%
Breast	699	229	133	41.5%
Pima	768	506	290	42.7%
Spiral	2250	138	75	45.7%
Spiral	3380	206	105	48.9%
Spiral	4500	298	133	55.4%
Spiral	9010	808	330	59.2%
Spiral	15760	1972	749	62.02%
Spiral	27010	7220	2626	63.63%
Spiral	33750	7285	2593	64.40%



五、并行HSC算法

- 问题

  大数据分类问题

  - 高于三维直接分类问题

- 并行HSC优势

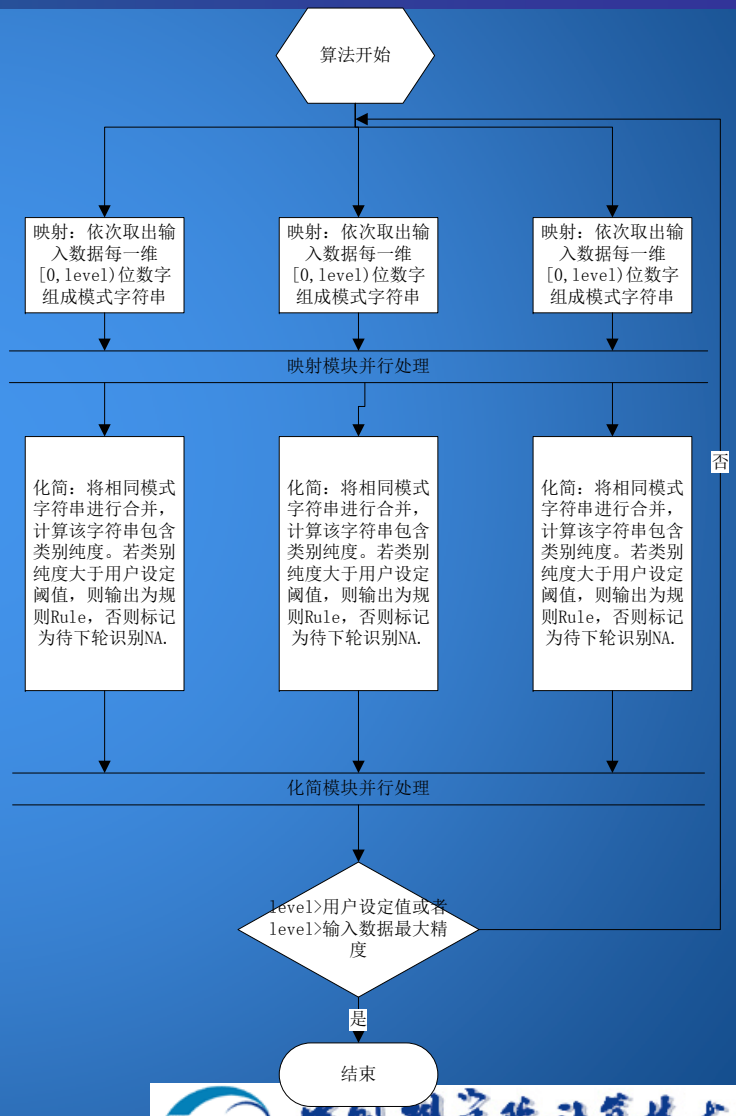
  - PHSC 基于Map/Reduce的sorting结构能处理大数据
  - PHSC 利用压缩有意义的表来存储规则
  - 对高于三维的分类问题理论上都能直接解决
  - 但超高维情况下会出现计算量过大问题

中国科学院

Chinese Academy of Sciences

并行HSC算法PHSC

- 并行的HSC算法
- 能直接有效处理高维大数据



六、基于超曲面的聚类算法

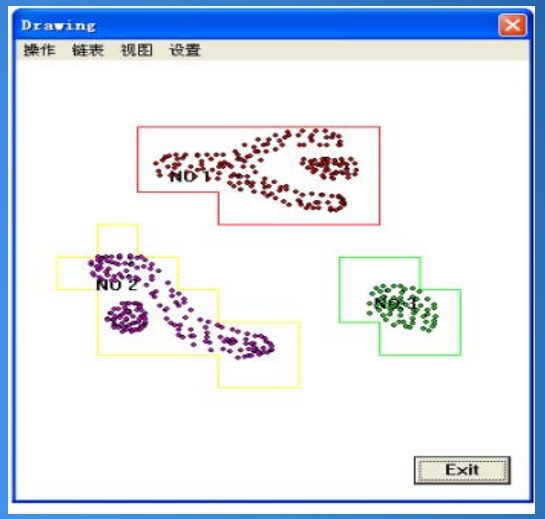
- 基于超曲面的聚类算法(Clustering based on Hyper Surface : CHS)思想
  - 所有的样本视为相同类别，并把该样本集作为HSC训练过程的输入
  - 划分到同一个超曲面片的样本视为一个聚簇
- 基于超曲面和最小支撑树的聚类算法CHSMST
  - 能够发现任意形状的聚簇
  - 抗噪性强
  - 时间性能优于DBSCAN，CLARANS和CLIQUE等经典聚类算法

中国科学院

Chinese Academy of Sciences

基于超曲面的聚类算法

示例数据集



CHS的聚类结果

- 存在问题
  - 发现每个簇内样本分布
  - 找出局部密集的子簇
  - 簇的细分

基于超曲面最小支撑树的聚类算法

- 基于超曲面和最小支撑树的聚类算法
- 步骤1 采用CHS算法进行初始的聚类
- 步骤2 用最小支撑树算法处理样本局部密集的区域
  - 需细分样本作为结点，样本间的距离作为连接权，构造一个完全图
  - 求得该完全图的最小支撑树
  - 删除最小支撑树中权较大的边，则得到更小的聚簇

基于超曲面最小支撑树的聚类算法

基于超曲面最小支撑树的聚类算法

测试数据集分布

CHSMST聚类结果

基于超曲面最小支撑树的聚类算法

算法运行时间比较(时间单位：秒)

Number of Points	1252	2503	3910	5213	6256
CHSMST	**0.09**	**0.29**	**0.65**	**1.12**	**1.57**
CLIQUE	0.15	0.42	1.28	2.70	4.11
DBSCAN	1.64	5.86	14.05	24.53	35.41
CLARANS	30.56	122.01	276.01	473.59	726.98
Number of Points	7820	8937	10426	12512	
CHSMST	**2.42**	**3.14**	**4.23**	**6.09**	
CLIQUE	5.77	7.35	11.38	16.33	
DBSCAN	55.02	72.39	98.34	144.70	
CLARANS	1202.68	1583.29	2441.16	3251.58	

七、基于超曲面的离群点检测

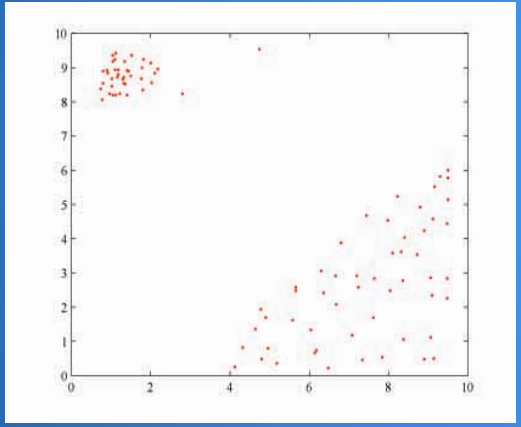
- 正常点与离群点的定义
  - 一个样本是*正常点*，当覆盖它的超曲面片或者最小支撑树包含较多样本超过总样本数的5%
  - 一个样本是*离群点*，当覆盖它的超曲面片或者最小支撑树仅仅包含一个样本，或者对于同曲面内的其它样本它分布在曲面的局部
- 无标签数据：基于超曲面的检测方法
  - 第一步，读入样本集，并使得所有样本点落在一个区域中
  - 第二步，把该区域划分成10×10×…10(10^d,d为维度)个单元格
  - 第三步，将包含一个或者多个样本的单元的边界作为一个字串保存下来
  - 第四步，根据第三步，合并相邻的单元。重新合并这些单元的边界并作为一个字串保存下来
  - 第五步，对于每个超曲面片，如果该片中样本分布较稀疏，则可判断其边界内的样本点为离群点

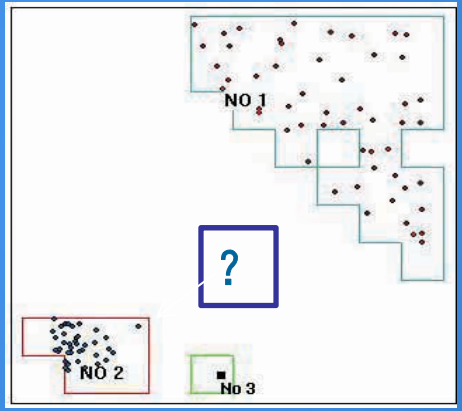
中国科学院

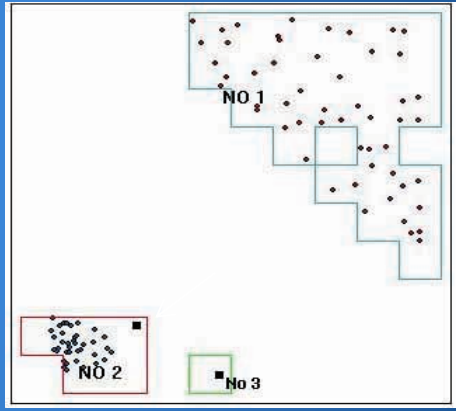
Chinese Academy of Sciences

基于超曲面的离群点检测

- 局部离群点





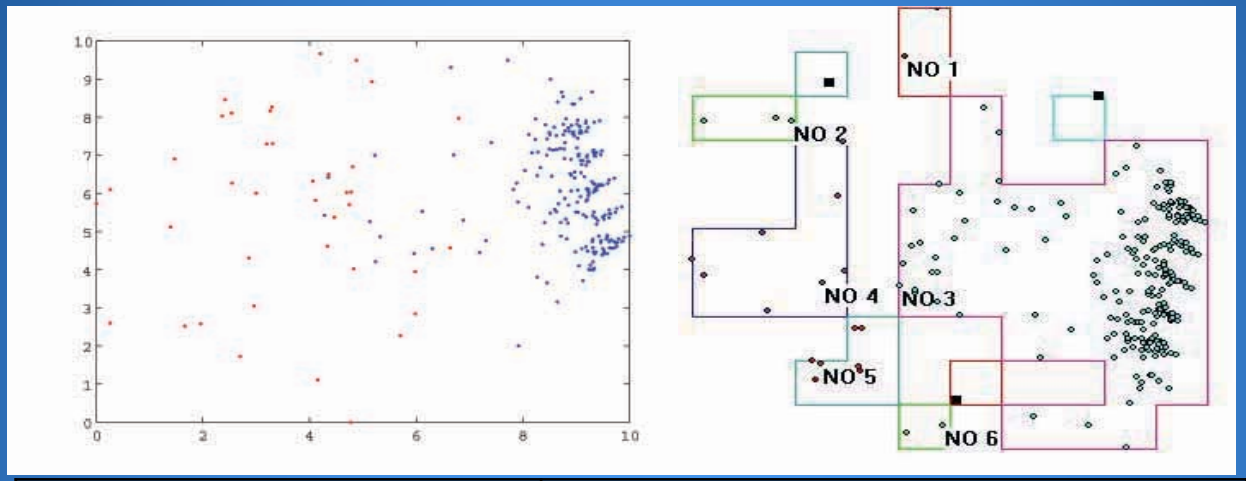


基于超曲面的离群点检测

- 局部离群点的检测策略
  - 最小支撑树：Kruskal算法，Prim算法
- 具体步骤：
  - 第一步，对数据集进行预处理并将所有样本视为同一类，利用无监督HSC算法构造分离超曲面
  - 第二步，根据领域知识，如果没有指定细分的超曲面，转到第四步；否则给定需细分的超曲面片的编号*Id*和细分程度*SubdivNum*
  - 第三步，在指定的超曲面内搜索最小支撑树并细分成*SubdivNum*个分支，转到第三步
  - 第四步，将分布较稀疏的分支或子树的样本点断定为离群点

中国科学院 Chinese Academy of Sciences

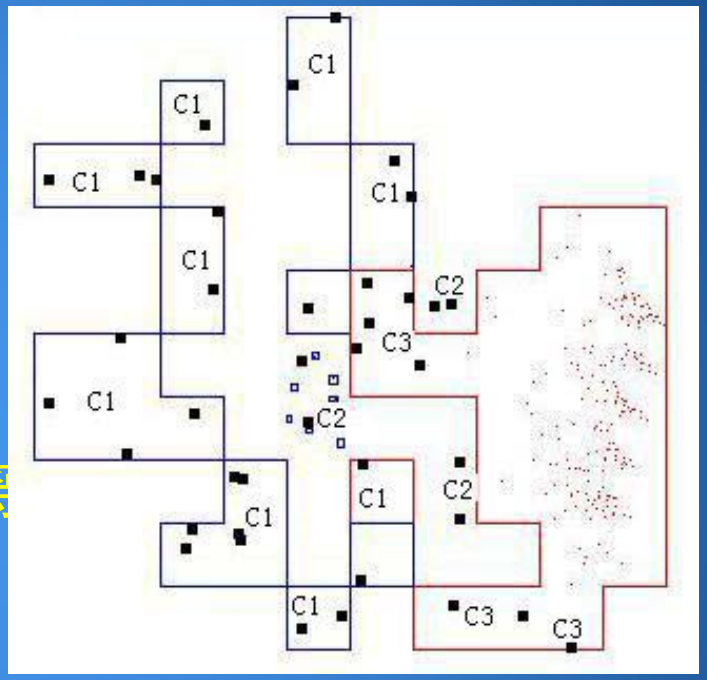
基于超曲面的离群点检测



Parameters		Result(Sample ID of the dataset)
HS.ID	*SubDivNum*	
1	2	169,397
2	3	108,285,409
4	8	141,171,42,125,133,150,262,470
5	7	36,62,67,94,158,297,431
6	6	34,385
3	20	16,29,64,77,160
…	…	……

有标签数据的离群点检测

- 需检测的三类离群点
  - C1：全局范围内稀疏且零散分布的样本点
  - C2：与正常样本点混合在一起分布的局部离群点
  - C3：超曲面片内的局部离群点



有标签数据的离群点检测

- 有标签数据的离群点检测
  - 采用HSC构造分离超曲面
  - 对于每个超曲面片，将分布较稀疏的超曲面片样本点断定为离群点
- 根据领域知识的局部离群点检测
  - 与无标签数据的局部检测思路一致

中国科学院

Chinese Academy of Sciences

关于HSC研究的历程

- 提出并实现了基于超曲面的覆盖分类学习算法并进行了理论分析
- 相继解决了二维两类分类、二维多类分类、三维两类分类、三维多类分类、高维多类分类问题
- 在相同的PC计算环境下，HSC准确分类的数据量达到了十的七次方，这是传统SVM达不到的
- 对于同数量级的相同数据，HSC比决策树算法快一倍,分类效果相当
- 在理论上提出了极小覆盖样本集概念并用于推广性的精确估计

中国科学院

Chinese Academy of Sciences

关于HSC研究的历程

- 分析了划分尺度与分类精度的关系
- 提出基于超曲面的聚类法
- 提出基于超曲面的离群点检测方法
- 实现基于超曲面的并行大数据分类算法

下一步工作

- 超高维数据直接求分类超曲面的快速方法

参考文献

- Qing He, Zhongzhi Shi, Lian Ren. The Classification Method Based on Hyper Surface，2002 IEEE International Joint Conference on Neural Networks, pp.1499-1503, March, 2002 (**EI**、**ISTP**、**INSPEC**).（二维两类）
- Qing He, Zhongzhi Shi, Lian Ren. The Multi-class Classification Method in Large Database Based on Hyper Surface. 2002 International Conference on Machine Learning and Application, CSREA, pp.164-169, June, 2002.（二维多类）

参考文献

- 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 (**INSPEC**). （二维一般连通区域）
- 任力安，何清，史忠植. HSC分类法及其在海量数据分类中的应用. 电子学报, 2002,30(12)：1870-1872(**EI, INSPEC**).（三维两类）
- 何清,史忠植,任力安. 基于超曲面的多类分类方法. 系统工程理论与实践, 2003.23（3）：92-99. （三维多类）
- Qing He, Zhong-Zhi Shi, Li-An Ren, E. S. Lee. A Novel Classification Method Based on HyperSurface. International Journal of Mathematical and Computer Modeling, 38(2003), 395-407(**SCI、INSPEC**).（二维，三维综合）

参考文献

- Xiurong Zhao, Qing He, Zhongzhi Shi. HyperSurface Classifier Ensemble for High Dimensional Data Sets, accepted by 3th International Symposium on Neural Networks (ISNN 2006)，LNCS，Springer(**SCI**,**EI**).. （高维集成）
- He, Qing; Zhao, Xiu-Rong; Shi, Zhong-Zhi. Sampling based on minimal consistent subset for hyper surface classification, Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007, 2007, p 12-18**(EI)**(HSC极小一致集抽样)
- He, Qing; Zhuang, Fu-Zhen; Zhao, Xiu-Rong; Shi, Zhong-Zhi. Enhanced algorithm performance for classification based on hyper surface using Bagging and AdaBoost, Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007, 2007, p 3624-3629**(EI)**(Bagging、AdaBoost与HSC性能提高)

参考文献

- He, Qing; Zhao, Xiu-Rong; Luo, Ping; Shi, Zhong-Zhi. Combination methodologies of multi-agent hyper surface classifiers: Design and implementation issues, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v 4476 LNAI, Autonomous Intelligent Systems: Agents and Data Mining - Second International Workshop, AIS-ADM 2007, Proceedings, 2007, p 100-113**(EI)**（HSC分类器融合）
- QING HE, XIU-RONG ZHAO and ZHONG-ZHI SHI. MINIMAL CONSISTENT SUBSET FOR HYPER SURFACE CLASSIFICATION METHOD, International Journal of Pattern Recognition and Artificial Intelligence,Vol. 22, No. 1 (2008) 95–108**(SCI,EI)**（HSC的极小一致集）

参考文献

- Qing He, Xiurong Zhao, Zhongzhi Shi. A KIND OF DIMENSION REDUCTION METHOD FOR CLASSIFICATION BASED ON HYPER SURFACE, International Conference on Machine Learning and Cybernetics, 18-21 August 2005(**EI**). （高维换维）
- F.Z, Zhuang, Q, He, Z.Z, Shi. Multi-Agent based Automatic Evaluation System for Classification Algorithms, in Proceedings of the 2008 IEEE International Conference on Information and Automation June 20 -23, 2008, pp264-269. (EI) .(HSC 分类与SVM等分类算法自动比较系统)

参考文献

- FU-ZHEN ZHUANG, QING HE, ZHONG-ZHI SHI. DATA FEATURE TRANSFORMATION FOR EFFICIENTLY IMPROVING PERFORMANCE OF HSC, ICMLC08, 423-428.(EI)(特征转换在HSC性能提高中的作用)
- Qing He, Fuzhen Zhuang, Zhongzhi Shi. The Data Selection Criteria for HSC and SVM Algorithms, ICNC08,pp384-388. (EI)(HSC与SVM数据选择判据比较)
- Qing He, Jincheng Li, and Zhongzhi Shi. Rule Extraction and Reduction for Hyper Surface Classification, in W. Yu, H. He, and N. Zhang (Eds.): ISNN 2009, Part II, LNCS 5552, pp. 88–98, 2009.(HSC分类规则抽取)

参考文献

- Qing He, Jincheng Li, Weizhong Zhao, Zhongzhi Shi. Unusual Pattern Detection Based on Hyper Surface and Minimum Spanning Tree, Proceedings of the 2009 IEEE International Conference on Information and Automation June 22 -25, 2009, pp. 1093-1098（基于超曲面和最小支撑树的异常模式发现）
- Li JC, He Q, Ma H, Shi ZZ. Detecting Unusual Pattern with Labeled Data in Two-Stage, 2009 Proceedings of Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Pages: 164-168, Published: Augest, 2009(两阶段带类标的数据异常模式发现)

参考文献

- He, Qing, Ma, Xu-Dong, Zhuang, Fu-Zhen,Shi, Zhong-Zhi. The effect of scale transformation for hyper surface classification method, 2009 International Conference on Machine Learning and Cybernetics, pp 1856-1860, 2009/7/12 (EI)(研究尺度变化对HSC的影响)
- 何清, 赵卫中, 史忠植. 分类超曲面算法复杂度研究, 计算机学报录用，2009（关于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）
- Qing He, Weizhong Zhao, Zhongzhi Shi. CHSMST: A clustering algorithm based on hyper surface and minimum spanning tree. Accepted by Soft Computing 2009(SCI源)(基于HSC和极小支撑树的聚类算法)

参考文献

- Qing He, Wenjuan Luo, Fuzhen Zhuang, and Zhongzhi Shi. Local Bayesian based Rejection Method for HSC Ensemble, ISNN2010.(基于局部贝叶斯的HSC集成中的拒识方法)
- Tingting Li, Fuzhen Zhuang, Qing He.A noise handling method for HSC, submited to ICNC2010.(HSC噪音处理方法)
- Qing He, Qing Tan, Xiurong Zhao, Zhongzhi Shi. A Visual Cognitive Method Based on Hyper Surface for Data Understanding, in Advances in Cognitive Informatics, Du Zhang, Yingxu Wang, and Witold Kinser, Co-Editors for Springer's book （数据理解中的基于超曲面的视觉认知方法）

参考文献

- Qing He, Haocheng Wang, Fuzhen Zhuang, Tianfeng Shang, Zhongzhi Shi. Parallel sampling from big data with uncertainty distribution, Fuzzy Sets and Systems 2014 accepted. http://dx.doi.org/10.1016/j.fss.2014.01.016 (并行大数据抽样分布发现算法)

中国科学院

Chinese Academy of Sciences

参考文献

- Qing He, Zhongzhi Shi, Lian Ren. The Classification Method Based on Hyper Surface，2002 IEEE International Joint Conference on Neural Networks, pp.1499-1503, March, 2002 (**EI**、**ISTP**、**INSPEC**).（二维两类）
- Qing He, Zhongzhi Shi, Lian Ren. The Multi-class Classification Method in Large Database Based on Hyper Surface. 2002 International Conference on Machine Learning and Application, CSREA, pp.164-169, June, 2002.（二维多类）

参考文献

- 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 (**INSPEC**). （二维一般连通区域）
- 任力安，何清，史忠植. HSC分类法及其在海量数据分类中的应用. 电子学报, 2002,30(12)：1870-1872(**EI，INSPEC**).（三维两类）
- 何清,史忠植,任力安. 基于超曲面的多类分类方法. 系统工程理论与实践, 2003.23（3）：92-99. （三维多类）
- Qing He, Zhong-Zhi Shi, Li-An Ren, E. S. Lee. A Novel Classification Method Based on HyperSurface. International Journal of Mathematical and Computer Modeling, 38(2003), 395-407(**SCI、INSPEC**).（二维，三维综合）

参考文献

- Xiurong Zhao, Qing He, Zhongzhi Shi. HyperSurface Classifier Ensemble for High Dimensional Data Sets, accepted by 3th International Symposium on Neural Networks (ISNN 2006)，LNCS，Springer(**SCI**,**EI**).. （高维集成）
- He, Qing; Zhao, Xiu-Rong; Shi, Zhong-Zhi. Sampling based on minimal consistent subset for hyper surface classification, Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007, 2007, p 12-18**(EI)**(HSC极小一致集抽样)
- He, Qing; Zhuang, Fu-Zhen; Zhao, Xiu-Rong; Shi, Zhong-Zhi. Enhanced algorithm performance for classification based on hyper surface using Bagging and AdaBoost, Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007, 2007, p 3624-3629**(EI)**(Bagging、AdaBoost与HSC性能提高)

参考文献

- He, Qing; Zhao, Xiu-Rong; Luo, Ping; Shi, Zhong-Zhi. Combination methodologies of multi-agent hyper surface classifiers: Design and implementation issues, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v 4476 LNAI, Autonomous Intelligent Systems: Agents and Data Mining - Second International Workshop, AIS-ADM 2007, Proceedings, 2007, p 100-113**(EI)**（HSC分类器融合）
- QING HE, XIU-RONG ZHAO and ZHONG-ZHI SHI. MINIMAL CONSISTENT SUBSET FOR HYPER SURFACE CLASSIFICATION METHOD, International Journal of Pattern Recognition and Artificial Intelligence,Vol. 22, No. 1 (2008) 95–108**(SCI,EI)**（HSC的极小一致集）

参考文献

- Qing He, Xiurong Zhao, Zhongzhi Shi. A KIND OF DIMENSION REDUCTION METHOD FOR CLASSIFICATION BASED ON HYPER SURFACE, International Conference on Machine Learning and Cybernetics, 18-21 August 2005(**EI**). （高维换维）
- F.Z, Zhuang, Q, He, Z.Z, Shi. Multi-Agent based Automatic Evaluation System for Classification Algorithms, in Proceedings of the 2008 IEEE International Conference on Information and Automation June 20 -23, 2008, pp264-269. (EI) .(HSC 分类与SVM等分类算法自动比较系统)

参考文献

- FU-ZHEN ZHUANG, QING HE, ZHONG-ZHI SHI. DATA FEATURE TRANSFORMATION FOR EFFICIENTLY IMPROVING PERFORMANCE OF HSC, ICMLC08, 423-428.(EI)(特征转换在HSC性能提高中的作用)
- Qing He, Fuzhen Zhuang, Zhongzhi Shi. The Data Selection Criteria for HSC and SVM Algorithms, ICNC08,pp384-388. (EI)(HSC与SVM数据选择判据比较)
- Qing He, Jincheng Li, and Zhongzhi Shi. Rule Extraction and Reduction for Hyper Surface Classification, in W. Yu, H. He, and N. Zhang (Eds.): ISNN 2009, Part II, LNCS 5552, pp. 88–98, 2009.(HSC分类规则抽取)

参考文献

- Qing He, Jincheng Li, Weizhong Zhao, Zhongzhi Shi. Unusual Pattern Detection Based on Hyper Surface and Minimum Spanning Tree, Proceedings of the 2009 IEEE International Conference on Information and Automation June 22 -25, 2009, pp. 1093-1098（基于超曲面和最小支撑树的异常模式发现）
- Li JC, He Q, Ma H, Shi ZZ. Detecting Unusual Pattern with Labeled Data in Two-Stage, 2009 Proceedings of Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Pages: 164-168, Published: Augest, 2009(两阶段带类标的数据异常模式发现)

参考文献

- He, Qing, Ma, Xu-Dong, Zhuang, Fu-Zhen,Shi, Zhong-Zhi. The effect of scale transformation for hyper surface classification method, 2009 International Conference on Machine Learning and Cybernetics, pp 1856-1860, 2009/7/12 (EI)(研究尺度变化对HSC的影响)
- 何清，赵卫中，史忠植. 分类超曲面算法复杂度研究, 计算机学报，2010,33(4)（关于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）
- He Qing, Zhao Weizhong, Shi Zhongzhi. CHSMST: A clustering algorithm based on hyper surface and minimum spanning tree, SOFT COMPUTING - A Fusion of Foundations, Methodologies and Applications Volume 15, Number 6 (2011), 1097-1103 (基于HSC和极小支撑树的聚类算法)

参考文献

- Qing He, Wenjuan Luo, Fuzhen Zhuang, and Zhongzhi Shi. Local Bayesian based Rejection Method for HSC Ensemble, ISNN2010.(基于局部贝叶斯的HSC集成中的拒识方法)
- Tingting Li, Fuzhen Zhuang, Qing He.A noise handling method for HSC, ICNC2010.(HSC噪音处理方法)
- Qing He, Qing Tan, Xiurong Zhao, Zhongzhi Shi. A Visual Cognitive Method Based on Hyper Surface for Data Understanding, in Advances in Cognitive Informatics, Du Zhang, Yingxu Wang, and Witold Kinser, Co-Editors for Springer's book （数据理解中的基于超曲面的视觉认知方法）

参考文献

- Qing He, Haocheng Wang, Fuzhen Zhuang, Tianfeng Shang, Zhongzhi Shi. Parallel sampling from big data with uncertainty distribution, Fuzzy Sets and Systems 258 (2015) 117–133, 2015
  http://dx.doi.org/10.1016/j.fss.2014.01.016
  (并行大数据抽样分布发现算法)

中国科学院

Chinese Academy of Sciences

中国科学院
CHINESE ACADEMY OF SCIENCES

欢迎批评指正！
谢谢！

heq@ics.ict.ac.cn

中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences