GSwMKnn:基于类别基尼系数子空间的加权互 K 近邻算法①

陈雪云[1,2], 卢伟胜[2]

[1](龙岩学院 数学与计算机科学学院, 龙岩 364012)
[2](福建师范大学 数学与计算机科学学院, 福州 350007)

摘 要: 在高维数据空间中, 存在大量冗余或无用的属性, 这使得在子空间中寻找目标类更为有效. 为此文章提出基于类别基尼系数子空间的加权互 k 近邻算法, 利用类别基尼系数求出其对应的软子空间并将待分类样本和训练样本投影到各个类别子空间中, 再在各软子空间中使用类别基尼系数加权距离互 k 近邻算法计算出待分类样本在各个子空间的投票权重并叠加, 最终得出待分类样本的类标签. 在公共数据集上的实验结果验证了该方法的有效性.

关键词: 类属性数据; 子空间; 互 k-近邻; 基尼系数

GSwMKnn: Weighted MKnn Algorithm Based on the Category's Gini Subspace

CHEN Xue-Yun[1,2], LU Wei-Sheng[2]

[1](School of Mathematics and Computer Science, Longyan University, Longyan 364012, China)
[2](School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China)

Abstract: In high-dimensional data spaces, there exists a large number of redundant or useless attributes, and therefore it might be more effective to find target class in their subspaces. A weighted MKnn algorithm based on the Category's Gini Coefficient subspace is proposed in this paper. Using the Category's Gini Coefficient, the algorithm firstly calculates the corresponding soft subspaces, and projects the training and testing samples onto each category subspaces. Secondly, it calculates the vote weights of unclassified samples on each subspace by the weighted MKnn algorithm and then accumulates them. Finally, it obtains the category labels of unclassified samples. The experimental results on some UCI public datasets demonstrate the effectiveness of the proposed method.

Key words: nominal data; subspace; mutual k-nearest neighbor; Gini index

数据挖掘研究领域涉及数据库和人工智能等学科, 是当前相当活跃的研究领域, 是指从大型数据库中, 挖掘潜在的, 未知模式的过程. 其中, 分类(Data Classification)在实际应用中得到广泛运用, 也是数据挖掘中非常重要的任务之一. 分类的目的是学会一个分类函数或者分类模型(也常常称为分类器), 该模型能把数据库中的数据项映射到给定类别中的某一个. 分类算法一般分为 Lazy 和 Eager 两种类型[1]. Lazy 分类算法, 每当对一个待分类样本进行分类时需要重新建立分类模型; 而 Eager 分类算法只需建立一次分类模型, 之后就可利用该分类模型对待分类样本进行分类. 许多传统的分类算法, 如 k 近邻(Knn), 朴素贝叶斯(NBC), 支持向量机(SVM)和决策树算法(C4.5)已被广泛应用于入侵检测、故障监测、信用卡欺诈分析等领域.

k 近邻(Knn)是由 Cover 和 Hart[2]提出的一个有效和强大的懒惰学习算法. 它的分类思想是: 给定一个待分类的样本 x, 首先找出与 x 最接近的或最相似的

① 基金项目:国家自然科学基金(61070062);福建高校产学合作科技重大项目(2010H6007);福建省教育厅 B 类项目(JB12201)
收稿时间:2013-07-08;收到修改稿时间:2013-09-09

k 个已知类别标签的训练集样本, 然后根据这k个训练样本的类别标签来确定待分类样本 x 的类别. 尽管 Knn 易于实现但是, 其性能很大程度上依赖于训练数据的质量. 由于许多复杂的实际应用, 各种噪声往往也存在于大型数据库中. 如何消除异常和提高数据的质量仍然是一个挑战. 为了缓解这一问题, Liu[3]等提出了一种通过判断是否互为近邻来选择 k 近邻, 并以此代替用于作数据表决的 k 近邻, 而丢弃训练样本可能的噪声数据的互 k 近邻(MKnn)算法, 它克服了传统 Knn 分类算法可能存在伪近邻的缺陷. 无论是 k 近邻方法还是互 k 近邻方法, 其近邻的选择依赖于相似性度量的选择, 而相似性度量性能与数据集中属性的类别有很大的关系, 如欧拉距离较适用于数值型属性, 然而 MKnn 对类属型数据的处理依然采用传统 Knn 的方法, 这在一定程度上限制了它在一些应用领域上的推广. 针对这个缺陷, GwMKnn 算法引进类别基尼系数的概念来处理类属性数据, 用基尼系数统计某一类属性中不同值分布对这个类的贡献度作为此类属性的权重, 并以此作为估算不同样本之间的相似性度量对 MKnn 进行优化, 扩宽了 MKnn 的使用面. 然而, 在数据空间中尤其是高维数据空间中, 常常有许多不相关的属性, 使得所寻找的目标类在某些子空间上更加有效, 而不同类别所关联的子空间往往是不一样的. 在利用互 K 最近邻进行噪声消除的过程中, 并没有将类别属性一起考虑进去, 则极其有可能把真实有效的数据当成噪声消除掉, 进而影响分类效果. 同时在许多实际运用中, 数据往往具有很高的维度, 同时不同类别的样本之间可能存在大量重叠.

本文提出基于类别基尼系数子空间的加权互 k 近邻算法, 简记为 GSwMKnn 算法, 先用类别基尼系数求出其对应的软子空间, 然后将待分类样本及训练样本投影到各软子空间中, 再使用类别基尼系数加权距离互 k 近邻算法计算出待分类样本在每个子空间的投票权重, 最后叠加各子空间的投票权重, 最终得出待分类样本的最后类标签.

1 相关工作与背景知识

Knn 分类算法因为其简单直观, 易于实现而被列为十大数据挖掘算法之一[4]. 但在实际应用中同时也存在以下问题: 难以设定较好的 k 值, 常用的相似性度量的距离函数不适用类属性数据. 很多研究人员针对 Knn 的不足做了一系列的改进, 其中 Guo[5,6]提出 KnnModel 算法, 该算法通过使用代表点建立分类模型, 并能在学习过程中自动确定 k 的取值; 陈黎飞等[7]提出多代表点的学习算法 MEC, 使用无监督的局部聚类算法学习优化的代表点集合以提高分类效率. 对于相似度度量上, 基本 Knn 的算法是基于欧拉距离来计算, 造成了 Knn 算法对噪声非常敏感. 孙彩堂等[8]提出了一种基于 RSW KNN 和 WMV 相结合的虹膜识别方法对特征和投票都赋予不同权重对 Knn 进行改进; Gao 等[9]则用互近邻(MNN)去提高移动对象的查询系统的效率. 针对类属型数据的相似性, Liu 和 Pan[10]等在蚁群算法用上基于熵的度量(EBM)而获得更快, 更准确的聚类结果; 而 Nie 和 Kambhampati[11]则利用基于频率的方法建立常被访问属性值的层并学习统计分类等等.

子空间聚类是一种寻找隐藏在不同低维子空间中的聚类技术, 根据加权方式的不同, 子空间聚类一般可分为两大类:硬子空间聚类和软子空间聚类. 硬子空间聚类方法主要用于识别不同类所在的精确子空间, 其属性权值为 0 或者 1. 而软子空间(以下简称为子空间)聚类则是给每类的特征赋予[0, 1]区间的不同权值, 用来表示聚类过程中各特征对此类别贡献的大小, 并以此确定类内的紧凑度. 子空间聚类因为其能有效减少冗余或不相关属性对聚类过程的干扰而提高高维数据集上的聚类效果, 而在近年来成为学术界的研究热点. 子空间概念最早的提出是为了聚类分析, 当然其应该也主要是在聚类上. 如: 单世民等[12]提出一种处理高维分类数据集的 FPSUB 子空间聚类算法, 利用频繁模式树将聚类问题转化为寻找属性值的频繁模式作为候选子空间, 再基于这些子空间进行聚类. 毕志升等[13]运用复合差分演化算法优化结合模糊加权类内相似性和界约束权值矩阵的新目标函数和搜索子空间中的聚类而提出 DESC 算法. Tatu, A.等[14]引入一个可视化的子空间聚类分析系统 ClustNails, 它集成了一些新的带用户交互设备的可视化技术于子空间聚类中. 陈黎飞等[15]在最小化子空间簇类的簇内紧凑度的同时, 最大化每个簇类所在的投影子空间推导出新的局部特征加权方式,并以此为基础提出一种自适应的 k-means 型软子空间聚类算法, 动态地计算最优的算法参数, 极大提高了聚类精度和聚类结果的稳定性. 子空间在分类上的应用相对较少, 如杨明等[16]提出一种基于局

部随机子空间的分类集成算法, 依据特征的贡献大小不同选取特征子空间并以此提高子分类器的性能且增强子分类器的多样性. 张健飞等[17]利用子空间模型簇构造分类模型, 有效分隔了不同样本在全空间中重叠的区域, 以提高分类性能. 李南等[18]提出利用子空间分类算法建立若干个底层分类器, 然后由这几个底层分类器组成集成分类模型的基分类器并且能够适应概念漂移数据流的分类算法. 本文则用基尼类别子空间对互 k 近邻进行改进以提高分类效率.

2 GSwMKnn: 基于类别基尼系数子空间的加权互K近邻算法

2.1 基本概念与定义

2.1.1 互 k 近邻(MKnn)

样本 A 为样本 B 的 k 近邻的同时样本 B 也为样本 A 的 k 近邻, 则称样本 A 和样本 B 互为 k 近邻, Liu[3]等则把这个思路应用于对 Knn 的改进. 其认为对给定一个有 n 个样本的数据集 D, 参数为 k, 则 X_t 的互 k 近邻 $M_k(X_t)$可以表示为: $M_k(X_t)=\{X_i \in D \mid X_i \in N_k(X_t) \wedge X_t \in N_k(X_i)\}$, 即当 X_i 属于另一个样本 X_t 的 k 近邻, 而 Xt 同时属于 Xi 的 k 近邻, 则样本 Xi 是 Xt 的互 k 近邻, 这里 $N_k(X_t)$是 X_t 的 k 近邻的集合.

2.1.2 基于类别基尼系数的互 k 近邻(GwMKnn)

在有 n 个样本的训练集 $X=\{(x_1,y_1),\cdots,(x_i,y_i),\cdots,(x_n,y_n)\}$中, 其中 x_i 是一个 D 维实例, $Y=\{c_1,c_2,\cdots,c_q\}$是训练集 X 的类别集合, y_i 属于 Y, 训练集 X 中属于 c 类的样本在 j 属性上的基尼系数即为类别基尼系数, 记为 Gini(c,j), $Gini(c,j)=1-\sum_{l=1}^{m}P_{jl}^2(c)$, 其中 $P_{jl}(c)$表示 X 中属性 j 上属于 c 类的样本的某种属性值1∈(1,2,⋯,m)出现的概率. 当 X 中属性 j 上属于 k 类的样本只出现一种值时, Gini(c,j)=0, 而当 X 中属性 j 上属于 k 类的样本出现的值为均匀分布时, 则 Gini(c, j)= $\frac{q-1}{q}$, 此时值最大, 即类别基尼系数的值域为[0, $\frac{q-1}{q}$]. 从类别基尼系数的定义可知: 类别基尼系数越大, 则表示属性 j 上属于此类的数据分布就越分散, 样本属于这个类别的可能性就小; 反之, 则属性 j 上属于该类的数据分布就越集中, 样本属于这个类别的可能性就大.

考虑到 MKnn 对距离的度量仍然采用 Knn 的欧拉距离. 这无法对类属型数据做很好的分类, GwMKnn 算法通过对类属性数据引进类别基尼系数的概念, 并以此作为估算不同样本之间的相似性度量方法对 MKnn 进行优化. GwMKnn 算法的基本思想是: 判断待分类样本 x_t 在属性 j 上的数据类型, 若数据类型为类属型且样本 x_t 与样本 x_i 在属性 j 上的值不相等, 则用 x_i 所在的类别基尼系数度量其相似性; 反之则用欧拉距离作为相似性度量.

2.1.3 基尼类别子空间

定义 1. 基尼类别子空间

$$SubSpace_l=(Class_l, Weight_l)$$

其中:

① $Class_l \in Y$; 当 $i=j$ 时, $Class_i \neq Class_j$;

② $Weight_l=\begin{pmatrix} w_{l1} & & & \\ & w_{l2} & & \\ & & \cdots & \\ & & & w_{lD} \end{pmatrix}$,

其中 $\sum_{d=1}^{D} w_{ld}=1; \forall d=1,2,\cdots,D: w_{ld} \geq 0;$

$Weight_l$ 为一个 D 阶的对角矩阵, 与类别为 $Class_l$ 的子空间相对应. 矩阵中的每一个元素代表子空间里某个维度的权重. 维度的权重越大, 表示该维度与类别的相关性越大. 反之, 权重越小则说明该维度与对应类别的相关性越小.

对于连续型数据而言, 在一个训练数据集中, 一般都是根据 q 个类别标号将训练集分成 q 个集合, 然后利用 FWKM 算法中特征权重计算公式计算每个类别对应子空间的权重矩阵 Weight. 因为 GwMKnn 主要处理的是类属型数据的问题. 而类别基尼系数恰恰体现类别与维度之间的关系, 故而本文直接用类别基尼系数计算每个类别对应子空间的权重. 即:

$$w_{ld}=1-\sum_{d=1}^{D}P_{ld}^2(l)$$

2.2 GSwMKnn 算法的基本思想

GSwMKnn 算法根据训练集中的类别信息求得类别基尼系数, 并以此作为基尼类别子空间的维度权重求出其对应的子空间; 再将待分类样本及训练样本投影到各个子空间以加强各样本与类别间的关联性, 使用类别基尼系数加权互 k 近邻算法, 计算出待分类样

本在每个子空间的互 k 近邻, 并统计各子空间的投票权重; 最后叠加各子空间的投票权重, 把其中投票权重最大的类别选出作为待分类样本的类标签.

在实验过程中, GSwMKnn 算法在计算互 k 近邻时所使用的相似性度量方法与 GwMKnn 相同. 还有考虑到混合数据类型的复杂性, GSwMKnn 算法先对连续型数据用分箱法进行离散化处理. 分箱是一种基于箱的指定个数自顶向下的分裂技术, 通过使用等宽或等频分箱, 然后用箱均值或中位数替换箱中的每个值将属性值离散化, 就像分别用箱的均值或箱的中位数光滑一样, 是一种非监督的离散化技术.

2.3 GSwMKnn 的算法实现

GSwMKnn 算法的基本步骤为:

输入: X_t: 待分类样本; K:初始选择的最近邻个数;T={t1,t2,⋯,tn}: 带有 n 个具有类标签样本的训练集合

输出: 待分类样本 Xt 的类别标签

Begin

Step1: 对连续型属性用分箱法进行离散化处理;

Step2: 用矩阵 Sc=diag($w_{l1},w_{l2},\cdots w_{lD}$)创建类别子空间;

$$w_{ld}=1-\sum_{d=1}^{D}P_{ld}^2(l)$$

;

$$\text{其中}\sum_{d=1}^{D}w_{ld}=1;\forall d=1,2,\cdots,D:w_{ld}\geq 0$$

Step3: 将 Xt 及训练样本投影到各子空间中. 用基于类别基尼系数的互 k 近邻(GwMKnn)算法找出 Xt 的互 k 近邻, 统计各个类别信息的投票权重;

Step4: 累加在各个子空间下的投票权重, 然后选择其中值最大的作为待分类样本 Xt 的类别标签;

Step5: 返回待分类样本 Xt 的类别标签;

End.

3 实验与结果分析

为了验证 GSwMKnn 的有效性. 本文选择 Knn、MKnn、GwMKnn 和 GSwMKnn 这 4 种基于最近邻思想的分类器作为对比算法. 由于估计 Knn 算法参数 k 的取值较困难, 实验中通过对比设定 k= 5, 分箱法中所用的分箱参数设为 10. 另外, 实验采用的计算机配置如下: CPUPentium (R)D CPU 2.40GHZ, 内存 1.0GB, 操作系统 Windows XP. 实验中的所有数据都是在上述配置的计算机上运行取得.

3.1 实验数据集

实验采用 10 个数据集对 4 种分类器分别进行测试, 因为本文所用方法主要是针对类属性数据的, 因此所选的 10 个数据集中, 类属型属性数目都比数值型属性数目多. 考虑到子空间方法主要针对高维数据的处理, 所以有 5 个数据集的维度超过 10 维, 两个接近 10 维. 有关这些数据集的基本信息见表 1. 它们均来自 UCI 机器学习公共数据集[19]. 为了在相同的实验环境中进行比较,我们自己设计编写所有的算法, 感兴趣的读者可以发邮件给作者. 为减少不同取值范围对相似度度量的影响, 所有数据集的类别基尼系数最后都经标准化处理变换到[0, 1]区间.

表 1　实验数据集的基本信息

数据集	实例个数	属性数目	数值型属性数目	类属型属性数目	类别数目	类分布
anneal	898	38	6	32	5	8:99:684:0:67:40
balloons	76	4	0	4	2	35:41
breast-cancer	286	9	0	9	2	201:85
hayes-roth	132	4	0	4	3	51:51:30
house	91	17	0	16	2	54:37
house-votes-84	435	16	0	16	2	168:267
monks	124	7	0	6	2	62:62
soybean	683	35	0	35	19	20:20:20:88:44:20:20:92:20:20:20:44:20:91:91:15:14:16:8
tic-tac-toe	958	9	0	9	2	332:626
zoo	101	17	1	16	7	41:20:5:13:4:8:10

3.2 实验结果

为了使实验结果客观稳定, 本次实验采用 10 次 10-折交叉验证方法, 在选择的 10 个数据集上进行训练、测试. 每一次 10-折交叉验证都随机抽样每个数据集并将其平均分为 10 个子集, 每次抽选 9 个子集为训练数据集, 剩余的 1 个子集为测试数据, 轮流选择 10 个子集, 重复 10 次后取 10 次结果的均值作为最终的平均分类准确率, 并且在每次验证中都保证每个算法具有相同的训练集和分类集.

从表 2 可以看出, GSwMKnn 算法在 10 个数据集的实验中有 8 个数据集表现较高的分类准确率. 特别在 anneal、breast-cancer、house、house-votes-84 和 soybean 这 5 个数据集的分类精度明显高于其它分类算法的分类精度, 而这 5 个数据集的维度都相对较高, 这在一定的程度上体现了 GSwMKnn 在处理高维数据上的有效性. 相比之下, 这 5 个数据集也因为其维度相对较高而在 GwMKnn 算法上表现出了较低的分类性能, 这也从另一方面体现了全空间下对高维数据处理能力较弱的特点. Balloons、hayes-roth 和 monks 三个数据集在 GwMKnn 算法上的不俗表现恰恰说明了类别基尼系数对低维分类属性的有效性. 从总体上看, 在 10 个数据集的平均结果中, MKnn、GwMKnn 和 GSwMKnn 的性能都能高过 Knn, 而这其中尤其以 GSwMKnn 最高.

表 2 4 种分类器的分类精度对比

数据集	Knn	MKnn	GwMKnn	GSwMKnn
anneal	97.32	97.88	55.26	**98.64**
balloons	73.68	**73.95**	**73.95**	**73.95**
breast-cancer	73.5	72.8	70.21	**74.2**
hayes-roth	61.14	62.8	**72.42**	61.06
house	95.27	95.6	95.16	**96.26**
house-votes-84	93.13	93.86	86.57	**94.11**
monks	79.19	79.76	**81.37**	78.47
soybean	91.33	92.66	82.12	**93.1**
tic-tac-toe	**91.46**	**91.46**	83.52	**91.46**
zoo	94.95	**95.35**	90.2	**95.35**
各数据集的平均分类时间	174.9	603.6	687.8	**5099.5**

4 结语

文中提出基于类别基尼系数子空间的加权互 K 近邻算法(GSwMKnn), 在互 k 近邻学习算法(MKnn)基础上引进软子空间方法(用基尼系数作为各子空间的权重), 再利用类属性数据加权的互 k 近邻算法求出各子空间的投票权重并叠加, 取出最大投票权重的类别作为类标签. 子空间的引入加深了待分类样本与各个近邻间的紧密性, 降低冗余或不相关属性对分类效果的影响, 提高了算法对高维数据的处理能力. 在 10 个 UCI 机器学习公共数据集上的实验结果表明, GSwMKnn 可有效进行互 k 近邻分类, 其分类精度优于 GwMKnn 算法, 更优于传统的 Knn 算法及其本文所提的其它改进算法, 并大幅提升分类效率. 下一步的工作重点是深入分析影响模型期望风险的其它因素, 提出更有效的优化方法, 以进一步提高分类性能.

致谢

本项目受国家自然科学基金(No. 61070062),福建高校产学合作科技重大项目(No.2010H6007)的资助; 部分受福建省教育厅 B 类项目(No.JB12201)的资助, 特此表示感谢.

参考文献

1 Mitchell TM. Machine Learning McGraw-Hill Companies Inc, 1997: 230–247.

2 Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE Trans. on Information Theory, 1967, J3(1): 21–27.

3 Liu HW, Zhang SC. Noisy data elimination using mutual k-nearest neighbor for classification mining. The Journal of Systems and Software, 2012, 85: 1067–1074.

4 Yang Q, Wu X. 10 challenging problems in data mining research. International Journal of Information Technology and Decision Making, 2006, 5(4): 597–604.

5 Guo GD, Wang H, Bell D,et al. KNN model-based approach in classification. Proc of the OTM Confederated International Conference on CoopIS, DOA, and OD BASE. Atania, Italy. 2003. 986–996.

6 Guo GD, Wang H, Bell D,et al. Using KNN model for automatic text categorization. Soft Computing: A Fusion of Foundations, Methodologies and Application, 2006, 10(5): 423–430.

7 陈黎飞,郭躬德.最近邻分类的多代表点学习算法.模式识别与人工智能,2011,24(6):883–888.

8 孙彩堂,张利彪,周春光,刘小华.加权 K 近邻和加权投票相

(下转第 132 页)

从表 2 中可以看出, 当采用 10 折交叉验证时, 此时每类别取出作为训练集的已知标签的样本数较多, KNN_Improved 算法可以达到比 KNN 算法更好的分类效果. 对于 Glass、Page-blocks 数据集这样的情况, KNN_Improved 算法的分类效果比对数据集 Iris、Liver 分类时提高得要明显.

4 结语

本文针对传统 KNN 算法的不足, 从已有标签的同类别样本的特点出发, 提出了一种结合 K 近邻样本点的类别平均距离对分类进行加权多数投票的方法. 在对 UCI 数据集的实验中, KNN_Improved 算法都与传统 KNN 算法的 F1 值相当或略优, 特别在 K 近邻样本中各类别的样本点总数不均衡或者某些类别样本点数量很少时, KNN_Improved 算法的优势更加明显.

参考文献

1 Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans. on Information Theory, 1967, 13: 21−27.

2 Hart P. The condensed nearest neighbor rule. IEEE Trans. on Information Theory, 1968, 14(3): 515−516.

3 Devijver P, Kittler J. Pattern Recognition: A Statistical Approach. Englewood Cliffs: PrenticeHall, 1982.

4 李荣陆,胡运发.基于密度 KNN 文本分类器训练样本裁剪方法.计算机研究与发展,2004,41(4):539−545.

5 Goldberger J, Roweis S, Hinton G, Salakhutdinov R. Neighborhood components analysis. Proc. of the Advances in Neural Information Processing Systems. Vancouver. Canada, MIT Press. 2004. 512−520.

6 Torresani L, Lee K. Large margin component analysis. Proc. of the Advances in Neural Information Processing Systems. Vancouver. Canada, MIT Press. 2007. 1385−1392.

7 崔正斌,汤光明.基于遗传算法和 KNN 的软件度量属性选择研究.计算机工程与应用,2010,46(30):57−60.

(上接第 141 页)

结合的虹膜识别算法.小型微型计算机系统,2010,1846−1849.

9 Gao Y, Zheng B, Chen G., Li Q, Chen C, Chen G.. Efficient mutual nearest neighbor query processing for moving object trajectories. Information Sciences, 2010, 180: 2170−2195.

10 Liu B, Pan J, McKay RI. Entropy-based metrics in swarm clustering. International Journal of Intelligent Systems, 2009,(24): 989−1011.

11 Nie Z, Kambhampati S. A Frequency-based Approach for Mining Coverage Statistics in Data Integration. http://www.public.asu.edu/~zaiqingn/freqbased.pdf.

12 单世民,王新艳,张宪超.高维分类属性的子空间聚类算法.小型微型计算机系统,2009,30(10):2016−2021.

13 毕志升,王甲海,印鉴.基于差分演化算法的软子空间聚类.计算机学报,2012,35(10):2116−2128.

14 Tatu A, Zhang LS, Bertini E, Schreck T, Keim D, Bremm S, von Landesberger T. ClustNails: Visual analysis of subspace clusters. Tsinghua Science and Technology, 2012, 17(4): 419−428.

15 陈黎飞,郭躬德,姜青山.自适应的软子空间聚类算法.软件学报,2010,21(10):2513−2523.

16 杨明,王飞.一种基于局部随机子空间的分类集成算法.模式识别与人工智能,2012,25(4):595−603.

17 张健飞,陈黎飞,郭躬德,李南.多代表点子空间分类算法.计算机科学与探索,2011,(11):1037−1048.

18 李南,郭躬德.基于子空间集成的概念漂移数据流分类算法.计算机系统应用,2011,20(12):241−248.

19 UCI Repository of Machine Learning Databases. http://repository.seasr.org/Datasets/UCI/arff/. 2012-12-12.