SMwKnn:基于类别子空间距离加权的互k近邻算法

卢伟胜 郭躬德 严宣辉 陈黎飞
(福建师范大学数学与计算机科学学院 福州 350007)

摘 要 互k最近邻算法(mKnnc)是k最近邻分类算法(Knn)的一种改进算法,该算法用互k最近邻原则对训练样本以及k最近邻进行噪声消除,从而提高算法的分类效果。然而在利用互k最近邻原则进行噪声消除时,并没有将类别属性考虑进去,因此有可能把真实有效的数据当成噪声消除掉,从而影响分类效果。基于类别子空间距离加权的互k最近邻算法考虑到近邻的距离权重,既能消除冗余或无用属性对最近邻分类算法依赖的相似性度量的影响,又能较好地消除邻居中的噪声点。最后在UCI公共数据集上的实验结果验证了该算法的有效性。

关键词 类别子空间,互k最近邻,距离加权,子空间
中图法分类号 TP391 **文献标识码** A

SMwKnn:Mutual k Nearest Neighbours Algorithm Based on Class Subspace and Distance-weighted

LU Wei-sheng GUO Gong-de YAN Xuan-hui CHEN Li-fei
(School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China)

Abstract Mknnc is an improved algorithm of the k nearest neighbours (KNN), which uses the mutual k nearest neighbours to eliminate anomalies in the training set and the k nearest neighbours. It has the better performance than KNN. However, the real and effective data may be eliminated as the noises so that influencing the efficiency of classification in the noise elimination stage without taking the class label into consideration. The mutual k nearest neighbours algorithm based on class subspace and distance-weighted (SMwKnn) taking distance-weighted into account can eliminate the influence of the redundant or useless attributes on the similarity measurement of the k nearest neighbours classification algorithm and eliminate the anomalies in the neighbours. The experimental results on the UCI public datasets verify the effectiveness of the proposed algorithm.

Keywords Class subspace, Mutual k nearest neighbour, Distance weighted, Subspace

1 引言

分类是机器学习中的一项重要技术,已被广泛应用于各个领域,比如分本分类领域。胡元等[1]提出了一种基于区域划分的Knn文本快速分类算法,用于提高Knn在文本分类上的分类效率。分类算法一般分为急切(eager)和懒惰(lazy)两种类型[2]。急切类型的分类算法只需建立一次分类模型就可对待分类样本进行分类,而懒惰类型的分类算法,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。目前比较常用的分类算法有k最近邻、决策树、神经网络、贝叶斯分类器、支撑向量机等。

在众多的分类算法中,Knn算法[3]是一种简单的具有较成熟的理论基础并且在分类、回归和模式识别等领域都有着广泛应用的懒分类算法(lazy classifier),属于数据挖掘十大经典算法之一[4]。Knn算法首先依据某种相似性度量(通常为欧几里德度量方法)从训练集中选出k个最相近(最相似)的邻居样本,然后再依据少数服从多数的投票原则(majority voting)从这些邻居中选出最具代表性的类别(邻居样本数量最多的类别)作为待分类样本的类别。考虑到距离待分类样本比较近的邻居样本应该比距离待分类样本比较远的邻居样本在投票中占有更大的分量,Dudani S. A提出了wKnn (The distance-weighted k-nearest neighbor rule)[5],并且证明了wKnn算法相比于原始Knn算法具有较低的错误率、更好的分类表现。不管是Knn还是wKnn,当k取值较小时容易受到噪声、离群点的影响,而当k取值较大时又容易受到来自其它类别样本的干扰。为了消除噪声数据,Liu Huawen等[6]提出了mKnnc算法,该算法用互k最近邻原则对训练样本以及k最近邻进行噪声消除,从而提高了算法的分类效果。然而在数据空间中特别是高维数据空间中,往往有许多不相关的属性,使得要寻找的目标类在某些子空间中是更有效的,而不同的类别其关联的子空间通常也是不一样的。在利用互k最近邻原则进行噪声消除时,并没有将类别属性考虑进去,有可能把真实有效的数据当成噪声消除掉,从而影响分类效果。同时在许多实际应用中,数据往往具有很高的维度,比如文本

到稿日期:2013-05-20 返修日期:2013-07-19 本文受国家自然科学基金(61070062,61175123),福建高校产学合作科技重大项目(2010H6007)资助。

卢伟胜(1990—),男,硕士生,主要研究方向为数据挖掘与人工智能,E-mail:lwsbox@qq.com;郭躬德(1965—),男,博士,教授,主要研究方向为数据挖掘、机器学习;严宣辉(1968—),男,副教授,主要研究方向为人工智能和网络安全;陈黎飞(1972—),男,博士,副教授,主要研究方向为数据挖掘、模式识别。

数据、基因数据等，同时不同类别的样本之间可能存在大量重叠。

本文在距离加权以及互 k 最近邻的基础上提出了一个基于类别子空间距离加权的互 k 最近邻算法 SMwKnn，该算法加强了样本与类别属性的关联度并且在高维数据集上有较好的表现。SMwKnn 算法分别将待分类样本以及训练样本投影到依据各个类别所生成的类别子空间中；然后在每个子空间中采用距离权重的互 k 最近邻方法分类，产生投票结果；最后，综合在每个类别子空间中得到的投票结果，从而产生最终的分类结果。从直观上，该算法加强类属性在训练数据中的作用，增加不同类别样本的区分度，降低了不相关属性在分类中的权重，从而提高了分类的准确性。

2 相关工作

Knn 算法存在两个主要问题，其一是分类速度较慢，并且要占用较大的内存空间；其二是 k 值难以确定，不同的 k 值会影响最终的分类精度，即分类准确率对 k 值的敏感性较强。目前，Knn 已被众多学者广泛研究，并且得到了不少的改进。在 NN(nearest neighbours)上的改进可以被分为 structure less 和 structure based 两类[7]，其中 structure based 主要是利用 kd Tree、PAT(principal axis trees)、CT(Center Line)等数据结构来对 Knn 算法进行改进，与之相对的便是 structure less 类型的。此外，也有学者将其它技术融合到 Knn 中，对传统的 Knn 方法加以改进。张孝飞[8]将 Knn 与聚类算法相结合，首先通过聚类把若干个相似的文档合并成一个中心文档，然后以这些中心文档作为代表去建立分类模型，以提高算法的执行效率。余鹰等[9]运用粗糙集上下近似的概念将各类训练样本划分为核心和边界区域，分类过程计算新样本与各类的近似程度，获取新样本的归属区域，减小分类代价，增强算法的鲁棒性。Guo G. D 提出基于模型的 Knn 算法(KNN-Model)[10]，该算法通过选择代表点建立分类模型，并且能够在学习过程中自动确定 k 的取值。Chen L. F 提出基于 KNN-Model 改进的多代表点学习算法[11]，它使用无监督的局部聚类算法来学习优化的代表点集合，以提高分类效率。Gou J. P 依据双重权重投票方法设计出 DWKNN 算法[12]，该算法在原始的距离权重上再乘上一个排序权重，从而降低了算法对 k 值的敏感性，并且在一个宽广的 k 的取值范围内有具有较好的鲁棒性。Ding C 和 He X. F[13]采用 k-互近邻来提高 K-means 算法的性能，并探讨用它完成聚类和孤立点检测任务。Chidananda K 和 Krishna G[14]利用互 k 最近邻概念对训练集进行预处理，以达到浓缩训练集的效果，进而提高算法的执行速度，降低内存占用率。

子空间聚类算法根据加权方式的差异，可分为硬子空间(hard subspace)和软子空间(soft subspace)聚类两种方法，后者给维度赋予[0,1]区间的权值，表示维度与对应划分之间“模糊的”关联度，是近年来较为活跃的一个研究方向[15]。相比于硬子空间，软件子空间具有较好的灵活性和可伸缩性。软子空间是聚类研究领域一个重要的分支和研究热点，其中 FWKM[16]是一种比较具有代表性的聚类算法。此外，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，并且它能够适应概念漂移数据流的分类算法。

3 背景知识

3.1 互 k 最近邻关系

一个带有 n 个样本的集合 $S=\{s_1,s_2,\cdots,s_n\}$，给定一个值 $k\in\{k|0<k<n,k\in Z\}$，其中每一个样本 s_i 都有一个相对应的 k 最近邻集合 $N_i=\{t_1,t_2,\cdots,t_k\}$，$N_i\in S$。若 s_i 与 s_j 两个样本为互 k 最近邻关系，则有 $s_i\in N_j$ 并且 $s_j\in N_i$。反之则称 s_i 与 s_j 不成互 k 最近邻关系。

对于拥有互 k 最近邻关系的 s_i 和 s_j，我们称 s_i 为 s_j 的互 k 最近邻，s_j 为 s_i 的互 k 最近邻。如果一个集合都是由样本 s 的互 k 最近邻组成的，则称这个集合为样本 s 的互 k 最近邻集合。

3.2 基于权重的互 k 最近邻算法(MwKnn)

MwKnn 算法主要是对已经从训练集中选出的待分类样本的 k 最近邻集合进行进一步处理，消除其中的“伪邻居”，使得待分类样本与邻居之间的关系变得更加紧密，最终得到待分类样本的互 k 最近邻集合，之后再利用距离加权方式的投票原则，算出每个类别在互 k 最近邻集合中占有的距离权重比，从中选出比例最大的类别作为待分类样本的类标签。

在本文的实验中，MwKnn 算法的相似度采用欧几里德距离(见式(1))，在最终的类别投票中使用式(2)作为距离权重计算公式。

$$dist(x_i,x_j)=\sqrt{\sum_{d=1}^{D}(x_{id}-x_{jd})^2} \tag{1}$$

其中，x_{id} 表示样本 x_i 在第 d 维上的属性值。

$$w_i=1/d_i \tag{2}$$

其中，d_i 为待分类样本与互 k 最近邻邻居 t_i 的欧几里德距离。

3.3 MwKnn 算法步骤

Input:

s:待分类样本

k:初始选择的最近邻个数

$T=\{t_1,t_2,\cdots,t_n\}$:带有 n 个具有类标签样本的训练集合

Output:

待分类样本 s 的类标签

Begin

Step1 利用式(1)，找出待分类样本 s 与训练集 T 的 k 个最近邻集合 $N=\{t_1,t_2,\cdots,t_k\}$；

Step2 分别判断待分类样本 s 是否为 t_i(i=1,2,…,k)的 k 个最近邻之一，然后将与样本 s 不成互 k 最近邻关系的 i 个邻居样本从 k 最近邻集合 N 中移除，于是得到筛选后的集合 $N=\{s_1,s_2,\cdots,s_j\}$其中 i+j=k。如果 N 为空集，则保留 N 中与待分类样本 s 最近的一个邻居；

Step3 对剩余的邻居采用距离加权的投票方法，即利用式(2)计算出每个类别所占的权重，最终得到每个类别的权重比；

Step4 选中其中权重最大的类别作为待分类样本 s 的类别标签，输出待分类样本 s 的类标签；

End

3.4 类别子空间

给定 n 个样本组成的训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$以及具有 $m(m>1)$个类别标签的集合 $Y=\{1,2,\cdots,m\}$。其中 x_i 表示训练集中的第 i 个训练样本，并且 x_i 是一个 D 维的空间向量，即 $x_i=\{x_{i1},x_{i2},\cdots,x_{id}\}$。$y_i\in Y$，表示 x_i 对应的类别标号。

定义 1 类别子空间

$SubSpace_l=(Class_l, Weight_l)$，其中：

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

②$Weight_l=\begin{pmatrix} w_{l1} & & & \\ & w_{l2} & & \\ & & \cdots & \\ & & & w_{lD} \end{pmatrix}$，$\sum_{d=1}^{D} w_{ld}=1$；$\forall d=1, 2, \cdots, D$：$w_{ld} \geqslant 0$；

$Weight_l$ 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；权重越小则说明该维度与对应类别的相关性越小。

对于一个训练数据集，我们可以根据类别标号将训练集分成 m 个集合。然后利用 FWKM 算法中特征权重计算公式求得每个类别对应子空间的权重矩阵 $Weight$。具体计算公式如下：

$$w_{ld}=\left[\sum_{t=1}^{D}\left(\frac{\sum_{i=1}^{n}((x_{id}-v_{ld})^2+\delta)}{\sum_{i=1}^{n}((x_{it}-v_{lt})^2+\delta)}\right)^{\frac{1}{\beta-1}}\right]^{-1} \tag{3}$$

其中，v_{ld} 为对应 $Class_l$ 在维度 d 上的中心，δ 为很小的一个值，主要是为了避免分母为 0，β 为加权参数。在后面的实验中 δ，β 分别取 $\delta=10^{-4}$，$\beta=1.5$。

4 基于类别子空间距离加权的互 k 近邻算法(SMwKnn)

4.1 SMwKnn 算法的基本思想

SMwKnn 算法依据类别求得各个子空间，增加不同类别的区分度，相比于单纯的互 k 最近邻选择，进一步增强了邻居之间的关系。SMwKnn 算法首先将训练集依据类别分组，然后分别计算出每个类别子空间的维度权重，再将待分类样本以及训练样本投影到各个子空间中，以便加强各个样本与类别之间的关联性。再在每个对应的子空间中，使用 MwKnn 算法求得各个类别的距离权重比。最后累计各个子空间中的距离权重，选择其中距离权重最大的类别作为待分类样本的类标签。

实验中，SMwKnn 算法采用式(4)作为距离权重计算公式，当样本集投影到对应子空间之后，各样本之间的距离度量为加权欧几里德距离，计算公式如下：

$$Dist_{w_l}(x_i, x_j)=\sqrt{\sum_{d=1}^{D} w_{ld}(x_{id}-x_{jd})^2} \tag{4}$$

4.2 SMwKnn 算法基本步骤

Input:

s:待分类样本

k:初始选择的最近邻个数

$T=\{t_1, t_2, \cdots, t_n\}$:带有 n 个具有类标签样本的训练集合

Output:

待分类样本 s 的类别标签

Begin

Step1 将 n 个训练样本依据对应的类别标签分为 m 个集合$\{T_1, T_2, \cdots, T_m\}$；

Step2 依据式(3) 计算出每个类别对应的特征权重矩阵 $Weight_i$，其中 $i=1, 2, \cdots, m$；

Step3 依据每个特征权重矩阵，使用 MwKnn 算法中的 Step1－Step3(其中以式(4)作为距离计算公式)，计算每个子空间下类别距离权重比值；

Step4 累加在各个子空间下类别的距离权重比值，然后选择其中权重比值最大的类别作为待分类样本 s 的类别标签；

Step5 输出待分类样本 s 的类别标签；

End

5 实验与结果分析

为了验证 SMwKnn 算法的有效性，在实验中加入了 wKnn、MwKnn 以及 SwKnn 算法作为对比参照。将 SMwKnn 算法中使用互 k 最近邻原则对 k 个最近邻进行“伪邻居”消除的步骤省去，即略去 MwKnn 中的 Step2，便可得到 SwKnn 算法。

5.1 实验环境

实验中所采用的机器为笔记本电脑，其详细配置为：CPU 为 Intel(R) Core(TM) i3-2350M 2.30GHz，内存 4GB；所使用的软件包括 Windows7 操作系统、Eclipse 开发平台、JDK1.7、WEKA3.6 的应用程序接口。

5.2 实验数据集

实验使用了 15 个数据集作为测试对象，这些数据集可以从 weka 官方网站上下载的 UCI repository 压缩包获得(http://www.cs.waikato.ac.nz/ml/weka/)，数据集为 arff 格式。为了保证数据集的多样性，数据集中的实例个数、属性数目、类别数目以及类分布情况都有所甄选，其中 Iris、Wine、Monks、Ecoli、Glass 等 15 个公共数据集的具体细节见表 1。

表 1 数据集相关信息

数据集	实例数	属性数	类别数	类分布
Iris	150	4	3	50:50:50
Wine	178	13	3	59:71:48
Monks	124	2	2	62:62
Ecoli	336	7	8	143:77:52:35:20:5:2:2
Glass	214	9	7	70:76:17:0:13:9:29
Heart-statlog	270	13	2	150:120
Diabetes	768	8	2	500:268
Vehicle	848	18	4	212:217:218:199
Ionosphere	351	34	2	126:225
liver-disorders	345	6	2	145:200
Balance-scale	625	4	3	288:49:288
mfeat-morphological	2000	6	10	200:200:200:200:200:200:200:200:200:200
clean1	476	166	2	269:207
Segment	2310	19	7	330:330:330:330:330:330:330
waveform-5000	5000	40	3	1692:1653:1655

5.3 实验结果

10-折交叉验证方法：通过随机抽样将训练集均分成 10 个子集，轮流选择其中的一个子集作为测试集，其余子集作为训练集，将这么一组数据分别作为各个算法的输入，直到做了 10 次之后，求出最终的平均分类准确率，以上为 1 次 10-折交叉验证。本次实验进行 10 次的 10-折交叉验证，取 10 次结果的均值作为最终实验的最终结果，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。

在实验过程中，k 值分别取 30 以内的奇数，表 2 中的数

据为各算法在15个不同 k 值中具有的最高的分类准确率，括弧里的整数表示对应 k 值。

表2 实验中各算法的最优分类准确率

数据集	wKnn	MwKnn	SwKnn	SMwKnn
Iris	96.47(07)	95.87(07)	96.00(05)	96.13(23)
Wine	97.70(23)	96.46(13)	98.26(01)	98.76(03)
Monks	78.63(09)	78.63(09)	78.71(09)	78.79(09)
Ecoli	87.08(09)	87.47(29)	86.82(07)	86.96(29)
Glass	70.33(03)	71.96(03)	68.32(07)	69.21(21)
Heart-statlog	82.93(29)	80.81(29)	82.74(27)	82.41(29)
Diabetes	75.04(29)	74.49(29)	76.91(29)	75.95(29)
Vehicle	71.29(07)	72.48(05)	72.16(07)	72.55(07)
Ionosphere	86.89(01)	89.09(27)	88.60(03)	91.85(29)
Liver-disorders	65.28(15)	62.70(25)	68.14(19)	65.86(23)
balance-scale	90.10(21)	90.18(17)	90.10(21)	90.26(17)
mfeat-morphological	73.00(09)	73.08(27)	74.58(11)	74.40(19)
Cleanl	85.88(01)	88.82(05)	89.96(01)	91.32(03)
Segment	97.14(01)	97.14(01)	97.17(03)	97.48(03)
waveform-5000	83.98(29)	79.71(29)	85.24(29)	82.26(29)

从表2的数据中可以看出，在15个数据集的实验中，SMwKnn算法在其中的7个数据集上具有最高的分类准确率。与MwKnn相比，SMwKnn在13个数据集上具有较高的分类准确率，所以将样本投影到各个类别子空间的方法有效地提高了MwKnn的分类准确率，能够得到关系更为紧密的邻居样本，降低了剔除较好邻居的风险。从SwKnn与wKnn分类准确率的对比中可以看出，SwKnn算法的分类准确率总体优于wKnn算法，这也说明了基于子空间的改进思想是有效的。

对于高维数据集，如vehicle、ionosphere、cleanl、segment等数据集，它们的属性数目都是在十几维以上。从实验结果可以得出，在高维数据集中SMwKnn算法都具有较好的表现，其中对于高维的waveform-5000数据集，综合MwKnn以及SwKnn的表现可以得出SMwKnn算法在对 k 最近邻进行噪声消除时可能把较好的邻居样本误判为"伪邻居"而删除掉，从而影响整体的分类准确率。由此得出，SMwKnn算法能够较好地处理高维的数据集。

图1所展示的是4个算法在15个数据集上的平均分类准确率。从图中可以看出，SMwKnn在总体上具有较好的分类准确率，其次为SwKnn算法。当 k 取值超过7时，SMwKnn算法的分类准确率趋于平稳，并不会因 k 值的变化而产生较大的波动，进而降低了对 k 值的依赖性。相反，原始的wKnn算法受 k 值的影响较大，随着 k 取值的增大，平均分类准确率下降明显，可见 k 值的影响因素很大。

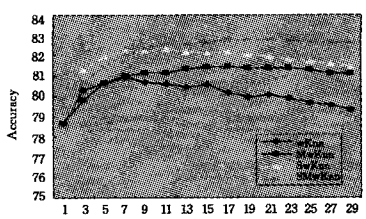

图1 实验中各算法平均分类准确率曲线图

综上，SMwKnn算法具有如下优点：1)分类准确率在总体上有了一定的提升；2)随着 k 值的增大，能够减少来自其它类别的样本的干扰，保持相对稳定的分类效果，降低对 k 值的依赖性，增强了算法的鲁棒性；3)能够较好地处理高维的数据集，降低不相关属性值对分类效果的影响；4)可结合性强，可以很容易地将该算法思想与其它Knn改进方法进行结合；5)改进算法简单，易于实现。

结束语 本文提出了基于类别子空间距离加权的SMwKnn算法，其以wKnn为基础，为样本集合生成若干个类别子空间，然后将样本投影到各个子空间进行分类，并且运用互 k 最近邻对 k 个最近邻进行筛选，进一步加深待分类样本与各个邻居间的紧密性，增强样本与类别之间的关联性，降低不相关属性对分类效果的影响，同时减少算法对于 k 值的依赖性，并且在总体上提高了分类准确率。在15个公共数据集上的实验结果证实，SMwKnn算法具有较好的分类效果，验证了SMwKnn算法的有效性。进一步的研究方向拟考虑将SMwKnn算法的思想与其它改进型的Knn算法进行结合，以求得更好的分类效果。

参考文献

[1] 胡元，石冰. 基于区域划分的kNN文本快速分类算法研究[J]. 计算机科学，2012，10：182-186

[2] Mitchell T M. Machine Learning[M]. McGraw-Hill Companies Inc，1997：230-247

[3] Cover T M，Hart P E. Nearest Neighbor Pattern Classification [J]. IEEE Trans on Informati on Theory，1967，J3(1)：21-27

[4] Wu X D，Kumar V，Quinlan J R，et al. Top 10 algorithms indata mining[J]. Knowl Inf Syst，2008，14：1-37

[5] Dudani S A. The Distance-weighted kNearest Neighbor Rule [J]. IEEE Transactions on System，Man and Cybernetics，1976，SMC-6(4)：325-327

[6] Liu H W，Zhang S C. Noisy data elimination using mutualk-nearest neighbor for classification mining[J]. The Journalof Systems and Software，2012(85)：1067-1074

[7] Bhatia N，Vandana. Survey of nearest neighbor techniques[J]. International Journal of Computer Science and Information Security，2008，8(2)：302-305

[8] 张孝飞，黄河燕. 一种采用聚类技术改进的KNN文本分类方法[J]. 模式识别与人工智能，2009，22(6)：936-940

[9] 余鹰，苗夺谦，刘财辉，等. 基于变精度粗糙集的KNN分类改进算法[J]. 模式识别与人工智能，2012，25(4)：618-623

[10] Guo G D，Wang H，Bell D，et al. KNN Model-Based Approach in Classification[J]. Proc of the OTM Confederated International Conference on CoopIS，DOA，and OD BASE. Catania，Italy，2003：986-996

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

[12] Gou J P，Xiong T S，Kuang Y. A Novel Weighted Votingfor K-Nearest Neighbor Rule[J]. Journal of Computers，2011，6(5)：833-840

[13] Ding C，He X F. K-nearest-neighbor consistency in data-clustering：incorporating local information into global optimization[C]// Proceedings of ACM Symposium on Applied Computing (SAC)，2004：584-589

[14] Chidananda K，Krishna G. The condensed nearest neighbor or rule using the concept of mutual nearest neighbor[J]. IEEE Trans on Information Theory，1979，IT-25：488-490

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

[16] Huang J Z，Ng M K，Rong H，et al. Automated variable weighting in k-means type clustering[J]. IEEE Trans on Pattern Analysis and Machine Intelligence，2005，27(5)：657-668

[17] 张健飞，陈黎飞，郭躬德，等. 多代表点子空间分类算法[J]. 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[J]. 计算机系统应用，2011，20(12)：241-248