基于 PCA 和 K-均值聚类的有监督分裂层次聚类方法*

浦路平[1,2]，赵鹏大[1]，胡光道[1]，张振飞[1]，夏庆霖[1]
(1.中国地质大学 遥感地质与数学地质所，武汉 430074; 2.桂林工学院 现代教育中心，广西 桂林 541004)

摘 要：提出了一种新的基于 PCA 和 K-均值聚类的有监督二叉分裂层次聚类方法 PCASHC，用 K-均值聚类进行逐次二叉聚簇分裂，选择 PCA 第一主成分相距最远样本点作为 K-均值聚类初始聚簇中心，解决了 K-均值聚类初始中心随机选择导致结果不确定的问题，用聚簇样本类别方差作为聚簇样本不纯度控制聚簇分裂水平，避免过拟合，可学习到合适的聚类数目。用四组 UCI 标准数据集对其进行了 10 折交叉验证分类误差检验，与另外七种分类器相比说明 PCASHC 有较高的分类精度。

关键词：数据挖掘；机器学习；有监督聚类；分裂层次聚类
中图分类号：TP301 **文献标志码**：A **文章编号**：1001-3695(2008)05-1412-03

PCA and K-means based supervised split hierarchy clustering method

PU Lu-ping[1,2], ZHAO Peng-da[1], HU Guang-dao[1], ZHANG Zhen-fei[1], XIA Qing-lin[1]
(1. *Research Institute of Mathematical Geology, China University of Geosciences, Wuhan* 430074, *China;* 2. *Modern Education Technical Center, Guilin University of Technology, Guilin Guangxi* 541004, *China*)

Abstract: The paper presented a new supervised bin-split hierarchy clustering method, PCASHC(PCA split supervised hierarchy clustering). The method bin-splited cluster by K-means clustering with initial centers undertaken by the samples of maximum and minimum of first principal component of principal component analysis of the cluster, which solve the problem of uncertain result as a result of the uncertain choice of initial centers. In the method, the variance of the classes of the samples in cluster was chose as measure of impurity of cluster samples class, which controls the slip level of cluster, avoid over-fitting and can find out the proper number of clusters. The method tested with 10-fold cross validation for classifying of 4 UCI datasets. It proves the method has excellent classifying accuracy rate comparing of the error rate of it to other 7 representative classifiers for classifying of same datasets with same test way.

Key words: data mining; machine learning; supervised clustering; split hierarchy clustering

0 引言

聚类分析依照物以类聚原理将研究对象分组，可以提供样本分布的结构信息，是一种重要数据挖掘方法，在自然科学和社会科学中得到广泛应用。经典聚类方法是无监督学习方法，要预先指定聚簇数目，如果聚簇数目不正确，无法得到正确聚类结果。因此正确的聚簇数目是很重要的聚类参数和样本结构信息，从样本特征数据中学习到合适的聚簇数目意义重大。

K-均值聚类方法和层次聚类方法都需要提供正确的聚簇数目。前人曾用逐步增加聚簇数目的 K-均值聚类或层次聚类方法寻找正确的聚簇数目，但拐点不明显时无法使用[1]。

为了通过数据挖掘从样本特征数据中学习到正确的聚簇数目，可以利用带有类别标签的样本进行有监督聚类。有监督聚类因有样本类别标签分布信息的教师监督信号，极大地降低了信息的不确定性，工作效率较高，分类结果为明确的真实类别，能反映出子类等样本分布结构。

有监督聚类的目的是找出划分样本为聚簇内样本纯度大而数量尽可能少的聚簇聚类方案。现有多种形式，如学习向量量化网络[2,3]、基于划分和增量的动态聚类方法[4,5]、支持向量机[5]等。学习向量量化网络在竞争学习网络中按分类结果对错进行奖惩来调整权值学习。基于划分和增量的动态聚类方法常用聚簇内类别不纯度惩罚指标最小化方法。支持向量机结合样本类别的约束信息，通过核函数非线性映射到高维希尔伯特空间，使其在新的空间中同类样本相聚一起，异类样本分离加大，可以用超平面划分，实现有监督聚类。这些方法在要求指定聚簇数目、学习及分类效率和提供显式的子类分布结构信息上各有长短。

K-均值聚类(又称 C-均值聚类)是一种普遍采用的基于划分的动态聚类方法，是在选定的相似性距离度量和评价聚类结果质量的准则函数基础上给定某个初始分类后，用迭代算法找出使准则函数取极值的最好聚类结果[1]。其最佳初始划分尚无解决良方，现多用随机方法，有较大不确定性。

非监督的增量逐次 K-均值聚类法有时可以学习聚簇数目。它是通过逐渐增加聚簇数目 K 和进行 K-均值聚类法，直

收稿日期：2007-02-30；修回日期：2007-07-30 基金项目：国家自然科学基金资助项目(402721122)；广西教育厅资助项目(桂教科研[2004]4 号)

作者简介：浦路平(1958-)，男，江苏南通人，博士，主要研究方向为矿产资源定量预测及勘察评价、GIS 应用和地学数据挖掘(puluping@sina.com)；赵鹏大(1931-)，男，辽宁清原人，中国科学院院士，教授，博导，主要研究方向为矿产普查与勘探学和数学地质学；胡光道(1945-)，男，北京人，教授，博导，主要研究方向为矿产资源定量预测及勘察评价、资源环境遥感研究、地质资源信息系统软件研究、GIS 及计算机应用软件开发等；张振飞(1961-)，男，教授，博士，主要研究方向为矿产勘察和矿床统计预测；夏庆霖(1968-)，男，副教授，博士，主要研究方向为矿产勘察和矿床统计预测.

到评价聚类结果质量的准则函数值对 K 的变化率达到一个拐点时停止,此时的 K 作为正确的聚类数目。如果没有明显的拐点,则此法失效。

层次聚类分析也是一种普遍采用的主要聚类方法[1,6,7],用指定的样本相似性距离度量和聚簇间相似性距离度量,用合并或分裂手段,把样本从每个样本自成一簇到所有样本全为一簇的多级层次聚簇树,但要靠人为指定聚簇数目等参数来将其划分为若干子聚簇。

合并层次聚类算法计算复杂度较大,为固定的 $O(N^2)$,只能用于中小样本学习;分裂层次聚类法可用 K-均值聚类法等基于划分的动态聚类方法进行分裂,计算复杂度随样本分布情况而变化,最好时与 K-均值聚类法相同,为 $O(N)$,多数近于 $O(N\log_2(N))$,极为罕见的极端分布最差时为 $O(N^2)$。因为是在已有聚簇基础上进行继续分裂,所以比每次从头开始的增量逐次 K-均值聚类法计算量要小。

用有监督逐步增加聚簇数目的 K-均值聚类或层次聚类方法可以找到正确的聚簇数目,但合并层次聚类方法计算复杂度较大。因 K-均值聚类初始化困难而多用随机初始化,带来了 K-均值聚类结果不确定问题。

为此本文提出了一种新的有监督聚类方法,即主成分有监督层次聚类方法(PCA supervised hierarchy clustering, PCASHC)。它用聚簇内样本不纯度作为停止分裂的准则函数进行逐次二叉层次分裂,以聚簇样本类别方差作为不纯度测度,聚簇分裂用两类 K-均值聚类方法,用 PCA 第一主成分进行确定性初始化的 K-均值聚类,消除了通常 K-均值聚类因随机初始化引起的聚类结果不确定性,可学习到合适的聚簇数目,学习效率较高。用多组 UCI 标准数据对其进行了检验,其结果与其他七种分类器比较,证明此方法有较高的分类精度。

1　原理和算法

1.1　原理

有监督聚类的目标是划分出类别不纯度最小的尽可能少的聚簇集合。分裂层次有监督聚类是从所有样本为一类开始不断分裂聚簇成多个子聚簇,直到聚簇样本类别不纯度小于指定阈值时停止。用 K-均值聚类分裂层次有监督聚类是用 K-均值聚类方法把聚簇分裂成两个或多个子聚簇。K-均值聚类的主要问题是初始化困难和随机初始化带来的结果不确定性问题,这可用主成分分析方法解决。

主成分分析(principal component analysis, PCA)是一种把原来由多个变量表示的样本转换为可用较少的互不相关的新综合变量表示的统计方法。新的综合变量由多个原有变量线性组合而成,称为主成分,可以通过计算特征值方法求得。然后在有用信息丢失最少的原则下保留特征值大的那部分主成分,舍弃那些仅含少量信息的主成分,从而达到降低维数的目的。其公式推导如下:

设有样本集合 $X=\{x_1, x_2, \cdots, x_n\}$, $x_i\in$瓗^d, $i=1,2,\cdots,N$。其均值向量 M 的分量为

$$m_j = N^{-1}\sum_{i=1}^{N} x_{ij} \tag{1}$$

如果用正交归一矩阵 U^{T} 定义一对线性变换:

$$V = U^{\mathrm{T}}(X-M)$$

$$X = UV + M \tag{2}$$

各样本向量到样本均值向量 M 的距离平方和的期望是

$$J = E(VV^{\mathrm{T}}) = \mathrm{E}(U^{\mathrm{T}}(X-M)(X-M)^{\mathrm{T}}U) =$$

$$E(UCU^{\mathrm{T}}) = 1/D\sum_{j=1}^{D} u_j^{\mathrm{T}} C^{\mathrm{T}} u_j$$

式中协方差矩阵　　$C=(X-M)(X-M)^{\mathrm{T}}$　　(3)

因为 U^{T} 是正交归一矩阵,所以其元素应满足

$$u_i^{\mathrm{T}}u_j = \begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases} \tag{4}$$

用拉格朗日乘子法可以求出在满足式(4)正交条件约束下,求 J 取最大极值的变换 U,所有点到这条直线的垂直距离的平方和最小,而在这条直线上的方差(投影分布区间)最大。令

$$G(U) = \sum_{j=1}^{D} u_j^{\mathrm{T}} C^{\mathrm{T}} u_j - \sum_{j=1}^{D}\lambda_j(u_j^{\mathrm{T}}u_j - 1);\ j=1,2,\cdots,D \tag{5}$$

将式(5)对 $u_j(j=1,2,\cdots,D)$ 求导数并令其等于 0:

$$\partial G(U)/\partial u_j = (C-\lambda_j I)u_j = 0;\ j=1,2,\cdots,D \tag{6}$$

解此方程可得协方差矩阵 C 的特征值 $\lambda_1, \lambda_2, \cdots, \lambda_D$。其中:$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_D$。

若最后 k 个特征值 $\lambda_{D-k+1}, \lambda_{D-k+2}, \cdots, \lambda_{D-1}, \lambda_D$ 为 0,则其相对应的特征向量和主成分也为 0:$u_j=0$; $v_j=0$; $j=D-k+1, D-k+2, \cdots, D-1, D$。即变换矩阵 U 降维成 $U_s\in D\times D_s$,式中余维数 $D_s=D-k$,对应的主成分 V 降维成 $V_s\in D_s\times N$。

主成分分析中最大特征值 λ_1 对应的第一主成分 u_1 在样本属性空间方差最大,延伸最长,变量载荷最大,拥有样本信息量最大。根据这个特点,可用相距最远的聚簇样本第一主成分最大值和最小值作为两个初始聚簇中心,进行两类 K-均值聚类。由于这是确定性过程,解决了 K-均值聚类方法分裂时其初始化聚簇难以确定和因初始值随机选取而产生的结果不确定问题。

在样本属性向量空间中,每个样本为一点。两个样本点之间距离表示其不相似的程度,相距越近越相似。聚类是把相似的样本划为一组。但相似是相对的,所以聚类可以有不同层次级别:从每个样本各自为一聚簇到所有样本全为一聚簇,如果后者为拟合不足的话,前者则可能是拟合过度了,一般是介于这两者之间的某个划分。如何判断是最佳拟合的聚簇划分呢?带类别样本的分布提供了从分类角度判断最佳聚簇划分的信息。

在不断分裂的层次聚类过程中可以通过类别不纯度及其阈值来控制拟合程度:当聚簇样本类别的不纯度小于阈值时,聚簇停止分裂;否则继续分裂成更小的子聚簇。

聚簇样本类别方差 var(y) 可以表示聚簇样本的类别不纯度,因此将其作为测度聚簇类别不纯度的指标。

设有样本属性向量及其类别集合 $D=(X,Y)=((x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n))$, $x_i\in\mathbb{R}^d$, $y_i\in\{1,2,\cdots,c\}$, $i=1,2,\cdots,N$。

$$m_y = N^{-1}\sum_{i=1}^{N} y_i$$

$$\mathrm{var}(y) = N^{-1}\sum_{i=1}^{N}(y_i-m_y)^2 \tag{7}$$

1.2　PCA 有监督分裂二叉层次聚类算法

输入:一组 n 个 m 维 c 类训练样本对象集合 $D=X\times Y=\{(x_i,y_i)\}_{i=1}^{N}$。其中:样本的属性 $x_i\in R^m$;样本的类别 $y_i\in\{1,2,\cdots,c\}$, $i=1,2,\cdots,N$;聚簇样本的类别不纯度阈值 $T\geq 0$。

输出:各样本对象所属的聚簇号集合 $Z=\{z_i\}$, $z_i\in\{1,2,\cdots,h\}$, $i=1,2,\cdots,n$。

方法:

a) 聚簇 $a\leftarrow${所有样本编号},聚簇集合 $A\leftarrow\{a\}$,聚簇集合

$E \leftarrow \{\}$。

b) 如果聚簇集合 A 为空结束, 转向 e); 否则继续。

c) 对聚簇集合 A 中每个聚簇 a_k 依次作:

(a) 从 A 取出聚簇 a_k 作为当前工作聚簇 B: $B \leftarrow a_k$, $A \leftarrow A - \{a_k\}$。

(b) 用式(7) 计算聚簇 B 的类别方差 var($y(B)$)。如果 var($y(B)$) $< T$, 则 $E \leftarrow E \cup \{B\}$。

(c) 用 PCA K-均值聚类法分裂聚簇 B。

①按式(3) 和(6) 分别计算聚簇 B 中样本 $x(B)$ 的均值向量 M 和 PCA 变换矩阵 U;

②用第一主成分变换向量 u_1 计算聚簇 B 内各样本第一主成分值 $v_1 = u_1^{\mathrm{T}}(x(B) - M)$;

③寻找聚簇 B 内第一主成分值的最小值和最大值 $v_{\min}$、$v_{\max}$对应的样本向量 $x_{\min}$、$x_{\max}$;

④以 $x_{\min}$、$x_{\max}$为二聚簇初始中心, 用 K-均值聚类把聚簇 B 分为 B_1 和 B_2 两个聚簇, $A \leftarrow A \cup \{B_1, B_2\}$。

d) 转向 b);

e) 为聚簇集合 E 中各样本 z_i 赋予所在聚簇编号, 返回 Z。

2 实验

实验所用程序为 MATLAB 6.5 编程实现的算法, 计算用的计算机 CPU 为 Pentium 1.73 GHz, 内存 512 MB, 操作系统为 Windows XP。

2.1 标准数据集测试

2.1.1 数据集

实验测试数据选择了四个 UCI 数据集: Iris、Glass、Wine 和 Ecoli[8], 所有属性均为连续数值型, 其各参数如表 1 下部所示。

表 1 七种分类器和 SHCC 对这四组 UCI 数据集作 10 折交叉验证分类误差率(除 SHCC 外的所有数据来源于文献[9])

分类器	Ecoli	Glass	Iris	Wine
libSVM-Linear	0.134	0.070	0.040	0.045
libSVM-Poly	0.172	0.687	0.467	0.117
libSVM-RBF	0.152	0.088	0.047	0.017
Na ve Bayes	0.140	0.159	0.053	0.028
C4.5	0.193	0.023	0.047	0.062
BP	0.143	0.047	0.027	0.039
Perceptron	0.238	0.389	0.327	0.140
PCASHC	0.128	0.032	0.023	0.019
样本集特征数	4	9	13	7
样本集类数	3	6	3	8
样本集样本数	150	214	178	336

2.1.2 PCASHC 分类方法

用 PCASHC 把已知类别训练样本聚类成若干聚簇, 用 MATLAB 统计工具箱的线性分类器 classify 和训练样本及其聚簇类把待测样本分类成聚簇类别, 按聚簇类原来的模式类别变换成模式类别。

2.1.3 测试方法

本实验以 10 组交叉验证的方式, 将样本材料随机分成 10 组, 每组轮流当测试样本, 其余为训练样本, 如此执行完 10 次后, 得到了 10 组分类误差率, 共进行 10 次, 把分类误差率作为该样本集的平均误差率。

2.1.4 实验结果分析

用目前具有一定代表性的七种分类器对此四个数据集进行分类的 10 折交叉验证结果进行了比较, 对比数据来源于网页[9]。

由表 1 和图 1 可见, 在分类器对 Ecoli、Glass、Iris 和 Wine 四种数据集的分类误差当中, PCASHC 有两项最好(其中对 Ecoli 数据集各种分类器误差都较大时 SHCC 误差最小), 两项第二, 说明 PCASHC 分类精度较高。

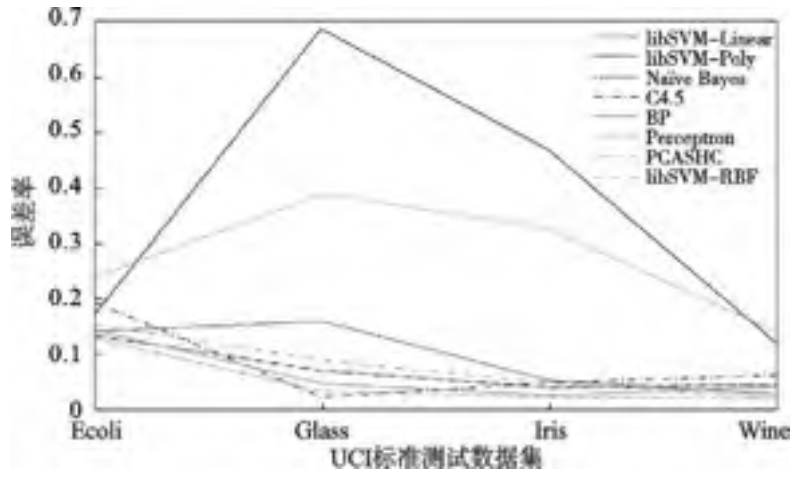

图 1 七种分类器和 PCASHC 对这四个数据集分类 10 折交叉验证的分类误差率

3 结束语

理论和实验说明: a) 基于 PCA 和 K-均值聚类的有监督二叉分裂层次聚类方法是一种性能良好的有监督学习方法, 可由样本类别分布信息自动学习到聚簇数目。b) 用 PCA 第一主成分相距最远的二极端值样本点作为初始聚簇中心来作二叉分裂 K-均值聚类可得到确定性结果, 避免了 K-均值聚类的初值不确定性。c) 用聚簇样本类别方差作为聚簇样本不纯度控制聚簇分裂水平, 使之达到最佳拟合, 可自动学习到有监督聚类的最优化聚簇划分, 无须人为指定聚簇间距离阈值、指定聚簇数目或划分最大树深度等划分聚簇的参数。聚簇样本类别方差阈值默认值为 0, 此时聚类结果为类别纯净的同类聚簇。d) PCASHC 与分类器组合成的有监督层次聚类分类器对四组 UCI 标准数据的 10 组交叉验证分类结果与七种其他代表性的分类器比较, 具有较高的分类精度。

参考文献:

[1] 边肇祺, 张学工. 模式识别[M]. 2 版. 北京: 清华大学出版社, 2000: 235-237.

[2] KOHONEN T. The self-learning map[J]. Proc of IEEE, 1990, 78: 1464-1480.

[3] 程剑锋, 徐俊艳. 基于 EM 算法的有监督 LVQ 神经网络及其应用[J]. 系统工程与电子技术, 2005, 27(1): 121-123.

[4] 宋彤, 宋保强. 一种新的监督聚类学习方法及其在故障诊断中的应用[J]. 计算机工程与科学, 2001, 23(5): 63-69.

[5] DETTLING M, BUHLMANN P. Supervised clustering of genes [C]//Proc of the 15th Conference in Computational Statistics. 2002.

[6] DUDA R O, HART P E, STORK D G. 模式分类 [M]. 北京: 机械工业出版社, 2003: 442-447.

[7] 赵鹏大, 胡旺亮, 李紫金. 矿床统计预测[M]. 北京: 地质出版社, 1983: 157-161.

[8] NEWMAN D J, HETTICH S, BLAKE C L, *et al*. UCI repository of machine learning databases[EB/OL]. (2006-10-06). http://www.ics.uci.edu/~mlearn/MLRepository.html.

[9] [EB/OL]. (2006-10-06). http://finalfantasyxi.inf.cs.cmu.edu/MATLABArsenal/MATLABArsenal.htm.

[10] FINLEY T, JOACHIMS T. Supervised clustering with support vector machines[C]//Proc of the 22nd International Conference on Machine Learning. Bonn: [s.n.], 2005.