基于中心距离比值准则的无监督特征选择算法

叶 菲,罗景青,俞志富
YE Fei,LUO Jing-qing,YU Zhi-fu

解放军电子工程学院,合肥 230037
PLA Electronic Engineering Institute,Hefei 230037,China
E-mail:yefeixyz@163.com

YE Fei,LUO Jing-qing,YU Zhi-fu.Unsupervised feature selection algorithm based on center distance ratio principle. Computer Engineering and Applications,2009,45(4):162-164.

Abstract: Feature selection is an important component of pattern recognition.For unknown class label samples set,an unsupervised feature selection algorithm based on center distance ratio principle is proposed.The algorithm uses the mountain method to get the range of clustering number and estimate original clustering centers,then K-means clustering algorithm is adopted to confirm the optimal classification number of feature subset,and then center distance ratio principle is used to measure the classification performance of feature subset,moreover the feature correlation is analyzed,so the features with good class effect and low correlation are selected.

Key words: feature selection;center distance ratio;correlation;clustering;unsupervised

摘 要:特征选择是模式识别中的一个重要组成部分。针对未知类标号的样本集,提出基于中心距离比值准则的无监督特征选择算法。该算法利用爬山法确定聚类数目范围和估计初始聚类中心,再通过K-均值聚类算法确定特征子集的最佳分类数,然后用中心距离比值准则来评价特征子集的分类性能,并通过特征间的相关性分析,从中选择出分类效果好,相关程度低的特征组成特征子集。

关键词:特征选择;中心距离比值;相关性;聚类;无监督
DOI:10.3778/j.issn.1002-8331.2009.04.046 **文章编号**:1002-8331(2009)04-0162-03 **文献标识码**:A **中图分类号**:TN957.51

1 引言

特征选择是模式识别、机器学习和数据挖掘等领域中的关键问题。在特征形成和提取阶段,为确保提供足够多的分类信息,原始特征数目一般比较多,维数也较高,其中不可避免地存在大量相关或冗余的信息[1]。因此,需要根据特征不同组合所形成的分类识别能力,选择出具有最大判别能力的特征组合,也就是进行特征选择。

本文针对未知类标号的样本集,提出一种基于中心距离比值准则的无监督学习特征选择方法,其基本思想是对每一个特征子集利用爬山法得到聚类数目的上限和初始聚类中心,再由K-均值聚类算法确定其最佳分类数,然后以中心距离比值准则设定一个判断函数用于特征选择,最后从选择出的特征子集中删除掉相关性较大的特征。

2 中心距离比值准则

在描述分类判定准则前,给出准则中涉及到的一些定义。

定义 1 某一类样本的平均值称为该类模式的中心 $\boldsymbol{c}$,已知样本向量集$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,那么其中心为:

$$\boldsymbol{c}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \tag{1}$$

定义 2 两个样本之间的特征差异称为样本距离,已知两个 N 维样本向量 $\boldsymbol{x}_1,\boldsymbol{x}_2$,其样本距离为:

$$d(\boldsymbol{x}_1,\boldsymbol{x}_2)=\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2=\sqrt{\sum_{i=1}^{N}(x_1^i-x_2^i)^2} \tag{2}$$

定义 3 样本的中心距离指的是各样本到模式中心的距离。假设有两类模式的 N 维样本向量集,分别为$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$和$\{\boldsymbol{x}_1',\boldsymbol{x}_2',\cdots,\boldsymbol{x}_m'\}$,其中心分别为 $\boldsymbol{c}_x$ 和 $\boldsymbol{c}_{x'}$,则样本的中心距离有4个,分别是 $\boldsymbol{x}$ 到 $\boldsymbol{c}_x$ 的距离 d_{xx},$\boldsymbol{x}$ 到 $\boldsymbol{c}_{x'}$的距离 $d_{xx'}$,$\boldsymbol{x}'$到 $\boldsymbol{c}_x$ 的距离 $d_{x'x}$,$\boldsymbol{x}'$到 $\boldsymbol{c}_{x'}$的距离 $d_{x'x'}$。其中 d_{xx} 和 $d_{x'x'}$称为样本的自中心距离,$d_{xx'}$和 $d_{x'x}$ 称为样本的互中心距离[2-3]。

$$d_{xx}(\boldsymbol{x},\boldsymbol{c}_x)=\|\boldsymbol{x}-\boldsymbol{c}_x\|_2=\sqrt{\sum_{i=1}^{N}(x^i-c_x^i)^2} \tag{3}$$

$$d_{xx'}(\boldsymbol{x},\boldsymbol{c}_{x'})=\|\boldsymbol{x}-\boldsymbol{c}_{x'}\|_2=\sqrt{\sum_{i=1}^{N}(x^i-c_{x'}^i)^2} \tag{4}$$

$$d_{x'x}(\boldsymbol{x}',\boldsymbol{c}_x)=\|\boldsymbol{x}'-\boldsymbol{c}_x\|_2=\sqrt{\sum_{i=1}^{N}((x')^i-c_x^i)^2} \tag{5}$$

作者简介:叶菲(1980-),女,在读博士生,主要研究方向为雷达信号处理、智能信息处理等;罗景青(1957-),男,博士,教授,博士生导师,863专家,主要研究方向为电子对抗、空间信号与信息处理、电磁环境中的数据融合技术等。

收稿日期:2008-01-08 **修回日期**:2008-04-16

$$d_{x'x'}(\boldsymbol{x}',\boldsymbol{c}_{x'})=\|\boldsymbol{x}'-\boldsymbol{c}_{x'}\|_2=\sqrt{\sum_{i=1}^{N}((x')^i-c_{x'}^i)^2} \tag{6}$$

定义 4 样本的中心距离比值：假设样本的分类数为 k，则某一样本 $\boldsymbol{x}$ 自中心距离与其互中心距离均值的比值称为该样本的中心距离比值。

$$Ratio_x=\frac{d_{xx}}{\frac{1}{k-1}\sum d_{xx'}} \tag{7}$$

定义 5 模式的中心距离比值：设 i 类模式的样本集为$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{n_i}\}$，则该模式中所有样本中心距离比值的均值为该模式的中心距离比值。

$$R_i=\frac{1}{n_i}\sum_{j=1}^{n_i}Ratio_{x_j} \tag{8}$$

定义 6 中心距离比值指标：设样本的分类数为 k，则 k 类模式中心距离比值的平均值为中心距离比值指标。

$$CR=\frac{1}{k}\sum_{i=1}^{k}R_i \tag{9}$$

中心距离比值准则是 CR 的值越小，说明分类的效果越好。

3　最佳聚类数的确定

3.1　爬山法确定聚类数目上限和初始聚类中心

给定样本集 X，样本数为 n，在没有给定任何样本分布信息的情况下，采用迭代的方法进行聚类，聚类数目 k 搜索范围的经验公式为：

$$2\leqslant k\leqslant\sqrt{n} \tag{10}$$

因此迭代算法可以在 2 到$\sqrt{n}$之间进行，但是当 n 较大时，搜索的范围仍然很大，故采用爬山法[4]确定聚类数目上限和估计初始聚类中心。

对于 N 维数据样本集合$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，定义样本点 $\boldsymbol{x}_i(i=1,2,\cdots,n)$处的势函数为：

$$P_i^{(0)}=\sum_{j=1}^{n}e^{-\alpha\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2} \tag{11}$$

$$\alpha=4/r_\alpha^2 \tag{12}$$

r_α 是一个正整数，表示邻域半径，在邻域半径 r_α 之外的数据点对势的计算影响很小。从式(11)可以看出，点 $\boldsymbol{x}_i$ 周围聚集的样本点越多，则点 $\boldsymbol{x}_i$ 的势就越高。令 $P_1^*=\max\{P_i^{(0)},i=1,2,\cdots,n\}$，同时取对应的 $\boldsymbol{x}_1^*$ 为第一个初始聚类中心位置，然后根据下式调整每个样本点的势：

$$P_i^{(1)}=P_i^{(0)}-P_1^*e^{-\beta\|\boldsymbol{x}_i-\boldsymbol{x}_1^*\|^2} \tag{13}$$

$$\beta=4/r_\beta^2 \tag{14}$$

r_β 是一个正常数。令 $P_2^*=\max\{P_i^{(1)},i=1,2,\cdots,n\}$，取相应的 $\boldsymbol{x}_2^*$ 为第二个初始聚类中心位置。势函数调整的一般关系式如下：

$$P_i^{(k)}=P_i^{(k-1)}-P_k^*e^{-\beta\|\boldsymbol{x}_i-\boldsymbol{x}_k^*\|^2} \tag{15}$$

其中，$P_k^*=\max\{P_i^{(k-1)},i=1,2,\cdots,n\}$，对应的样本点 $\boldsymbol{x}_k^*$ 取为第 k 个初始聚类中心位置。

邻域半径可以采用以下两种形式：

$$r_f=\frac{1}{2}\min\{\max\{(\|\boldsymbol{x}_i-\boldsymbol{x}_j\|),i=1,2,\cdots,n\},j=1,2,\cdots,n\} \tag{16}$$

$$r_m=\frac{1}{2}\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2} \tag{17}$$

其中 n 是数据集合的样本个数，$\max\{\cdot\}$、$\min\{\cdot\}$是集合求大求小函数。在具体应用中，可以令 $r_\alpha=r_\beta=r_f$ 或 $r_\alpha=r_\beta=r_m$。

聚类数目可以用下式确定：

$$\frac{P_{k+1}^*}{P_1^*}<\delta \tag{18}$$

当式(18)成立时的 k 值即为聚类中心数目。其中 $\delta<1$ 是事先给定参数，此参数决定了最终产生的初始化聚类中心数目，δ 越小，则产生的聚类数越多。通过实验发现 $\delta\geqslant0.5$ 会得到比较合理的聚类数目，而在 $\delta\geqslant0.5$ 这一范围，又以 $\delta=0.5$ 时取得的聚类数目 $k_{\max}$ 最多，故可以将 $k_{\max}$ 作为合理聚类数目的上限，由此将数据样本的聚类数限制在$[2,k_{\max}]$上。具体做法是：

(1)令 $r_\alpha=r_\beta=r_f$ 或 $r_\alpha=r_\beta=r_m$，$\delta=0.5$，$k=1$；

(2)根据式(11)计算 $P_i^{(0)}$，得到第一个初始聚类中心位置 $\boldsymbol{x}_1^*$ 和 P_1^*；

(3)根据式(15)计算 $P_i^{(k)}$，得到 $\boldsymbol{x}_{k+1}^*$ 和 P_{k+1}^*；

(4)如果 $\frac{P_{k+1}^*}{P_1^*}<\delta$，算法结束，$k_{\max}=k$ 即为聚类数目的上限，$\boldsymbol{x}_i^*$ $(i=1,2,\cdots,k_{\max})$为初始化聚类的中心位置；

(5)否则，令 $k=k+1$，返回步骤(3)。

3.2　k–均值聚类算法确定最佳聚类数

在确定聚类数的范围后，对每一个特征子集利用 k–均值聚类算法对样本进行聚类以确定对应的最佳聚类数，并使用中心距离比值准则作为聚类有效性的判断[5-6]。k–平均聚类算法以 k 为参数，把 n 个对象分为 k 个类，并使类内具有较高的相似度，类间具有较低的相似度。通常其处理流程如下：首先，随机地选择 k 个对象，每个对象初始地代表一个类的平均值或中心。对剩余的每个对象，根据其与各个类中心的距离，将它赋给最近的类。然后重新计算每个类的平均值，这个过程不断地重复，直到准则函数收敛为止。聚类数的确定过程如下：

(1)初始化，设定门限 α 的值；令聚类的类个数 $k=2$，是迭代变量；中心距离比值指标 $CR_{F_i}^*=\infty$，表示最小的 CR 值；$k_{F_i}=1$，表示特征子集 F_i 最佳的分类数；

(2)在聚类个数为 k 时，利用 k–均值聚类算法，根据特征子集 F_i 对样本进行聚类，并根据聚类结果计算出中心距离比值指标 $CR_{F_i}(k)$；

(3)在聚类个数为 $k+1$ 时，利用 k–均值聚类算法，根据特征子集 F_i 对样本进行聚类，并根据聚类结果计算出中心距离比值指标 $CR_{F_i}(k+1)$；

(4)如果 $CR_{F_i}(k)=\min(CR_{F_i}^*,CR_{F_i}(k),CR_{F_i}(k+1))$，则 $k_{F_i}=k$，$CR_{F_i}^*=CR_{F_i}(k)$；如果 $CR_{F_i}(k+1)=\min(CR_{F_i}^*,CR_{F_i}(k),CR_{F_i}(k+1))$，则 $k_{F_i}=k+1$，$CR_{F_i}^*=CR_{F_i}(k+1)$；

(5)令判定函数为：

$$d(i)=\frac{|CR_{F_i}(k+1)-CR_{F_i}(k)|}{CR_{F_i}(k)} \tag{19}$$

如果 $d(i)\leqslant\alpha$，聚类数目确定结束；

(6)如果 $d(i)>\alpha$,令 $k=k+1$,若 $k\leqslant k_{\max}$,回到步骤(3);否则聚类数目确定结束。

4　基于中心距离比值准则的特征选择算法

4.1　选择特征子集的判定准则

假设特征子集 F_i 对应的最佳分类数为 k_i,其判定函数为:

$$cirt(F_i,k_i)=CR_{F_i} \tag{20}$$

则特征子集的选择就是选择使式(20)最小的 F_i。对于两个特征子集 F_i 和 F_j,对应的特征不是完全相同的,其包含的特征个数分别为 N_i 和 N_j。N_i 可能等于 N_j,也可能不等于 N_j。如果 $N_i\neq N_j$,对于不同特征子集 F_i 和 F_j 求得的中心距离比值指标 CR_{F_i} 和 CR_{F_j} 没有直接的可比性,因而需要将判定函数进行处理。定义一个标准的判定函数如下:

$$normcrit(F_i)=\frac{1}{N_i}crit(F_i,k_i) \tag{21}$$

此时,特征子集的选择即为使式(21)最小的特征子集 F_i。

4.2　特征间的相关性分析

假设样本集数目为 n,x 和 y 为两种特征,u_x、u_y 分别为特征 x 和 y 的均值,则定义相关系数如下:

$$\gamma_{xy}=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-u_x)(y_i-u_y)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-u_x)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-u_y)^2}} \tag{22}$$

$-1\leqslant\gamma_{xy}\leqslant 1$,其绝对值大小表示特征 x 和 y 相关程度的高低,γ_{xy} 绝对值越高,表示 x 和 y 相关程度越高。

4.3　特征选择算法的基本步骤

在算法中采用序贯删除法进行特征子集的搜索。设 F 是待选择的特征集,特征个数为 N,令 $t=N$,$count=1$,$normal=\infty$,其中 t 是特征子集包含的特征个数,$count$ 记录算法执行的次数,$normal$ 用于保存前一次选择的最佳特征子集的 $normcrit$ 值。算法的基本步骤如下:

(1)从待选择的特征集 F 中依次删除一个特征 f_i,得到 t 个特征子集 $F_i(i=1,2,\cdots,t)$,对于特征子集分别采用 k-平均聚类算法求其对应的最佳分类数 k_i,以及在最佳分类数下的中心距离比值指标 CR_{F_i};

(2)采用选择特征子集的判定准则,选择使式(21)最小的特征子集 F_i,令 $t=t-1$,$F=F_i$;

(3)如果 $|normal-normcrit(F_i)|>\beta$,$\beta$ 是事先设定的门限,并且 $count\leqslant N$,则 $count=count+1$,回到(1);

(4)利用式(22)对选择的特征子集 F_i 中的特征进行相关性分析,如果两个特征的相关系数大于 γ(γ 为门限),则删除其中的一个特征。

5　仿真分析

仿真实验 1　仿真一个数据集,有 3 个类,每个类均含 100 个样本,数据集共有 300 个样本,每个类的中心依次为(3,0,0),(0,3,0)和(0,0,3),每一类样本皆由类中心加上服从多元正态分布 $N(0,1)$ 的数据点构成,如图 1 所示。

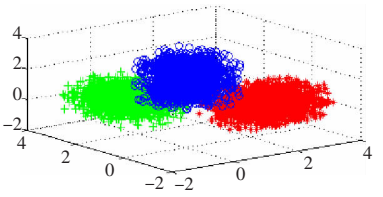

图 1　仿真的数据样本集

爬山算法在此数据样本集上得到 5 个聚类中心,分别为(2.762 7,-0.237 8,0.238 9)、(0.157 9,0.220 9,-0.225 6)、(-0.371 5,3.002 4,3.001 3)、(-0.368 6,-0.369 7,3.029 0)、(0.263 3,3.071 4,0.369 2)。即聚类数目的上限为 5,远远小于聚类数目的经验取值 $\sqrt{300}$,故用爬山算法确定聚类数目上限的方法,可以将聚类数目的上限限定在一个合理范围,从而提高查找到最佳聚类数的效率。

仿真实验 2　采用的数据是 11 种雷达辐射源信号的 20 种特征样本值,这 11 种雷达信号为:固定频率、线性频率调制、V 型频率调制、非对称的双线性频率调制、正切调频、反正切调频、双曲线调频、等级数调频、线性步进频率调制、二相编码和四相编码信号,对信号进行 4 级小波分解[7],得到 4 个盒维数,4 个信息维数,4 个关联维数,4 个多重分形熵和 4 个缝隙尺寸变化率,故 20 种特征分别为:D_{B1},D_{B2},D_{B3},D_{B4},D_{I1},D_{I2},D_{I3},D_{I4},D_{C1},D_{C2},D_{C3},D_{C4},H_{D_q1},H_{D_q2},H_{D_q3},H_{D_q4},$E_{\Lambda1}$,$E_{\Lambda2}$,$E_{\Lambda3}$,$E_{\Lambda4}$。在进行特征选择时,各类信号的样本数为 150,100 个作为训练样本,50 个作为测试样本,将实验重复 10 次,则得到算法所需时间、特征子集、识别的平均结果如表 1所示。

表 1　特征选择算法的计算结果

方法	计算时间/s	特征子集维数	特征子集组成	识别率/(%)
本文算法	$1.584\,3\times10^4$	6	$D_{B3},D_{I1},D_{C1},D_{I3},D_{C4},H_{D_q2}$	90.91

从实验结果可以看出,基于中心距离比值准则的无监督特征选择算法从 20 维的原始特征集中选出了 6 维的特征子集,大大降低维数,识别率也较高,达到了 90%以上。但是该算法需要的时间较长,不能实现实时处理。

6　结论

特征选择是模式识别方法中的难点之一,特别是无监督学习的特征选择问题。本文提出了一种基于中心距离比值准则的无监督特征选择算法,适合分类数不确定的情况。该算法能找出较好分类区分度的特征子集,从而实现降维。但是该算法复杂性较高,在特征维数高、样本数多的情况下,计算量较大。

参考文献:

[1] 何志文,李夕海,刘代志,等.基于相关性分析的特征选择方法研究[J].核电子学与探测技术,2005,25(6):729-732.

[2] 焦李成,张莉,周伟达.支撑矢量预选取的中心距离比值法[J].电子学报,2001,29(3):383-386.

[3] 孔波,刘小茂,张钧.基于中心距离比值的增量支持向量机[J].计算机应用,2006,26(6):1434-1436.

[4] 裴继红,范九伦,谢维信.聚类中心的初始化方法[J].电子科学学刊,1999,21(3):320-325.

[5] 张莉,孙钢,郭军.基于 k-均值聚类的无监督的特征选择方法[J].计算机应用研究,2005(3):23-24.

[6] Han Jia-wei,Kamber M.数据挖掘概念与技术[M].范明,孟小峰,译.北京:机械工业出版社,2001.

[7] Ye Fei,Luo Jing-qing,Lv Jiu-ming.Radar emitter signal fractal feature based on wavelet transform[C]//Proceedings of 2006 CIE International Conference on Radar,Shanghai,2006:1546-1549.