

# 基于 K-均值聚类的无监督的特征选择方法\*

张莉, 孙钢, 郭军

(北京邮电大学 信息工程学院, 北京 100876)

**摘要:** 模式识别方法首先要解决的一个问题就是特征选择, 目前许多方法考虑了有监督学习的特征选择问题, 对无监督学习的特征选择问题却涉及得很少。依据特征对分类结果的影响和特征之间相关性分析两个方面提出了一种基于 K-均值聚类方法的特征选择算法, 用于无监督学习的特征选择问题。

**关键词:** 特征选择; 相关性分析; 无监督学习; 聚类

中图法分类号: TP391.4 文献标识码: A 文章编号: 1001-3695(2005)03-0023-02

## Unsupervised Feature Selection Method Based on K-means Clustering

ZHANG Li, SUN Gang, GUO Jun

(School of Information Engineering, Beijing University of Posts & Telecommunications, Beijing 100876, China)

**Abstract:** The first problem need to be solved in pattern recognition method is feature selection. Now many methods think more about supervised feature selection problem, but involve little about unsupervised feature selection problem. In this paper, a feature selection algorithm based on K-means clustering method is proposed involving classification capabilities of feature vectors and correlation analysis between two features. This method can be used in unsupervised feature selection problem.

**Key words:** Feature Selection; Correlation Analysis; Unsupervised Learning; Clustering

### 1 引言

模式识别的主要任务是利用从样本中提取的特征将样本划分为相应的模式类别, 特征提取与选择是模式识别中的关键技术之一。一般情况下, 只有在特征向量中包含了足够的类别信息, 才能通过分类器实现正确分类, 而特征中是否包含足够的类别信息却很难确定。为了提高识别率, 总是最大限度地提取特征信息, 结果不仅使特征维数增大, 而且其中可能存在较大的相关性和冗余, 因而选择合适的特征来描述模式对模式识别的精度、需要的训练时间和需要的实例等许多方面都影响很大, 并且对分类器的构造也起着非常重要的作用。目前已有不少文献中提出了有监督学习的特征选择算法<sup>[1~4]</sup>, 但对于无监督学习的特征选择问题却涉及较少。无监督学习的特征选择问题就是依据一定的判断准则, 选择一个特征子集能够最好地覆盖数据的自然分类。目前的方法有基于遗传算法的特征选择方法<sup>[5]</sup>、基于模式相似性判断的特征选择方法<sup>[6]</sup>和信息增益的特征选择方法<sup>[7]</sup>, 这几种方法没有考虑特征之间的相关性和特征对分类的影响。文献[8]提出了一种无监督的特征选择方法, 基本思想是: 首先用竞争学习算法对样本进行分类, 确定分类数; 然后将原始特征集划分成多个特征子集, 在每一个特征子集计算判断函数  $J = \text{trace}((C + S)^{-1} S)$  (其中  $C$ ,  $S$  分别表示类内平均离散度和类间平均距离) 的值, 选择使判断函数值最大的特征子集, 从而确定相应的候选特征; 最后计算候选特征和已选择的特征之间的相关系数, 若相关系数大于 0.75 则放弃候选特征。但是由于特征数或特征不同,

不同的特征子集对应的自然分类可能也不同, 因而对不同的特征子集使用相同的分类结果, 不能有效地描述特征对样本自然分类的影响。本文依据特征对分类结果的影响和特征之间相关分析两个方面提出了一种基于 K-均值聚类的特征选择方法, 用于无监督学习的特征选择问题。其基本思想是对每一个特征子集利用 K-均值聚类算法确定其最佳分类数, 然后以 DB Index 准则设定一个判断函数用于特征选择, 最后从选择的特征子集中删除掉相关性较大的特征之一。

### 2 相关的背景知识

#### 2.1 聚类有效性的判断规则

类内离散度和类间距离常被用来判断聚类的有效性, DB Index 准则同时使用了类间距离和类内离散度, 因而在本文中采用 DB Index 准则<sup>[11]</sup>作为分类有效性的判断准则。DB Index 准则基本内容如下:

$$(1) \text{ 类内平均离散度 } S_i = \frac{1}{|C_i|} \sum_{x \in C_i} \|X - Z_i\|^2 \quad (1)$$

其中,  $Z_i$  是  $C_i$  类的类中心;  $|C_i|$  表示  $C_i$  类样本数。

$$(2) \text{ 类间距离 } d_{ij} = \|Z_i - Z_j\| \quad (2)$$

即用两个类中心的距离表示类间距离。

$$(3) \text{ DB Index } DB_k = \frac{1}{k} \sum_{i=1}^k R_i \quad (3)$$

其中  $R_i = \max_{j=1, \dots, k, j \neq i} \frac{S_i + S_j}{d_{ij}}$ ,  $k$  是分类数目。

DB Index 准则是  $DB_k$  的值越小, 说明分类的效果越好。

#### 2.2 特征之间的相关性分析

本文用式(4)计算两个特征之间的相关系数。相关系数的绝对值大小表示特征  $x, y$  相关程度的高低,  $r_{xy}$  绝对值越

大,表示相关程度越高。

$$r_{ij} = \frac{\sum_{p=1}^n (x_{pi} - Z_i)(x_{pj} - Z_j)}{\sqrt{\sum_{p=1}^n (x_{pi} - Z_i)^2 \sum_{p=1}^n (x_{pj} - Z_j)^2}} \quad (4)$$

### 3 特征选择算法

#### 3.1 聚类数的确定

对每一个特征子集  $F_i$  我们利用 K-均值聚类算法进行对样本进行聚类并确定对应的聚类数  $k_i$ , 使用 DB Index 准则作为聚类有效性判断。给定一个数据集  $X$ , 在没有给定任何样本分布信息的情况下进行聚类, 我们采用迭代的方法。一般情况下, 最佳的聚类数不会超过  $k_{max} = \lfloor n^{0.9} \rfloor$ 。因而迭代算法可以在  $k_{min} = 2$  到  $n$  之间进行, 并且我们可以根据具体的应用设定一个远小于  $n$  的  $k_{max}$  值, 聚类数  $k_i$  的确定过程如下:

(1) 初始化,  $C=2, DB^* = \infty, k_i=1$ 。其中,  $C$  为类的个数迭代变量,  $k_i$  表示最佳的分类个数,  $DB^*$  表示最小的 DB 值。

(2) 利用 K-均值聚类算法对样本进行聚类, 我们建立如式 (5) 所示的判断函数, 当  $d_j(i) \leq \delta$  时 ( $\delta$  是设定的门限), 聚类结束, 并且  $DB_c = DB_c(i)$ 。

$$d_j(i) = \frac{|DB_c(i+1) - DB_c(i)|}{DB_c(i)} \quad (5)$$

其中,  $DB_c(i)$  表示聚类数为  $C$  的第  $i$  次聚类 DB 的值。

(3) 若  $DB^* < DB_c$ , 则  $DB^* = DB_c, k_i = C$

(4)  $C=C+1$ , 若  $C < k_{max}$  则转 (2), 否则聚类结束。  $k_i$  即是第  $i$  个特征子集对应的最佳分类数。

#### 3.2 选择特征子集的判断规则

两个特征子集  $F_i, F_j (i=1 \dots t, j=1 \dots t, i \neq j, t$  是特征子集的个数) 对应的特征不是完全相同的, 所以对于不同的特征子集  $F_i, F_j$  求得的  $DB_i, DB_j$  的值没有直接的可比性, 因而我们需要将判断规则进行标准化处理。假设  $F_i$  对应的分类结果  $C_i$ , 则判断函数为

$$\text{crit}(F_i, C_i) = DB_i \quad (6)$$

在  $F_i$  特征子集中使用分类结果  $C_i$ , 求得相应 DB 的值, 则  $\text{crit}(F_j, C_j) = DB_j$ , 然后定义一个标准的判断函数如式 (7) 所示, 特征子集的选择就是要选择使式 (7) 最小的  $F_i$ 。

$$\text{normalizedcrit}(F_i) = \frac{1}{t} \sum_{p=1}^t \text{crit}(F_i, C_i) \quad (7)$$

#### 3.3 基于 K-均值聚类方法的无监督的特征选择算法

在文献 [10, 11] 中提出选择最好的特征子集比选择最好的特征组成特征子集更好, 因而在算法中我们利用序贯删除法进行特征子集的搜索。设  $F$  是原始特征集, 特征维数  $m$ , 令  $t=m, \text{count}=1, \text{normal}=0$ , 其中  $t$  记录特征子集的个数,  $\text{count}$  记录算法执行次数,  $\text{normal}$  保存前一次选择的最佳特征子集的  $\text{normalizedcrit}$  的值。算法基本步骤如下:

(1) 从  $F$  中依次删除一个特征  $x_i$ , 得到  $t$  个特征子集  $F_i, i=1 \dots t$ , 对这些特征子集分别采用 3.1 节中的方法求其对应的最佳分类数  $k_i$ 。

(2) 采用 3.2 节中描述的选择特征子集的判断规则, 选择使式 (7) 最小的  $F_i, t=t-1, F=F_i$ 。

(3) 若  $|\text{normalizedcrit}(F_i) - \text{normal}| > \delta$  ( $\delta$  事先设定的门

限) 并且  $\text{count} = m$  则  $\text{normal} = \text{normalizedcrit}(F_i), \text{count} = \text{count} + 1$ 。转 (1)

(4) 对选择的特征子集  $F_i$  利用式 (4) 进行特征相关性分析, 若两个特征的相关系数大于  $\delta$  ( $\delta$  为门限), 则删除其中的一个特征。

### 4 实验结果

对于有监督学习情况, 特征选择算法的有效性可以通过分类的准确度来评估, 但对无监督学习特征选择算法的有效性的评估不能采用这种方法。我们在验证算法时进行了两个实验, 首先选择两个维数较少的人工数据集 Wine, Pima-Diabetes 进行第一个实验(表 1)。这几个人工数据集已知分类数和每一个样本所属类别, 因为这两个数据集的特征维数较少, 我们在实验结果中给出了全部特征重要性的降序排序, 并列出了采用 Relief-F<sup>[10]</sup> 算法得到的特征顺序(表 2)。图 1 描述了利用本文的算法和 Relief-F 算法选择的特征进行分类的错误率。然后我们采用由哥伦比亚大学完成数据预处理的 KDD Cup 1999 Data 中的网络入侵检测的数据进行第二个实验。该数据集提供了从一个模拟的局域网上采集来的九个星期的网络连接数据, 数据集中的每条记录包含了 41 维特征, 并标注了每条记录所属类别(Normal, Dos, U2R, Probing, R2L)。我们从训练集中抽取了 16 645 条记录用于特征选择, 在实验结果中给出了完成相关性分析后的前 24 维特征(表 2)。实验时取  $\delta = 0.0001, \epsilon = 0.0001, \alpha = 0.75$ , 并用 BP 神经网络和 SVM 两个分类器对测试数据用选择的属性进行了测试(表 3)。

表 1 数据集基本信息

数据集	数据类型	特征维数	样本数	分类数
Wine	Continuous	13	178	3
Pima-Diabetes	Continuous	8	768	2
KDD Cup Data	Continuous + Nominal	41	16 645	5

表 2 本文的算法和 Relief-F 算法的特征选择结果

数据集	特征重要性的降序序列 (本文方法)	特征重要性的降序序列 (Relief-F)
Wine	6, 7, 12, 9, 11, 10, 5, 13, 1, 4, 3, 8, 2	6, 9, 1, 11, 5, 7, 10, 4, 12, 2, 13, 3, 8
Pima-Diabetes	8, 4, 3, 1, 2, 6, 7, 5	8, 1, 2, 5, 6, 4, 7, 3
KDD Cup Data	6, 5, 1, 34, 33, 36, 32, 8, 27, 29, 28, 30, 26, 38, 39, 35, 13, 24, 23, 11, 3, 10, 12, 4	6, 26, 12, 9, 13, 27, 5, 23, 31, 17, 39, 21, 29, 33, 20, 14, 34, 37, 15, 35, 36, 18, 28, 22

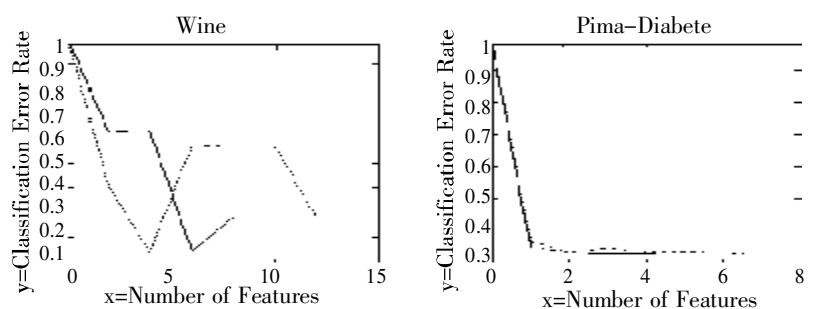


图 1 特征数和分类错误率(虚线表示的是 Relief-F 算法)

表 3 入侵检测测试数据集实验结果

分类器	Relief-F 算法	本文算法
BP Network	0.1993	0.1017
SVM	0.1020	0.056

### 5 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。本文从特征对分类的影响和特征相关性分析两方面出发, 提出了一种基于 K-均值的无 (下转第 42 页)