一种特征加权的聚类算法框架*)

高 滢 刘大有 徐 益
(吉林大学计算机科学与技术学院 长春 130012)
(吉林大学符号计算与知识工程教育部重点实验室 长春 130012)

摘 要 为了考虑数据各维特征对聚类的不同贡献,并把有监督特征评价方法应用到无监督分类问题中,提出一种特征加权的聚类算法框架。该框架首先通过某种聚类算法对数据聚类,然后,根据聚类结果,采用有监督特征评价方法学习各维特征的权值,再根据特征权值重新聚类,之后再次学习特征权值,该过程反复迭代,直至算法收敛或达到指定的迭代次数。欧几里德空间内基于距离、基于密度的聚类算法均适用于本框架。基于本框架,采用模糊C均值聚类算法(FCM)、密度聚类算法(DBSCAN),并通过信息增益特征评价、ReliefF特征评价方法,对多个UCI数据集进行了实验,验证了该框架的有效性。

关键词 聚类算法框架,特征加权,基于距离的聚类,基于密度的聚类

Framework of Feature Weighted Clustering Algorithm

GAO Ying LIU Da-you XU Yi
(College of Computer Science and Technology, Jilin University, Changchun 130012, China)
(Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China)

Abstract To consider the particular contributions of different features and apply supervised feature ranking methods to unsupervised classification, a framework of feature weighted clustering algorithm was proposed, which executes a clustering algorithm firstly, and then according to the results of clustering, learns feature weights using supervised feature ranking methods, and according the new feature weights executes the clustering algorithm again, this procedure iterates until convergence or maximum iteration times. Distance-based and density-based clustering algorithms in Euclidean space can be used in this framework. Based on this framework, fuzzy C-means clustering (FCM) and density-based spatial clustering of applications with noise (DBSCAN), and information gained and reliefF feature ranking algorithms are used to the experiments on several UCI machine learning databases, and validate the effectiveness of the framework.

Keywords Clustering algorithm framework, Feature weighted, Distance-based clustering, Density-based clustering

1 引言

聚类分析是将一个数据集划分为若干个类,使得类内相似性尽可能大且类间相似性尽可能小[1]。这一过程没有教师指导,是一种无监督的分类。聚类分析已被广泛地应用到许多领域,如模式识别、数据分析、图像处理以及市场研究等。

迄今为止,众多学者对聚类分析技术进行了研究,提出了多种聚类分析算法。根据聚类标准的不同,聚类算法可分为基于距离的聚类、基于密度的聚类和基于链接的聚类[2]。基于距离的聚类和基于密度的聚类,通常应用于欧几里德空间的数据。欧几里德空间为球形空间,即假设每一特征(也称为属性)在聚类过程中的重要性均相同,而实际情况并非如此。

为了考虑各个特征对聚类的不同贡献,一些研究者采用特征加权的方法:文献[3]提出WFCM算法。该算法通过极小化属性评价函数 $E(w)$ 为每个属性学习权重,构造加权的欧氏距离;文献[4]在文献[3]的基础上提出CF-WFCM算法,通过梯度递减算法极小化属性评价函数 CFuzziness(w),为每个属性赋予一个权重,将属性权重应用于模糊C均值聚类算法;文献[5]利用ReliefF技术对特征进行加权选择,并进行模糊聚类。实验结果表明,特征加权的聚类算法,与传统聚类算法相比,聚类质量有所提高。在特征加权的聚类算法中,特征权值的学习是影响聚类效果的重要因素。

特征权值的学习可以认为是特征选择问题的泛化[6]。特征选择过程可分为两个步骤:特征搜索和特征评价,其中的特征评价方法可以用来学习特征权值。现有的特征评价方法有信息增益方法、Relief/ReliefF方法、相关性方法、一致性方法和Wrapper子集方法等,其中前两种方法用于评价单个特征,而后三种方法用于评价特征子集[7]。特征选择在分类问题上已经非常成功。在聚类问题上,特征选择的研究与应用相对较少,其原因是聚类没有像分类一样的训练数据。所以,在缺乏类信息的条件下,无监督的特征选择很难选择出最具类区分力的特征。

为了将已有的有监督特征评价方法应用到无监督分类,本文提出一种特征加权的聚类算法框架。该框架首先使用某

*)国家自然科学基金重大项目(60496321),国家自然科学基金项目(60573073,60773099),国家863高技术研究发展计划项目(2006AA10Z245,2006AA10A309),吉林省科技发展计划项目(20030523),欧盟项目TH/Asia Link/010(111084)。高 滢 讲师,博士研究生,从事数据挖掘、统计关系学习的研究;刘大有 教授,博士生导师,从事知识工程与专家系统、多Agent系统、不确定性推理、数据挖掘、算法与数据结构、空间推理与GIS应用的研究;徐 益 硕士研究生,从事数据挖掘研究。

种聚类算法对数据聚类，然后把聚类结果看作训练样本，进行特征权值的学习，这样把聚类中的无监督特征权值学习问题转化为有监督特征权值学习问题，然后根据学得的权值，再次聚类。该过程反复迭代，直到算法收敛或达到指定的迭代次数。在欧几里德空间内基于距离、基于密度的聚类算法，及各种有监督特征评价方法均可应用于本框架。基于本框架，本文使用模糊 C 均值聚类算法（FCM）、密度聚类算法（DBSCAN），分别对信息增益特征评价、ReliefF 特征评价方法，在多个 UCI 数据集上进行实验，并给出了对比实验结果，验证框架的有效性。

本文安排如下：第 2 节详细介绍特征加权的聚类算法框架；第 3 节在多个 UCI 机器学习数据集上进行实验，给出并分析实验结果；最后总结全文。

2 特征加权的聚类算法框架

2.1 符号约定

$X=\{x_1,x_2,\cdots,x_N\}$是待聚类分析的对象的全体，N 为对象总数，其中 $x_i=[x_{i1},x_{i2},\cdots,x_{iM}]$表示第 i 个对象的 M 个特征值；

$W=\{w_1,w_2,\cdots,w_M\}$为特征权值，其中 w_i 表示第 i 维特征的权值；

C 为类簇数；

$V=\{v_1,v_2,\cdots,v_C\}$是各类中心，其中 $v_i=[v_{i1},v_{i2},\cdots,v_{iM}]$表示第 i 类的聚类中心；

$d(v_i,x_j)$表示对象 x_j 与聚类中心 v_i 的距离；

$U=\{u_{ij}\mid 1\leqslant i\leqslant C,1\leqslant j\leqslant N\}$为 $C\times N$ 的矩阵，u_{ij} 表示对象 x_j 对聚类中心 v_i 的隶属度。对于硬划分，U 需满足下列条件式(1)；对于模糊聚类，U 需满足条件式(2)。

$$\begin{cases} u_{ij}\in\{0,1\} \\ \sum_{i=1}^{C}u_{ij}=1,\text{对任意 } j=1,2,\cdots,N \\ \sum_{j=1}^{N}u_{ij}>0,\text{对任意 } i=1,2,\cdots,C \end{cases} \tag{1}$$

$$\begin{cases} u_{ij}\in[0,1] \\ \sum_{i=1}^{C}u_{ij}=1,\text{对任意 } j=1,2,\cdots,N \\ \sum_{j=1}^{N}u_{ij}>0,\text{对任意 } i=1,2,\cdots,C \end{cases} \tag{2}$$

2.2 框架描述

特征加权的聚类算法框架描述如下：

(1)数据标准化。数据标准化的主要目的是消除数据量纲的影响，即映射所有的属性值到特定范围内（通常为[0,1]）。

(2)定义样本间的相异度。在欧几里德空间内，样本间的相异度通过样本间的距离（如欧几里德距离、曼哈顿距离等）来度量，本框架使用加权的距离。

(3)使用某种聚类算法，对数据聚类。欧几里德空间内基于距离、基于密度的聚类算法，均可应用于本框架。

(4)根据聚类结果，运用有监督特征评价方法，学习各维特征的权值。

(5)利用求得的特征权值，采用特征加权的方法再次聚类。

(6)步骤(4)、(5)反复迭代，直到收敛或达到指定的迭代次数。

(7)分析评价聚类结果。

```
FeatureWeightedClusteringFramework(X,MT)
  //X为待聚类分析的对象的全体，MT为最大迭代次数
BEGIN
1.    X←Standardization(X)        //数据标准化
2.    W[]←1,T←0,F←0               //特征权值、迭代次数、目标函数初始化
3.    do
4.    {CX←Clustering(X,W)         //使用某种聚类算法聚类
5.     W←LearningWeight(CX)       //学习特征权值
6.     F'←F
7.     F←ObjFunction(CX)          //计算新的目标函数值
8.     T←T+1 }
9.    while ((T<MT)||(|F-F'|=0))
END
```

图 1 特征加权的聚类算法框架

2.3 复杂性和收敛性说明

该框架下算法的时间复杂性与其中采用的聚类算法和特征权值学习算法的时间复杂性密切相关。若聚类算法的时间复杂性为 $O(T_1)$，特征权值学习算法的时间复杂性为 $O(T_2)$，指定的最大迭代次数为 MT，则该框架下算法的时间复杂性为 $O(MT*(T_1+T_2))$。

该框架下算法的收敛性未能从理论上加以证明。但从迭代过程来看，每次迭代使特征权值趋向于与最终聚类结果一致的方向。本文下节的实验结果，给出了算法的迭代次数及收敛情况。

3 实验

3.1 聚类算法

对于欧几里德空间内的数据，基于距离的聚类算法（如 C-MEANS,CURE,BIRCH 等）和基于密度的聚类算法（如 DBSCAN,OPTICS 等）都可应用于本框架。本文选择模糊 C-均值聚类算法（FCM）[8]，DBSCAN[9] 这两个代表性的算法进行实验。

3.2 特征权值学习

在现有特征评价方法中，本文选择信息增益（IG）方法、Relief/ReliefF[10] 方法学习特征权值。

在 IG 特征评价中，若 X 为某个特征，Y 为类变量，用 Hall 等人[11] 提出的对称不确定性（symmetrical uncertainty）计算特征权值，公式如下：

$$SU(X,Y)=2\times\left[\frac{H(X)+H(Y)-H(X,Y)}{H(X)+H(Y)}\right] \tag{3}$$

其中，$H(X)$，$H(Y)$，$H(X,Y)$表示熵。

3.3 实验数据集

利用特征加权的聚类算法框架，采用 FCM 和 DBSCAN 聚类算法，使用 IG 和 ReliefF 特征权值学习方法，对 UCI 机器学习数据库[12] 中的 5 个数据集进行实验，各数据集的特征描述见表 1。

表 1 数据集特征

数据集名称	数据个数	属性个数	聚类个数
glass	345	6	6
ionosphere	351	34	2
iris	150	4	3
pima	768	8	2
wine	178	13	3

3.4 聚类结果评价

目前，常用聚类有效性函数来评价不同聚类算法的结果以及同一算法在不同参数情况下得到的聚类结果。

对于硬聚类，本文使用常用的紧致分离函数作为聚类的

有效性函数：

$$D=\frac{\underset{x_i\in u_k,x_j\in u_k}{avg}(dis(x_i,x_j))}{\underset{x_i\in u_p,x_j\in u_q,p\neq q}{avg}(dis(x_i,x_j))} \tag{4}$$

D 为同类内对象平均距离与异类间对象平均距离的比值，D 越小，表明聚类效果越好。

对于模糊聚类，本文使用 Xie 和 Beni[13] 定义的有效性函数：

$$S=\frac{\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^2d^2(v_i,x_j)}{N\times(\min\limits_{i\neq k}\{d^2(v_i,v_k)\})} \tag{5}$$

函数值 S 越小，说明聚类效果越好。

3.5 实验结果

实验中各种聚类算法及特征权值学习方法的有效性评价函数值及迭代次数见表 2、表 3。

表 2 基于 FCM 聚类算法的实验结果

数据库	FCM	FCM+IG		FCM+ReliefF	
	S 值	S 值	迭代次数	S 值	迭代次数
glass	3.321	0.404	MT	0.736	6
ionosphere	0.744	0.537	2	0.569	3
iris	0.176	0.182	3	0.181	3
pima	1.168	0.557	3	0.608	MT
wine	0.405	0.355	3	0.378	4

表 3 基于 DBSCAN 聚类算法的实验结果

数据库	DBSCAN	DBSCAN+IG		DBSCAN+ReliefF	
	D 值	D 值	迭代次数	D 值	迭代次数
glass	0.476	0.459	2	0.477	2
ionosphere	0.802	0.763	2	0.775	1
iris	0.356	0.343	1	0.343	1
pima	1.044	0.680	2	0.680	4
wine	0.700	0.710	2	0.641	3

3.6 实验结果分析及说明

通过上面实验对加权前后聚类结果的比较，可以看出：

(1)学习特征权值可以普遍提高聚类质量。

(2)学习特征权值使无关属性的影响尽量减小，甚至权值可以为零。因此，不仅提高聚类质量，还可以减少特征维数。

(3)学习特征权值改善聚类质量的程度依赖于具体的数据库和具体特征。

(4)学习特征权值对聚类结果的改善是以多次迭代为代价的。聚类算法所需的时间随迭代次数的增加而增加，但并没有从阶上增加算法的时间复杂度。

(5)在特征加权的聚类算法框架中，每种聚类算法保持了原有特征，即各种聚类算法所适用的数据特征、算法复杂度、处理噪音能力等方面特征没有变化。

(6)通过实验中的迭代次数可以看出，本实验中算法未能保证收敛性，但大多数情况下是收敛的。

结束语 为了考虑各维特征对聚类的不同贡献，并把有监督分类问题中的特征评价方法应用到聚类分析的特征权值学习中，本文提出了一种特征加权的聚类算法框架，该框架首先采用某种聚类算法对数据聚类，然后按照聚类结果，采用有监督特征评价方法学习各维特征的权值，再根据学得的权值再次聚类。该过程反复迭代，直到算法收敛或达到指定的迭代次数。在欧几里德空间内基于距离、基于密度的聚类算法，及用于有监督分类的特征评价方法，均可应用于该框架。本文把经典的基于距离的聚类算法 FCM、基于密度的聚类算法 DBSCAN，以及信息增益、ReliefF 特征评价方法应用到框架中，并通过对多个 UCI 机器学习数据库进行实验，验证了该框架的有效性。需要注意的是，聚类算法以及特征权值学习方法是本框架的关键，因此，研究高效的聚类算法和特征权值学习方法至关重要。

参 考 文 献

[1] Hartigan J A. Clustering Algorithms. Wiley, 1975

[2] Qian Wei-ning, Zhou Ao-ying. Analyzing Popular Clustering Algorithms from Different Viewpoints. Journal of Software, 2002, 13(8): 1382-1394

[3] Wang Xizhao, Wang Yadong, Wang Lijuan. Improving fuzzy c-means clustering based on feature weight learning. Pattern Recognition Letters, 2004 (25): 1123-1132

[4] 王丽娟，关守义，王晓龙，等. 基于属性权重的 Fuzzy C Mean 算法. 计算机学报，2006，29(10)：1797-1803

[5] 李洁，高新波，焦李成. 基于特征加权的模糊聚类新算法. 电子学报，2006，34(1)：89-92

[6] Wettschereck D, Aha D W, Mohri T. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artificial Intelligence Review, 1997, 11: 273-314

[7] Qu Guang Zhi, Yousif M. A New Dependency and Correlation Analysis for Features. IEEE Transactions on Knowledge and Data Engineer, 2005, 17(9): 1199-1207

[8] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981

[9] Ester M, Kriegel H P, Sander J, et al. A Density-Based Algorithm for Discovering Cluster in Large Spatial Databases with Noise // Proceedings 2nd International Conference on Knowledge Discovery and Data Mining. Portland, OR, 1996: 226-231

[10] Kononenko I. Estimating Attributes: Analysis and Extensions of RELIEF // Proceedings of the 7th European Conference on Machine Learning. Berlin: Springer, 1994: 171-182

[11] Hall M A. Correlation-based feature selection for discrete and numerical class machine learning // Proc. of Intl. Conf. on Machine Learning. 2000

[12] Asuncion A, Newman D J. UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: Department of Information and Computer Science, University of California

[13] Xie X L, Beni G. A validity measure for fuzzy clustering. Pattern Analysis and Machine Intelligence, 1991, 13(8): 841-847