基于主动学习策略的半监督聚类算法研究*

芦世丹，崔荣一†
（延边大学 计算机科学与技术学科 智能信息处理研究室，吉林 延吉 133002）

摘　要：提出一种选择最富信息数据并予以标记的基于主动学习策略的半监督聚类算法。首先，采用传统 K-均值聚类算法对数据集进行粗聚类；其次，根据粗聚类结果计算出每个数据隶属于每个类簇的隶属度，筛选出满足最大与次大隶属度差值小于阈值的候选数据，并从中选择差值较小的数据作为最富信息的数据进行标记；最后，将候选数据集合中未标记数据分组到与每类已被标记数据平均距离最小的类簇中。实验表明，提出的主动学习策略能够很好地学习到最富信息数据，基于该学习策略的半监督聚类算法在测试不同数据集时均获得了较高的准确率。

关键词：K-均值算法；主动学习策略；半监督学习；聚类
中图分类号：TP181；TP301.6　　**文献标志码**：A　　**文章编号**：1001-3695(2013)06-1718-03
doi:10.3969/j.issn.1001-3695.2013.06.030

Semi-supervised clustering method based on active learning strategy

LU Shi-dan, CUI Rong-yi†
(*Intelligent Information Processing Laboratory*, *Dept. of Computer Science & Technology*, *Yanbian University*, *Yanji Jilin* 133002, *China*)

Abstract: By employing active learning strategy to learn informative dataset to be labeled, this paper proposed a semi-supervised clustering method based on active learning strategy. Firstly, it employed traditional K-means algorithm to make coarse clustering for unlabeled dataset. And furthermore, based on the result of coarse clustering, it calculated the membership degree of each data belonging to each cluster, then screened out alternative data of which the difference between maximum and the second maximum membership degree was lower than threshold, then the partial data would be labeled if the difference of which was relatively small, i.e., the data were informative samples. Finally, they grouped each selected unlabeled data to corresponding labeled cluster which acquired minimum average distances. The experimental results show that the proposed active learning strategy is very powerful to learn informative data, and the semi-supervised clustering method based on active learning strategy is quite accurate with regards to various dataset.

Key words: K-means algorithm; active learning strategy; semi-supervised learning; clustering

0　引言

有监督学习是通过训练大量带标记的数据来提高泛化能力，但是对数据进行人工标记的代价往往很高。无监督学习属于无任何监督信息的自动学习，虽然不需要带标记数据，但所得模型却不够精确。因此，将少量带标记数据和大量无标记数据结合的半监督学习成为机器学习的研究热点[1]。按照学习任务的不同，半监督学习分为半监督聚类和半监督分类[2]。

现有的半监督聚类算法大致可分为三类，包括基于约束的半监督聚类算法（constraint-based semi-supervised clustering method，CBSSC）、基于距离的半监督聚类算法（distance-based semi-supervised clustering method，DBSSC）、集成了约束与距离的半监督聚类算法（constraint and distance based semi-supervised clustering method，CDBSSC）[3]。以上三类算法尽管利用约束来指导聚类，但在求解过程中常常遇到成对约束的违反问题。例如，当通过利用 must-link 和 cannot-link 成对约束来学习一个距离函数以提高聚类效果时，这类算法不能保证 must-link 点总能被分到同一个类簇中，而相反 cannot-link 中的一部分点对却被分配到同一个类簇内，产生了约束违反问题。另一方面，如果通过添加惩罚项来解决成对约束违反问题，又会出现选择合适的惩罚因子的难题。

本文提出了一种基于主动学习策略的半监督聚类算法，在传统 K-均值聚类算法的基础上，主动学习出最富信息的数据进行标记，充分利用标记数据来调节未标记数据的分组情况，不断改善聚类效果。实验结果表明，提出的主动学习策略能够很好地学习到富信息数据，并且基于该主动学习策略的半监督聚类算法获得了较高的准确率。

1　主动学习策略

在半监督学习过程中，需要花费很多的时间评价和标记数据的类别。为了解决选择标记数据带来的困难，主动学习已经成为机器学习和模式识别领域的研究热点。在主动学习中，学习器不再是被动地接收由用户提供的训练数据，而是主动要求用户对那些最有价值的数据进行标记[4]。理论上，在获得相近准确率的情况下，主动样本选择相对于随机选择可以显著地减少所需要的样本数[5]。

收稿日期：2012-09-14；**修回日期**：2012-10-31　**基金项目**：吉林省教育厅“十二五”科学技术研究项目(2011-8)；吉林省科技发展计划项目(20050703-1)

作者简介：芦世丹(1987-)，女(朝鲜族)，黑龙江绥化人，硕士研究生，主要研究方向为语种识别、模式识别；崔荣一(1962-)，男(朝鲜族)(通信作者)，吉林延吉人，教授，硕导，博士，主要研究方向为模式识别、智能计算(cuirongyi@ybu.edu.cn).

近年来,研究者提出了很多主动学习方法,包括基于 SVM 的主动学习方法、基于委员会的主动学习方法、基于熵的主动学习方法等。基于 SVM 的主动学习方法是采样最靠近决策面的样本,因为该样本被认为是最富信息的,但是该方法只适用于二分类问题。基于委员会的主动学习方法是选择分类结果最不一致的样本。基于熵的主动学习方法是计算每个样本属于每个类别的概率,根据概率计算每个样本的熵,熵越大表示该样本的分类不确定性越高,信息量越大,但是熵只能代表整体的平均不确定性。针对这个问题,Joshi 等人[6] 提出了一种基于最优标号和次优标号的主动学习方法。

2　半监督聚类算法及评价

典型的基于主动学习的半监督聚类算法流程要先采用简单的聚类算法进行粗聚类,然后根据粗聚类的结果主动学习到最富信息的待标记数据并进行标记,以不断指导聚类过程并改善聚类结果,具体框架如图 1 所示。

2.1　半监督聚类算法

传统 K-均值聚类方法对初始聚类中心的选择具有很大的随机性,因而直接影响到聚类的结果。本文在传统 K-均值的基础上,采用主动学习策略学习到最富信息的待标记数据并予以标记,利用标记数据来改善未标记数据的聚类结果。

Frossyniotis 等人在文献[7]中通过式(1)来度量数据 i 对类簇 j 的隶属度 $h_{i,j}$:

$$h_{i,j}=\frac{1}{\sum_{k=1}^{C}\frac{d(x_i,\mu_j)}{d(x_i,\mu_k)}} \tag{1}$$

其中:x_i 表示第 i 个数据属性值;μ_j 表示第 j 个类簇的聚类中心;函数 $d(\cdot,\cdot)$ 表示距离度量;C 为类簇总数。为了避免数据和聚类中心重合时出现分母为 0 的情况,本文对式(1)进行了如下修改:

$$h_{i,j}=\frac{1}{\sum_{k=1}^{C}\frac{d(x_i,\mu_j)+c}{d(x_i,\mu_k)+c}} \tag{2}$$

其中:c 为常数,本文采用 0.1。

利用隶属度 $h_{i,j}$ 可以定义误差函数如式(3):

$$\varepsilon_t=\frac{1}{2N}\sum_{i=1}^{N}(1-h_{i,\mathrm{good}}^t+h_{i,\mathrm{bad}}^t) \tag{3}$$

其中:N 是数据总数;t 表示迭代次数;$h_{i,\mathrm{good}}^t$ 和 $h_{i,\mathrm{bad}}^t$ 分别表示第 i 个数据在第 t 次计算中隶属某类簇的最大隶属度和最小隶属度。

利用隶属度和误差按下述方式更新聚类结果[7]:

$$H_{ag}^t=\arg\max_{k=1,\cdots,C}\sum_{\tau=1}^{t}\left[\frac{\log(\beta_\tau)}{\sum_{j=1}^{t}\log(\beta_j)}h_{i,k}^\tau\right] \tag{4}$$

其中:

$$\beta_t=\frac{1-\varepsilon_t}{\varepsilon_t} \tag{5}$$

基于主动学习策略的半监督聚类算法步骤如下:

输入:无标记的数据集$(x_1,\cdots,x_N)$,聚类数 C。

输出:聚类结果 H_{ag}^t。

a)初始化最大误差 $\varepsilon_{\max}=0$、阈值 V、计算隶属度或误差的总次数 T。

b)采用 K-均值算法对无标记数据集聚类成 C 类,给每个类簇一个标号,如 $1,2,\cdots,C$,并记录粗聚类结束时的聚类中心。

c)利用式(2)计算每个数据隶属于每个类簇的隶属度 $h_{i,j}$。

d)利用式(3)进行误差计算。

e)选出所有满足条件 $h_{i,\mathrm{good}}-h_{i,\mathrm{secondgood}}<V$ 的未标记数据作为候选数据集,并按差值从小到大排序,选择前 $m\%$ 的数据作标记。

f)计算候选数据中未被标记的$(1-m\%)$的数据到每类被标记数据的平均距离,把该未标记数据分组到与标记数据平均距离最小的类簇中,并用候选数据的聚类结果计算出局部聚类中心。

g)根据 f)得到的局部聚类中心,利用式(2)(3)重新计算候选数据的隶属度和误差。

h)利用式(4)更新聚类结果。

考虑到数据标记代价和误分组纠正效率,在确定标记数据个数时需要满足的条件为被标记的数据个数不超过初始聚类时被误分组的数据个数。因此,可以将此条件用于 e)以确定 m 的值。

2.2　半监督聚类评价指标

聚类结果常用聚类准确性来评价[8]。在纯粹的无监督聚类问题中,没有被标记的数据,所以得到的聚类正确率等于被正确分组的数据个数与总数据个数 N 之比,即 $R=(N-E)/N=1-E/N$,其中 E 表示被误分组的数据个数。而在半监督聚类中,存在标记数据,当不考虑标记样本对聚类过程起指导作用时,聚类正确率等于 $R=(N-E+L)/N=1-(E-L)/N$,其中 L 为标记数据个数且均来自误分组的 E 个数据。但由于受主动学习策略的限制,被标记的 L 个数据不一定均来自初始聚类误分组的 E 个数据。

一般情况下,在标记数据中,部分标记数据是针对已正确分组数据进行的,这些标记对聚类性能的改善不提供任何有效信息,称之为无效标记;而对其余数据的标记能够纠正初始聚类的误分组,称这些标记为有效标记,如图 2 所示。显然,从数据标记代价和错误纠正效率考虑,人们希望无效标记数据个数尽可能地小。

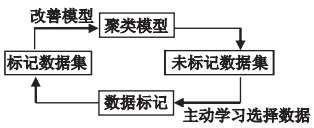

图1　典型的基于主动学习的半监督聚类算法框架

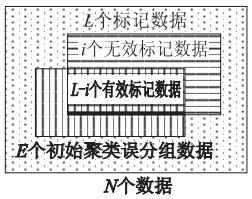

图2　标记数据分布情况

假设在被标记的 L 个样本中无效标记数据个数为 i,则有效标记数据个数为 $L-i$,错误分组数据个数为 $E-(L-i)$,因此聚类正确率为

$$R=(N-E+L-i)/N=1-(E-L+i)/N\quad i=0,1,\cdots,L \tag{6}$$

由于当标记数据个数 L 大于初始聚类的误分组数据个数 E 时,一定会存在无效标记,因此数据标记个数必须满足

$$L\leqslant E\leqslant N \tag{7}$$

此时,最理想的情况是无效标记数据个数 $i=0$,即此时正确率为

$$R_{\max}=1-(E-L)/N \tag{8}$$

而最糟糕的情况是无效标记数据个数 $i=L$,此时正确率为

$$R_{\min}=1-E/N \tag{9}$$

这相当于无监督的纯聚类情况。

考虑无效标记数据是等概率出现的情况,即 i 以等概率取$[0,L]$内的整数,则无效标记数据个数的期望值为

$$I=\sum_{i=0}^{L}\frac{1}{L+1}i=\frac{L}{2} \tag{10}$$

因此,根据式(6)(10),本文可以得出:当标记数据个数 L 不超过初始聚类误分组数据个数 E,并且假设无效数据标记为等概率分布,则聚类正确率的期望值为

$$R_{avg} = 1 - (E - L/2)/N \tag{11}$$

实际上,在半监督聚类中,由于存在少量的标记数据,在聚类过程中若不考虑标记数据对未标记数据的指导作用,聚类结果的正确率会因这些少量的标记数据而有所提高,且平均结果应趋于 R_{avg};如果考虑标记数据对聚类过程的指导作用,好的半监督聚类算法的正确率应该得到更进一步的提高,且高于 R_{avg},因此可以用 R_{avg} 来衡量标记样本对半监督聚类算法的指导作用。

3 半监督聚类算法及评价

为了验证基于主动学习策略的半监督聚类算法,分别对四个数据集进行了测试。其中的两个数据集是 UCI 机器学习数据库(http://archive.ics.uci.edu/ml/datasets.html)中的 iris 数据集和 wine 数据集,其他两个数据集分别是某高校调查统计的男女生特征数据集(包括身高、体重、年龄)和人工生成的香蕉形数据集。各数据集的具体描述信息如表 1 所示。

表 1 数据集描述

dataset	No. attributes	No. classes	No. instances
iris	4	3	150
wine	13	3	178
gender	3	2	64
banana	2	2	1 000

3.1 主动学习策略的实验

本文在用传统 K-均值算法粗聚类的基础上,通过采用主动学习策略学习出那些最富信息的候选数据,并从中选择部分数据进行标记。因此,能否全面地选择出最富信息的数据,能否在保证聚类正确率的情况下避免不必要的标记负担是衡量主动学习策略的主要标准。

图 3 是针对 iris 数据集,当 $m = 20$ 时,随着 V 值的变化主动学习出的候选数据的个数、被标记数据的个数、被标记数据中包括被粗聚类误分组的数据个数之间的对比关系曲线。从曲线的变化趋势可以看出,通过主动学习得到的标记数据中至少有一半的数据来自于粗聚类时被误分组的富信息数据。

3.2 基于主动学习策略的半监督聚类算法实验

针对表 1 所列出的四个数据集,将本文算法同传统 K-均值算法、文献[9]提出的基于分类的半监督聚类方法(K-means-Guider)以及 R_{avg} 获得的聚类正确率进行比较。表 2 所示为当 $V = 0.3$ 时,运行 20 次获得的平均正确率。可见,本文算法在无须人工选择标记数据的同时,保证了聚类正确率。

表 2 各数据集的测试结果

dataset	accuracy/%			
	traditional K-means	R_{avg}	K-means-Guider	proposed algorithm
iris	88.67	91.67	94.81	96.67
wine	70.22	72.75	87.04	74.72
gender	85.94	89.06	84.48	93.75
banana	75.00	77.50	82.45	85.90

图 4、5 是针对 iris 数据集,当固定 $m = 20$ 时聚类正确率与 V 的关系、当固定 $V = 0.3$ 时聚类正确率与 m 值的关系。从图 4 可以看出,当 m 值不变时,随着 V 值的增大正确率也随着提高,且正确率均高于理论上通过标记获得的平均正确率。从图 5 可以看出,当 V 固定不变时,随着 m 值的增大正确率虽然也有所提高,但是实际正确率与理论上通过标记数据获得的平均正确率的差距越来越小,这也充分证明过分提高标记数据的比例,有时付出的人工标记代价与获得的效果不成正比。

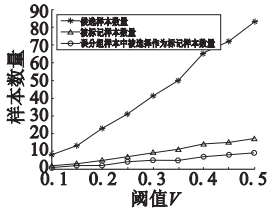

图3 三种情况下样本数量关系曲线

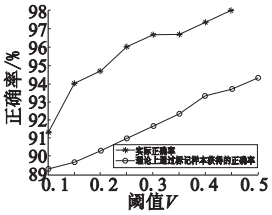

图4 聚类正确率与V值的关系

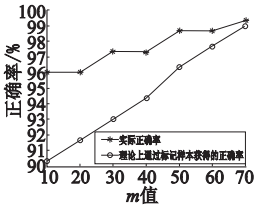

图5 聚类正确率与m值的关系

4 结束语

本文主要提出了基于主动学习策略的半监督聚类算法。针对 UCI 机器学习数据库提供的 iris 数据集、wine 数据集,以及实际调查统计的男女生特征数据集和人工生成的香蕉形数据集,首先采用 K-均值算法进行粗聚类,然后采用主动学习策略学习出富信息数据,最后利用富信息数据改善聚类结果。实验结果表明,提出的算法在无须人工选择标记数据的同时,获得了较好的正确率。

由于本文算法以传统 K-均值为基础,因此聚类结果会受数据集分布特点和初始粗聚类结果的影响,这也正是本文算法的局限性。针对这一问题,研究如何利用富信息数据提高不同形状数据集的聚类正确率和如何增强聚类算法鲁棒性是下一阶段工作重点。

参考文献:

[1] 邓超, 郭茂祖. 基于 Tri-Training 和数据剪辑的半监督聚类算法[J]. 软件学报, 2008,19(3): 663-673.

[2] OLIVER C, BERNHARD S, ALEXANDER Z. Semi-supervised learning[M]. London: MIT Press,2006: 13-101.

[3] 尹学松,胡恩良,陈松灿. 基于成对约束的判别型半监督聚类分析[J]. 软件学报,2008,19(11): 2791-2802.

[4] 陈荣,曹永锋,孙洪. 基于主动学习和半监督学习的多类图像分类[J]. 自动化学报,2011,37(8): 954-962.

[5] DASGUPTA S. Coarse sample complexity bounds for active learning[C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2006: 235-242.

[6] JOSHI A J, PORIKLI F, PAPANIKOLOPOULOS N. Multi-class active learning for image classification[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]. IEEE Computer Society,2009:2372-2379.

[7] FROSSYNIOTIS D, LIKAS A, STAFYLOPATIS A. A clustering method based on boosting[J]. Pattern Recognition Letters,2004, 25(6): 641-654.

[8] 高滢,刘大有,齐红,等. 一种半监督 K-均值多关系数据聚类算法[J]. 软件学报,2008,19(11):2814-2821.

[9] 李杉, 张化祥. 基于分类的半监督聚类方法[J]. 计算机工程与应用,2011, 47(3):132-134.