基于主动数据选取的半监督聚类算法*

文 平[1]，冷明伟[1,2]，陈晓云[1]
（1. 兰州大学 信息科学与工程学院，兰州 730000；2. 上饶师范学院 数学与计算机学院，江西 上饶 334001）

摘 要：利用少量标签数据获得较高聚类精度的半监督聚类技术是近年来数据挖掘和机器学习领域的研究热点。但是现有的半监督聚类算法在处理极少量标签数据和多密度不平衡数据集时的聚类精度比较低。基于主动学习技术研究标签数据选取，提出了一个新的半监督聚类算法。该算法结合最小生成树聚类和主动学习思想，选取包含信息较多的数据点作为标签数据，使用类 KNN 思想对类标签进行传播。通过在 UCI 标准数据集和模拟数据集上的测试，结果表明提出的算法比其他算法在处理多密度、不平衡数据集时有更高精度且稳定的聚类结果。

关键词：数据挖掘；半监督聚类；主动学习；标签数据；数据选取；最小生成树；多密度数据集；不平衡数据集

中图分类号：TP301.6　　**文献标志码**：A　　**文章编号**：1001-3695(2012)08-2841-04

doi:10.3969/j.issn.1001-3695.2012.08.010

Novel semi-supervised clustering algorithm based on active data selection

WEN Ping[1], LENG Ming-wei[1,2], CHEN Xiao-yun[1]
(1. *School of Information Science & Engineering, Lanzhou University, Lanzhou* 730000, *China*; 2. *School of Mathematics & Computer Science, Shangrao Normal University, Shangrao Jiangxi* 334001, *China*)

Abstract: Semi-supervised clustering, which aims to significantly improve the clustering results using limited supervision, has inevitably been the research focus in data mining and machine learning in recent years. But the accuracy of existing semi-clustering algorithms is low when dealing with the datasets with little labeled data or the multi-density and unbalanced datasets. Based on the active learning, this paper studied the data selection and presented a novel semi-supervised clustering algorithm. It selected information-rich data as labeled data by combining the ideas of minimum spanning tree clustering and active learning, and then used the KNN-like technology to propagate labels. Evaluating on several UCI standard datasets and synthetic datasets, the results show that the proposed method has manifest higher accuracy and stable performance in comparison with others, even when the datasets are multi-density and unbalanced.

Key words: data mining; semi-supervised clustering; active learning; labeled data; data selection; minimum spanning tree; multi-density dataset; unbalanced dataset

0 引言

利用少量标签数据对聚类过程进行指导的半监督聚类算法是数据挖掘和机器学习领域的重要研究方向。国内外的相关研究主要是对 K-means 算法[1~3]和 DBSCAN 算法[4,5]的变种或改进，constrained-K-means[2]和 SSDBSCAN 是比较典型的两种算法[5]。但是它们都要求标签数据分布到每个类，如果不能够保证从每个类中选取到标签数据，那么聚类结果将会很差。当选取的标签数据只分布到如图 1 中的 A、C 两个类中时，使用 constrained-K-means 对图 1 中的数据进行聚类，则把图 1 中的 A、B 这两个类聚成了一个类，因此如何在不平衡数据集中尽可能地保证从每个类中选取到标签数据是半监督聚类中亟需解决的问题之一。同时在实际应用中，数据的密度可能如图 2 所示，即数据集是多密度和不平衡的。对于类似于图 2 中的数据集，SSDBSCAN 这类基于密度的半监督聚类算法显然并不合适，同时由于类 C 中数据的数量相对于类 A 中数据数量而言很小，盲目地进行从数据集中选取少量数据作为标签数据，类 C 中的数据被选取的几率很小。所以使用类似于 constrained-K-means 半监督聚类算法对这种多密度不平衡数据集进行聚类，效果同样较差。

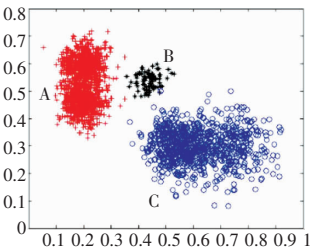

图 1　模拟不平衡数据集

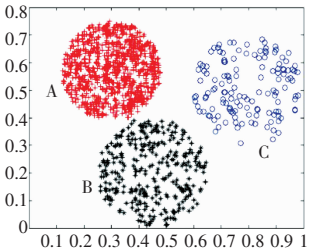

图 2　模拟多密度数据集

主动学习[6]从未标注样本集中选择包含信息量较大的样本交由专家进行标注以形成标签数据集，该方法可以有效缩小样本规模而达到较高的分类精度，被广泛应用于数据挖掘和机器学习等领域[7]。主动学习技术近几年来发展迅速，主要是为了克服传统学习技术的被动学习的状况，现今主要应用于分类，应用于聚类的较少。Basu 等人[8]基于 K-means 聚类算法

收稿日期：2012-01-15；**修回日期**：2012-02-21　　**基金项目**：江西省教育厅科技课题资助项目(GJJ11609)

作者简介：文平(1986-)，男，四川泸州人，硕士，主要研究方向为聚类分析(wenping2005@foxmail.com)；冷明伟(1979-)，男，山东菏泽人，副教授，博士研究生，主要研究方向为聚类分析、机器学习；陈晓云(1954-)，女，教授，博导，主要研究方向为数据挖掘、数据仓库、数据库.

提出了一种两阶段的主动学习模式;Huang 等人[9]提出了一种使用聚类中心来表示每个类的半监督文本聚类算法;Grira 等人[10]通过使用主动机制选取候选约束提出了一种主动模糊约束的聚类算法。文献[11]中提出了一种使用已知标签数据和成对点约束对近邻传播算法的相似度矩阵进行调整的半监督聚类算法。文献[12]中把属性子集选择的思想引入到主动学习应用于图像分类,提出了一个选择标签图像的主动学习算法。文献[13]中分析了使用主动学习的抽样复杂度问题,并得出使用主动学习的抽样比一些传统方法有更好的效果。文献[14]中从多个角度分析并对比了多个主动学习方法,指出应用主动学习可减少样本以训练出更好的分类器。

针对在多密度、不平衡数据集上进行聚类时存在的问题,本文借鉴主动学习的思想,利用最小生成树聚类的结果,从每个类中主动选取包含信息量较大的样本数据作为标签数据,同时基于标签数据所在类的标签扩展阈值进行类标号的传播以达到聚类目的。

1 基于主动数据选取的半监督聚类算法

基于主动数据选取的半监督聚类算法分为两步:a)利用主动学习技术选取标签数据,目的是选取信息量较大的样本数据作为标签数据集;b)基于少量标签数据的半监督聚类,利用选取的标签数据集通过标签传播扩展的方式进行聚类。

1.1 主动学习选取标签数据

基于主动学习的标签数据选取算法,以主动学习思想为指导,使用最小生成树的聚类算法[15]对数据集进行预聚类,然后选取每个类中密度最大的点作为待标签数据。在最小生成树中,本文按边的不一致权值以从大到小顺序删除边,直到产生给定标签数据个数的类为止,然后从每个类中选取密度最大的一个点作为待标号的数据。最小生成树 MST 中,某一边 e 的不一致权值 f_e 定义为

$$f_e = \frac{w_e}{\max(w_{E_1}, w_{E_2})} \tag{1}$$

其中:w_e 表示边 e 的权值;v_1 和 v_2 是 e 的两个端点;E_1 和 E_2 分别是从 v_1 和 v_2 出发,搜索深度为 d(本文中 $d=3$)的所有边的集合,但不包括 e 所在路径上的边;w_{E_1} 和 w_{E_2} 是 E_1 和 E_2 中边权值的均值。本文使用 v_1、v_2 之间的距离表示边 e 的权值 w_e。

选取标签数据时,为了避免将噪声数据或孤立点聚成一个单独的类,影响选取数据的质量,本文对每个类中包含数据的最小个数定为 3。如果某个类中数据数量少于 3 则将该类数据从最小生成树中删除,而且把只有一个数据的类标记为噪声。对于每个类中的数据点来说,密度最大的点应该是最能够代表该类中数据的特征,所以从每一个类中选取密度最大的点作为待标号的数据。

算法 1 基于主动学习的标签数据选取

输入:数据集 $D(x_1, x_2, \cdots, x_n)$,要选取的标签数据个数 m。

输出:标签数据集 D_L。

a)计算数据集 D 的最小生成树 MST,计算 MST 所有边的不一致权值 f;

b)按 f 值从大到小的顺序删除 MST 中的边,直到产生 m 个类为止,以 $S_1, S_2, \cdots, S_m$ 来标记这 m 个类;

c)计算 $S_1, S_2, \cdots, S_m$ 中的每个数据点在所属类内的密度,以计算 KNN 的方法从每个类中找出一个密度最大点,记为 $y_1, y_2, \cdots, y_m$;

d)将 $y_1, y_2, \cdots, y_m$ 这 m 个数据点交由领域专家进行类别标号,用 D_L 表示这 m 个数据组成的标签数据集并输出。

算法 1 中,使用 Prime 算法构建数据集 D 的最小生成树 MST,其时间复杂度为 $O(n^2)$,其中 n 是数据集 D 的大小。使用邻接表存储 MST,计算其 $n-1$ 条边的 f 值时,使用广度优先搜索的方式,搜索深度为 3,则可以把计算 f 的时间复杂度当做是 $O(n)$。切割 MST 构成 m 个类,然后计算各个类中每个对象在其类内部的 KNN,求出密度最大的对象,则计算 KNN 的时间复杂度可认为是 $O(p^2)$,其中 $p=n/m$。因此,基于主动学习的标签数据选取算法时间复杂度为 $O(n^2+n+p^2)$ 即 $O(n^2)$。

1.2 基于少量标签数据的聚类

为了对本文提出的基于少量标签数据的半监督聚类进行更好的说明,本小节首先给出几个符号定义。

定义 1 $k\_\mathrm{dst}(D)$。给定数据集 D 和其中的某个数据点 p, $k\_\mathrm{dst}(p)$ 表示点 p 到它的第 k 个最近邻的距离(在整个数据集 D 中计算 p 的 k 个最近邻),则 $k\_\mathrm{dst}(D)$ 定义为

$$k\_\mathrm{dst}(D) = \{k\_\mathrm{dst}(p) \mid p \in D\} \tag{2}$$

定义 2 $k\_\mathrm{dst}(C_t)$。给定数据集 D 中的某个类 C_t,$k\_\mathrm{dst}(C_t)$ 定义为

$$k\_\mathrm{dst}(C_t) = \{k\_\mathrm{dst}(p) \mid p \in C_t, C_t \subseteq D\} \tag{3}$$

定义 3 $k\_\mathrm{avgdst}(C_t)$。给定数据集 D 中的某个类 C_t,$k\_\mathrm{avgdst}(C_t)$ 定义为

$$k\_\mathrm{avgdst}(C_t) = \frac{\sum_{p \in C_t} k\_\mathrm{dst}(p)}{|C_t|}, \text{其中 } C_t \subseteq D \tag{4}$$

在整个数据空间内,密度较大的数据点和其 k 个最近邻中的多数邻居应该属于同一个类(把 $k\_\mathrm{dst}(p)$ 较小的点 p 称做密度比较大的数据点)。在同一个类中,密度较大点的数据能更好地代表本类的数据特征。基于少量标签数据的半监督聚类算法首先对标签数据按照其密度进行排序;然后从中取出密度最大的标签数据 x(假设 $x \in C_t$),以 x 所在类的平均距离 $k\_\mathrm{avgdst}(C_t)$ 为扩展阈值,对 x 在数据集 D 中的 k 个最近邻进行扩展并标号;其次将扩展后的数据按照 $k\_\mathrm{dst}(p)$ 的值将其插入到标签数据列表 dlst 中,同时更新 $k\_\mathrm{avgdst}(C_t)$;最后将 x 从 dlst 中删除,重复上述过程直到 dlst 为空。在对标签数据进行扩展的过程中,采用的扩展阈值是根据标签数据所属类别自动计算的阈值 $k\_\mathrm{avgdst}(C_t)$。如果两个类的密度不同,则对该类进行标签扩展过程中采用的阈值不同,实现了针对数据密度的扩展技术。

算法 2 基于少量标签数据集的聚类

输入:数据集 $D(x_1, x_2, \cdots, x_n)$,近邻阈值 k 和标签数据集 $D_L(x_1, x_2, \cdots, x_m)$。

输出:聚类结果。

a)基于 D_L 的类标签传播扩展。

(a)计算数据集 D 中所有数据的 KNN。

(b)把 D_L 中的数据按照其类标号分成 q 个类:$C_1, C_2, \cdots, C_q$,计算每个类的 $k\_\mathrm{avgdst}(C_t)$ $(1 \leqslant t \leqslant q)$。

(c)把 D_L 中的数据按照其各自的 $k\_\mathrm{dst}(p)$ 升序排列组成标签数据列表 dlst。

(d)扩展类标签。

①从 dlst 中取出第一个(即 $k\_\mathrm{dst}(p)$ 最小)数据 x_i $(1 \leqslant i \leqslant n)$ 且 $x_i \in C_t$ $(1 \leqslant t \leqslant q)$;

②如果 x_i 的第 j 个最近邻未标号且满足:$j\_\mathrm{dst}(x_i) \leqslant k\_\mathrm{avgdst}(C_t)$,则用 x_i 的类标号标记该最近邻,且加入到 C_t 中,按 $k\_\mathrm{dst}(p)$ 升序插入到 dlst 中,更新 $k\_\mathrm{avgdst}(C_t)$;

③循环步骤①②直至 dlst 为空。

b)把 D 中剩余未标号的数据按 KNN 分类算法进行标号。

c)输出聚类结果。

算法 2 中,计算 KNN 的时间复杂度为 $O(kn^2)$;在排序或更新 dlst 时,采用二分插入排序,其时间复杂度为 $O(m^2)$;在扩展类标签是一个循环、更新 dlst 和更新 $k\_avgdst(C_i)$ 的过程,因为一趟插入排序的时间复杂度可认为是 $O(n)$,所以类标签扩展的时间复杂度可近似计算成 $O(n^2)$。因而整个聚类算法的时间复杂度为 $O(kn^2+m^2+n^2)$,即 $O(n^2)$,m 是选取的标签数据个数。

2　实验与分析

为了验证本文算法,实验中使用标准数据集 UCI[16] 中的 Iris、Wine 和 Ecoli 数据集,模拟不平衡数据集和模拟多密度不平衡数据集对算法进行测试。对每个数据集选取 1% ~10% 的数据作为标签数据。实验中,把本文提出的算法与 KNN、DBSCAN、SSDBSCAN、MST 和 constrained-K-means 聚类算法进行了比较。对每个数据集,随机选取 60% 的数据作为 KNN 算法的标签数据集。当使用基于 MST 的算法聚类时,使用式(1)的定义来计算边的不一致权值。Constrained-K-means 算法聚类时,使用随机选取的方法选取数据并标号作为标签数据集。

2.1　Iris 数据集

Iris 数据集中一共有 150 个数据,包含 3 个类,每个类包含 50 个数据。选取 1% ~10% 的数据作为标签数据,选取的标签数据个数分别为 3、4、5、6、8、9、11、12、14 和 15,在 $k=6$ 的情况下,聚类结果如图 3 所示。从图 3 可以看出,本文算法在选取的标签数据超过 3%(4 个数据)以后,其聚类精度明显高于其他算法。但是选取的标签数为 2%(3 个数据)时,本文算法的聚类精度不高。深入分析发现,这是因为 IRIS 的第 2、3 个类的边界很模糊,当选取 2% 的标签数据时,只能选取到第 1、2 个类中的数据;当标签数据超过 3% 以后,可以保证每个类中至少有一个数据被选取。本文的半监督聚类算法的精度达到一个比较稳定的状态,验证了本文的半监督聚类算法在极少量标签数据时就有一个较高的聚类精度,甚至高于 KNN 分类算法的精度。因为在 KNN 分类算法中,即使选取 60% 的标签数据,由于标签数据的分布很难与整个数据空间一致,导致了 KNN 分类算法只有 95% 的精度。

2.2　Wine 数据集

Wine 数据集一共有 178 个数据,每个数据有 13 个数值属性,包含 3 个类,第 1 个类含有 59 个数据,第 2 个类含有 70 个数据,第 3 个类含有 79 个数据。选取的标签数据个数为 3、4、5、7、9、11、12、14、16 和 18,给定 $k=5$,聚类结果如图 4 所示。

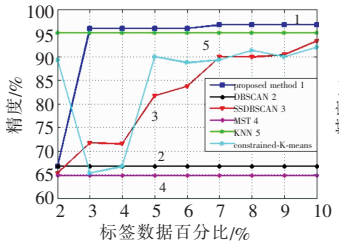

图 3　Iris 数据集聚类结果

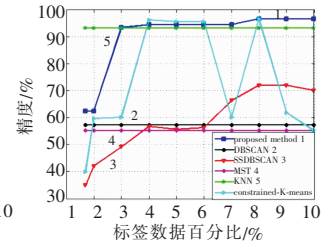

图 4　Wine 数据集聚类结果

Wine 数据集和 Iris 类似,也是第 2、3 个类的边界比较模糊,甚至两个类中的数据出现重叠现象,当标签数据个数很少时(3(1.7%)、4(2.2%)),本文的选取算法只能从第 1、2 个类中选出标签数据,所以这时的聚类精度只有 62.35%。当选取的标签数据超过 5(2.8%)时,聚类精度明显比 DBSCAN、SSDBSCAN 和 MST 聚类算法的高,且略微高于 KNN 分类的精度,同时聚类精度达到一个稳定状态。虽说当选取 4% ~6% 的标签数据时,半监督距离算法 constrained-K-means 的精度略高于本文的聚类算法的精度,但 constrained-K-means 算法不稳定,其精度过重地依赖于标签数据集,当标签数据集不同时,精度相差较大。

2.3　Ecoli 数据集

Ecoli 数据集中有 336 个数据,每个数据有 7 个属性,一共有 8 个类,分别包含的数据个数为 143、77、52、35、20、5、2 和 2。从 Ecoli 中选取的标签数据个数为 8、10、13、17、20、24、27、30 和 34。$k=6$ 时,聚类结果如图 5 所示。

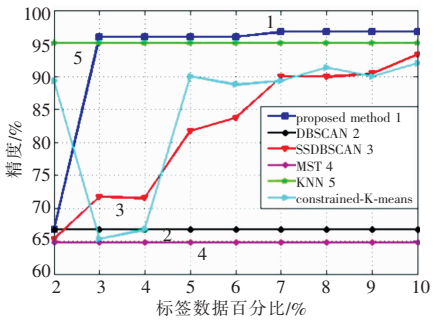

图 5　Ecoli 数据集聚类结果

Ecoli 数据集实验室中,给定 $k=5$。该数据集中的第 6、7、8 个类都比较小,而第 1、2 个类都相对比较大,因而可以把 Ecoli 当做是类倾斜的数据集。从 Ecoli 中选取的标签数据集只覆盖了前面 6 个类,从图 5 可知,本文算法聚类精度明显高于 DBSCAN、SSDBSCAN、constrained-K-means 和 MST,当选取的标签数据超过 5% 后,本文聚类算法达到稳定状态,并且聚类精度明显高于 KNN 算法的分类精度。

2.4　模拟不平衡数据集

模拟的不平衡数据集如图 1 所示。其中一共有 1 880 个数据,每个数据包含 2 个属性,包含 3 个类,大小分别为 1000、80 和 800。选取的标签数据的个数为 19、38、56、75、94、113、132、150、169 和 188,给定 $k=6$,聚类结果如图 6 所示。

从图 6 可以看到本文算法精度已接近 100%,这说明对于类边界较明显的不平衡数据集,选取的标签数据能够分布到每个类。在类边界比较明显的不平衡数据集上,虽说多数算法都有一个比较高的精度,但是本文算法的精度基本可以达到 100%。而对于半监督聚类算法 constrained-K-means 来说,由于不能保证每个类中都有标签数据被选取,导致聚类精度较差。

2.5　模拟多密度数据集

本文模拟的多密度数据集其分布如图 2 所示,该数据集是一个的多密度、不平衡的数据集。其中一共有 1 245 个数据,每个数据包含 2 个属性,包含 3 个类,大小分别为 791、315 和 139。选取的标签数据个数分别为 12、25、37、50、62、75、87、100、112 和 124,给定 $k=6$,聚类结果如图 7 所示。

从图 7 可以看出,对该多密度、不平衡数据集,虽说 KNN 算法的分类精度是 100%,但在多数情况下本文算法也很接近 100%,DBSCAN、MST、SSDBSCAN 相对来说比较低。虽说 constrained-K-means 算法有时也能达到 100%,但是其精度受标签数据集影响较大,导致算法精度波动比较大。

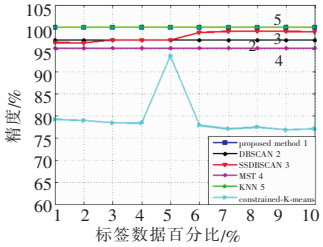

图 6 模拟不平衡数据集聚类结果

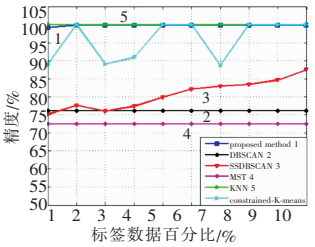

图 7 模拟多密度数据集聚类结果

观察对三个 UCI 和两个模拟数据集的实验结果可以看出,本文算法的聚类精度是优于其他几个聚类算法的,且有时到达或超过了 KNN 分类的精度。因为在选取标签数据时,选取密度最大数据作为标签数据,所以本文提出的聚类算法精度高且稳定。Constrained-K-means 使用随机选取标签数据的方法,所以其聚类精度较低;另一方面,实验中使用的数据集有的是多密度不平衡的,且类是任意形状的,所以 constrained-K-means 在聚类这些数据时精度也不高。但是本文提出的算法使用类 KNN 的思想且不断更新类标签传播阈值,所以能够适应多密度不平衡数据集的聚类。

3 结束语

在本文中,将最小生成树聚类和主动学习思想引入到标签数据的选择中,同时基于标签扩展的方法提出了一个基于主动数据选取的半监督聚类算法。该算法能够使用较小标签数据获得较高的聚类精度,但在处理类边界比较模糊的数据集时需要较大的标签数据集。由于本文的半监督聚类算法在预聚类结果中选取密度较大的数据作为标签数据,标签数据能够很好地代表其所在类中多数数据的特性,当标签数据超过一定数量时可以达到一个比较稳定的聚类结果;当标签数据的数量少于数据集中类的数量或者标签数据不能包含所有类时,本文的半监督聚类算法不能够对新类进行检测。因此,怎样从类边界模糊的数据集中选取尽可能少的标签数据指导聚类,以及发现数据集中的新类是笔者今后研究的重点。

参考文献:

[1] WAGSTAFF K, CARDIE C, ROGERS S, *et al*. Constrained K-means clustering with background knowledge[C]//Proc of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 2001: 577-584.

[2] BASU S, BANERJEE A, MOONEY R. Semi-supervised clustering by seeding[C]//Proc of the 19th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 2002: 27-34.

[3] DANG Yan-zhong, XUAN Zhao-guo, RONG Li-li, *et al*. A novel initialization method for semi-supervised clustering[C]//Proc of the 4th International Conference on Knowledge Science, Engineering and Management. Berlin: Springer-Verlag, 2010: 317-328.

[4] RUIZ C, SPILIOPOULOU M, MENASALVAS E. Density-based semi-supervised clustering[J]. Data Mining and Knowledge Discovery, 2010, 21(3): 345-370.

[5] LELIS L, SANDER J. Semi-supervised density-based clustering[C]// Proc of the 9th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2009: 842-847.

[6] LEWIS D D, GALE W A. A sequential algorithm for training text classifiers[C]//Proc of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: Springer-Verlag, 1994: 3-12.

[7] 张春阳,周继恩,钱权,等. 抽样在数据挖掘中的应用研究[J]. 计算机科学, 2004, 31(2): 126-128, 141.

[8] BASU S, BANERJEE A, MOONEY R J. Active semi-supervision for pairwise constrained clustering[C]//Proc of the 4th SIAM International Conference on Data Mining. 2004: 333-344.

[9] HUANG Rui-zhang, LAM W, ZHANG Zhi-gang. Active learning of constraints for semi-supervised text clustering[C]//Proc of the 7th SIAM International Conference on Data Mining. 2007: 113-124.

[10] GRIRA N, CRUCIANU M, BOUJEMAA N. Active semi-supervised fuzzy clustering[J]. Pattern Recognition, 2008, 41(5): 1834-1844.

[11] BALCAN M F, HANNEKE S, VAUGHAN J W. The true sample complexity of active learning[J]. Machine Learning, 2010, 80(2-3): 111-139.

[12] ZHANG Qing-jiu, SUN Shi-liang. Multiple-view multiple-learner active learning[J]. Pattern Recognition, 2010, 43(9): 3113-3119.

[13] ZAHN C T. Graph-theoretical methods for detecting and describing gestalt clusters[J]. IEEE Trans on Computers, 1971, 20(1): 68-86.

[14] ASUNCION A, NEWMAN D. UCI machine learning repository[EB/OL]. http://archive.ics.uci.edu.cn/ml/datasets.html.

[15] 肖宇,于剑. 基于近邻传播算法的半监督聚类[J]. 软件学报, 2008, 19(11): 2803-2813.

[16] SHEN Jian-feng, JU Bin, JIANG Tao, *et al*. Column subset selection for active learning in image classification[J]. Neurocomputing, 2011, 74(18): 3785-3792.