一种符号型增量数据标签算法

李艳红 李德玉 王素格
(山西大学计算机与信息技术学院 太原 030006)
(计算智能与中文信息处理教育部重点实验室 太原 030006)

摘 要 数据标签是一种提高增量数据聚类效率的简单而有效的方法。数据标签就是分配每个新增数据点到与之最相似的簇的过程。符号数据分析的难点之一在于缺少一种恰当的方法来定义数据点与数据簇之间的相似性。为此,将簇代表定义为簇中所有属性的属性值及其在簇中的频率构成的列表,用信息熵的变化来定义"点-簇"不相似性。基于此不相似性度量,设计了一个符号型增量数据标签算法来分配无标记数据到恰当的簇。在公开数据集和文本语料上的对比实验表明,该数据标签算法不但数据标记精度高、时间开销小,而且有较好的可伸缩性。

关键词 聚类,数据标签,增量数据,符号数据,信息熵
中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.047

Categorical Incremental Data Labeling Algorithm

LI Yan-hong LI De-yu WANG Su-ge
(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)
(Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006, China)

Abstract Data labeling has become a simple but efficient solution to improve the efficiency of incremental data clustering. This process of data labeling is performed by assigning each new coming data point to some cluster that is closest to the new data point. One of the main difficulties in categorical data analysis is, however, lacking an appropriate way to define the similarity between data point and cluster. To overcome this difficulty, in this paper, we defined the representative of a cluster as a list of all attribute values with their frequencies in each attribute domain of the cluster, and then, defined the point-cluster dissimilarity measure by means of the change of information entropy. Based on the dissimilarity measure, we designed a categorical incremental data labeling algorithm, to allocate each unlabeled data point into the appropriate cluster. Comparative experiments on several public data sets and a text corpus show that the proposed algorithm has not only the higher labeling accuracy and the less execution time, but also better scalability.

Keywords Clustering, Data labeling, Incremental data, Categorical data, Information entropy

1 引言

聚类是数据分析的一项重要任务,其目的在于将一组无标签的数据划分为一些有意义的簇,使得相似的数据划分在同一个簇中,不相似的数据划分在不同的簇中[1]。增量数据的聚类是目前聚类算法的主要挑战之一。增量数据指的是不断有新的数据累加到初始数据集中,因而数据集是不断增长的而不是静止不变的。在初始数据集上应用聚类算法可以得到每个数据的类别标签,但是随后增加到数据集中的数据是没有类别标签的。可以通过将扩充后的整个数据集重新聚类来解决这个问题,但是这样做的效率通常很低。如果能够利用已有的聚类结果,只为这些新增加的数据打上标签,将会有效地提高聚类效率,这种方法称为数据标签技术。

研究人员在数值数据上已经开展了一些数据标签研究工作。在文献[1,2]中,使用簇中心作为簇的代表,利用数据点到簇中心的距离来度量数据点和簇的相似性,新来的数据点被分配到与簇中心距离最小的簇中。在标签算法 CURE[3] 中,使用簇中的多个随机点作为簇的代表,定义了一个数据点与多个代表点的距离,将数据点分配到距离簇代表最小的簇中。这样做的好处是当簇的形状不是球形或者簇的大小变化时,依然可以正确地对数据打标签。

由于符号数据的数值特性难以刻画,从而缺少恰当的方法来定义符号数据点和符号数据簇之间的相似性,因此在符号数据领域,数值型数据上的标签算法是不适用的。最近,已经有研究者关注符号型数据的标签技术[4-6]。在文献[4]中,一个簇的代表由簇中每个属性的出现频率最高的属性值构成。这种簇代表的定义相对简单,但忽视了属性中出现频率接近最高值的那些属性值的代表性。为了完善符号数据簇代

到稿日期:2014-09-14 返修日期:2014-12-09 本文受国家自然科学基金(61272095,61175067,61303091,61202365,61100138,61403238),山西省自然科学基金(2012061015),山西省科技攻关项目(20110321027-02),山西省回国留学人员科研项目(2013-014)资助。
李艳红(1977—),女,博士生,讲师,主要研究方向为数据挖掘与机器学习,E-mail:liyh@sxu.edu.cn;李德玉(1965—),男,博士,教授,博士生导师,主要研究方向为智能计算与数据挖掘;王素格(1964—),女,博士,教授,主要研究方向为自然语言处理与文本挖掘。

表的定义，研究者提出了一些改进算法。在文献[5]中，一个簇的代表是簇中所有属性的属性值和对应权重构成的列表，权重是属性值在簇中的频率和属性值在所有簇中的分布的函数。一个属性值对应的权重越高，该属性对簇的代表性就越强。在文献[6]中，一个簇的代表也是簇中所有属性的属性值和对应权重构成的列表，不同的是权重是属性值在簇中的频率和属性值在所有簇中的频率的函数。尽管这些算法对文献[4]中的方法有了不同程度的改进，但是符号数据标签算法仍然面临算法复杂、效率低的问题，算法性能有待进一步提高。

国内已经有一些针对增量数据异常发现[7]、聚类[8]和规则获取[9,10]的研究工作。孟等人[7]面向动态增量数据库环境，对传统 DBSCAN 算法进行改进，提出了一种基于聚类和快速计算的异常数据挖掘算法。刘等人[8]在信息源变化的数据挖掘体系结构下，利用一群特殊的智能代理增量修改知识模型，提出了群体智能聚类模型的构建方法及增量模型维护算法。此外，在混合数据标签的研究方面，李等人[11]利用 K-prototypes 算法对用户兴趣数据聚类构建用户兴趣域，基于用户兴趣域和“数据-用户兴趣域”隶属度的概念，提出了一种基于用户兴趣的混合数据聚类标签算法。

本文主要研究符号型增量数据标签技术，使用簇中不同属性值在簇中的出现频率作为簇的代表，利用信息熵的变化定义了符号数据点和符号数据簇之间的相似性度量，并提出了一种符号型增量数据的标签算法。该算法利用初始数据集中数据的分布情况确定“点-簇”相似性阈值的大小，从而很好地解决了参数的选择问题。通过在几个 UCI 公开数据集和一个文本语料库上的对比实验，验证了算法的标记精度、时间效率和可伸缩性。

2 问题定义及数据标签框架

2.1 问题的形式化定义

由符号数据构成的初始数据集记作 D，$\forall \boldsymbol{x}_i \in D$ 是一个 d 维向量，即 $\boldsymbol{x}_i=(x_i^1,x_i^2,\cdots,x_i^d)$，分量 $x_i^j(1\leqslant j\leqslant d)$ 从第 j 个属性的值域 V_j 中取值。使用 k-modes[12] 聚类算法对数据集 D 进行聚类并得到聚类结果 $C=\{c_1,c_2,\cdots,c_k\}$，$c_m(1\leqslant m\leqslant k)$ 表示聚类结果中的第 m 个簇，其簇标签为 l_m。记 r_m 为 c_m 的簇代表。新增到 D 中的数据构成的增量数据集记作 E，E 中的元素是没有类别标签的。对于任意的数据 $\boldsymbol{x}\in E$，利用下文所提出的“点-簇”不相似性度量来寻找 C 中与 $\boldsymbol{x}$ 不相似性最小(即最相似)的数据簇 c_{m^*} $(1\leqslant m^*\leqslant k)$。如果该最小不相似性大于所选取的阈值，则将 $\boldsymbol{x}$ 标记为 c_{m^*} 的标签 l_{m^*}，否则将 $\boldsymbol{x}$ 标记为异常点。

2.2 数据标签框架

符号型增量数据标签的精度取决于两个方面。第一，从簇中提取能有效刻画簇结构的簇代表；第二，定义一个恰当的数据点和簇之间的相似性度量。本文所提出的符号型增量数据标签的框架如图 1 所示，数据标签过程分为两个阶段：初始化阶段和数据标签阶段。

初始化阶段，首先通过调用符号型数据聚类算法对初始数据库进行聚类，得到部分聚类结果；然后对数据簇进行分析，得到每个簇的簇代表，并确定一个衡量新增数据点与数据簇是否相似的阈值。

数据标签阶段，根据“点-簇”不相似性度量来确定每个新增数据点与所有簇的不相似性，并将新数据点标记为与之不相似性最小的簇的标签。但是，如果新数据点与最相似的簇的不相似性大于所选阈值，说明该新增数据点与已知数据簇的不相似性都很高，此时将该新增数据点标识为一个异常点。在数据标签的过程中当异常点达到一定数量时，需要必要的异常处理办法来为其打标签。本文采用重聚类的方式来为异常点打标签。

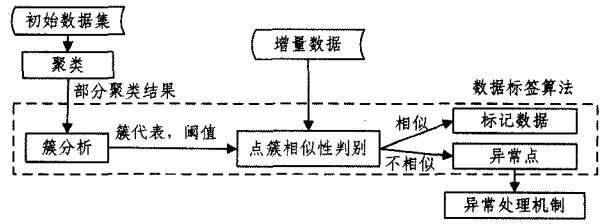


图 1 符号型增量数据标签框架

3 数据点与簇的相似性定义

本节将介绍符号型增量数据标签算法的核心，即如何提取有效的簇代表，以及如何定义未标记数据与初始数据聚类结果中的簇的相似性度量。

3.1 符号数据簇代表

定义 1 初始数据集 D 的聚类结果为 $C=\{c_1,c_2,\cdots,c_k\}$，将属性 $x^j(1\leqslant j\leqslant d)$ 看作一个离散随机变量，其值域为 V_j。针对数据簇 $c\in C$，可以根据式(1)构造一个离散概率分布：

$$p(x^j=v)|_{v\in V_j}=\frac{|\{\boldsymbol{x}\in c;x^j(\boldsymbol{x})=v\}|}{|c|} \tag{1}$$

其中，$x^j(\boldsymbol{x})$ 表示样本 $\boldsymbol{x}$ 在属性值 x^j 下的取值。将该离散概率分布确定的随机变量记作 c^j，由此可以得到一个 d 维的离散随机向量 $(c^1,c^2,\cdots,c^d)$。在本文提出的符号型增量数据标签算法中，将该离散随机向量作为数据簇 c 的簇代表，记作 r。该簇代表由簇中所有属性的属性值和属性值在簇中的频率构成。

3.2 信息熵与“点-簇”不相似性

符号数据分析的难点在于缺少恰当的方法衡量数据之间的相似性。对于符号型增量数据标签算法而言，需要一种有效的方法来衡量符号数据点与符号数据簇之间的相似性。

一个符号数据簇的结构是由每个属性的属性值的分布决定的。在信息论里，信息熵是对不确定性的度量。一个符号数据簇的信息熵越高，则信息量越多；反之，信息量越少[13,14]。在一个簇中加入一个和簇中大部分数据点都相似的数据点将不会显著改变簇的分布，信息熵的变化相对较小；反之，在一个簇中加入一个和簇中大部分数据点都不相似的数据点，将会显著改变簇的分布，信息熵的变化相对较大。因此，尽管无法用信息熵直接度量一个符号数据点与一个符号数据簇的相似程度，但是可以利用将一个数据点加入一个簇前后信息熵的改变来衡量数据点和数据簇之间的相似程度，信息熵变化的程度越小，表明数据点和数据簇越相似，反之越不相似。

定义 2 令 x 是一个离散随机变量，分别以概率 $p_1,p_2,\cdots,p_n$ 取值 $v_1,v_2,\cdots,v_n$，并且 $p_i\geqslant 0(i=1,2,\cdots,n)$，$\sum_{i=1}^{n}p_i=1$。

那么离散随机变量 x 的信息熵 $H(x)$ 定义为：$H(x)=-\sum_{i=1}^{n} p_i \log_2 p_i$。

定义 3 令 $X=(x^1,x^2,\cdots,x^d)$ 是一个离散随机向量，x^j $(1\leqslant j\leqslant d)$ 的值域为有限集 V_j。用 $p(x^j=v)$ 表示事件 $x^j=v$ $(v\in V_j)$ 的概率。假设随机变量 x^j $(1\leqslant j\leqslant d)$ 是相互独立的，则 X 的信息熵 $H(X)$ 定义为：$H(X)=\sum_{i=1}^{d} H(x^j)=-\sum_{j=1}^{n}\sum_{v\in V_j} p(x^j=v)\log_2 p(x^j=v)$。

下面，利用信息熵来定义一个数据点与一个簇的不相似性。

定义 4 数据点 $\boldsymbol{x}$ 与簇 c 的不相似性的定义如式(2)所示：

$$\begin{aligned} d(\boldsymbol{x},c)&=|\{\boldsymbol{x}\}\cup c|H(\{\boldsymbol{x}\}\cup c)-|c|H(c)\\ &=|\{\boldsymbol{x}\}\cup c|H(r')-|c|H(r) \end{aligned} \tag{2}$$

其中，r 为簇 c 的代表，r' 为簇 $\{\boldsymbol{x}\}\cup c$ 的代表。

性质 1 对于 $\forall \boldsymbol{x}\in E, \forall c\in C$，有下面的性质：

(1) $d(\boldsymbol{x},c)$ 取得最大值当且仅当

$$x^j(\boldsymbol{x})=\begin{cases} v\in V_j\backslash x^j(c), & V_j\backslash x^j(c)\neq\varnothing \\ v\in\{\arg\min\limits_{v'\in V_j}|(x^j)_c^{-1}(v')|\}, & \text{otherwise} \end{cases}$$

其中，$x^j(c)=\{x^j(\boldsymbol{x})\in V_j:\boldsymbol{x}\in c\}$，$(x^j)_c^{-1}(v')=\{\boldsymbol{x}\in c:x^j(\boldsymbol{x})=v'\}$。

(2) $d(\boldsymbol{x},c)$ 取得最小值当且仅当 $x^j(\boldsymbol{x})\in\{\arg\max\limits_{v'\in V_j}|(x^j)_c^{-1}(v')|\}$。

(3) $d(\boldsymbol{x},c)=0$ 当且仅当 $c=\{\boldsymbol{x}'=\boldsymbol{x}|\boldsymbol{x}'\in c\}$，即 c 中的所有的数据点都与 $\boldsymbol{x}$ 相同。

4 符号型增量数据标签算法

4.1 阈值确定

为了提高增量数据的聚类效率，本文在初始数据库聚类结果的基础上，利用定义 4 给出的“点-簇”不相似性来度量增量数据库中的数据 $\boldsymbol{x}$ 与初始数据库的聚类结果中的簇的相似性。如果 $\boldsymbol{x}$ 和与 $\boldsymbol{x}$ 最相似的簇 c_m 之间的不相似性不大于阈值 σ，则认为 $\boldsymbol{x}$ 的标签为 l_m，否则认为 $\boldsymbol{x}$ 属于新的数据簇，将 $\boldsymbol{x}$ 标记为异常点。当异常点的数量达到 M 时，对所有异常点进行聚类。

衡量 $\boldsymbol{x}$ 是否属于一个簇 c_m $(1\leqslant m\leqslant k)$ 的主要难点在于阈值 σ 的确定。本文提出的数据标签算法使用 c_m 中的数据点来确定阈值 σ_m 的大小，从而为每个簇指定不同的阈值 σ_m。考虑到阈值 σ_m 与数据簇 c_m 中最边缘的数据点有内在关联，本文选定阈值 σ_m 为 c_m 中的样本与 c_m 的最大不相似性的线性函数(即 $\sigma_m=(1+\lambda|E|/|D|)\max\limits_{\boldsymbol{x}\in c_m} d(\boldsymbol{x},c_m)$，其中 $\max\limits_{\boldsymbol{x}\in c_m} d(\boldsymbol{x},c_m)$)为 c_m 中的样本与 c_m 的不相似性的最大值。此外，本文所给出的阈值确定办法考虑了新增数据集和原始数据集的元素个数之比，其目的是在数据标签的过程中允许簇的边界有一定的扩展，从而适应增量数据动态变化的要求。λ 为调节因子，在[0,1]上取值。

4.2 算法描述

基于图 1 所示的符号型增量数据标签框架，本文设计了符号型增量数据标签算法 CIDL，具体描述见算法 1。

算法 1 符号型增量数据标签算法 CIDL

输入：初始数据库 D，增量数据集 E

输出：增量数据库 E 中数据的类别标签

初始化阶段：

步骤 1 调用 k-modes 聚类算法对初始数据集 D 聚类，得到聚类结果 $C=\{c_1,c_2,\cdots,c_k\}$；

步骤 2 计算 c_m $(1\leqslant m\leqslant k)$ 的簇代表 r_m；

步骤 3 计算每个簇的相似性阈值 $\sigma_m=(1+\lambda|E|/|D|)\max\limits_{\boldsymbol{x}\in c_m} d(\boldsymbol{x},c_m)$。

数据标签阶段：

对于 $\forall \boldsymbol{x}\in E$，进行如下操作：

步骤 1 如果 $\min\limits_{1\leqslant m\leqslant k} d(\boldsymbol{x},c_m)\leqslant\sigma_m$，则 $\boldsymbol{x}$ 的标签为 c_m，将 $\boldsymbol{x}$ 加入数据簇 c_m，即 $c_m=c_m\cup\{\boldsymbol{x}\}$，更新 r_m，转步骤 4；

步骤 2 如果 $\min\limits_{1\leqslant m\leqslant k} d(\boldsymbol{x},c_m)>\sigma_m$，则 $\boldsymbol{x}$ 被标记为异常点；

步骤 3 如果异常点数目达到 M，则调用 k-modes 聚类算法对所有的异常点聚类，得到聚类结果 $C'=\{c_1',c_2',\cdots,c_{k'}'\}$，计算每个簇的簇代表及相似性阈值，$C=C\cup C'$；

步骤 4 新增数据 $\boldsymbol{x}$ 处理完成。

5 实验及结果分析

5.1 测试环境与数据集

文中所有实验均在 PC 机上完成，配置为 Intel Pentium 2.66GHz 处理器，3.37GB 内存，Windows XP SP3 操作系统。

实验使用两类数据集来测试算法的性能，第一类为来自公开数据集 UCI (University of California at Irvine)[15] 的 Dermatology、Car、DNA、Mushroom 和 Connect-4 数据集；第二类为第一届和第二届中文情感分析评测语料库的文本数据集，为了方便起见，本文称之为 Subject text。

Dermatology 数据集包含的 366 个对象、6 个类别、33 个属性均为符号型数据。Car 数据集包含的 1728 个对象、4 个类别、6 个属性均为符号型数据。DNA 数据集包含的 3190 个对象、3 个类别、60 个属性均为符号型数据。Mushroom 数据集包含的 8124 个对象、2 个类别、22 个属性均为符号型数据。Connect-4 数据集包含的 67557 个对象、3 个类别、42 个属性均为符号型数据。Subject text 文本数据集包含的 1280 个文本数据、属于 8 个类别、271 个属性均为符号型数据。这些数据集的主要特征如表 1 所列。

表 1 数据的主要特征

数据集	对象数	属性数	类别数
Dermatology	366	33	6
Car	1728	6	4
DNA	3190	60	3
Mushroom	8124	22	2
Connect-4	67557	42	3
Subject text	1280	271	8

5.2 算法性能评价

分别在 Dermatology、Car、DNA、Mushroom 和 Subject text 这 5 个数据集上将本文提出的算法 CIDL 与 3 个已有的符号型数据标签算法做对比实验。具体的实验方法为：从数据集中随机选取一定比例的数据作为初始数据集 D，数据集中剩余的样本被看作是新增数据集 E。由于实验的目的是比较不同数据标签算法在已知聚类划分前提下对增量数据集中

的数据的标记能力，为了不受聚类结果优劣的影响，在实验中采用的是初始数据集 D 中的数据的真实类别标签。采样的百分比记为 θ，实验中 θ 从 0.7 到 0.97 取值，步长为 0.03。调节因子 λ 取值为 0.5。

为了评价符号型增量数据标签算法的标记精度，精度指标 AC 定义为：$AC=a/|E|$，其中，a 是 E 中被正确标记的样本数。

图 2—图 6 对比了本文提出的标签算法 CIDL 与其它 3 种算法（分别简称为 Frequency[6]、NIR[7] 和 RMF[8]）在不同数据集上的标记精度和时间开销。图中的实验结果均为 20 次实验的平均值。

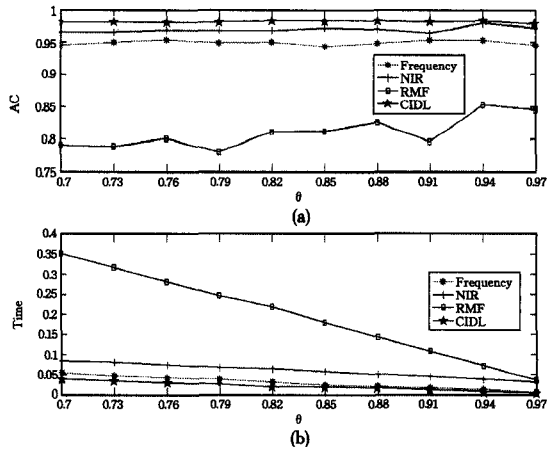


图 2　Dermatology 上的标记精度和时间开销

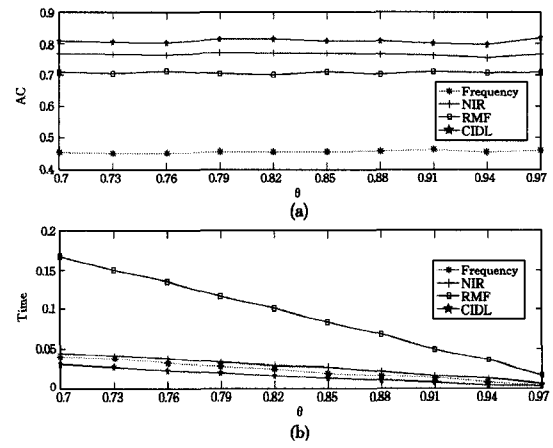


图 3　Car 上的标记精度和时间开销

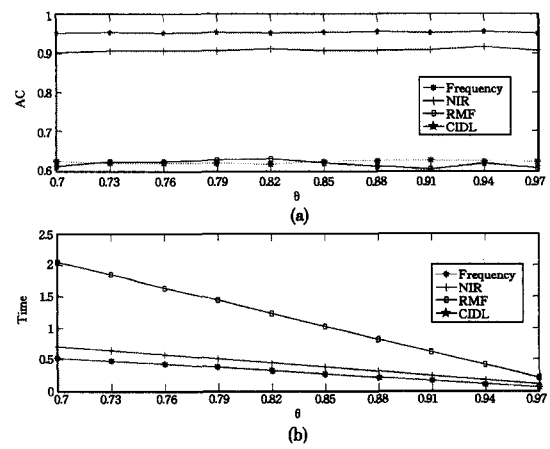


图 4　DNA 上的标记精度和时间开销

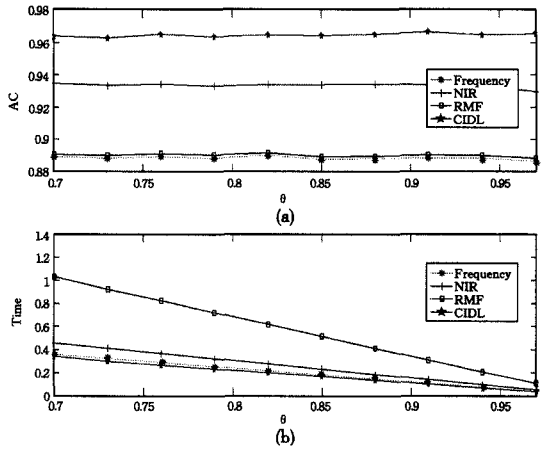


图 5　Mushroom 上的标记精度和时间开销

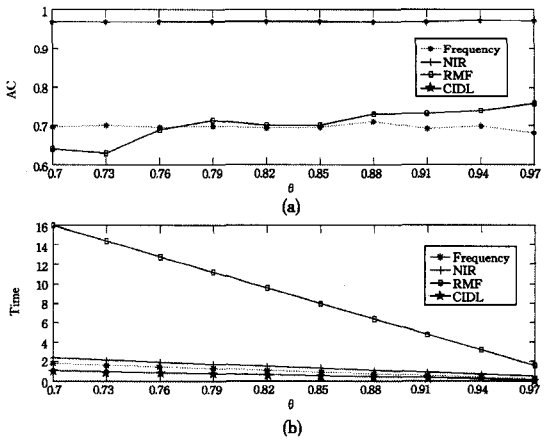


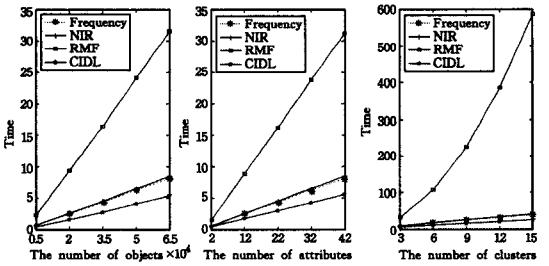
图 6　Subject text 上的标记精度和时间开销

从图 2—图 6 中可以看到，对于 5 个不同的数据集以及不同的 θ 取值，本文提出的算法 CIDL 和算法 NIR 的标记精度高于其它 2 个算法的精度。在除了 Subject text 的其它数据集上，算法 CIDL 标记精度略高于算法 NIR。总体而言，随着 θ 的增长，4 个算法在不同数据集上的时间开销均呈递减趋势。实验表明算法 RMF 是相对耗时的，算法 CIDL 标签时间略少于算法 NIR 和 Freguency。

5.3　算法可伸缩性评价

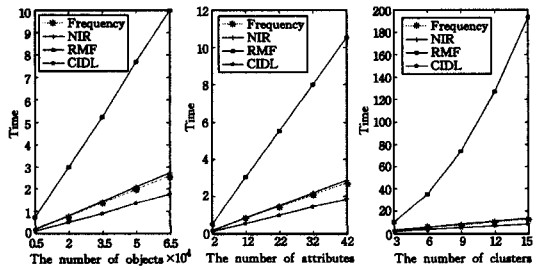
为了分析算法 CIDL 的可伸缩性，在数据集 connect-4 上对比了本文提出的算法 CIDL 与文献[6-8]中的算法的可伸缩性。所有的实验结果均为 5 次实验的平均值。

图 7 和图 8 分别展示了 $\theta=0.7$ 和 $\theta=0.9$ 时 4 种算法在数据集 connect-4 上的时间开销。图 7(a)和图 8(a)为对象数增长而属性的个数和簇的个数固定（分别为 42 和 3）时 4 种算法的计算时间。图 7(b)和图 8(b)为属性的个数增长而簇的个数和对象的个数固定（分别为 3 和 67557）时 4 种算法的计算时间。图 7(c)和图 8(c)为簇的个数增长而属性的个数和对象的个数固定（分别为 42 和 67557）时 4 种算法的计算时间。由图 7 和图 8 可知，4 种标签算法均为可伸缩的，即：算法的时间开销随着对象数、属性个数和簇的个数的增长而呈线性增长。并且，本文提出的算法 CIDL 的时间开销要比其它算法小。而 RMF 算法的时间开销要比其它算法大，这是因为该算法在数据标签过程中需要计算符号属性值在所有簇中的出现频率，而其它算法不需要计算该频率。



(a)不同的对象数　(b)不同的属性数　(c)不同的簇数

图 7　θ=0.7 时 connect-4 上的时间开销



(a)不同的对象数　(b)不同的属性数　(c)不同的簇数

图 8　θ=0.9 时 connect-4 上的时间开销

结束语　为了提高符号型增量数据的聚类效率，本文提出了一种有效的数据标签技术，利用数据簇中各属性值的分布定义簇代表元，利用信息熵的变化定义了“点-簇”不相似性度量，有效刻画了符号数据点与符号数据簇的距离的数值特性。基于此度量，设计了一个针对符号型增量数据的数据标签算法。在一些公开数据集和文本数据集上的对比实验表明，该算法能够很好地度量数据点与数据簇之间的距离，因而具有较高的标记精度。实验还表明，本文提出的算法也具有效率高、可伸缩性好的优点。算法引入了异常处理机制，通过对异常点集合的聚类可以引入新的数据簇，从而能够适应数据的动态变化。

参 考 文 献

[1] Jain A K, Murty M N, Flynn P J. Data clustering: A review[J]. ACM Computing Surveys, 1999, 31(3): 264-323

[2] Zhang T, Ramakrishnan R, Livny M. Birch: An efficient data clustering method for very large databases[C]// Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data. Montreal: ACM Press, 1996: 103-114

[3] Guha S, Rastogi R, Shim K. Cure: An efficient clustering algorithm for large databases[C]// Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. Seattle: ACM Press, 1998: 73-84

[4] Cao F, Liang J, Bai L, et al. A framework for clustering categorical time-evolving data[J]. IEEE Transactions on Fuzzy Systems, 2010, 18(5): 872-882

[5] Chen H L, Chen M S, Lin S C. Catching the trend: A framework for clustering concept-drifting categorical data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(5): 652-665

[6] Cao F, Liang J. A data labeling method for clustering categorical data[J]. Expert Systems with Applications, 2011, 38(3): 2381-2385

[7] 孟静，吴锡生. 一种基于聚类和快速计算的异常数据挖掘算法[J]. 计算机工程，2013，39(8)：60-63，68
Meng Jing, Wu Xi-sheng. An Outlier Data Mining Algorithm Based on Clustering and Rapid Calculation[J]. Computer Engineering, 2013, 39(8): 60-63, 68

[8] 刘波，潘久辉. 基于群体智能的增量数据挖掘方法研究[J]. 计算机工程与设计，2006，27(11)：180-186
Liu Bo, Pan Jiu-hui. Research of incremental data mining based on swarm intelligence[J]. Computer Engineering and Design, 2006, 27(11): 180-186

[9] 胡开明，陈建华. 一种改进的增量数据挖掘算法[J]. 计算机应用与软件，2011，28(8)：260-264
Hu Kai-ming, Chen Jian-hua. An improved algorithm for incremental data mining[J]. Computer Applications and Software, 2011, 28(8): 260-264

[10] 宋中山，成林辉，吴立峰. 一种基于关联规则的增量数据挖掘算法[J]. 湖北大学学报，2006，28(3)：240-243
Song Zhong-shan, Cheng Lin-hui, WU Li-feng. The incremental data mining algorithm based on association rules[J]. Journal of Hubei University, 2006, 28(3): 240-243

[11] 李德玉，翁小奎，李艳红. 基于用户兴趣域的混合数据聚类标签算法[J]. 山西大学学报，2013，36(2)：180-186
Li De-yu, Weng Xiao-kui, Li Yan-hong. Mixed Data clustering label algorithm based on user's interest domain[J]. Journal of Shanxi University, 2013, 36(2): 180-186

[12] Huang Zhe-xue. Extensions to the k-means algorithm for clustering large data sets with categorical values[J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304

[13] Cover T M, Thomas J A. Elements of Information Theory(2nd Edition)[M]. Hoboken: Wiley, 2006: 13-30

[14] 赵志刚，吴鑫，洪丹枫，等. 基于信息熵的 GLBP 掌纹识别算法[J]. 计算机科学，2014，41(8)：293-296
Zhao Zhi-gang, Wu Xin, Hong Dan-feng, et al. Palmprint Recognition Method Based on Energy Spectrum of GLBP [J]. Computer Science, 2006, 27(11): 180-186

[15] Frank A, Asuncion A. UCI Machine Learning Repository[OL]. 2010. http://archive.ics.uci.edu/ml

(上接第 219 页)

[12] 易培. 基于 S 型函数的倒谱域加权缺失数据处理方法[D]. 北京：清华大学，2010
Yi Pei. Cepstrum domain weighted methods for handling missing data based on S function [D]. Beijing: Tsinghua University, 2010

[13] Kuandykov L, Sokolov M. Impact of social neighborhood on diffusion of innovation S-curve[J]. Decision Support Systems, 2010, 48(4): 531-535

[14] 黄国庆，王国良，臧青松. 参数可调的战斗机空战效能评估系统研究[J]. 电光与控制，2013，20(2)：33-36
Huang Guo-qing, Wang Guo-liang, Zang Qing-song. An Air Combat Effectiveness Evaluation System with Adjustable Parameters for Fighters[J]. Electronics Optics & Control, 2013, 20(2): 33-36