基于平均互信息的混合条件属性聚类算法

刘晋胜
(广东石油化工学院计算机与电子信息学院 茂名 525000)

摘 要 混合条件属性参数间的距离值存在较大的差异,导致仅聚合距离数量级较大、较规律的数值条件属性对象,而忽视数量级较小、混沌,但类别特征更加明显的分类条件属性对象。提出了一种基于平均互信息的聚类算法。通过熵量化参数类别特性的大小,再根据熵的平均互信息计算方法衡量数据对象间类别的相同、相异特征量,统一数值和分类条件属性参数间距离的数量级,最后通过优化迭代自适应过程得到最终聚类结果。实验结果表明,该算法具有良好的聚类质量和自适应性。

关键词 混合条件属性,平均互信息,聚类
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.3.054

Clustering with Mixed Condition Attributes Based on Average Mutual Information

LIU Jin-sheng
(College of Computer and Electronic Information, Guangdong University of Petrochemical Technology, Maoming 525000, China)

Abstract There is a great difference between the distances of mixed condition attributes parameter. The numeric condition attributes object with larger and law magnitude tends to be clustered only. With small and chaos magnitude, the categorical condition attributes object which has obvious category characteristics will be ignored. A clustering algorithm based on average mutual information was proposed. First, the size of parameter category characteristics is quantified through entropy. Then, the similarity and the difference between category characteristics are measured according average mutual information of entropy. The magnitude between distances of numeric and categorical condition attributes parameter is unified. At last, the final clustering result is got by optimizing iterative adaptive process. The experimental results show that the proposed algorithm was high clustering quality and good adaptability.

Keywords Mixed condition attributes, Average mutual information, Clustering

1 引言

混合数据集是指在同一个数据集中,既包括数值型条件属性(长度、面积、重量等),又包括分类条件属性(天气、颜色、职业等)[1]。随着网络技术的日趋成熟,数据流特征的混合性已广泛、深入地存在于实际应用中,如网络入侵检测、天气预报、垃圾邮件识别、金融等。如何有效地从混合属性数据集中提取出有价值的类别信息显得非常重要。

聚类分析是根据数据中发现的描述对象及其关系的信息,将相近似的对象聚成一类,使得类内对象相似性尽量大,类间对象相似性尽量小,属于无监督学习[2]。其算法可分为层次法[2]、基于密度[2]、模型法[2]等。目前,大多数聚类算法都只能单一地处理数值或分类型条件属性数据集,少量针对混合条件属性的方法也仅能将分类条件属性粗糙地数值量化,聚类效果较差。因此,提出基于熵的平均互信息的聚类算法。

本文第1节介绍混合数据集合的生成背景及聚类算法针对混合型数据集所存在的问题;第2节介绍相关研究;第3节提出平均互信息聚类算法;第4节给出实验及结果分析;最后总结全文。

2 相关研究

对于数值型条件属性,MacQueen[3]首次提出k均值(k-means)聚类算法,基于欧氏距离度量方式的局限,其仅能对数值条件属性的特征点进行操作。Clustream[4]是最具代表性的算法,它是由线层(初级聚类、生成微簇)和离线层(生成最终聚类结果)两层框架组成,其拓展性强,有良好的聚类效果。HPStream[5],DeStream[6]等算法大都是在此基础上的拓展研究。

对于分类型条件属性,Huang提出k-modes[7]算法,其使用基于频率的方式促使聚类效果代价函数至最小,但其仅与分类型条件属性相关。文献[8]为分类条件属性参数设置间隔时间段,生成初始化聚类微簇。该方法在文本数据流中得到了广泛应用。另外,Coolact[9]、Roct[10]都是仅能对分类条件属性进行处理的算法。

为克服k-means算法仅适合数值型条件属性的局限,

到稿日期:2014-04-22 返修日期:2014-07-15 本文受广东省教育部产学研结合项目(2011A090200088),广东省茂名市科技计划项目(2012B009),广东省石化装备故障诊断重点实验室资助。

刘晋胜 男,硕士,高级实验师,主要研究方向为模式识别、信号与信息处理,E-mail:liujinsheng11@126.com。

Ralambondrainy[11]提出将多值分类条件属性换成0和1二元属性，并将二元属性作为k-modes算法中的数值条件属性进行处理。但特征参数类别特性简单地量化为0和1，并不能很好地展示聚类特征，且大量二元组处理大大增加了算法的计算与空间代价。k-prototypes[12]用欧氏距离衡量数值条件属性参数间的相似性，用匹配算法计算分类条件属性参数的相异度，并设置两种不同类型属性间的失衡权重，相比之下，聚类性能优于文献[11]，且其运行效率是目前处理混合条件属性数据聚类算法中较为出色的，是混合条件属性聚类的经典。文献[13]针对混合条件属性大型数据集，提出使用CF树进行预聚类，将预处理结果产生的密集对象区域作为单一的叶子结点，并在区域内采用k-prototypes算法进行重聚类。文献[14]介绍了将squeezer算法应用于混合条件属性数据的伸缩性聚类方法。文献[15]在Clustream的基础上，提出对分类条件属性各参数进行自适应匹配并筛选以提取类别特性的方法。文献[16]分析数值、分类条件属性下参数之间的相关性，给出它们有效的取值范围，并采用层次树形的方式来计算混合条件属性参数间的距离。上述算法为保证聚类泛化性，参数量化过程中都存在假设对象类别相关性、最大相似度或最小距离判断阈值的情况，甚至丢弃分类条件属性的参数，当有效值范围优化性不高时，会使得聚类效果较差。D_kHD-SC[17]结合KNN算法，完整保留了数值和分类条件属性的k个近邻(根据迭代过程中数值、条件属性的个数筛选)，由k个近邻条件属性经过初始化宏聚类、优化宏聚类的过程实现最终聚类，其分类属性量化采用参数原始值排序、位序赋值的方法；条件属性各维间的相似性描述是绝对值计算。文献[18]提出了以双重近邻图搜索策略为基础的聚类算法，在规整有序和无序条件属性后，有序参数值采用欧氏距离，无序属性参数采用0和1二元设置的方法计算相异度。这类算法研究重心倾向聚类的优化过程，虽聚类质量高于k-modes，但针对混合条件属性参数的特性量化显得过于单一，会为最终聚类的优化埋下"迭代误差"隐患。文献[19]提出了基于熵论的检索最优聚类数目的算法。文献[20]采用距离度量数值条件属性参数的相似性，采用熵度量分类条件属性参数的相似性。这类算法的弊端在于数值距离值与分类熵值(值越小，越能体现参数在类别识别过程中的重要性)的数字级数不在同一范围内。例如UCI中Adult数据集，取前100条记录，按文献[18]参数的量化方法，数据集中前两个样本在第二个条件属性"workclass"(分类条件属性)下参数的类别期望熵为0.561、0.673，若采用绝对值距离衡量相似性，它们的距离为0.112。而这两个样本在第三个条件属性"fnlwgt"(数值条件属性)下的参数分别为77516和83311，采用同样距离度量，它们间的相似性值为5795。显然，距离值的大数会无限扩大数值条件属性的权重，压制甚至忽略掉分类熵值对应的分类条件属性所体现的类别特征，从而导致聚类质量的下降。

到目前为止，没有一种算法对所有混合类型的数据集都是有效的[18]。本文提出的C-AMIE(Clustering data with Average Mutual Information of Entropy)聚类方法，首先通过特征期望熵量化混合条件属性各参数体现类别特征的大小，再根据熵的平均互信息计算方法衡量数据间类别的相同、相异特性量，消除数值和分类条件属性距离的差异数量级，最后通过迭代自适应过程优化得到最终聚类结果。

3 算法介绍

3.1 平均互信息距离

设S是m个数据样本的集合，其矩阵表示为$S=\{X_i=x_{ij},i=1,\cdots m,j=1,\cdots,n\}$，其中$i$和$j$分别表示样本数据的序号和属性标号，$[1,n-1]$表示混合条件属性，第$n$维为类决策属性，若$S$分为$t$个类$C_1,\cdots,C_z,\cdots,C_t$，$C_t$是决策属性下的特征参数值。$p(x_{ij})=|x_{ij}(C_z)|/|x_{ij}|$，且$\sum_{i=1}^{m}p(x_{ij})=1$，其中$|x_{ij}(C_z)|$表示在第$j$维条件属性并属于$C_z$类背景下$x_{ij}$的实例个数，$|x_{ij}|$则表示$x_{ij}$在第$j$维条件属性下的总数，$x_{ij}$体现$t$个类的特征的量：

$$H(x_{ij})=-\sum_{z=1}^{t}\left(\frac{|x_{ij}(C_z)|}{|x_{ij}|}\cdot\ln\frac{|x_{ij}(C_z)|}{|x_{ij}|}\right)=-\sum_{z=1}^{t}p(x_{ij})\cdot\ln p(x_{ij}) \tag{1}$$

对于数值条件属性参数，在检索实例个数过程中，根据实际情况，规划数量级，同一数量级的数值表示相同的特征值。$H(x_{ij})$越小，突出x_{ij}的类别特征与t个类的类别特征关联越大；反之越小。$H(x_{ij})$转换是客观特征参数特征层面上类别不确定性的度量，不涉及数值条件属性参数的定性，也不涉及分类条件属性的定量，统一了两者数据间特征的表示方法。若X_i由n个特征参数x_{ij}组成，X_i能体现t个类的特征的量：

$$H(X_i)=-\sum_{j=1}^{n-1}H(x_{ij}) \tag{2}$$

针对t个类，X_p(隶属类C_z，$p=1,\cdots,m_{c_z}$，m_{c_z}是类C_z中样本的个数)相对X_q(隶属类C_z'，$p=1,\cdots,m_{c_z'}$，$m_{c_z'}$是类C_z'中样本的个数)的平均互信息量$I(X_p;X_q)$，即X_p和X_q的距离：

$$\begin{aligned}d(X_p,X_q)&=I(X_p;X_q)\\&=H(X_p)+H(X_q)-H(X_pX_q)\\&=\sum_{j=1}^{n-1}H(x_{pj})+\sum_{j=1}^{n-1}H(x_{qj})-\sum_{j=1}^{n-1}H(x_{pj}x_{qj})\end{aligned} \tag{3}$$

式中：

$$\begin{aligned}H(X_pX_q)&=\sum_{j=1}^{n-1}H(x_{pj}x_{qj})\\&=-\sum_{j=1}^{n-1}\left[\sum_{z=1}^{t}\frac{|x_{pj}(C_z)|}{|x_{pj}|+|x_{qj}|}\cdot\ln\left(\frac{|x_{pj}(C_z)|}{|x_{pj}|+|x_{qj}|}\right)+\sum_{z=1}^{t}\frac{|x_{qj}(C_z)|}{|x_{pj}|+|x_{qj}|}\cdot\ln\left(\frac{|x_{qj}(C_z)|}{|x_{pj}|+|x_{qj}|}\right)\right]\end{aligned} \tag{4}$$

x_{pj}、x_{qj}分别是X_p和X_q的特征参数值，$H(x_{pj}x_{qj})$是联合熵的形式，表示x_{pj}、x_{qj}一起出现时，它们共同包含t个类相同和相异特征量的大小。类别特征量的度量与熵理论互信息的关系如图1所示，阴影部分面积越大，即$I(X_p;X_q)$越大，说明X_p和X_q包含t个类的相同类别特征的量越大，X_p和X_q归属为同一个类的可能性越大，X_p和X_q的距离越接近。同时，结合图1，$H(X_pX_q)$具备如下性质：

$$H(X_p/X_q)=H(X_pX_q)-H(X_q) \tag{5}$$

$$H(X_q/X_p)=H(X_pX_q)-H(X_p) \tag{6}$$

由式(3)、式(5)和式(6)可得：

$$d(X_p,X_q)=I(X_p;X_q)=H(X_p)-H(X_p/X_q) \tag{7}$$

$$d(X_p,X_q)=I(X_p;X_q)=H(X_q)-H(X_q/X_p) \tag{8}$$

其中，$H(X_p/X_q)$和$H(X_q/X_p)$的定义如下：

$$\begin{aligned}H(X_p/X_q)&=\sum_{j=1}^{n-1}H(x_{pj}/x_{qj})\\&=-\sum_{j=1}^{n-1}\sum_{z=1}^{t}\left(\frac{|x_{pj}(C_z)|+|x_{qj}(C_z)|}{|x_{pj}|+|x_{qj}|}\cdot\right.\end{aligned}$$

$$\frac{|x_{pj}(C_z)|}{|x_{pj}|})\ln(\frac{|x_{pj}(C_z)|+|x_{qj}(C_z)|}{|x_{pj}|+|x_{qj}|}\cdot\frac{|x_{pj}(C_z)|}{|x_{pj}|}) \tag{9}$$

$$H(X_q/X_p)=\sum_{j=1}^{n-1}H(x_{qj}/x_{pj})=-\sum_{j=1}^{n-1}\sum_{z=1}^{t}(\frac{|x_{pj}(C_z)|+|x_{qj}(C_z)|}{|x_{pj}|+|x_{qj}|}\cdot\frac{|x_{qj}(C_z)|}{|x_{qj}|})\ln(\frac{|x_{pj}(C_z)|+|x_{qj}(C_z)|}{|x_{pj}|+|x_{qj}|}\cdot\frac{|x_{qj}(C_z)|}{|x_{qj}|}) \tag{10}$$

式(9)、式(10)中，$H(x_{pj}/x_{qj})$是条件熵的形式，以 x_{pj}、x_{qj} 一起出现时包含 t 个类的特征总量为参照，在已知 x_{qj} 所能体现的类别特征量的前提下，$H(x_{pj}/x_{qj})$表示基于 x_{qj} 的 x_{pj} 相异特征量的大小，$H(x_{qj}/x_{pj})$ 表示基于 x_{pj} 的 x_{qj} 相异特征量的大小。根据上述分析，在后面的实验中验证 $I(X_p;X_q)$、$H(X_pX_q)$与 $H(X_p/X_q)$、$H(X_q/X_p)$之间的关系，进一步说明熵平均互信息理论映射在聚类距离计算中的合理性。

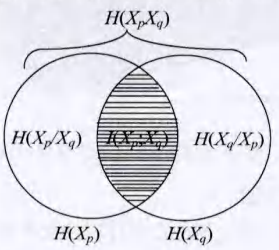

图1 平均互信息距离几何意义

因此，任两个类间的互信息距离如下：

$$D(C_z,C_z')=\sum_{p=1}^{m_{C_z}}[\sum_{q=1}^{m_{C_z'}}d(X_p,X_q)] \tag{11}$$

由式(3)和式(4)可知，平均互信息距离是特征参数特性点相同、相异量的完整规划。面对混合条件属性数据，不用考虑数值或类别条件属性计算时权重的设置方案，不需要对分类条件属性特征进行假设性的数值化定性转换，不需要在训练模型层中描述类别间的相似距离，即使在数据分布不平衡的状态下，都能在很大程度上消除类似两个大类倾向合并，两个小类倾向不合并的累计迭代聚类误差现象。

3.2 聚类过程

Step1 将 m 个数据样本随机分成 t 个类，设定初始化任一类 C_z 只有一个样本 X_p，并根据实际情况划分数值条件属性参数的量级范围。

Step2 由式(3)、式(4)和式(10)计算 t 个类中任意两类 C_z 和 C_z'间的平均互信息距离，共 $(1/2)\times t\times(t-1)$ 个距离。取其中最小 $D_{\min}(C_z,C_z')$，与聚分限制值 $limit_r$ 比较，其中，

$$limit_r=\mathrm{MIN}(D(C_z,\overline{C_zC_z'}),D(C_z',\overline{C_zC_z'})) \tag{12}$$

由式(12)可知，$limit_r$ 用于衡量待合并的 C_z 和 C_z'在 t 个类中预计损失的相似类别特征量的大小。随着迭代次数的增加，$limit_r$ 会随着各类中样本数目、参数特征特性的更新而自适应地变化。式中，$\overline{C_zC_z'}$表示 t 个类中除去 C_z 和 C_z'的所有类，即 $m_{\overline{C_zC_z'}}=t-m_{C_z}-m_{C_z'}$，$C_z\notin\overline{C_zC_z'}$，$C_z'\notin\overline{C_zC_z'}$，$r$ 是迭代的次数。若 $D_{\min}(C_z,C_z')<limit_r$ 则进入 Step3；否则进入 Step4。

Step3 $D_{\min}(C_z,C_z')$是平均互信息距离最小的两个类，将其合并为同一类 $C_{zz'}$，$m_{C_{zz'}}=m_{C_z}+m_{C_z'}$，$t=t-1$ 类别数减1。由式(1)更新 $C_{zz'}$类中特征参数的 $H(x_{ij})$，$i=1,\cdots,m_{C_{zz'}}$，随着迭代次数的增加，$H(x_{ij})$会由于类别的总数、数据聚分情况的改变而变化。返回 Step2 继续。

Step4 $C_1,\cdots,C_z,\cdots,C_t$ 是最终聚类结果，t 是最终类别总数。

3.3 算法收敛性分析

基于平均互信息距离聚类算法的关键在于距离计算方法的收敛性，且 $0\leqslant p(x_{pj})\leqslant 1$，$0\leqslant p(x_{qj})\leqslant 1$，$0\leqslant p(x_{pj})\cdot p(x_{qj})\leqslant 1$，故对式(3)进行[0,1]区间内的收敛性分析。

对 $H(X_p)$在[0,1]区间求定积分：

$$\begin{aligned}&\int_0^1 H(x_{pj})\\&=-\sum_{z=1}^{t}\int_0^1\frac{|x_{pj}(C_z)|}{|x_{pj}|}\ln\frac{|x_{pj}(C_z)|}{|x_{pj}|}\mathrm{d}(\frac{|x_{pj}(C_z)|}{|x_{pj}|})\\&=-\sum_{z=1}^{t}\int_0^1\ln(\frac{|x_{pj}(C_z)|}{|x_{pj}|})\mathrm{d}(\frac{|x_{pj}(C_z)|}{|x_{pj}|})^2\\&=\frac{1}{2}(\frac{|x_{pj}(C_z)|}{|x_{pj}|})^2\ln(\frac{|x_{pj}(C_z)|}{|x_{pj}|})-\frac{1}{2}\int_0^1\frac{|x_{pj}(C_z)|}{|x_{pj}|}\\&\quad\mathrm{d}(\frac{|x_{pj}(C_z)|}{|x_{pj}|})\\&=\frac{1}{2}(\frac{|x_{pj}(C_z)|}{|x_{pj}|})^2\ln(\frac{|x_{pj}(C_z)|}{|x_{pj}|})-\frac{1}{4}(\frac{|x_{pj}(C_z)|}{|x_{pj}|})^2\end{aligned}$$

$\because\lim_{p(x_{pj})\to 1}p(x_{pj})\ln p(x_{pj})=0$ 存在，$\therefore\int_0^1 H(x_{pj})=\frac{1}{4}$，以此类推，得 $\int_0^1 H(x_{qj})=\frac{1}{4}$，$\int_0^1 H(x_{pj},x_{qj})=\frac{1}{4}$，因此 $\int_0^1\mathrm{d}(X_p,X_q)=\frac{1}{4}$，收敛。

3.4 时间及空间复杂度分析

由3.2节可知，本文算法在迭代过程中主要是进行数据体现类别特征量的熵转换预处理、类间平均互信息距离两个阶段。设数据集具有 n 个混合条件属性，每个条件属性下具有 m 个特征参数，并任意划分为 t 个类。在预处理阶段，若每个特征参数可取的不同值的平均数是 r，$r<n\times m$，则检索每个特征参数值在任一个条件属性下隶属 t 个类的实例个数的计算时间复杂度为 $O(t\log_2(rn))$。接着，统计每个特征参数在任一个条件属性下出现的总次数，时间复杂度为 $O(rn)$。则计算任意 $H(x_{pj})$ 和 $H(x_{pj}\,x_{qj})$ 的时间开销均为 $O(t\log_2(rn))+O(rn)$。

在类间平均互信息计算阶段，任意 C_z 和 C_z'两个类中任一条件属性下的特征参数个数平均为 m_{C_z} 和 $m_{C_z'}$，则 t 个类间的平均互信息距离计算的时间复杂度为 $\frac{t(t-1)}{2}O(m_{C_z}\cdot m_{C_z'}\cdot n)$。另外，若存在 w 次迭代，聚分限制值计算的时间复杂度为 $wn\cdot(O(m_{C_z}\cdot m_{\overline{C_zC_z}})+O(m_{C_z'}\cdot m_{\overline{C_zC_z}}))$，在迭代过程中取最小值的时间复杂度为 $O(W)$。

本文算法的空间开销主要是由熵值转换引起的，其数据结构为 $\{R,L,COUNT\}$。其中 $R=\{H(x_{ij}),i=1,\cdots,m,j=1,\cdots,n\}$是二维向量，使用空间为 $O(nm)$；L 是一个链表集，$L=\{|x_{ij}(C_z)|,i=1,\cdots,m,j=1,\cdots,n,z=1,\cdots,t\}$，其使用空间为 $O(nmt)$；根据数据集的实际最大特征参数数量，$COUNT=\{|x_{ij}|,i=1,\cdots,m,j=1,\cdots,n\}$，使用空间最大为 $O(nm)$。因此，本文算法的空间复杂为 $O(2nm+nmt)$。

4 实验及结果分析

所有实验均在一台硬件配置为英特尔酷睿 2，四核 2.66G，320G 硬盘，4G 内存上进行，操作系统为 Win7，采用 MATLAB结合 VC6.0 实现。

实验数据集为 UCI 的 Adult 和 Credit Approval。Adult 含 48842 个样本，14 个条件属性，分 2 类，数值条件属性集合 nca-set：{AGE (10)，FNLWGT (100000)，EDUCATION-NUM，CAPITAL-GAIN，CAPITAL-LOSS，HOUR-PER-WEEK}，共 6 个，由式(1)计算特征参数体现类别特征量的过程中，AGE 以 10 为一个量级，FNLWGT 以 100000 为一个量级，其它的以参数值直接进行匹配统计；分类条件属性集合 cca-set：{WORKCLASS，EDUCATION，MARITAL-STATUS，OCCUPATION，RELATIONSHIP，RACE，SEX，NATIVE-COUNTRY}，共 8 个。Chess 数据集包含 28056 个样本，6 个条件属性，分 16 类，其中 3 个数值型、3 个分类型，数值型条件属性参数值均直接进行匹配统计。

4.1 可行性

可行性实验主要是验证熵平均互信息理论映射在聚类距离计算中的合理性。从 Adult 中随机抽取 N 次 1000 个样本，根据 3.2 节 C-AMIE 聚类过程步骤，分别用式(3)和式(7)两种方式计算每次迭代产生的所有平均互信息距离。取前 15 次迭代计算结果，且 $N=10$，采用平均互信息的平均值作为聚类可行性的评价指标。实验结果如图 2 所示。从图中可以看出，两种平均互信息计算方式生成的距离值基本一致。第 14 次迭代是两种结果的最大差距，误差不超过 3%，而两者在第 9、11 次迭代产生的距离值误差仅相差 0.5 个百分点。这说明熵平均互信息论在聚类距离计算中的应用是合理的、可行的。

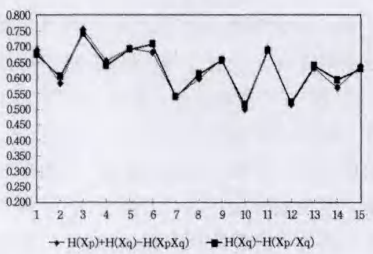

图 2 平均互信息距离 $I(X_p;X_q)$计算方式对比

4.2 自适应性

采用聚类效果作为实验结果的评价指标，定义为：设聚类算法最终将数据集 S 划分为 k 个类，第 i 个类 C_i 包含有 $|C_i|$ 个样本，$|C_i|\_tab$ 表示实际决策属性为 C_i 的样本个数，$|S|$ 表示数据集样本个数，则聚类效果为式(13)，且 $CE(k)$ 越小，聚类效果越好。

$$CE(k)=\frac{1}{\ln|S|}\sum_{i=1}^{k}\frac{|C_i|\_tab}{|C_i|}\ln\frac{|C_i|\_tab}{|C_i|} \tag{13}$$

自适应实验的主要目的是测试数值条件属性和分类条件属性个数的变化对聚类性能的影响。采用式(3)计算类间平均互信息距离。实验结果如图 3 所示。数值条件属性按 nca-set 集合中元素顺序依次增加。分类条件属性按 cca-set 集合元素顺序依次增加。由图 3 显示的实验结果可知，当数据含数值条件属性和分类条件属性，两者仅其一的比例占主导时，聚类效果较差。特别是数值条件属性占最大比例数时，$CE(k)$值超过了 0.3，分类效果要差于分类条件属性最大比例数的情况。随着数值条件属性个数和分类条件属性个数的逐渐增加，数据信息愈来愈完善，条件属性个数的增长与 $CE(k)$值基本呈线性关系。但在分类条件属性个数为 5 时，即增加分类条件属性 RACE 后，曲面有明显的陡度下滑，随后逐步趋于平稳，这说明，一方面数值和分类条件属性个数接近时，开始获得较高质量的聚类效果；另一方面，条件属性的特征权重、数据聚分限制值在聚类计算过程中均可自适应形成，且 RACE 分类条件属性类别相关性相对较高。

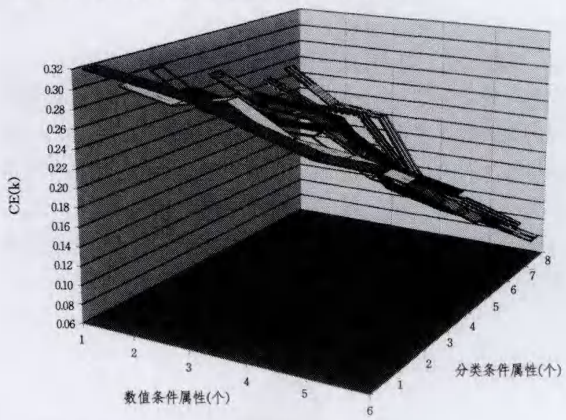

图 3 C-AMIE 算法对混合条件属性的自适应性

4.3 聚类效果比较

k-prototypes 是处理混合条件属性聚类成功的算法，因此将其与 C-AMIE 算法进行对比。采用式(3)计算类间平均互信息距离，以 $CE(k)$值作为聚类评价指标。实验结果如表 1 所列。由比较结果可知，C-AMIE 的聚类泛化性能更高。主要原因在于 C-AMIE 在保证数据对象类别信息完整性的前提下，互信息的计算方式统一了混合条件属性下各参数相似值的数量级，且其聚类自适应方法也是迭代优化剩余类与已聚分类互信息距离的过程。而 k-prototypes 分类条件属性的二元(0,1)参数匹配方法，损失的数据类别信息量较大，且与数值条件属性参数的相似性计算方式也不统一。这使得聚类效果没有 C-AMIE 好。另外，在时间性能上，k-prototypes 距离计算的复杂性和需要随机选中 k 中心簇的特点均对聚类速度产生了较大的影响。而 C-AMIE 在参数量化的预处理阶段，花费的时间复杂度较大。因此，两种算法的聚类速度差别不大。

表 1 算法聚类效果对比

比较		算法名称	
		k-prototypes	C-AMIE
Adult	CE(k)	0.187	0.096
	时间(s)	98	96
Chess	CE(k)	0.194	0.085
	时间(s)	112	108

结束语 本文提出了基于平均互信息的聚类算法 C-AMIE。该算法将熵平均互信息映射为数据对象间相似性计算的方法。C-AMIE 统一了数值与分类条件属性相似性度量值的数量级单位，使其在保证聚类效率的同时，具有较高的聚类质量。

参 考 文 献

[1] Tang Pang-ning, Michael Steinbach, Vipin Kumar. Introduction

to data mining [M]. Beijing: Post&Telecom Press, 2006
[2] 史忠植. 高级人工智能[M]. 北京:科学出版社, 2006
[3] Jain A K. Data clustering: 50 years beyond k-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666
[4] Aggarwal C C, Han J, Wang J, et al. A framework for clustering evolving data streams[C]// Proc of VLDB. 2003: 81-92
[5] Aggarwal C C, Han J, Wang J, et al. A framework for projected clustering of high dimensional data streams [C] // Proc. of VLDB. 2004: 852-863
[6] Cao F, Estery M, Qian W, et al. Density-based clustering overran evolving data stream with noise[C] // Proc of the SIAM Conference on Data Mining (SDM). 2006: 326-337
[7] Huang Z. Extension to K-means algorithm for clustering large datasets with categorical values[J]. Data Mining and Knowledge Discovery II, 1998(2): 283-304
[8] Aggarwal C C, Yu P S. A framework for clustering massive text and categorical data streams[C]// Proc of 6th Siam IntConf on Data Mining. Bethesda, 2006: 477-481
[9] Guha S, Rastogi R, Shim K. ROCK: a robust clustering algorithm for categorical attributes[C]// Proc of ICDE. 1999: 512-521
[10] Barbara D, Couto J, Yi L. COOLCAT: an entropy-based algorithm for categorical clustering[C]// Proc of CIKM. 2002: 582-589
[11] Ralambondrainy H. A conceptual version of the k-means algorithm[J]. Pattern Recognition Letters, 1995: 1147-1157
[12] Huang Z. Clustering large data sets with mixed numeric and categorical values[C]// Pro of 1th Pacific-Asic Conf. 1997: 21-34
[13] Yin Jian, Tan Zhi-fang, Ren Jiang-tao, et al. An efficient clustering algorithm for mixed type attributes in large dataset[J]. IEEE transactions on Machine Learning and Cybernetics, 2005, 8(3): 1611-1614
[14] He Zeng-you, Xu Xiao-fei, Deng Sheng-chun. Scalable algorithms for clustering large datasets with mixed type attributes [J]. International Journal of Intelligent Systems, 2005, 20(10): 1077-1089
[15] 杨春宇,周杰. 一种混合属性数据流聚类算法[J]. 计算机学报, 2007, 30(8): 1364-1372
[16] Hsu C C, Huang Y. Incremental clustering of mixed data based on distance hierarchy[J]. Expert Systems with Applications, 2008, 35(3): 1177-1185
[17] 黄德才,沈仙桥,陆亿红. 混合属性数据流的二重 k 近邻聚类算法[J]. 计算机科学, 2013, 40(10): 226-230
[18] 陈新泉. 面向混合属性数据集的双重聚类方法[J]. 计算机工程与科学, 2013, 35(2): 127-132
[19] Liang Ji-ye, Zhao Xing-wang, Li De-yu, et al. Detemining the number of clusters using information entropy for minxed data [J]. Pattern Recognition, 2012, 45(6): 2251-2265
[20] 王述云,胡运发,范颖捷,等. 基于距离和熵的混合属性数据流聚类算法[J]. 小型微型计算机系统, 2012, 12(12): 2365-2371

(上接第 255 页)

他算法的 MAE 值波动较大。说明本文算法在冷启动数据集下,可以较好地提高推荐的准确率。

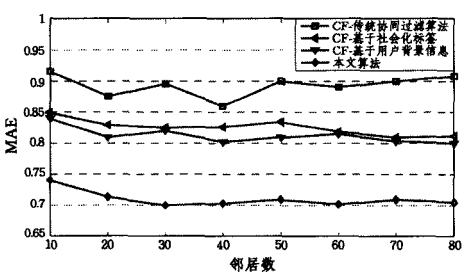

图 3 在冷启动数据集下算法的 MAE 比较

结束语 推荐系统作为解决信息过载的一种有效手段,如何提高其推荐精度已成为非常重要的研究问题。人们对社会标签标注的时间变化体现了其社会行为的动态性。考虑了社会标签的时间属性并集成了用户背景,本文提出的方法相对于传统的协同过滤(CF)等推荐算法能够有效地提高算法的推荐精度和改善数据稀疏及冷启动的问题。目前,社交网络平台与电子商务平台多数是独立运行的,难以获得同时具有电子商务和社交网络的标准数据集,本文验证则另辟途径,先计算两个数据集用户评分间的相似性得出一个用户相似类,再将两个数据集整合成本文所需要的数据集来进行验证,但存在一定的不足。未来工作是集成电商和社交网络平台数据,对本文的工作给予进一步的验证,并考虑其它社交行为对推荐质量的影响。

参 考 文 献

[1] 于洪,李转运. 基于遗忘曲线的协同过滤推荐算法[J]. 南京大学学报:自然科学版, 2010, 46(5): 520-527
[2] 冯勇,李军平,徐红艳,等. 基于社会网络分析的协同过滤推荐方法改进[J]. 计算机应用, 2013, 33(3): 841-844
[3] Eleftherios T, Yannis M. Product recommendation and rating prediction based on multi-modal social networks[C] // Poceedings of the 5th ACM Conference on Recommender Systems. New York: ACM Press, 2011: 61-68
[4] 贾大文,曾承,彭智勇,等. 一种基于用户偏好自动分类的社会媒体共享和推荐方法[J]. 计算机学报, 2012, 35(11): 2381-2391
[5] 顾亦然,陈敏. 一种三部图网络中标签时间加权的推荐方法[J]. 计算机科学, 2012, 39(8): 96-98, 129
[6] Cheng Yuan, Qiu Guang, Bu Jia-jun, et al. Model bloggers' interests based on forgetting mechanism[C]// Proceedings of the 17th International Conference on World Wide Web. New York: ACM Press, 2008: 1129- 1130
[7] 庄景明,王明文,叶茂盛. 基于内容过滤的农业信息推荐系统[J]. 计算机工程, 2012, 38(11): 38-41
[8] Zhou Tao, Jiang L L, Su R Q, et al. Effect of initial configuration on network-based recommendation[J]. Europhys Lett, 2008, 81: 58004
[9] 赵琴琴,鲁凯,王斌. SPCF:一种基于内存的传播史协同过滤推荐算法[J]. 计算机学报, 2013, 36(3): 671-676