

## 基于集成学习的无监督离散化算法

徐盈盈\*, 钟才明

(1. 宁波大学 信息科学与工程学院, 浙江 宁波 315210; 2. 宁波大学 科学技术学院, 浙江 宁波 315210)

(\* 通信作者电子邮箱 xuyingying1227@sina.com)

**摘要:**模式识别与机器学习的一些算法只能处理离散属性值,而在现实生活中的很多数据具有连续的属性值,针对数据离散化的问题提出了一种无监督的方法。首先,使用 *K*-means 方法将数据集进行划分得到类别信息;然后,应用有监督的离散化方法对划分后的数据离散化,重复上述过程以得到多个离散化的结果,再将这些结果进行集成;最后,将集成得到的最小子区间进行合并,这里根据数据间的邻居关系选择优先合并的维度及相邻区间。其中,通过数据间的近邻关系自动寻求子区间数目,尽可能保持其内在结构关系不变。将离散后的数据应用于聚类算法,如谱聚类算法,并对聚类后的效果进行评价。实验结果表明,该算法聚类精确度比其他 4 种方法平均提高约 33%,表明了该算法的可行性和有效性。通过该算法得到的离散化数据可应用于一些数据挖掘算法,如 ID3 决策树算法。

**关键词:**无监督离散化;集成学习;分类数据;相似性;谱聚类

**中图分类号:** TP391; TP18 **文献标志码:** A

### Unsupervised discretization algorithm based on ensemble learning

XU Yingying\*, ZHONG Caiming

(1. College of Information Science and Engineering, Ningbo University, Ningbo Zhejiang 315210, China;

2. College of Science and Technology, Ningbo University, Ningbo Zhejiang 315210, China)

**Abstract:** Some algorithms in pattern recognition and machine learning can only deal with discrete attribute values, while in real world many data sets consist of continuous data values. An unsupervised method was proposed according to the question of discretization. First, *K*-means method was employed to partition the data set into multiple subgroups to acquire label information, and then a supervised discretization algorithm was applied to the divided data set. When the process was repeatedly executed, multiple discrete results were obtained. These results were then integrated with an ensemble technique. Finally, the minimum sub-intervals were merged after priority dimensions and adjacent intervals were determined according to the neighbor relationship of data, where the number of sub-intervals was automatically estimated by preserving the correlation so that the intrinsic structure of the data set was maintained. The experimental results of applying categorical clustering algorithms such as spectral clustering demonstrate the feasibility and effectiveness of the proposed method. For example, its clustering accuracy improves by about 33% on average than other four methods. Discrete data attained can be used for some data mining algorithm, such as ID3 decision tree algorithm.

**Key words:** unsupervised discretization; ensemble learning; categorical data; similarity; spectral clustering

## 0 引言

现实世界的的数据很多都是呈连续属性,但是大部分机器学习算法(如决策树算法),只适合处理离散属性值。为了能够很好地应用这些算法,需要对连续属性值的数据进行离散化处理,使其转变为离散属性。因此,离散化是一种重要的预处理技术,尤其在频繁项集发现应用中很重要<sup>[1]</sup>,离散化是指将数值的值域划分为若干区间,将这些区间标号变为离散属性。离散化算法要求能识别连续属性与离散属性的对应关系。对训练样本进行离散化处理,有如下优点:1)离散化可以降低机器学习的复杂度,对 ID3 学习算法的训练样本进行离散化处理后,其学习率比动态算法提高了,这点在将连续属性转变为离散属性<sup>[2]</sup>中得到验证。2)可将连续数值划分为可被人类理解的结果,如可将学生成绩分为及格、中等、优秀

等。

根据离散化过程中区间的划分可分为自底向上(bottom-up)和自顶向下(top-down)方法:前者是将初始的每一个属性值作为一个区间,然后迭代地合并相邻区间,利用停止准则结束离散化,最终将连续属性离散成数目少、有实际意义有代表性的若干个区间;而后者是个相反的过程,将最小与最大的属性值作为一个区间,然后逐步进行划分得到最终的离散结果。有监督算法和无监督离散化算法,可根据数据是否具有类别信息来区分。静态和动态的离散化:前者仅仅考虑单个属性,执行过程独立于学习算法;后者考虑属性之间的联系,并在分类器被建立的同时执行,如 ID3、C4.5 决策树。单属性和多属性离散化:前者是在离散每一个连续属性时,均以一个个独立于其他属性的方式对属性值进行合并或分割;后者还会考虑到属性与属性之间的相关性,如基于主成分分析的无监督关系

收稿日期:2014-04-08;修回日期:2014-05-08。 基金项目:国家自然科学基金资助项目(61175054)。

作者简介:徐盈盈(1990-),女,安徽桐城人,硕士研究生,主要研究方向:机器学习、模式识别;钟才明(1970-),男,浙江宁波人,副教授,博士,主要研究方向:模式识别、机器学习。

保持离散化方法<sup>[3]</sup>。

## 1 相关研究

典型的有监督的算法如 ChiMerge<sup>[4]</sup>,通过测试相邻区间的卡方值来确定是否合并,但要设定参数阈值会产生很多缺陷;改进的 Chi2<sup>[5]</sup>通过数据间的不一致率来作为区间合并的停止准则,但会降低原始数据的可信度,产生分类错误。Yang等<sup>[6]</sup>首先提出了成比例的 $k$ 区间离散化方法,通过按比例地修改区间大小和区间数来调整离散化偏差和方差,以适应 Naive 贝叶斯分类器,该算法是小样本的优化问题,不适合于大数据的处理。基于区间距离的离散化算法<sup>[7]</sup>需要用户定义区间数目。

针对无标签的数据进行离散化,即无监督的算法,在 Dougherty 等<sup>[8]</sup>提出的算法中最简单的为等宽与等频率的算法,虽然都易于实现,但都忽视了数据分布信息,因而区间边界的确定不具有代表性; $K$ -means 离散化方法,对于数值型的离散化而言,采用欧几里得距离作为区间划分的依据缺乏理论根据。此外,该算法依靠用户来指定区间数目,不能自动确定区间数;保持关系的离散化方法,考虑属性间的相关性通过主成分分析(Principal Component Analysis, PCA)降维的方法来离散,对于高维非线性可分的数据离散效果不佳;基于混合概率模型的无监督离散化方法<sup>[9]</sup>,将数值属性的值域划分为若干子区间,再通过贝叶斯信息准则自动的寻求子区间数目和划分方法,在离散化过程中针对不同的属性离散化时间可能相差较大。

目前应用最广泛的有:监督算法是类属性关系最大化(Class-Attribute Interdependence Maximization, CAIM)算法<sup>[10]</sup>,综合考虑类与属性之间的相关性,通过最大化相互依赖性来选择合适的切断点,能很好地保持数据的内在结构,可能会导致划分的区间数目与类数之间过拟合;以及后来提出的基于类属性应变系数(Class-Attribute Contingency Coefficient, CACC)的离散化算法<sup>[11]</sup>,即类属性相关系数的离散化算法。

无监督离散化方法分为基于树的无监督离散化方法<sup>[12]</sup>和基于核函数的无监督离散化方法<sup>[13]</sup>。前者是建立树的模型,切断点的位置选择在左边和右边的对数似然最大化;后者是计算区间中点的得分函数来选择切断点的位置。最后都通过交叉验证的方法自动寻求划分停止的位置,这种自上而下的方法可能导致区间数目过多,丢失的信息也比较多。

## 2 基于集成的无监督离散化算法

无监督离散化算法针对无类标签的数据,即不知道数据内在分布信息,但是离散化后希望尽可能地保持数据结构关系,可选用这样的有监督算法——CAIM,首先可用  $K$ -means 算法得到类别标签,即可得到离散区间,但是仅仅一次的划分不具有代表性,因此可将将这些结果通过集成学习的方法即可得到最终的区间。算法流程如图1所示。

### 2.1 数据集的划分

给定无类别标签的数据集  $X = \{x_1, x_2, \dots, x_N\}$ ,使用  $K$ -means 重复划分,由于  $K$ -means 比较简单易于实现,而且该算法将相似性大的数据划分到一类,因此能很好地保持数据的内在结构。 $K = \sqrt{N}$ ,初始中心点随机选取。确定  $K$  为  $\sqrt{N}$  有

如下的两个原因:

1) 当  $K$  较大时,聚类得到的类则较小,由于  $K$ -means 总将相对近邻的数据划分成一个类,这样可使类中的数据基本保持同质性,这符合统计机器学习中的一个假设,即近邻数据通常具有相同的类别条件概率。

2) 符合聚类分析中关于类个数的通常假设,即聚类的个数一般不大于  $\sqrt{N}$ 。

当然,  $K$  较大时的负面影响是  $K$ -means 的效率变低,因为其时间复杂度为  $O(I d N K)$ ,其中:  $I$  为循环次数,  $d$  为数据的维数,即达  $O(N^{1.5})$ 。

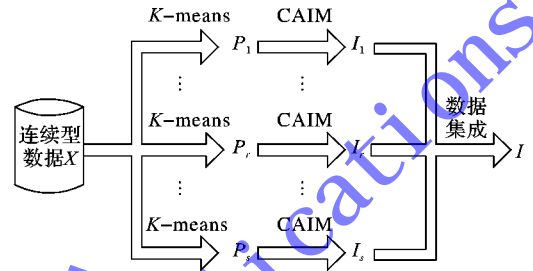


图1 算法流程

### 2.2 离散化结果的集成

#### 2.2.1 划分最小子区间

这里所谓的最小子区间是指在离散化时不能继续被切断的区间,这也是集成学习的思想在本文的体现。正如在上文中所述,  $K$ -means 取较大的聚类数目是为了尽量使同一类中的数据是同质的。那么,将  $K$ -means 划分所得到的类别标签作为 CAIM 的输入,可以假设 CAIM 的输出即离散化的区间具有一定的合理性。根据集成学习的思想,多个弱学习结果的集成将较大地提高学习精确性与鲁棒性,那么将  $K$ -means 的多次划分输入 CAIM 得到多个离散结果,然后将这些结果集成。

既然集成是为了综合考虑、合并多个弱学习的结果,那么本文集成多个离散化的过程简单地设为合并一个维度的多个区间,且称此时得到的区间为最小子区间,其过程如下。

设  $K$ -means 重复划分数据集的次数为  $s$ ,即  $P_1, P_2, \dots, P_s$ ,若对其的某一维利用 CAIM 算法离散化为区间  $I_1 = \langle c_{11}, c_{21}, \dots, c_{k11} \rangle, \dots, I_s = \langle c_{1s}, c_{2s}, \dots, c_{ks} \rangle$ ,其中  $c_{ij}$  是一个切断点。那么最小子区间的形成即为:

$$I = I_1 \cup I_2 \cup \dots \cup I_s \quad (1)$$

例如,某一维度数据取值范围为  $[1, 10]$ ,两次离散化的区间分别为  $I_1 = \langle 1, 3, 4, 6, 10 \rangle, I_2 = \langle 1, 2, 5, 8, 10 \rangle$ ,那么合并后的区间为  $I = \langle 1, 2, 3, 4, 5, 6, 8, 10 \rangle$ ,其中  $[1, 2], (2, 3], (3, 4], (4, 5], (5, 6], (6, 8], (8, 10]$  成为最小子区间。

#### 2.2.2 最小子区间的合并

获得最小子区间后,需将这些区间进行合并以得到最终的离散化区间。合并过程分为两个步骤:1) 确定要合并的维度;2) 确定该维度上要合并的区间。传统的算法在合并或者划分某一维度时,并不考虑其他维度的影响,这丢失了维度之间的相关性信息。而本文每合并一个区间,都将判断这对相邻区间来自哪一个维度。

1) 确定某一维度上要合并的区间。显然,要合并的两个区间一定是相邻的,且这一对区间的相似性最大。定义一对相邻区间的相似性如下。

$P_1, P_2, \dots, P_s$  为  $K$ -means 在  $X$  上的  $s$  个划分,  $A$  是一个  $N \times N$  的关联矩阵, 其元素如式(2)所示:

$$a_{ij} = \frac{1}{s} \sum_{m=1}^s \sum_{l=1}^K f(i, j, p_{ml}) \quad (2)$$

其中:

$$f(i, j, p_{ml}) = \begin{cases} 1, & \mathbf{x}_i \in p_{ml} \text{ 且 } \mathbf{x}_j \in p_{ml} \\ 0, & \text{其他} \end{cases} \quad (3)$$

且  $p_{ml} \in P_m$ 。

考虑某一维上的两个相邻区间  $[c_{i-1}, c_i]$  与  $[c_i, c_{i+1}]$  之间的相似性。设  $C_1$  与  $C_2$  分别是  $X$  落入上述两个区间的数据点的集合, 那么其相似性定义为式(4):

$$R(C_1, C_2) = \frac{1}{|C_1| + |C_2|} \sum_{i=1}^{|C_1| + |C_2|} A(\mathbf{x}_i, \mathbf{x}_j) \quad (4)$$

其中  $A(\mathbf{x}_i, \mathbf{x}_j)$  是  $\mathbf{x}_i, \mathbf{x}_j$  在关联矩阵  $A$  中对应的元素,  $R$  值越大表明相邻的两区间的相似性大, 应优先将其合并, 因此当  $R(C_1, C_2)$  最大时合并。

2) 选择优先合并哪一个维度上的相邻区间。实际上, 对于每一次合并, 由于  $X$  有  $d$  维, 那么有  $d$  个选择。本文选择的准则是  $d$  个选项中能最大保持近邻特征的一对。所谓保持近邻特征, 是指合并一对相邻区间后, 在离散空间数据点的  $k$  近邻与原始连续空间的  $k$  近邻的差异度最小。那么如何来定义这一差异度呢?

原始连续数据的邻居可通过欧氏距离计算得到,  $dist(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}, \mathbf{y})^T (\mathbf{x}, \mathbf{y})}$ 。而在离散空间, 两个点之间的相似性选择 OF (Occurrence Frequency) 测度<sup>[14]</sup>:

$$OF(x_i, y_i) = \begin{cases} 1, & x_i = y_i \\ \frac{1}{1 + \lg(N/f_i(x_i)) \times \lg(N/f_i(y_i))}, & \text{其他} \end{cases} \quad (5)$$

其中:  $x_i$  与  $y_i$  分别是数据点  $\mathbf{x}$  与  $\mathbf{y}$  第  $i$  维的离散值;  $f_i(x_i)$ ,  $f_i(y_i)$  分别是  $X$  落入区间  $x_i$  与  $y_i$  的元素个数, 那么其距离为式(6):

$$dist(\mathbf{x}, \mathbf{y}) = \frac{1}{d} \sum_{i=1}^d OF(x_i, y_i) \quad (6)$$

设邻居数为  $k$ , 点  $\mathbf{x}_i$  在原始空间将数据集划分为两个子集: 邻居集合与非邻居集合, 称此划分为  $S_i$ , 同样在离散空间对应的划分为  $T_i$ , 则根据 ARI (Adjusted Rand Index)<sup>[15]</sup> 指标可估计此划分的相似性为  $ARI(S_i, T_i)$ 。那么, 合并一对相邻区间后, 其近邻保持度可定义为:

$$PerS = \frac{1}{N} \sum_{i=1}^N ARI(S_i, T_i) \quad (7)$$

$PerS$  值越大, 表明离散后数据间的邻居关系越能保持, 或差异度越小。

3) 合并停止准则。当区间数目较多的时候, 数据的邻居数比较少, 合并后邻居增加了, 但非邻居的数目也增加了, 因此在合并的过程中一定有一个最佳区间数, 使原始空间与离散空间的邻居相似性最大, 这就可以作为离散化算法的停止条件。图2是 Iris 数据离散化过程中, 反复合并相邻区间时, 原始空间与离散空间邻居的  $PerS$  值变化, 合并一个维度时可能邻居增加的数目没有非邻居多, 甚至邻居没有变化但是非邻居增加了, 因此在合并的过程是一个上下波动的过程, 当其最大时离散化停止。

用基于树或者核函数算法离散化, 每切断一次都通过交叉验证的方法使对数似然达到最大值时, 停止离散化过程, 但是这样自上而下的划分可能会导致区间数目比较多, 图3与图4分别是用这两种方法对 Iris 数据一个维度的对数似然值变化。

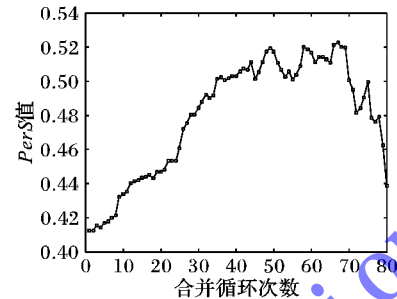


图2 合并准则对 Iris 数据的离散化过程中  $PerS$  值的变化

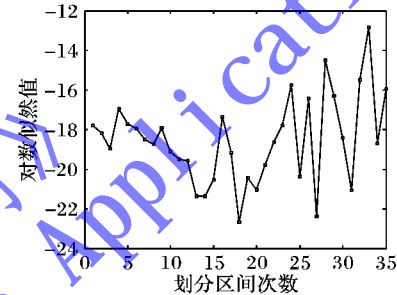


图3 基于树对 Iris 离散化过程中对数似然值的变化

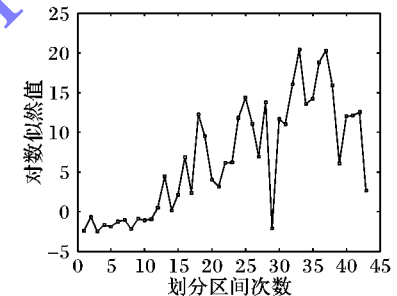


图4 基于核对 Iris 离散化过程中对数似然值的变化

在实验中将进一步比较本文的合并停止准则与文献[12-13]中的停止准则。

根据上述讨论, 描述本文算法如下:

输入 连续属性值的数据集  $X$ 。

输出 每维属性值的划分区间。

- 1) 对原始数据  $K$ -means 聚类将数据划分好标签  $C_{11}, C_{12}, \dots, C_{1k}$ ;
- 2) 用 CAIM 算法对其离散化, 得到离散区间  $I_1$ ;
- 3) 重复进行第1)~2)步操作, 得到不同的离散子区间  $I_1, I_2, \dots, I_s$ ;
- 4) 应用集成方法划分为最小子区间  $I$ ;
- 5) 计算原始空间与离散空间的邻居, 并计算其邻居的相似性;
- 6) 通过各个维度的  $ARI$  值选择优先合并的维度;
- 7) 根据相邻区间的关联度  $R$ , 选择最佳的合并区间;
- 8) 依据停止条件, 当  $PerS$  最大时离散化停止。

### 3 实验结果

本文的算法将与 CAIM<sup>[10]</sup>、基于树<sup>[12]</sup>、基于核函数算法离



散化以及一次  $K$ -means 后离散化的方法进行比较,其中:CAIM 是有监督的离散化方法,其余的则是无监督的方法。由于 CAIM 要使用类别标签,选择 UCI (University of California Irvine) 中具有标签的 6 个数据集 Iris、Wine、Pima、Wdbc、Statlog 和 Shuttle 来测试。此 6 个数据集的概要信息如表 1 所示。

表 1 6 个数据集的概要信息

数据集	数值属性数目	类别数目	样本数目
Iris	4	3	150
Wine	13	3	178
Pima	8	2	768
Wdbc	30	2	569
Statlog	19	7	2310
Shuttle	9	7	14500

既然本文讨论无监督的离散化方法,在实验中使用聚类算法对离散化的结果进行测试。由于 Spectral Clustering<sup>[16]</sup> 具有较好的精确性和鲁棒性,因此选择该方法作为测试算法。在传统的 Spectral Clustering 使用欧氏距离测度来建立相似性矩阵,而对离散化的数据并不直接实用,那么使用 OF<sup>[14]</sup> 测度建立离散化后的数据集的相似矩阵,再应用到 Spectral Clustering 算法。聚类的结果的质量使用 ARI、Rand、Jaccard 和 FM<sup>[17]</sup> 这 4 个指标来度量。

上述 6 个数据集离散化后的聚类结果如表 2 所示。

表 2 6 个数据集离散化后 Spectral Clustering 的聚类结果

数据集	离散化算法	ARI	Rand	Jaccard	FM
Iris	CAIM 算法	0.8508	0.9341	0.8180	0.8999
	基于集成的离散化算法	0.8860	0.9495	0.8578	0.9234
	基于树的离散化算法	0.4669	0.7415	0.4991	0.6603
	基于核函数离散化算法	0.5309	0.7921	0.5225	0.6864
	单次 $K$ -means 离散化算法	0.7570	0.8922	0.7205	0.8376
Wine	CAIM 算法	0.8332	0.9125	0.8012	0.8803
	基于集成的离散化算法	0.8515	0.9339	0.8202	0.9012
	基于树的离散化算法	0.4084	0.7229	0.4238	0.6015
	基于核函数离散化算法	0.5237	0.7758	0.5106	0.6731
	单次 $K$ -means 离散化算法	0.8024	0.9120	0.7676	0.8686
Pima	CAIM 算法	0.8083	0.9124	0.7887	0.8654
	基于集成的离散化算法	0.7821	0.8993	0.7467	0.8519
	基于树的离散化算法	0.4123	0.7356	0.4324	0.6031
	基于核函数离散化算法	0.5235	0.7857	0.5162	0.6793
	单次 $K$ -means 离散化算法	0.7309	0.8653	0.6927	0.8012
Wdbc	CAIM 算法	0.7552	0.8782	0.7952	0.8859
	基于集成的离散化算法	0.7871	0.8994	0.8196	0.8903
	基于树的离散化算法	0.4815	0.7454	0.6433	0.7863
	基于核函数离散化算法	0.5433	0.7622	0.6816	0.8115
	单次 $K$ -means 离散化	0.5933	0.7973	0.6753	0.8063
Statlog	CAIM 算法	0.5365	0.8768	0.4377	0.6117
	基于集成的离散化算法	0.5219	0.8686	0.4276	0.6002
	基于树的离散化算法	0.3556	0.7119	0.3528	0.5278
	基于核函数离散化算法	0.4460	0.8578	0.3905	0.5603
	单次 $K$ -means 离散化算法	0.5033	0.8492	0.4153	0.6015
Shuttle	CAIM 算法	0.7725	0.8964	0.8203	0.9018
	基于集成的离散化算法	0.7511	0.8778	0.8069	0.8632
	基于树的离散化算法	0.6306	0.8101	0.6528	0.7278
	基于核函数离散化算法	0.5415	0.8007	0.5312	0.6949
	单次 $K$ -means 离散化算法	0.7119	0.8232	0.7603	0.8347

从表 2 可看出,本文提出的基于集成学习的离散化结果在 6 个数据集上的结果明显优于基于树和基于核函数以及单次  $K$ -means 后离散化的结果。而本文的算法对于小数据集聚类效果也是很好的,对于大数据集虽然聚类效果比 CAIM 算法差,因为 CAIM 算法应用数据标签信息,聚类效果会更精确。

离散化后的区间数目是离散化的另一个性能指标。区间数越少,数据压缩率越低,但过少将导致原始数据集信息丢失过多,对以后的分类学习效果不佳,因此合适的离散区间数非常重要。

图 5 描述了 5 种离散化方法在 6 个数据集上的离散化区间数目。本文提出的基于集成学习的离散化方法生成的区间数远小于基于树与基于核函数算法离散化的区间数。

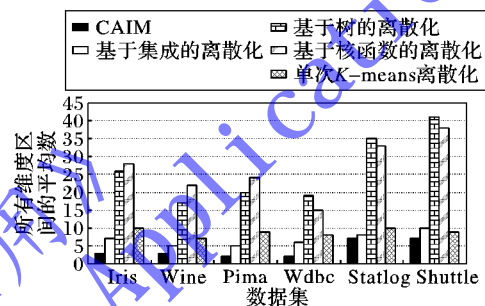


图 5 不同离散化方法生成的平均区间数对比

## 4 结语

本文提出了一个基于集成学习的无监督离散化方法,其新颖之处有两点:1) 利用集成学习的方法创建最小区间;2) 根据数据集原始空间与离散空间的邻居相似性进行区间的合并,以尽可能保持数据内在结构。实验表明,所提算法的离散化过程能较好地保持数据集的内在结构,且离散化后的平均区间数较小。

但是本文的方法有一个缺点,即其计算复杂性较高。当  $K$ -means 划分数据集  $K = \sqrt{N}$  时, $K$ -means 的计算时间为  $O(N^{1.5})$ ,CAIM 的计算复杂度则为  $O(N^{2.5})$ ,合并的过程则为  $O(dN^2)$ ,所以本文算法的计算复杂度为  $O(N^{2.5})$ 。如何降低其计算复杂度是未来需要完成的工作。

## 参考文献:

- [1] SRIKANT R, AGRAWAL R. Mining quantitative association rules in large relational tables [C]// Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 1996: 1-12.
- [2] CATLETT J. On changing continuous attributes into ordered discrete attributes [C]// Proceedings of the European Working Session on Learning on Machine Learning, LNCS 482. Berlin: Springer, 1991: 164-178.
- [3] MEHTA S, PARTHASARATHY S, YANG H. Toward unsupervised correlation preserving discretization [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(9): 1174-1185.
- [4] KERBER R. ChiMerge: discretization of numeric attributes [C]// Proceedings of the Tenth National Conference on Artificial Intelligence. Menlo Park: AAAI Press, 1992: 123-128.

本文提出的自动标注方法的效果依赖于关键词和类别标签词,需要进一步研究针对不同语料的关键词和类别标签词选取方法和新词扩充方法。在标注方法集成方面还有许多值得深入研究的问题,诸如每个标注器(每种标注方法称为标注器)的产生、选择和可信度研究以及集成方法研究等。

#### 参考文献:

- [1] YANG A, ZHOU Y, LIN J. A method of Chinese texts sentiment classification based on Bayesian algorithm [J]. *Applied Mechanics and Materials*, 2012, 263/264/265/266: 2185–2190.
  - [2] YANG A, LIN J, ZHOU Y, *et al.* Research on building a Chinese sentiment lexicon based on SO-PMI [J]. *Applied Mechanics and Materials*, 2012, 263/264/265/266: 1688–1693.
  - [3] CUI G, CHENG Y. Corpus annotation in the corpus [J]. *Journal of Tsinghua University: Philosophy and Social Sciences*, 2000(1): 89–94. (崔刚, 盛永梅. 语料库中语料的标注[J]. 清华大学学报: 哲学社会科学版, 2000(1): 89–94.)
  - [4] LI S. Sentiment classification of micro-blogs corpus based on automatic annotation training set [D]. Shenyang: Northeast Normal University, 2013. (李圣楠. 基于自动标注训练集的微博语料情感分类的研究[D]. 沈阳: 东北师范大学, 2013.)
  - [5] XU L, LIN H, ZHAO J. Construction and analysis of emotional corpus [J]. *Journal of Chinese Information Processing*, 2008, 22(1): 116–122. (徐琳宏, 林鸿飞, 赵晶. 情感语料库的构建和分析[J]. 中文信息学报, 2008, 22(1): 116–122.)
  - [6] PANG L, LI S, ZHOU G. Sentiment classification method of Chinese micro-blog based on emotional knowledge [J]. *Computer Engineering*, 2012, 38(13): 156–158. (庞磊, 李寿山, 周国栋. 基于情绪知识的中文微博情感分类方法[J]. 计算机工程, 2012, 38(13): 156–158.)
  - [7] HAN Z, ZHANG Y, ZHANG H, *et al.* On effective short text tendency classification algorithm for Chinese microblogging [J]. *Computer Applications and Software*, 2012, 29(10): 89–93. (韩忠明, 张玉沙, 张慧, 等. 有效的中文微博短文本倾向性分类算法[J]. 计算机应用与软件, 2012, 29(10): 89–93.)
  - [8] YANG A. Fuzzy classification models and ensemble methods [M]. Beijing: Science Press, 2008. (阳爱民. 模糊分类模型及其集成方法[M]. 北京: 科学出版社, 2008.)
  - [9] China Computer Federation. Test data for evaluation [EB/OL]. [2013-12-10]. [http://tcci.ccf.org.cn/conference/2013/pages/page04\\_tdata.html](http://tcci.ccf.org.cn/conference/2013/pages/page04_tdata.html). (中国计算机学会. 评测测试数据[EB/OL]. [2013-12-10]. [http://tcci.ccf.org.cn/conference/2013/pages/page04\\_tdata.html](http://tcci.ccf.org.cn/conference/2013/pages/page04_tdata.html).)
  - [10] Information Retrieval Laboratory, Dalian University of Technology. Emotional vocabulary ontology database [EB/OL]. [2014-01-18]. [http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx?utm\\_source=weibolife](http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx?utm_source=weibolife). (大连理工大学信息检索研究室. 情感词汇本体库[EB/OL]. [2014-01-18]. [http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx?utm\\_source=weibolife](http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx?utm_source=weibolife).)
  - [11] JIANG F, ZHANG H, LIU Y, *et al.* THUIR-Senti at Chinese microblog mood analysis evaluation [EB/OL]. [2013-12-02]. <http://tcci.ccf.org.cn/conference/2013/dldoc/evrpt02.rar>. (姜飞, 张辉, 刘奕群, 等. THUIR-Senti 中文微博情绪分析评测报告[EB/OL]. [2013-12-02]. <http://tcci.ccf.org.cn/conference/2013/dldoc/evrpt02.rar>.)
  - [12] SUN X, YE J, TANG C, *et al.* Multi-granularity based Chinese microblog sentiment analysis [EB/OL]. [2013-12-02]. <http://tcci.ccf.org.cn/conference/2013/dldoc/evrpt02.rar>. (孙晓, 叶嘉琪, 唐诚意, 等. 基于多粒度模型的中文微博情感分析[EB/OL]. [2013-12-02]. <http://tcci.ccf.org.cn/conference/2013/dldoc/evrpt02.rar>.)
- 
- (上接第 2187 页)
- [5] LIU H, SETIONO R. Chi2: feature selection and discretization of numeric attributes [C]// *Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence*. Washington, DC: IEEE Computer Society, 1995: 388–391.
  - [6] YANG Y, WEBB G I. Discretization for naive-Bayes learning: managing discretization bias and variance [J]. *Machine Learning*, 2009, 74(1): 39–74.
  - [7] RUIZ F J, ANGULO C, AGELL N. IDD: a supervised interval distance-based method for discretization [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2008, 20(9): 1230–1238.
  - [8] DOUGHERTY J, KOHAVI R, SAHAMI M. Supervised and unsupervised discretization of continuous features [C]// *Proceedings of the Twelfth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann Publishers, 1995: 194–202.
  - [9] LI G. An unsupervised discretization algorithm based on mixture probabilistic model [J]. *Chinese Journal of Computers*, 2002, 25(2): 158–164. (李刚. 基于混合概率模型的无监督离散化算法[J]. 计算机学报, 2002, 25(2): 158–164.)
  - [10] KURGAN L A, CIOS K J. CAIM discretization algorithm [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2004, 16(2): 145–153.
  - [11] TSAI C J, LEE C I, YANG W. A discretization algorithm based on class-attribute contingency coefficient [J]. *Information Sciences*, 2008, 178(3): 714–731.
  - [12] SCHMIDBERGER G, FRANK E. Unsupervised discretization using tree-based density estimation [C]// *PKDD 2005: Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, LNCS 3721. Berlin: Springer, 2005: 240–251.
  - [13] BIBA M, ESPOSITO F, FERILLI S, *et al.* Unsupervised discretization using kernel density estimation [C]// *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. San Francisco: Morgan Kaufmann Publishers, 2007: 697–701.
  - [14] BORIAH S, CHANDOLA V, KUMAR V. Similarity measures for categorical data: a comparative evaluation [C]// *Proceedings of the 8th SIAM International Conference on Data Mining*. Philadelphia: SIAM, 2008: 243–254.
  - [15] ZHANG S, WONG H S, SHEN Y. Generalized adjusted rand indices for cluster ensembles [J]. *Pattern Recognition*, 2012, 45(6): 2214–2226.
  - [16] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm [C]// *Advances in Neural Information Processing Systems*. Cambridge: MIT Press, 2002: 849–856.
  - [17] THEODORIDIS S, KOUTROUMBAS K. *Pattern recognition* [M]. Waltham: Academic Press, 2003.