Journal of Computer Applications
计算机应用,2014,34(8):2188-2191
ISSN 1001-9081
CODEN JYIIDU
2014-08-10
http://www.joca.cn

文章编号:1001-9081(2014)08-2188-04
doi:10.11772/j.issn.1001-9081.2014.08.2188

# 中文微博语料情感类别自动标注方法

阳爱民[1*],周咏梅[1],周剑峰[2]
(1.广东外语外贸大学 思科信息学院,广州 510006; 2.广东外语外贸大学 图书馆,广州 510006)
(*通信作者电子邮箱 amyang18@163.com)

**摘 要**:针对大规模微博语料手动标注困难的问题,提出了中文微博语料情感类别自动标注的方法,包括基于关键词的、基于概率求和的和基于概率乘积的3种自动标注方法和一种集成标注方法。自动标注时首先分别使用3种标注方法进行标注,得到3种标注结果;然后,采用标注方法集成的策略,对3种标注的结果通过投票的方式决定最终的标注结果。通过设计自动标注实验系统进行实验,实验结果验证了所提方法的可行性和有效性。实验结果表明,单个标注方法的准确率均在70%以上,投票方法的准确率达90%以上。

**关键词**:中文微博;微博情感;情感分类;自动标注;准确率
**中图分类号**:TP301.6; TP391 **文献标志码**:A

# Automatic annotation methods for Chinese micro-blog corpus with sentiment class

YANG Aimin[1*], ZHOU Yongmei[1], ZHOU Jianfeng[2]
(1. *Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou Guangdong* 510006, *China*;
2. *Library, Guangdong University of Foreign Studies, Guangzhou Guangdong* 510006, *China*)

**Abstract:** For the difficulty of manual annotation on large-scale micro-blog corpus, three automatic annotation methods and an integrated annotation method by voting for Chinese micro-blog corpus were proposed. Three automatic annotation methods included keywords-based annotation method, probability-summation-based annotation method and probability-product-based annotation method. During the process of automatic annotation, firstly, micro-blog corpus were annotated by three annotation methods respectively, and three results were obtained, then the final annotation results were determined by voting method with the integrated strategy. By designing automatic annotation experiment system, experimental results verify the feasibility and effectiveness of the proposed methods, and show that the accuracy of the single annotation method is more than 70%, and it is more than 90% for the voting method.

**Key words:** Chinese micro-blog; micro-blog sentiment; micro-blog sentiment classification; automatic annotation; accuracy

## 0 引言

微博是一种流行的网络社交平台,用户通过操作手机就可以轻松实现获取、分享和转发微博平台上海量的微信息。对微博进行情感分析在市场分析预测、民意调查、智能导购、信息安全等诸多领域有着广阔的应用空间和发展前景[1-2]。微博语料库是进行微博情感分析研究的重要基础,要提高语料的利用价值,关键在于语料的标注,所谓标注[3]就是对语料库中的原始语料进行加工,把各种表示语言特征的附码标注在相应的语言成分上,以便于计算机的识读。然而,规模庞大的微博文本给通过人工标注工作带来非常大的困难,当前中文微博情感分析研究领域没有标准的语料库,这在一定程度上影响了该领域的研究。为了减轻标注人员的负担,提高标注的效率和精确度,减少标注的错误率,非常有必要研究自动标注方法,以便协助标注人员的工作。因此,探索研究微博情感类别自动标注方法是一项非常重要的工作。

在语料库情感自动标注研究领域,李圣楠[4]提出一种无人工干预的微博语料库自动标注方法,采用表情符号及情感词对微博语料进行筛选标注,在特定语料集情况下其标注准确率达到了约83%;徐琳宏等[5]介绍了情感语料库构建中情感标注方面的相关成果,提出了相应的情感标注体系和规范,并对语料库中的情感分布进行了介绍,这有助于进行语料自动标注的研究;庞磊等[6]提出利用情绪词和表情图片两种情绪知识对大规模微博非标注语料进行筛选并自动标注,其用于电影及手机评论语料,标注准确率达到约87%;韩忠明等[7]以HowNet的情感词典为基础,提出一个微博新词发现算法,构建微博情感词典,通过一个自动机来计算短文本情感倾向性,实验对比了基于Hownet的方法和基于支持向量机(Support Vector Machine, SVM)的方法,表明文献提出的方法在短文本分类准确率及分析效率上更胜一筹。

本文以中文微博的"正向"和"负向"情感作为标注的两个类别,提出基于多种方法集成[8]的中文微博情感自动标注

**收稿日期**:2014-04-09;**修回日期**:2014-05-09。
**基金项目**:国家社会科学基金资助项目(12BYY045);教育部新世纪优秀人才支持计划项目(NCET-12-0939)。
**作者简介**:阳爱民(1970-),男,湖南永州人,教授,博士,CCF高级会员,主要研究方向:机器学习、文本情感分析; 周咏梅(1971-),女,湖南永州人,教授,CCF会员,主要研究方向:自然语言处理、文本情感分析; 周剑峰(1986-),男,湖南株洲人,助理馆员,硕士,主要研究方向:机器学习、文本情感分析。

方法,将基于关键词的标注方法、基于概率乘积的标注方法和基于概率求和的标注方法集成起来,采用投票机制确定标注结果,实现中文微博语料情感类别自动标注。

## 1　中文微博情感类别自动标注基本思路

本文提出的中文微博情感类别自动标注基本思路如图 1 所示。

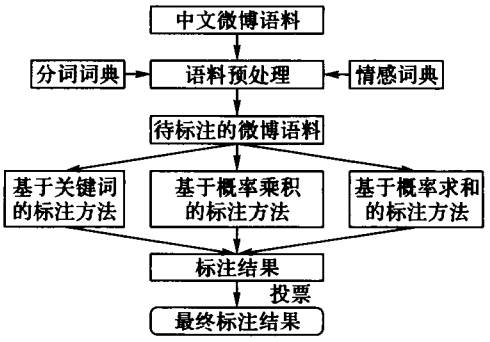


图 1　微博情感类别自动标注过程

从图 1 可以看出:首先,采集后的中文微博语料经过语料预处理过程得到待标注的语料;然后,待标注的语料分别经 3 种方法进行自动标注,并分别产生标注结果;最后,采用投票的方式把 3 种标注方法集成起来,确定最终的标注结果。本文的自动标注类别为“正向”和“负向”两种类别。图 1 中的语料预处理主要进行以下几方面的工作:1)利用情感词典(含表情符号)过滤不包含情感词或表情符号的微博,在这里假设不包含情感词或表情符号的微博没有情感倾向;2)对微博语料进行分词处理;3)在分词的基础上对微博语料进行词频统计、权重计算等工作。图 1 中的基于关键词的标注方法、基于概率乘积的标注方法和基于概率求和的标注方法将在第 2 章详细介绍。

## 2　中文微博语料情感类别自动标注方法

本文提出了 3 种文本情感语料自动标注方法,主要针对中文微博语料,语料中除文字外,还要考虑表情符、表情符链接、标点符号等。本文的研究将微博情感标注为正向情感和负向情感两种类别。

为了方便介绍自动标注方法,在此先约定相关符号:设 $D=\{d_1,d_2,\cdots,d_L,d_{L+1}\cdots,d_N\}$ 表示中文微博语料集,$N$ 表示语料集中的微博条数。$D_L=\{d_1,d_2,\cdots,d_L\}$ 表示语料集 $D$ 中已标注的语料集,$D_U=\{d_{L+1},d_{L+2},\cdots,d_N\}$ 表示 $D$ 中需要标注的语料集。$C=\{c_1,c_2,\cdots,c_K\}$ 表示语料集的类别标志,$K$ 表示类别数,本文假设仅有两个类别,$K=2$。$S=\{l_1,l_2,\cdots,l_L\}$ 表示 $D_L$ 的类别标志集合,$l_j$ 表示语料 $d_j$ 的类别标志,其中 $l_j\in C$。$W=\{w_1,w_2,\cdots,w_K\}$ 表示类别关键词(情感词和情感符号)及其权重,其中:$w_k=\{wd_k,wt_k\}$,$wd_k=\{wd_{1k},wd_{2k},\cdots\}$ 表示第 $k$ 类的关键词,$wt_k=\{wt_{1k},wt_{2k},\cdots\}$ 表示第 $k$ 类关键词词的权重值。$W$ 可以单独来自基础情感词典,也可以从标注类别的语料统计和基础情感词典相结合得到。$\boldsymbol{Z}=[z_1,z_2,\cdots,z_{N_u}]^{\mathrm{T}}$ 表示待标注语料的类别矩阵,其中 $z_j=(z_{j1},z_{j2},\cdots,z_{jK})$。

### 2.1　基于关键词的自动标注方法

基于关键词的自动标注方法的基本思路是根据微博语料中各类别关键词出现的频率来计算语料所属的类别,其中各类关键词的获取是通过情感词典(包含表情符号)和对微博语料词频来确定关键词,根据情感词典的极性和强度确定每个类别的关键词及其权重。具体的算法如方法 1 所示。

方法 1　基于关键词的自动标注方法。

输入　微博语料集 $D$(由 $D_L$ 和 $D_U$ 组成)、$K(K=2)$ 和情感词典 $SL$(含表情符号)。

输出　微博语料集 $D_U$ 的类别。

步骤 1　生成 $W$,对微博语料进行词频(含表情符号)的统计,并与情感词典 $SL$ 比较确定各类别关键词及其权重。

步骤 2　计算矩阵 $\boldsymbol{Z}$,按式(1)计算并生成矩阵 $\boldsymbol{Z}$:

$$z_{jk}=\left(\sum_{v\in wd_k}wdn_{v_j}*wt_{v_j}\right)/wd_k \tag{1}$$

其中:$wdn_{v_j}$ 表示待标注语料 $d_j$ 中关键词 $v$(属于第 $k$ 类)的频率(次数),$wt_{v_j}$ 表示其权重,$wd_k$ 为当前语料中属于第 $k$ 类的关键词的个数。

步骤 3　归一化矩阵 $\boldsymbol{Z}$,满足式(2)要求:

$$\sum_{k=1}^{K}z_{jk}=1;\quad j=1,2,\cdots,N_u \tag{2}$$

步骤 4　判定所属类别,如果 $z_{jk}>0.5+\frac{1}{k+1}$,则把微博 $d_j$ 的类别定为 $k$。

步骤 5　方法结束。

### 2.2　基于概率乘积的自动标注方法

基于概率乘积的自动标注方法主要通过统计微博语料中各类别标签词的数目,分别计算微博属于“正面”和“负向”类别的概率,以微博所包含类别标签词概率乘积确定微博的得分,根据分数高低判定微博情感倾向所属类别。类别标签词借鉴方法 1 关键词确定方法来产生。具体见方法 2。

方法 2　基于概率乘积的自动标注方法。

输入　微博语料集 $D$(由 $D_L$ 和 $D_U$ 组成)、$K(K=2)$ 和情感词典 $SL$(含表情符号)。

输出　微博语料集 $D_U$ 的类别。

步骤 1　生成 $W$,对微博语料进行词频(含表情符号)的统计,并与情感词典 $SL$ 比较确定各类别标签词及其权重。

步骤 2　计算类别标签词分数,按式(3)计算待标注微博相对类别标签词的分数:

$$Score(d_j,wd_k)=\prod_{v\in wd_k,v\in d_j}p(v\mid wd_k)^{wdn_{v_j}} \tag{3}$$

其中:$Score(d_j,wd_k)$ 表示语料 $d_j$ 相对标签词的得分数;$p(v\mid wd_k)$ 表示语料 $d_j$ 中标签词 $v$ 相对于第 $k$ 类的概率;$wdn_{v_j}$ 表示待标注语料 $d_j$ 中关键词 $v$(属于第 $k$ 类)的频率(次数),$\sum_{i=1}^{|v|}p(v\mid wd_k)=1$。

通过最大似然估计法估值 $p(v\mid wd_k)$,按式(4)估算:

$$p(v\mid wd_k)=wdn_{v_j}/|wd_k| \tag{4}$$

其中 $|wd_k|$ 表示第 $k$ 类标签词的数目。

步骤 3　判定所属类别,根据式(5)判定微博所属类别:

$$d_j\in \arg\max_{1\le k\le K}Score(d_j,wd_k) \tag{5}$$

步骤 4　方法结束。

### 2.3　基于概率求和的自动标注方法

基于概率求和的自动标注方法主要通过统计微博语料中各类别标签词的数目,分别计算微博属于“正面”和“负向”类

别的概率,以微博所包含类别标签词概率之和确定微博的得分,根据分数高低判定微博情感倾向所属类别。类别标签词借鉴方法 1 关键词确定方法来产生。具体见方法 3。

方法 3 基于概率求和的自动标注方法。

输入 微博语料集 $D$(由 $D_L$ 和 $D_U$ 组成)、$K(K=2)$ 和情感词典 $SL$(含表情符号)。

输出 微博语料集 $D_U$ 的类别。

步骤 1 生成 $W$,对微博语料进行词频(含表情符号)的统计,并与情感词典 $SL$ 比较确定各类别标签词及其权重。

步骤 2 计算类别标签词分数,按式(6)计算计算待标注文档相对类别标签词的分数:

$$Score(d_j,wd_k)=\sum_{v\in wd_k}p(v\mid d_j)\lg p(v\mid wd_k) \tag{6}$$

其中:$Score(d_j,wd_k)$ 表示语料 $d_j$ 相对标签词的得分数;$p(v\mid wd_k)$ 表示语料 $d_j$ 中标签词 $v$ 相对于第 $k$ 类的概率,$\sum_{i=1}^{|V|}p(v\mid wd_k)=1$。

通过最大似然估计法估值 $p(v\mid d_j)$,按式(7)估算,$p(v\mid wd_k)$ 按式(8)估算:

$$p(v\mid d_j)=wdn_{v_j}/|d_j| \tag{7}$$

$$p(v\mid wd_k)=wdn_{v_j}/|wd_k| \tag{8}$$

其中:$|d_j|$表示第 $j$ 条微博包含的类别标签词的数目,$wdn_{v_j}$ 表示待标注语料 $d_j$ 中关键词 $v$(属于第 $k$ 类)的频率。

步骤 3 判定所属类别,根据式(9)判定微博所属类别:

$$d_j\in \arg\max_{1\le k\le K}Score(d_j,wd_k) \tag{9}$$

步骤 4 方法结束。

在本文研究中借鉴分类器集成的思想,将相互之间具有独立决策能力的分类器联合起来的方式就称作分类器集成,事实证明,通常情况下集成分类器的预测能力要比单个分类器的预测能力好得多[8]。基于上述思想和 3 种方法标注的结果,采用投票方式,以简单多数票的原则确定微博最终所属的类别。

## 3 实验及其结果分析

### 3.1 实验数据

实验语料库主要来自于第 2 届自然语言处理与中文计算会评测当中的微博语料[9]。这些语料已标注了情感类别,本文用来作为验证提出方法的有效性。微博语料共有12 895 条,其中:语料表中“正向”类的语料有 4 819 条,“负向”类的语料有 8 076 条。实验中选用了大连理工大学的词汇本体库[10]作为所使用的情感词典,词典总情感词 21 957 个,其中:正向词语 11 198 个,负向词语 10 759,并在词典中增加如表 1 所示表情图片(表情符号词汇本体库已包含一些)。

**表 1 表情图片示例**

| 正向表情 | | 负向表情 | | 正向表情 | | 负向表情 | |
|---|---|---|---|---|---|---|---|
| 图片 | 含义 | 图片 | 含义 | 图片 | 含义 | 图片 | 含义 |
| | 微笑 | | 衰 | | 爱心 | | 惊恐 |
| | 嘻嘻 | | 吐 | | 抱拳 | | 咒骂 |
| | 鼓掌 | | 愤怒 | | 赞 | | 泪 |
| | 强 | | 鄙视 | | 色 | | 哭 |
| | 胜利 | | 弱 | | | | |

### 3.2 实验结果及分析

12 000 多条微博语料经预处理后,大约 8 000 多条微博进行自动标注实验。使用第 2 章介绍的方法设计软件程序分别进行自动标注实验。图 2 是设计的中文微博自动标注实验系统界面。

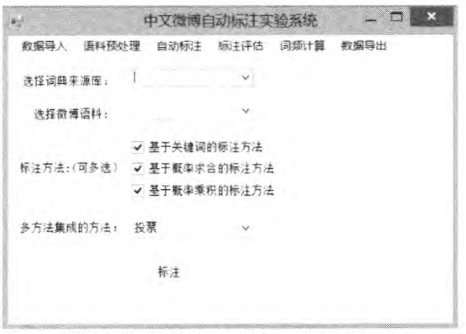


图 2 微博自动标注实验系统界面

使用设计的实验系统,对选定的中文微博语料进行自动标注实验,实验结果如表 2 所示。表 2 中,“*Pos_acc*”表示正向微博语料标注的准确率,“*Neg_acc*”表示负向微博语料标注的准确率,“*All_acc*”表示微博语料标注总的准确率。

**表 2 自动标注准确率** %

| 方法 | *Pos_acc* | *Neg_acc* | *All_acc* |
|---|---|---|---|
| 方法 1 | 75.14 | 82.05 | 78.18 |
| 方法 2 | 78.75 | 86.50 | 82.16 |
| 方法 3 | 81.61 | 89.91 | 85.26 |
| 投票法 | 88.89 | 95.82 | 91.78 |

从表 2 中可以看出,第 3 种方法自动标注的准确率比其他两种方法高,结合 3 种方法的投票法自动标注的准确率比单个方法高,表明将 3 种方法集成起来的投票方法是非常有效的。

文献[11-12]与本文的研究均采用文献[9]的数据来评测提出的方法。其中:文献[11]使用了多种方法,使用的 SVM 方法在闭环(Close)的条件下分类准确率为 66.50%,在开环(Open)条件下分类准确率为 72%;使用的隐含狄利克雷分配(Latent Dirichlet Allocation,LDA)模型,在闭环(Close)的条件下分类准确率为 57.47%。文献[12]使用 SVM 方法的分类准确率为 49.55%。本文研究的目标与文献[11-12]不同,本文的研究目标是研究获得有效的自动标注方法和自动标注语料的类别,为分类器提供有效的训练样本和测试样本。因此,本文的研究可以对数据集进行预处理,对不能标注的语料可以过滤掉。这样本文方法的准确率自然就比文献[11-12]准确率要高。文献[11-12]的方法是构建分类器,其目的是分类未知语料的类别,分类准确率高的分类器才有实用的价值。实验结果与上述文献对比分析表明,本文提出的自动标注方法是有效的。

## 4 结语

本文对中文微博语料情感类别自动标注方法进行了研究,提出了 3 种自动标注方法,并采用投票方式集成 3 种方法,使得标注准确率大幅度提高。根据提出的方法设计了实验系统,实验结果和相关参考文献分析表明了本文提出的方法是可行的和有效的。

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. Sensiment 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. (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.)

[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. (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.)

[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.