

不确定噪声下海量文本数据的模糊挖掘算法研究

潘大胜

(百色学院 信息工程学院, 广西 百色 533000)

摘 要: 针对传统的数据挖掘方法一直存在挖掘精度低、运行时间长的问题, 提出基于小波变换与关联规则的不确定噪声下海量文本数据的模糊数据挖掘算法, 首先利用小波变换对不确定噪声下海量文本数据的模糊数据进行预处理, 将模糊海量文本数据时间序列转换至频谱空间中, 获得频谱空间内距离最小、类间聚类最大的变换基系数, 并将其作为海量文本模糊数据特征系数, 利用数据特征系数计算出其从属于各类别的隶属度, 确定模糊文本数据集的关联规则, 依据多维海量数据集之间的相关程度进行区间划分, 由此实现对不确定噪声下海量文本数据的有效挖掘. 实验结果证明, 所提算法能有效提高海量文本数据挖掘精度, 且挖掘效率较高.

关键词: 不确定噪声; 海量文本数据; 模糊数据挖掘算法; 特征系数; 关联规则

中图分类号: TN391

文献标识码: A

文章编号: 1000-7180(2017)09-0129-04

Research on Fuzzy Mining Algorithm for Massive Text Data Under Uncertain Noise

PAN Da-sheng

(School of Information Engineering, Baise University, Baise 533000, China)

Abstract: According to the traditional data mining methods have been mining of low precision, long running time, the wavelet transform and the uncertain fuzzy association rules data text data mining algorithm based on noise, firstly using wavelet transform for uncertain fuzzy data of massive text data noise preprocessing, fuzzy time massive text data sequence conversion to spectrum space, get the distance transform based clustering coefficient, the maximum of the minimum inter class spectrum space, and as a massive text data feature data using fuzzy coefficient, calculate the feature coefficients from membership belonging to the respective categories, determine the fuzzy association rules text data sets, interval division basis the multidimensional degree between the massive data sets, thus the uncertainty of effective mining of massive text data noise. The experimental results show that the proposed algorithm can effectively improve the accuracy of massive text data mining, and the mining efficiency is high.

Key words: uncertain noise; massive text data; the fuzzy data mining algorithm; characteristics of the coefficient of; association rules

1 引言

近年来,随着科学技术的不断发展,数据挖掘的应用逐渐广泛,多应用在零售业、制造业、财务金融保险业、通讯业以及医疗服务等多个领域^[1]. 现在企业内部由于商业运作涌现越来越多的海量文本数据,对不确定噪声下的海量文本模糊数据进行分析、挖掘成为有待解决的重要问题^[2], 现有的海量文本

模糊数据挖掘算法,是从海量文本数据中发现相对稀疏和孤立的模糊数据,对其进行挖掘,但对海量文本数据的挖掘精度不高^[3-4]. 为了从不确定噪声下已知的海量文本数据中统计出详细的数据,发现潜在的规则,对海量文本数据进行归类整理,需要对海量文本数据的模糊数据挖掘算法进行研究^[5-7]. 越来越多的国内外学者开始了对不确定噪声下海量文本数据的模糊数据挖掘的探索与研究,提出了很多算法.

文献[8]提出基于统计的不确定噪声下海量文本数据的模糊数据挖掘算法,但在不确定噪声下对海量文本模糊数据的挖掘精度不高.文献[9]提出基于分层递阶约的不确定噪声下海量文本数据的模糊数据挖掘算法,但在不确定噪声下对海量文本模糊数据挖掘的运行时间较长,挖掘效率不高.文献[10]提出基于神经网络的不确定噪声下海量文本数据的模糊数据挖掘算法,但对海量文本模糊数据的挖掘精度不高,实用范围受限.

针对上述方法存在的问题,提出基于小波变换与关联规则的不确定噪声下海量文本数据的模糊数据挖掘算法,并进行实验对比分析.实验结果证明,所提算法能有效提高海量文本数据挖掘精度,且挖掘效率较高,实用性广.

2 不确定噪声下海量文本数据的模糊数据挖掘

2.1 基于小波变换的海量文本模糊数据预处理

小波变换是用一簇模糊数据小波函数系去逼近一个模糊数据信号,其中小波具有模糊数据傅里叶变换中正弦和余弦函数的作用,而海量文本模糊数据小波函数系是通过一个基本海量文本模糊数据小波函数经过伸缩和平移后构成的,计算公式如下:

$$\phi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right), a, b \in R, a \neq 0 \quad (1)$$

式中, a 表示模糊数据伸缩因子; b 表示模糊数据平移因子; ψ 表示模糊数据母小波; t 表示模糊数据尺度因子; R 表示模糊数据平移与伸缩因子合集.对于模糊数据任意函数 $f \in L^2(R)$,关于 $f(t)$ 的连续小波变换公式如下:

$$\begin{aligned} W_{\psi} f(a, b) &= \langle f(t), \phi_{a,b}(t) \rangle \\ &= \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} f(t) \psi\left(\frac{t-b}{a}\right) dt \end{aligned} \quad (2)$$

为了减少计算量,将上述变换进行模糊数据二进制离散化处理如下:

$$\phi_{j,k} = \phi_{a,b}(t) = 2^{\frac{j}{2}} \psi(2^j t - k) \quad (3)$$

式中, j 表示模糊数据小波分解系数; k 表示模糊数据小波离散系数.则模糊数据连续小波变换 $f(t)$ 可以表示成如下模糊数据小波序列:

$$f(t) = \sum_{j,k=-\infty}^{+\infty} c_{j,k} \phi_{j,k}(t) \quad (4)$$

则海量文本模糊数据离散二进小波变换过程如下:

$$c_{j,k} = \langle f(t), \phi_{j,k}(t) \rangle = 2^{\frac{j}{2}} \int_{-\infty}^{+\infty} f(t) \psi(2^j t - k) dt \quad (5)$$

海量文本数据的模糊数据尺度函数和小波函数满足如下模糊数据的双尺度方程:

$$\begin{cases} \varphi(t) = \sqrt{2} \sum_{k=0}^{L-1} h_k \varphi(2t - k) \\ \psi(t) = \sqrt{2} \sum_{k=0}^{L-1} g_k \psi(2t - k) \end{cases} \quad (6)$$

式中, $\psi(t)$ 与 $\varphi(t)$ 分别表示模糊数据的尺度函数与小波母函数; L 表示海量文本的模糊数据滤波器长度; h_k 表示海量文本模糊数据的低通滤波器系数; g_k 表示海量文本模糊数据的高通滤波器系数.假设模糊数据小波是正交的,则模糊数据小波重构公式如下:

$$g_k = (-1)^k h_{L-k-1} \quad (7)$$

模糊数据离散小波分解和重构能够进行分层处理,对模糊数据每一层的平滑数据信号进行分解,分成第一层的海量文本模糊数据的平滑信号和细节信号.

2.2 海量文本数据的转换与提取

不确定噪声下海量文本数据父波形式与母波形式分别为如下:

$$\varphi(t) = \begin{cases} 1t \in [0 \quad 0 \quad 5] \\ -1t \in [0 \quad 5 \quad 0] \\ 0t \notin [0, 1] \end{cases} \quad (8)$$

海量文本数据父波形式的提取时间序列的平均值信息,称为海量文本数据低频信息;海量文本数据母波形式提取的是时间序列的变化信息,称为海量文本数据高频信息.通过对海量文本数据的父波与母波形式进行偏移和缩放,得到两个海量文本数据函数族,分别表示如下:

$$\varphi_i^j(t) = \varphi(2^j t - i) \quad (9)$$

式中, $j = 0, 1, \dots; i = 0, 1, \dots, 2^j - 1$; i 表示海量文本数据缩放尺度,这两个海量文本数据函数族构成了海量文本数据的变换基函数向量空间:

$$W^j = sp \{ \varphi_i^j \}_{i=0,1,\dots,2^j-1} \quad (10)$$

在上述公式计算基础上对海量文本数据训练集成员的小波系数进行评价,找出有助于海量文本数据的数据挖掘系数位置及其数值作为该海量文本数据的特征系数.

假设将两个类别的海量文本数据训练集分别标记为 A 和 B ,将海量文本数据 A 类成员 32 个系数记为 $a_1, a_2, \dots, a_{32}$, 海量文本数据均值为 $\bar{a}_1, \bar{a}_2, \dots$,

$\bar{a}_{32}$;海量文本数据标准差为 $v(a_1), v(a_2), \dots, v(a_{32})$;与海量文本数据 A 类同理, B 类成员 32 个系数则标记为 $b_1, b_2, \dots, b_{32}$;海量文本数据的均值为 $\bar{b}_1, \bar{b}_2, \dots, \bar{b}_{32}$;海量文本数据标准差为 $v(b_1), v(b_2), \dots, v(b_{32})$; v 表示海量文本数据挖掘因子;海量文本数据特征系数集合则记为

$$S = (p_1, p_2, \dots, p_k) \quad (11)$$

式中, $p_{1,2,\dots,k}$ 代表海量文本数据系数位置. 初始状态下, S 集合为空集, 按顺序考察海量文本数据训练集各个系数位置, 完成对海量文本模糊数据的提取.

2.3 海量文本数据特性挖掘

在计算出海量文本数据特征系数的基础上, 假设海量文本数据样本是分布于多维子空间, 在同一多维子空间内的两个海量文本数据样本的关联度越大, 则说明具有强相关性, 反之则具有弱相关性. 对同一多维子空间的海量文本数据样本进行划分时, 需要根据海量文本数据样本的关联程度制定相应的挖掘规则. 在对不确定噪声下海量数据文本的模糊数据进行去噪、转换、提取的预处理后对海量文本数据进行挖掘.

假设海量文本数据多维子空间的维度为 d , 先挖掘处于海量文本数据位于不同子空间的不同模糊数据集, 其中模糊数据子空间用矩阵 M 表示, 其公式定义如下:

$$M = \begin{bmatrix} M^1 \\ \vdots \\ M^d \end{bmatrix} d \leq n \quad (12)$$

假设位于两个不同文本数据的子空间 $M^i (i \leq d)$ 和 $M^k (k \leq d)$ 的两个文本数据集分别为 V_i 和 V_k ; $D(i, k)$ 表示两个文本数据子空间的欧几里得距离, $d(i, k)$ 表示两个文本数据集的欧几里得距离. 则对不同文本数据子空间的两个文本数据集的挖掘公式如下:

$$W(M^i, M^k) = \frac{\sigma}{2} \begin{bmatrix} M^1 \\ \vdots \\ M^d \end{bmatrix} p(V_i) p(V_k) \times \log_2 \sqrt{D(i, k)^2 + d(i, k)} \quad (13)$$

式中, σ 代表海量文本数据子空间挖掘因子, 海量文本数据集 V_i 和 V_k 的挖掘频率分别为 $P(V_i)$ 和 $P(V_k)$.

对于同一海量文本数据子空间的不同模糊数据集的挖掘, 通过不同海量文本数据集之间的关联程度进行划分. 计算出在同一空间下海量文本数据集

V_i 和 V_k 的关联因子:

$$g(i, k) = \sqrt{\frac{K_1}{K_2} \begin{bmatrix} x_{1i} \cdots x_{mi} \\ \vdots \\ x_{ni} \cdots x_{li} \end{bmatrix}} d(i, k) - \ln 2 \quad (14)$$

$$\frac{1}{m} \left[\frac{K_2}{K_1} \begin{bmatrix} x_{1k} \cdots x_{mk} \\ \vdots \\ x_{nk} \cdots x_{lk} \end{bmatrix} \right]$$

式中, K_1 表示海量文本数据之间关联程度的平均值; K_2 表示海量文本数据之间差异度的平均值; V_i 和 V_k 表示两个数据集:

$$V_i = (x_{1i}, x_{2i} \cdots x_{mi}) V_k = (x_{1k}, x_{2i} \cdots x_{mk}) \quad (15)$$

m 表示海量文本数据的关联语义. 得到海量文本数据集 V_i 和 V_k 的关联因子 $g(i, k)$ 后, 可以得到相同海量文本数据子空间下这两个海量文本数据集的挖掘公式为

$$W(V^i, V^k) = (P(V_i) - P(V_k)) g(i, k) d(i, k) \times e^{g(i, k)} \begin{bmatrix} x_{1k} \cdots x_{mk} \\ \vdots \\ x_{nk} \cdots x_{lk} \end{bmatrix} \begin{bmatrix} x_{1i} \cdots x_{mi} \\ \vdots \\ x_{ni} \cdots x_{li} \end{bmatrix} \quad (16)$$

假设在同一海量文本数据空间 M_i 下海量文本数据集之间的关联程度限定阈值 $T(V)$, 当 $T(V)$ 小于海量文本数据集之间的关联因子时, 则这两个海量文本数据集具有强相关性, 其区分公式表示如下:

$$f_{M^i}(V_i) = \begin{bmatrix} V_1 \cdots V_n \\ \vdots \\ V_n \cdots V_1 \end{bmatrix} W(V^i, V^k) - \ln \left(\sqrt{\sum_{j=1}^n (P(V_i) - P(V_j))} \right) \quad (17)$$

当 $T(V)$ 大于海量文本数据集之间的关联因子时, 则这两个海量文本数据集具有弱相关性, 其海量文本数据集区分公式表示如下:

$$f_{M^i}(V_i) = \frac{\begin{bmatrix} V_1 \cdots V_n \\ \vdots \\ V_n \cdots V_1 \end{bmatrix} W(V^i, V^k)}{(\pi e^2 - 1)} + \frac{e}{2} \sum_{j=1}^n (P(V_i) - P(V_j)) \quad (18)$$

3 实验结果与分析

为了有效说明基于小波变换与关联规则映射的不确定噪声下海量文本数据的模糊数据挖掘算法的优越性, 能提高海量文本数据的模糊数据挖掘精度.

实验仿真硬件平台为 IBM 的 PC 机,主频为 2.3 GHz CPU,操作系统为 Windows 7,内存为 4 GB. 软件仿真平台为 MATLAB 7.0,在实验中准备了随机真实的海量文本数据集,包括零售数据集、制造数据集、通讯数据集等 1000 个数据集. 引用模糊数据查全率和查准确率来衡量所提算法的有效性.

实验选取文献[7]算法、文献[9]算法、与本文算法在零售业、制造业、财务金融保险业、通讯业以及医疗服务等不同行业的海量文本数据的模糊数据挖掘精度进行对比分析. 实验结果如表 1 所示.

表 1 海量文本模糊数据挖掘精度对比 %

类别	文献[7]		文献[9]		本文算法	
	P	R	P	R	P	R
零售	96.8	93.8	95.3	92.8	98.5	97.9
制造	95.4	87.5	80.5	72.6	98.8	96.7
金融	93.2	85.7	83.7	71.2	95.6	97.3
通讯	87.6	96.8	83.9	76.9	90.9	93.5
医疗	96.9	91.2	87.8	87.6	94.0	94.2
平均	93.4	92.1	83.7	85.4	94.9	95.8

由表 1 可以看出,本文算法在随机真实的各个行业文本数据集的准确率与召回率均高于其他两种对比算法. 实验结果证明,本文算法对海量文本数据的挖掘精度高于对比算法.

为了验证本文算法能够提高文本数据的挖掘效率,抽样选取海量文本数据中不同数量的海量文本数据集数,测试其挖掘精度(%)与挖掘运行时间,实验结果如图 1、图 2 所示.

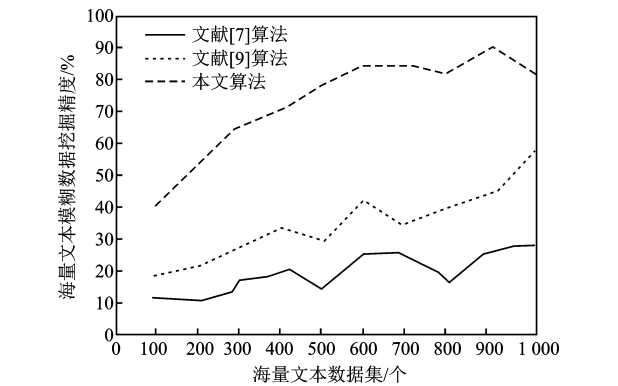


图 1 海量文本数据不同挖掘算法的挖掘精度

通过图 1、图 2 可以看出,本文算法在不确定噪声下海量文本数据的模糊数据挖掘过程中,挖掘精度高于其他两种对比算法;而且在对海量文本数据挖掘过程中的运行时间较短,挖掘效率较高.

4 结束语

采用当前算法对不确定噪声下海量文本数据的模

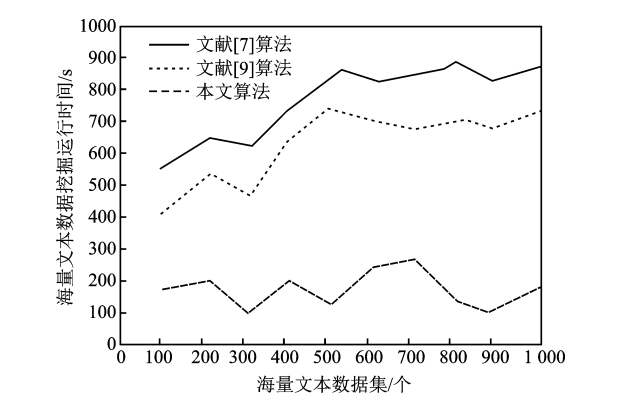


图 2 模糊数据不同挖掘算法的运行时间

糊数据进行挖掘的过程中,存在挖掘精度不高且运行时间较长,实用性不好的问题. 提出一种基于小波变换与关联规则的不确定噪声下海量文本数据的模糊数据挖掘算法. 经仿真实验证明,所提算法无论是在挖掘精度上,还是在挖掘效率上均有所提高,实用性较好.

参考文献:

[1] 王元卓,贾岩涛,刘大伟,等. 基于开放网络知识的信息检索与数据挖掘[J]. 计算机研究与发展,2015,52(2): 456-474.

[2] 李静. 基于数据挖掘技术的电子商务 CRM 研究[J]. 现代电子技术,2015,38(11):126-128.

[3] 柳萌萌,赵书良,韩玉辉,等. 多尺度数据挖掘方法[J]. 软件学报,2016,27(12):3030-3050.

[4] 梁吉业. 大数据挖掘面临的挑战与思考[J]. 计算机科学,2016,43(7):1-2.

[5] 张健. 基于供给侧改革和数据挖掘的茶叶在线销售策略探析[J]. 商业经济研究,2016,35(12):53-54.

[6] 杨莎,余伟,李石君,等. 基于 Web 大数据挖掘的证券价格波动实时影响研究[J]. 计算机科学,2015,42(4): 166-171.

[7] 汪荆琪,徐林莉. 一种基于多视图数据的半监督特征选择和聚类算法[J]. 数据采集与处理,2015,30(1): 106-116.

[8] 陈小燕. 机器学习算法在数据挖掘中的应用[J]. 现代电子技术,2015,38(20):11-14.

[9] 张春生. 大数据环境下相容数据集的关联规则数据挖掘[J]. 微电子学与计算机,2016,33(8):34-39.

[10] 刘静静,智淑敏. 一种传感器网络不确定感知数据挖掘方法研究[J]. 电子设计工程,2016,24(13):73-76.

作者简介:

潘大胜 男(壮族),(1975-),硕士,副教授. 研究方向为数据挖掘技术. E-mail:bspandsh@163.com.