基于特征模糊贴近的数据库约束挖掘算法

王 勇[1] 邹盛荣[2]
(江苏科技大学电气与信息工程学院 张家港 215600)[1] (扬州大学信息工程学院 扬州 225009)[2]

摘 要 传统的关联规则算法,只考虑了类内的关联性,忽略了类间的相似性特征、高开销的分类过程、耗时的关联过程。提出了数据内间特征模糊贴近分类的数据库约束挖掘算法,其通过数据模糊集间的贴近度描述数据间的一致度,在传统的神经网络挖掘技术中,引入数据融合技术,对类间数据进行分类处理后,对原始挖掘数据的动态特征进行分析获取新的挖掘模型,以在大规模数据库中准确查询目标数据。仿真实验结果表明,算法挖掘稀疏数据集和密集数据集的效率都优于传统的关联规则算法,极大提高了数据库的挖掘效率。

关键词 模糊贴近,数据挖掘,神经网络
中图法分类号 TP391.9 **文献标识码** A

Mining Algorithm of Database Constraints Based on Characteristics Fuzzy Closer

WANG Yong[1] ZOU Sheng-rong[2]
(Department of Electrical and Information Engineering, Jiangsu University of Science and Technology, Zhangjiagang 215600, China)[1]
(College of Information Engineering, Yangzhou University, Yangzhou 225009, China)[2]

Abstract Traditional association rules algorithm only considers the class of close contact, ignores similarity features of the kind, high overhead classification process, time consuming association process. This paper proposed a mining algorithm of database constraints based on characteristics fuzzy closer, which describes the consistency between data through the closer between data fuzzy sets, introduces data fusion techniques into the traditional mining technology of neural network, after classifying and processing the data, analyzes the dynamic characteristics of the original mining data and gets new mining model, in order to accurately query target data in the large-scale database. The simulation experimental results show that the efficiency for the algorithm to mine the sparse data sets and dense data sets is superior to the traditional association rules algorithm, and it greatly improves the efficiency of database mining.

Keywords Fuzzy closer, Data mining, Neural network

1 引言

随着科学技术的进步,网络中的信息量呈现大幅度增加趋势,作为信息载体的互联网的规模也逐渐增大。当前企业信息化程度不断发展,企业管理网络中的应用系统和数据对于总体网络具有重要的应用价值,通过不同类型的应用软件能够对相关的数据执行有效的处理,进而对相应的数据进行查询、存储和传输等操作,应用软件增强了企业的生产和管理效率,确保企业具有较强的竞争力。如今的互联网广泛应用在不同的行业中,导致 Web 站点数量以及网站范围逐渐增大,同时不同 Web 站点的网络数据库信息量以及网络拓扑结构的复杂度越来越大,因而寻求有效的方法从数据库中挖掘相关的信息成为相关学者研究的热点问题。

文献[1]的算法在计算支持数和产生候选频繁项目集时,均通过二进制的布尔运算来完成,虽然能够快速方便地产生候选频繁项目集和计算其支持数,提高挖掘算法效率,但它们忽略了冗余问题。文献[2]虽然在关联规则上取得了一定的进步,但是在计算支持数时会多次重复扫描数据库,同时冗余候选频繁项目集也大量存在,制约着算法效率的提高。文献[3]中对数据的分类过程过于繁琐,造成算法耗时。传统的关联规则算法中,只考虑了类内的关联性,忽略了类间的相似性特征,造成分类过程开销高、关联过程耗时[4,5]。

为了解决传统方法存在的问题,本文提出了基于特征模糊贴近的数据库约束挖掘算法,实现了在大规模数据库中,高效查询相应的目标数据。

2 传统挖掘方法的原理及其弊端分析

2.1 数据预处理

在挖掘过程中,首先需要对数据进行预处理。以 Web 信息为例,Web 服务器上记录了大量详细的用户信息,并不是所有的信息都能与用户信息形成直接的关联性连接,一些干扰信息的存在会对信息中的有效信息形成干扰,尤其是大量的冗余信息的存在会对数据分类过程中的数据中心聚类形成干扰,因此,需要对其中一些干扰性信息进行"清洗"。这么做

到稿日期:2013-02-05 返修日期:2013-05-08 本文受国家自然科学基金(61105071)资助。
王 勇(1975—),男,硕士,讲师,主要研究方向为计算机软件与理论,E-mail:jkd_wangyong@126.com;邹盛荣(1964—),男,教授,博士生导师,主要研究方向为计算机软件与理论。

的目的是最大程度地摆脱相似性的干扰，去除与文章无关的内容，保证后期的数据挖掘过程高效。

2.2 数据特征的模糊距离表述

一般的挖掘过程需要对数据的特征进行距离特征转换描述，详细的步骤为：假设 $e(x,y)$ 表示信息的特征，$e(x,y)$ 同时含有已完成聚类的信息。通过信息的特征值 $e(x,y)$，能够得到信息的特征值以及挖掘路线。

采用特征距离描述信息的差异特征能够准确描述信息的相关特征。设信息特征距离为 $k(x,y)$，如果 x 以及 y 具有相同的含义，则 $k(x,y)$ 中的变量会随着特征的变化而变化。要求 x 以及 y 不能同时为变量，则 $k(x,y)$ 会形成一定的模糊化描述，通过以上分析可得信息的最优距离，如式(1)所示：

$$k(x,y)=[\sum_{i=1}^{p}|x_i-y_i|^l]^{1/l} \tag{1}$$

如果式(1)获取的最优距离位于相似模糊范围内，则获取的距离差会产生反向特征值。此时，应通过技术绝对值方法，获取相应的距离大小：

$$k(x,y)=\sum_{i=1}^{p}|x_i-y_i| \tag{2}$$

可将式(2)变换为标准的欧式几何距离，如式(3)所示：

$$k(x,y)=[\sum_{i=1}^{p}|x_i-y_i|^2]^{1/2} \tag{3}$$

通过特征距离描述能够获取高相像度的特征值。

通过以上方法分析可以看出，传统的挖掘方法中，需要将数据特征转换成相应的数学距离，虽然在转换过程中加入了一定的模糊规则，一定程度上解决了数据相似性的问题，但是这种方法没有跳出基于数据差异特征进行划分的传统理论，一定特征过于接近，这种模糊距离也不能很好地表达这种差异性，造成数据挖掘效果不好。

3 基于特征模糊贴近的数据库约束挖掘算法

3.1 模糊贴近分类技术的引入

数据库中的数据在首次分类后，类之间具有某种相似性，这种相似性可以运用数据类特征间的模糊贴近分析数据类间的可分类程度，运用这种可分类程度，提出一种基于特征模糊贴近的数据库约束挖掘算法[6-9]。挖掘数据库中的数据时，充分分析类间分类相关的影响因素，通过分析数据库中数据特征的模糊贴近度来描述数据间的一致度，从而构建数据挖掘模型。

算法能够将一定关联的类间群数据相应的关联方式进行连接、分类，最终对数据进行相关的挖掘。该方法具有较强的自学习性能，并能适应复杂的挖掘环境。数据类间特征的倾斜一致性的计算方法如下：t 时刻，第 i 个类间挖掘的数据用 $x_i(t)$，$i=1,2,\cdots,n$ 表示，若 $x_i(t)$、$x_j(t)$ 具有较大的差异，则该不同分类采集的数据具有较低的统一性，高度统一的数据能够确保数据评估模型具有较高的精度，通过数据模糊集间的贴近度表示数据间的一致度，计算方法如下：

$$s_{ij}(t)=e^{(-\lambda(x_i(t)-x_j(t)))^2} \tag{4}$$

t 时刻的倾斜度的数据统一性矩阵是：

$$S_{(t)}=\begin{bmatrix}1 & s_{12}(t) & \cdots & s_{1n}(t)\\ s_{21}(t) & 1 & \cdots & s_{2n}(t)\\ \cdots & \cdots & \cdots & \cdots\\ s_{n1}(t) & s_{n2}(t) & \cdots & 1\end{bmatrix} \tag{5}$$

倾斜度的可靠矩阵具有空间和时间两个维度的信息可靠性，并且在 t 时刻挖掘相同数据的倾斜度表达式是：

$$p_i(t)=\frac{1}{n}\sum_{i=1}^{n}s_{ij}(t) \tag{6}$$

依据式(6)能够得到可靠信息，采用具有最高统一性的数据构建数据评估模型，进而对无效数据进行分析。为了加强这种类间倾斜度的分类能力，通常采用 BP 神经网络对倾斜度计算的方法进行合理的分析，在神经网络中引入最小二乘法，以确保神经网络实际输出的真实倾斜度的值同期望输出值的误差均方值的最小化，保证倾斜度计算的准确性。

设置数据库中的类间数据样本是：(X_K,Y_K)，其中 $K=1,2,\cdots,m$，X_K 表示输入的样本，且有 $X_K^T=(x_{1k},x_{2k},\cdots,x_{nk})$，$n$ 用于描述数据样本的维度，数据库的神经网络模型用图 1 描述。

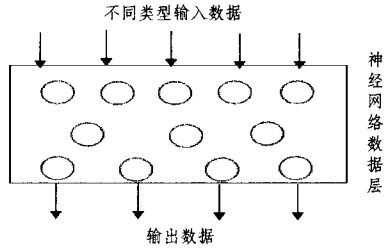

图 1 数据库神经网络模型

若数据库神经网络的隐层数为 L，则引入倾斜度计算后，第 i 个输出为：

$$\begin{aligned}&\hat{y}_{ik}=\sigma_0(\bar{y}_{ik})\\&\bar{y}_{ik}=W_{ik}^{(o)T}\hat{H}_k^{(l)}=\sum_{j=1}^{nl}w_{ij}^{(o)}h_{jk}^{(l)}\end{aligned} \tag{7}$$

采用式(8)所示的倾斜度标准对倾斜度权值矩阵第 P 行进行微分变换：

$$\begin{aligned}\frac{\partial E_K}{\partial W_{PK}^{(o)}}&=\frac{\partial E_K}{\partial e_{pk}}\cdot\frac{\partial e_{pk}}{\partial \hat{y}_{pk}}\cdot\frac{\partial \hat{y}_{pk}}{\partial y_{pk}}\cdot\frac{\partial \overline{y_{pk}}}{\partial \hat{y}_{pk}^{(o)}}\\&=-e_{pk}\cdot\sigma_0'(\bar{y}_{pk})\cdot\hat{H}_{pk}^{(l)}\end{aligned} \tag{8}$$

$$\Delta W_{PK}^{(o)}=W_{PK}^{(o)}-W_{pk-1}^{(o)}=-\alpha\frac{\partial E_k}{\partial W_{pk}^{(o)}} \tag{9}$$

对 BP 网络从输出层向输入层进行反向递推，能够获取第 r 层的倾斜度权值变换方式是：

$$\begin{aligned}&\Delta W_{pk}^{(r)}=W_{pk}^{(r)}-W_{pk-1}^{(r)}=\alpha\cdot\varepsilon_{pk}^{(r)}\cdot\hat{H}_{pk}^{(r+1)}\\&\varepsilon_{pk}^{(r)}=\sigma_r'(\bar{h}_{pk}^{(r)})\cdot\sum_{i=1}^{n_r-1}\varepsilon_{ik}^{(r-1)}w_{ip}^{(r-1)}\end{aligned} \tag{10}$$

通过上述分析可知神经网络输出数据的局部误差同输出误差以及该层偏导数具有较强的关联性，运算时局部误差的干扰主要是从后向前进行。可按照数据的种类数量选择合理的网络层数，最终实现对数据库数据的类间数据倾斜度的分类，这种分类过程可以有效地解决数据之间的相似性干扰问题，为高效、准确地挖掘后期数据打下良好的基础。

3.2 数据库的数据约束挖掘过程

大规模数据库具有多样性和随机性特征，在引入倾斜度分类后，还需要对其进行约束，使得挖掘数据的过程更为高效。对数据的动态特征进行分析，可以用数据中的特征均值描述数据库的动态特征，计算数据库中各数据的跳变率的线性关联性，用式(11)描述：

$$W_{avg}=\lim_{T\to+\infty}\frac{\sum_{t=1}^{T}\sum_{k=1}^{\mu}\{(h_k[(t-1)]\oplus h_k[t])\cdot H(k)\}}{2T}$$
$$=\lim_{T\to+\infty}\frac{\sum_{g=1}^{\zeta}\{\sum_{t=1}^{T}\{(f_g[(t-1)]\oplus f_g[t])\cdot F(g)\}\}}{2T}$$
$$=\sum_{k=1}^{\mu}\{\lim_{T\to+\infty}\frac{\sum_{t=1}^{T}[(h_k[(t-1)]\oplus h_k[t])]}{2T}\cdot H(k)\}$$
$$=\sum_{k=1}^{\mu}[\frac{q_k(T)}{2T}\cdot H(k)]=\sum_{k=1}^{\mu}[B_k\cdot H(k)] \quad (11)$$

式中，$q_k(T)$用于描述数据 k 在$(T,-T)$时期的类间特征变数，$h_k[t]$表示数据 k 在 t 时刻的情况。采集数据库中各数据的活度，能够获取数据库的平均动态特征值，表达式为：

$$Drip_{avg}=\lim_{U\to+\infty}\frac{\sum_{i=1}^{U}Drip(Z_i)}{U}$$
$$=\lim_{U\to+\infty}\frac{\sum_{i=1}^{U}\sum_{j=1}^{Q}I(T_j,Z_{j,i},R_j)}{U}$$
$$=\sum_{i=1}^{Q}\lim_{U\to+\infty}\frac{\sum_{i=1}^{U}I(T_i,Z_{i,j},R_i)}{U}$$
$$=\sum_{i=1}^{N}\lim_{L\to+\infty}\frac{\sum_{i=1}^{2^{R_i}}n_{i,j}I(T_i,S_j,M_i)}{L}$$
$$=\sum_{i=1}^{N}\sum_{j=1}^{2^{M_i}}\{\lim_{L\to+\infty}\frac{n_{i,j}}{L}I(T_i,S_j,M_i)\}$$
$$=\sum_{i=1}^{N}\sum_{j=1}^{2^{M_i}}\{prob_{i,j}I(T_i,S_j,M_i)\} \quad (12)$$

若数据库中的不同 z_i 相互独立，则可通过数据库的输入端的约束模型，反映输入数据向量的特征，约束用式(13)描述：

$$E=E_{avg}+\sum_{i=1}^{I_{num}}\Delta\beta(x_{x_i})W_{\beta(x_i)} \quad (13)$$

式中，E_{avg}表示数据库的平均特征，通过上述分析的度量式可得到$W_{\beta(x_i)}$，$\Delta\beta(x_i)$用于描述数据库采集的平均特征 $\beta(x_i)$和输入端的数据活度差值。

最终通过数据库的挖掘模型，能获取数据库的平均动态特征和静态特征约束。如果数据库中的数据特征大量存在于控制位置，则可得到相应数据的挖掘模型，进而塑造新的数据特征，最终能够实现在大规模数据库中，准确查询相应的目标数据。

4 实验结果及其分析

为了验证本文分析算法的有效性，应进行相关的实验。实验将传统的关联规则挖掘算法与本文提出的方法进行对比，实验环境如表 1 所列。

表 1 实验环境

实验环境	
CPU	InterP42.4GHz
内存	DDR1.0GB
操作系统	Windows XP
开发环境	Matlab 7.1
数据库	Oracle 10g

为了验证本文算法对于不同特征数据库的挖掘性能，实验采用两个数据集，分别是稀疏的数据集与相对密集的数据集，两种数据集可通过相关资料获取，并且两种数据集能够检测不同算法的数据挖掘性能，传统的关联规则算法以及本文算法对不同特征数据集的挖掘效果用图 2—图 4 描述。

实验数据先采用 Gazelle，相应的数据集来自于 Bule Martini 数据库公司，其是一个稀疏的数集，在较小支持度条件下不同算法能够挖掘出相应的长模式数据，该数据集存在 28258 个序列、86435 个事件以及 1312 个项。传统的关联规则算法、模糊距离关联挖掘算法以及本文算法对于该稀疏数据集的挖掘结果用图 2 描述。

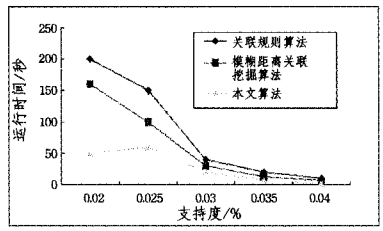

图 2 3 种算法挖掘稀疏数据集运行时间对比结果

分析图 2 可知，采用本文算法挖掘稀疏数据集时，在不同支持度下的运行时间始终低于传统的关联规则算法。如果支持度达到 0.03%左右，关联规则算法的挖掘效率大幅度降低，当支持度趋于 0.025%时关联规则算法无法终止相关的挖掘，模糊关联挖掘的效果比传统的关联挖掘有了较为明显的提高，但是与本文提出的方法相比，效果也明显变差，而本文算法可以挖掘出支持度更低的序列。

为了进一步验证本文算法的性能，实验采用了生物数据集 Snake，并分别采用关联规则算法、模糊关联挖掘算法以及本文算法对其进行挖掘。数据集 Snake 是一种密集数据集，包括 164 个 Toxin—Snake 蛋白质序列，共计 20 个项，平均序列长度是 56，最大序列长度是 110。3 种算法挖掘该密集数据集的性能对比效果用图 3 描述。

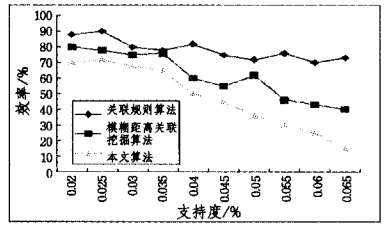

图 3 3 种算法挖掘密集数据集的挖掘效率对比结果

分析图 3 可知，随着支持度的增加，本文算法的挖掘效率逐渐高于关联规则算法，并且关联规则算法的效率呈现大幅度下降的趋势，主要因为传统的关联规则算法在挖掘高密度的数据集时，挖掘出的规则集数量巨大，对 CPU 时间的开销高、I/O 使用率大，并且容易出现局部最优解问题，最终导致算法的挖掘效率降低，模糊关联规则随着支持度的增加，算法的收敛性也呈现了明显的下降，这是因为随着数据相似性的增加，模糊距离也呈现了一致性的趋势，无法准确地分割数据。

对 3 种不同的算法进行伸缩度的比较，结果如图 4 所示。

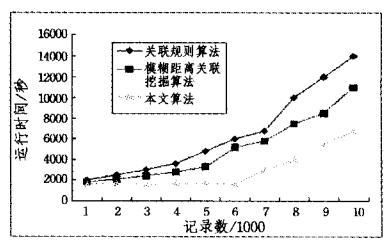

图 4 3 种算法伸缩性比较

(下转第 227 页)

从原子服务顺序调整前后的表示情况看，处于同层的复合服务，其向量表示与调整前相比，“1”的排列情况显示出了较好的聚合特性。在特定的服务领域，如果用户所发布的服务具有较好的聚类特性，则从每个服务向量中“1”的分布情况，即可粗略地估算出服务间功能的相似程度，并初步实现服务基于功能的聚类。本文由于表述问题，采用的本体比较简单，在本体中所包含的原子服务更多的情况下，这种聚合特性将表现得更加明显。

结束语 随着 Web 服务数量的急剧增加和应用的日益普及，如何快速、准确和高效地发现满足用户需求的服务已经成为制约 Web 服务发展的瓶颈之一。从用户查找服务的实际操作过程出发，采用显式地描述服务功能语义、有效地滤除与请求功能无关的 Web 服务是简化服务检索计算量、提高服务发现准确率的有效方法之一。本文针对现有常用 Web 服务描述语言（如 OWL-S），不能显式地描述服务功能语义的问题，通过定义服务功能描述模型，构建领域功能本体，提出了一种基于功能语义的 Web 服务描述方法，并对 OWL-S 语言进行扩充，在此基础之上，为简化服务滤除操作，设计了一种服务映射方法及原子服务排序算法，使生成的服务向量能更好地反映原子服务间的相关特性，最后给出了一种服务预检索方法。分析表明，该方法可有效地支持用户实现基于功能语义服务检索，滤除无关服务，降低精确匹配计算量，提高服务查找效率。同时由于本描述方法是通过扩充 OWL-S 语言实现的，可以更好地继承现有基于 OWL-S 语义服务描述语言获取的研究成果。

参 考 文 献

[1] D'Mello D A, Ananthanarayana V S. A Review of Dynamic Web Service Description and Discovery Techniques[C]// 2010 First International Conference on Integrated Intelligent Computing. 2010:246-251

[2] D'Mello D A, Ananthanarayana V S. A Tree Structure for Efficient Web Service Discovery[C]// Second International Conference on Emerging Trends in Engineering and Technology. 2009:826-831

[3] Paliwal A V, Shafiq B, Vaidya J, et al. Semantics-Based Automated Service Discovery [J]. Services Computing, IEEE Transactions on, 2012, 5(2):260-275

[4] 于晓浩，罗雪山，胡丹，等. 基于行为约束的军事信息服务描述及过滤方法[J]. 国防科技大学学报，2011，33(4)：157-162

[5] 万长林，史忠植，胡宏，等. 基于本体的语义 Web 服务 QoS 描述和发现[J]. 计算机研究与发展，2011，48(6)：1059-1066

[6] Jean S, Losaviot F, Matteot A, et al. An Extension of OWL-S with Quality Standards[C]// Fourth International Conference on Research Challenges in Information Science. IEEE, 2010:483-494

[7] Yousefipour A, Ne A G, Mohsenzadeh M, et al. An Ontology-based Approach for Ranking Suggested Semantic Web Services [C]// Sixth International Conference on Advanced Information Management and Service. 2010:17-23

[8] D'Mello D A, Ananthanarayana V S. Effective Web Service Discovery Based on Functional Semantics[C]// International Conference on Advances in Computing, Control, and Telecommunication Technologies. 2009:1-3

[9] Ye Lei, Zhang Bin. Web Service Discovery Based on Functional Semantics[C]// Second International Conference on Semantics, Knowledge, and Grid. 2006:57-60

[10] Amorim R, Claro D B, Lopes D, et al. Improving Web service discovery by a functional and structural approach[C]// International Conference on Web Services. 2011:411-418

（上接第 210 页）

图 4 描述了最小支持度为 0.4 时，3 种算法的运行时间对比结果。分析时间结果能够得出，同关联规则算法、模糊关联规则算法对比，随着数据库规模的不断增加，本文算法的运行时间越来越小于关联规则算法，具有较高的运行效率，更加适合于大型数据库的挖掘。

结束语 本文提出了基于特征模糊贴近的数据库约束挖掘算法，其通过数据模糊集间的贴近度描述数据间的一致度，采用神经网络数据融合技术对数据进行一致性处理后获取初始挖掘数据，对原始挖掘数据的动态特征进行分析获取新的评估模型，以实现在大规模数据库中，准确查询相应的目标数据。仿真实验结果表明，本文提出的算法挖掘稀疏数据集和密集数据集的效率都优于传统的关联规则算法，极大提高了数据库的挖掘效率，具有重要的应用价值。

参 考 文 献

[1] 何娟，涂中英，牛玉刚. 一种遗传蚁群算法的机器人路径规划方法[J]. 计算机仿真，2010，27(3)：170-175

[2] Abiteboul S, Bnueman P, Suciu D. Data on the Web[Z]. Morgan Kaufmann Publishers. 2010:89-93

[3] 丁艳辉，王洪国，高明，等. 一种基于矩阵的关联规则挖掘新算法[J]. 计算机科学，2006，33(4)：188-189

[4] 郑泉，王建东. 基于 FP 树挖掘大数据库的方法及算法 PCM[J]. 计算机工程与应用，2009，40(7)：182-184

[5] Kitts B, Freed D, Vrieze M. Cross-sell: A Fast Promotion Tunable Customer-Item Recommendation Method Based on Conditionally Independent Probabilities[C]// Proceedings of KDD. Based on ACM Press, 2010:437-446

[6] 胡刚，刘建鑫，郭炜. 基于 PIC18F4550 的环境监测数据采集系统设计[J]. 科技通报，2012，10(28)：17-18

[7] Gauch S, Chaffee J. Ontology-based personalized search and browsing[J]. Web Intelligence and Agent Systems, 2011, 1(3): 77-79

[8] 赵国锋，李兵，徐川，等. 移动社交网的生命周期评估模型研究[J]. 计算机学报，2013(4)：727-737

[9] 赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报，2012，21(8)：1834-1848