数据挖掘中关联弱化问题的解决方法分析

杨泽民 郭显娥 王文军
(山西大同大学数学与计算机科学学院 大同 037009)

摘 要 当前的支持向量机和均值聚类等数据挖掘算法中,几乎都是依靠数据之间的关联性来完成数据匹配。一旦数据库中含有大量的冗余数据,将造成数据之间的相关性降低,关联性被破坏,导致传统的数据挖掘算法效率降低。为了避免上述缺陷,提出了一种弱化关联规则修补挖掘算法。利用弱聚类方法,在数据选择过程中,不将所有的元素都进行初始分类处理,只计算某一元素属于某一个类别的概率,确定多个弱聚类中心,计算不同数据之间的弱聚类关联性,从而实现关联规则较弱的冗余环境下准确的数据挖掘。实验结果表明,这种算法能够有效提高海量冗余环境下的数据挖掘效率,取得了令人满意的效果。

关键词 海量冗余,数据挖掘,关联规则
中图法分类号 F127 **文献标识码** A

Research on Solution to Association Weakening Problem in Data Mining

YANG Ze-min GUO Xian-e WANG Wen-jun
(School of Mathematics and Computer Science, Shanxi Datong University, Datong 037009, China)

Abstract The support vector machine (SVM) and mean cluster data mining algorithm, almost all rely on the correlation between data, complete data matching. Once the database contains a large amount of redundancy data, the correlation between data will be reduced, and relevance is destroyed, resulting in traditional data mining algorithm efficiency lower. In order to avoid the above defects, this paper proposed a weakening association rules repair mining algorithm. In the data selection process, the method will not make initial classification processing for all elements only calculates probability that one element belongs to a category, and determines multiple weak clustering center, calculates weak clustering relevance between different data, so as to realize the association rules weaker redundancy environment accurate data mining. The experimental results show that this algorithm can effectively improve the massive redundant environment data mining efficiency, has made the satisfactory effect.

Keywords Mass redundancy, Data mining, Association rules

1 引言

随着计算机信息处理技术的快速发展,利用数据挖掘算法进行信息搜索已经成为一种主要的信息获取方式[1]。数据挖掘方法是计算机信息提取技术的核心内容。利用这种方法,能够从大量数据中提取出最有价值的信息,从而实现知识发现[2]。这种数据挖掘方法在人工智能领域、数据库领域和计算机决策领域都有着比较重要的价值,引起了诸多专家的广泛重视[3]。因此,数据挖掘算法已经成为信息领域研究的热点问题。现阶段,主要的数据挖掘算法包括支持向量机数据挖掘算法、K均值聚类数据挖掘算法和信息增益数据挖掘算法[4]。其中,最常用的是支持向量机数据挖掘算法。数据挖掘算法由于应用范围比较广泛,因此受到了诸多学者的重视,有很大的发展空间和应用前景[5]。

在数据挖掘过程中,数据库中含有大量的冗余数据,将造成数据之间的相关性降低,导致数据挖掘效率降低[6,7]。为了避免上述缺陷,提出了一种弱化关联规则挖掘算法。利用弱聚类方法,对数据库中的数据进行聚类处理,以保证在冗余环境下的高效数据挖掘。

2 弱化关联规则下的弱聚类方法

数据挖掘方法,是信息领域需要研究的核心问题。利用传统算法进行数据挖掘,无法避免由于数据库中的冗余数据较多造成的数据之间相关性较差的缺陷,从而降低了数据挖掘的效率。为此,提出了一种弱化关联规则下的弱聚类挖掘算法。

2.1 关联数据弱聚类处理

设置数据库中的元素构成的数据集合能够用$U=\{u_1, u_2, \cdots, u_p\}$进行描述,其中$u_k$是该数据集合中的第$k$个元素,上述数据的属性能用$J=\{j_1, j_2, \cdots, j_n\}$进行描述,$u_k[j_l]$是第$k$个元素在属性$j_l$上的取值。利用弱聚类方法能够将属性元素进行分类,将数量型元素变换为类别型。

到稿日期:2012-10-25 返修日期:2013-03-26 本文受国家自然科学基金(11171112)资助。
杨泽民(1974—),男,硕士,副教授,主要研究方向为数据挖掘(关联规则),E-mail:yzm1212@yahoo.com.cn;郭显娥(1964—),女,硕士,主要研究方向为数据挖掘;王文军(1977—),男,硕士,讲师,主要研究方向为数据挖掘。

设置样本空间是 $Y=\{y_1,y_2,\cdots,y_p\}$，与传统的聚类过程不同，利用弱聚类方法能够将其分为 d 个不同的类别，其中的任意元素 $y_j\in Y$。在聚类过程中，不再将所有的元素都进行准确的分类处理，因此需要计算某一元素 j 属于第 k 个类别的概率 x_{jk}。利用下述公式能够描述对样本空间进行弱分类的矩阵：

$$X=(x_{jk}) \tag{1}$$

其中，x_{jk} 是样本空间中第 j 个元素属于第 k 个类别的概率，具备下述特性：

$$\begin{aligned}&x_{jk}\in[0,1](j=1,2,\cdots,p;k=1,2,\cdots,d)\\&\sum_{k=1}^{d}x_{jk}=1(j=1,2,\cdots,p)\\&0\leqslant\sum_{j=1}^{p}x_{jk}\leqslant p(k=1,2,\cdots,d)\end{aligned} \tag{2}$$

利用下述公式能够描述海量冗余环境下的数据分类目标：

$$K_n(X,A)=\sum_{j=1}^{p}\sum_{k=1}^{d}x_{jk}^{n}e_{jk}^{2}(y_j,a_k) \tag{3}$$

式中，$A=(a_1,a_2,\cdots,a_p)$，a_k 是第 k 个属性类别的聚类中心，x_{jk}^{n} 是与其对应的权值系数。利用下述公式能够计算样本到聚类中心之间的长度：

$$e_{jk}(y_j-a_k)=|y_k-a_k| \tag{4}$$

设置海量冗余环境中数据相关参数 $d,p,n,c=1$，弱聚类中心是 $A_{(c)}=(a_1,a_2,\cdots,a_d)$，利用下述公式，能够进行数据更新：

$$\begin{aligned}&x_{jk}=\frac{1}{\sum_{l=1}^{d}\left[\frac{e_{jk}}{e_{jl}}\right]^{\frac{2}{n-1}}}\text{当 }e_{jk}\neq1\\&x_{jk}=0\text{ 当 }d_{jk}=0,k\neq l\\&x_{jk}=1\text{ 当 }d_{jk}=0\end{aligned} \tag{5}$$

利用下述公式能够计算样本均值参数：

$$a_k=\frac{\sum_{i=1}^{p}x_{jk}}{\sum_{j=1}^{p}x_{kl}} \tag{6}$$

将 a_c 与 $a_{(c+1)}$ 的取值进行对比，如符合下述公式的要求，则完成聚类处理；如不符合下述公式的要求，则继续进行聚类分析：

$$|a_c-a_{(c+1)}|\leqslant\phi \tag{7}$$

在上述迭代处理过程中，目标函数的结果是不断减小的。利用这种方法进行聚类处理，能够避免聚类处理陷入局部极小的缺陷中，从而获取符合要求的聚类结果。根据上面阐述的方法，能够将数据库中的全部数据进行聚类处理，从而为海量冗余环境下的弱化关联规则挖掘提供准确的数据基础。

2.2 弱化规则下的挖掘过程

利用弱化关联规则方法，能够对聚类处理后的数据进行挖掘，其步骤如下所述：

设置需要进行挖掘的数据集合 Y，其中元素数目是 p，数据属性数目是 n，模糊数据属性构成的集合是 $C=\{c_1,c_2,\cdots,c_q\}$，利用下述公式能够计算不同属性数据的支持系数：

$$G(C)=\sum_{k=1}^{p}\prod\upsilon(c_j,y_k) \tag{8}$$

式中，p 是数据库中全部数据的数量，υ 是与其对应的隶属度，$\upsilon(c_j,y_k)$ 是第 k 个元素在 c_j 上的隶属度。

利用下述公式能够计算任意元素的置信度：

$$G'=\frac{\sum_{i=1}^{q}[\prod\upsilon(e_j,y_k)]}{p} \tag{9}$$

利用下述公式能够计算数据权值系数：

$$\upsilon_{jk}=\frac{e(d_j,y_k)}{\sum_{k=1}^{M}e(d_j,y_j)} \tag{10}$$

利用下述公式能够计算数据相关性系数：

$$x_{jk}=\frac{\exp(\frac{-\alpha}{\eta})}{\sum_{j=1}^{n}\exp(\frac{-\alpha}{\eta})} \tag{11}$$

利用下述公式能够计算不同数据之间的关联性系数：

$$d_j=\frac{\sum_{k=1}^{p}\upsilon_j y_{kl}}{\sum_{k=1}^{p}\upsilon_j} \tag{12}$$

设置数据关联性阈值是 χ，利用下述公式能够实现数据挖掘：

$$\begin{cases}d_j>\chi, & \text{该数据是需要获取的信息}\\d_j\leqslant\chi, & \text{该数据不是需要获取的信息}\end{cases} \tag{13}$$

根据上面阐述的方法，能够利用模糊聚类方法，对数据库中的数据进行聚类处理，为数据挖掘提供准确的数据基础。利用弱化关联规则方法，计算不同数据之间的关联性，从而实现海量冗余环境下的数据挖掘。

3 实验结果分析

为了验证本文算法的有效性，需要进行一次实验，实验环境是 Visual C++6.0[8]。设置数据库中包含的数据数目是 N，全部数据种类数目是 n，全部数据构成的数据集合是 $\{a_1,a_2,\cdots,a_N\}$，全部数据属性构成的数据集合是 $\{b_1,b_2,\cdots,b_n\}$，数据 a_i 属于种类 b_j 的概率是 η。

利用下述公式能够计算数据挖掘的效率：

$$\epsilon=\frac{\sqrt{a_i-a_{i-1}^{2}}}{b_j-1} \tag{14}$$

根据数据挖掘的效率，能够准确衡量数据挖掘方法的性能。设置数据库中全部数据的数量是 1000。数据属性种类数目是 29。从上述数据中随机选取 15 个不同属性的数据，则选取数据的详细情况如表 1 所列。

表 1 不同属性数据表

数据属性序号	数据属性	数据量
1	长度数据	13
2	空间数据	7
3	医疗数据	5
4	公路数据	15
5	交通数据	10
6	教学数据	14
7	养老数据	32
8	工资数据	21
9	企业数据	6
10	图书数据	15
11	计算机数据	31
12	网络数据	29
13	消费数据	17
14	房产数据	29
15	人口数据	31

将表 1 中前 10 项数据的分布情况进行整理，能够得到图 1。

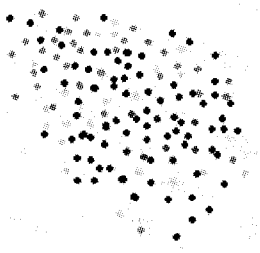

图 1　不同属性数据分布图

在图 1 中，每一种颜色代表一个数据属性。

在数据库中含有少量冗余数据的情况下，分别利用不同方法进行数据挖掘，得到的数据挖掘结果用图 2 进行描述。

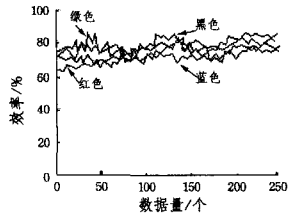

图 2　冗余数据较少时的数据挖掘结果

其中，黑色曲线是利用本文算法进行数据挖掘的结果，红色曲线是利用支持向量机挖掘算法进行数据挖掘的结果，蓝色曲线是利用 K 均值聚类挖掘算法进行数据挖掘的结果，绿色曲线是利用信息增益挖掘算法进行数据挖掘的结果。根据上图能够得知，在冗余数据较少的情况下，利用本文算法进行数据挖掘的效率与传统算法基本一致。

在数据库中含有海量冗余数据的情况下，分别利用不同方法进行数据挖掘，得到的结果用图 3 进行描述。

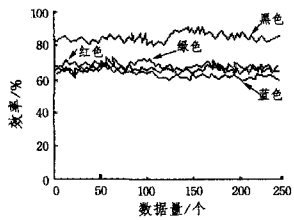

图 3　冗余数据较多时的数据挖掘结果

表 2　冗余数据较少时不同属性数据表

序号	支持向量机算法挖掘效率（%）	K 均值聚类算法挖掘效率（%）	信息增益算法挖掘效率（%）	弱化关联规则算法挖掘效率（%）
1	0.83	0.81	0.81	0.84
2	0.79	0.81	0.78	0.79
3	0.84	0.79	0.82	0.85
4	0.74	0.73	0.76	0.75
5	0.85	0.83	0.83	0.82
6	0.82	0.81	0.81	0.83
7	0.72	0.73	0.72	0.72
8	0.84	0.81	0.84	0.83
9	0.81	0.82	0.84	0.82
10	0.81	0.83	0.85	0.79

根据图 3 能够得知，在冗余数据较多的情况下，利用传统算法进行数据挖掘的效率较低，利用本文算法进行数据挖掘的效率是最高的，比支持向量机挖掘算法的数据挖掘效率高 11%，比 K 均值聚类挖掘算法的数据挖掘效率高 13%，比信息增益挖掘算法的数据挖掘效率高 14%，这充分展示了本文算法在冗余数据较多的情况下进行数据挖掘的优越性。

在冗余数据较少的情况下，利用不同算法进行 10 次数据挖掘，将实验结果进行整理分析，能够得到如表 2 所列的结果。

在冗余数据较多的情况下，利用不同算法进行 10 次数据挖掘，将实验结果进行整理分析，能够得到如表 3 所列的结果。

表 3　冗余数据较多时不同属性数据表

序号	支持向量机算法挖掘效率（%）	K 均值聚类算法挖掘效率（%）	信息增益算法挖掘效率（%）	弱化关联规则算法挖掘效率（%）
1	0.81	0.75	0.76	0.69
2	0.83	0.73	0.75	0.75
3	0.82	0.74	0.71	0.71
4	0.81	0.75	0.78	0.73
5	0.84	0.74	0.75	0.71
6	0.82	0.73	0.71	0.72
7	0.85	0.72	0.69	0.68
8	0.83	0.71	0.74	0.77
9	0.82	0.73	0.68	0.78
10	0.81	0.72	0.71	0.74

通过上述实验能够得知，利用本文算法进行数据挖掘，能够避免由于数据库中的冗余数据较多造成的数据之间相关性较差的缺陷，从而提高了数据挖掘的效率。

结束语　本文提出了一种弱化关联规则挖掘算法。利用模糊聚类方法，对数据库中的数据进行聚类处理，为数据挖掘提供准确的数据基础。利用弱化关联规则方法，计算不同数据之间的关联性，从而实现数据挖掘。实验结果表明，这种算法能够有效提高海量冗余环境下数据挖掘的效率，取得了令人满意的效果。

参 考 文 献

[1] 崔建，李强，杨龙坡. 基于垂直数据分布的大型稠密数据库快速关联规则挖掘算法[J]. 计算机科学，2011(4)：216-219

[2] Tojanovic Z, Dahanayake A. Service-Oriented Software System Engineering: Challenges and Practices [J]. Idea Group Publishing, 2011: 1-47

[3] Tasi T, Zhang D, Chen Y, et al. A software reliability model for Web services [C] // 8th IASTED International Conference on Software Engineering and Applications. Cambridge, MA, USA, 2011: 144-149

[4] 穆肇南，张健. 数据挖掘技术在经济预测中的应用[J]. 计算机仿真，2012(6)：347-350

[5] 王晟，赵壁芳. 基于模糊数据挖掘和遗传算法的网络入侵检测技术[J]. 计算机测量与控制，2012(3)：660-663

[6] Xu Yue, Li Yue-feng. Mining non-redundant association rules based on concise bases[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2007, 21(4): 659-675

[7] Loglisci C, Malerba D. Mining multiple level non-redundant association rules through two-fold pruning of redundancies[C] // Proceedings of MLDM. 2009: 251-265

[8] Cheng J, ke Y P, Ng W. Effective elimination of redundant association rules[J]. Data mining and knowledge discovery, 2008, 16(2): 221-249