

一种有效的多关系贝叶斯分类算法

邓左祥^{1,2,3}, 涂 芳⁴

(1 广西科技大学 计算机科学与通信工程学院, 广西 柳州 545006;

2 广西科技大学 智能计算与分布式信息处理重点实验室培育基地, 广西 柳州 545006;

3 上海交通大学 电子信息与电气工程学院, 上海 200240;

4 上汽通用五菱汽车股份有限公司 技术中心产品认证科, 广西 柳州 545007)

摘 要: 传统数据挖掘算法在处理多关系时, 需要物理连接, 因此存在效率不高的问题. 为了解决这一问题, 研究多关系数据挖掘中的分类问题, 提出一种有效的多关系贝叶斯分类算法 EMBC. EMBC 算法的目标是提高分类的准确率, 并且降低运行时间. EMBC 算法利用元组 ID 传播的思想, 结合朴素贝叶斯分类算法, 可以直接在多关系上进行分类. 实验表明, EMBC 算法提高分类的准确率, 并且显著降低运行时间.

关键词: 数据挖掘; 多关系; 分类; 元组 ID 传播; 贝叶斯

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2017)07-0123-05

Efficient Multi-relational Bayesian Classification Algorithm

DENG Zuo-xiang^{1, 2, 3}, TU Fang⁴

(1 School of Computer Science and Communication Engineering, Guangxi University of Science and Technology, Liuzhou 545006, China; 2 Key Laboratory of Intelligent Computing and Distributed Information Processing, Guangxi University of Science and Technology, Liuzhou 545006, China; 3 School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; 4 Technical Department Center Test & Validation Section, SAIC GM Wuling Automobile CO., Ltd., Liuzhou 545007, China)

Abstract: While dealing with multi-relation, traditional data mining algorithms used physical join, thus it had the problem of low efficiency. In order to solve this problem, the problem of classification in multi-relational data mining was investigated, and an efficient multi-relational Bayesian classification algorithm called EMBC was proposed. EMBC aims at increasing the accuracy of classification, and decreasing running time. By taking advantage of tuple ID propagation approach, and combined with naive Bayesian classification algorithm, EMBC can directly classify in multi-relation. Performance results demonstrate that, EMBC increases the accuracy of classification, and significantly decreases running time.

Key words: data mining; multi-relation; classification; tuple ID propagation; Bayesian

1 引言

多关系数据挖掘, 是数据挖掘领域热门的研究方向之一^[1]. 在现实生活中, 多关系数据挖掘在许多方面都有应用, 研究多关系数据挖掘, 具有积极的现实意义和潜在的应用价值.

当面对多关系时, 传统数据挖掘算法, 不得不首

先将多关系通过物理连接的方法, 合成到一张表中, 这使得传统数据挖掘算法在处理多关系时, 存在效率低、准确度低等缺点. 因此, 研究有效的多关系数据挖掘算法, 避免物理连接, 直接的多关系中进行挖掘, 是一个非常具有研究意义的问题.

分类是数据挖掘的主要任务之一, 它找出描述和区分数据类或概念的模式, 以便能够使用模型来

预测类标号未知的对象^[2]. 多关系分类是多关系数据挖掘的主要任务之一,它是在多关系数据库上进行分类. 多关系数据库包含多张表,每张表都描述着不同的实体或实体之间的联系. 用户可以选择某个实体作为多关系分类的研究对象. 用户选择的实体所对应的表,就是用户进行分类分析所要研究的表,称为目标关系. 在目标关系中,有一个属性是类标号属性. 目标关系的元组称为目标元组,每个目标元组都与一个类标号相关联(其中,有些目标元组的类标号是已知的,作为训练数据构造分类模型;另一些目标元组的类标号是未知的,需要对这些元组进行分类). 在多关系数据库中,除了目标关系之外的其余关系,称为非目标关系.

本文借助元组 ID 传播的思想^[3],结合贝叶斯分类,提出一种有效的多关系贝叶斯分类算法 EM-BC,可以避免物理连接,减少分类所需的时间,并且提高分类的准确率.

2 相关工作

多关系数据挖掘的主要任务包括多关系聚类、多关系关联规则、多关系分类^[4].

在多关系聚类算法中,比较著名的研究成果是 Crossclus^[5]. Crossclus 是利用元组 ID 传播思想的一种多关系聚类算法,它的聚类过程是:首先,用户指定目标关系中的某一个属性;接着,利用元组 ID 传播的思想,Crossclus 在多关系中搜索与用户指定属性相关联的属性,通过这些属性计算目标关系中对象之间的相似度;最后,根据相似度,Crossclus 对目标关系中的对象进行聚类. 多关系关联规则挖掘分为两步:第一步,在多关系中挖掘频繁模式;第二步,通过多关系频繁模式,产生多关系关联规则.

前人已经提出多种分类算法,比如决策树分类、贝叶斯分类、基于规则的分类等. 分类算法包含两个步骤:第一步,通过对训练数据进行分析,分类算法构造一个分类器. 第二步,利用构造出的分类器,对类标号未知的对象进行类标号预测. 在多关系分类中,主要采用基于规则的分类,以及贝叶斯分类^[6].

文献^[6]提出一种基于语义关系优化的多关系朴素贝叶斯分类算法 Graph-NB. Graph-NB 利用多关系之间存在的不同语义关系,对表进行剪裁,达到优化语义关系图、消除无关表对分类影响的目的,提高分类的效率. 然而,Graph-NB 存在不足,不足之处在于该算法对表进行剪裁的思想,很容易造成一些关键信息的丢失,从而影响分类准确率. 文献^[7]

利用元组 ID 传播思想,提出一种基于特征加权的多关系朴素贝叶斯分类算法,但是该算法考虑属性权值,而现实生活中多关系的属性不一定都有权值.

3 元组 ID 传播

元组 ID 传播(也称为元组标识传播)的定义如下.

定义 1 元组 ID 传播(tuple ID propagation): 在多关系数据库中,假设目标关系 R_1 的主码是整数型属性,这些整数代表每个目标元组的 ID(如果没有整数型属性的主码,就人工创建一个整数型属性的主码). 假设目标关系 R_1 和关系 R_2 可以按属性 $R_1.B$ 和 $R_2.B$ 连接(假设 R_2 的主码 $R_2.B$ 在 R_1 中作为外码, R_1 的主码用 $R_1.A$ 来表示). 那么,在 R_2 中增加一列属性 IDset. 对于 R_2 的每一个元组 t ,它在 IDset 属性上的值,为 R_1 中所有可以与 t 连接的元组在 $R_1.A$ 上的值的并集,用公式表示为 $IDset = \bigcup_{u \in R_1, u.B=t.B} R_1.A(u)$,其中 $R_1.A(u)$ 代表元组 u 在 $R_1.A$ 上的值.

元组 ID 传播的一个例子如表 1 所示. 在表 1 中,多关系数据库包含两张表,分别是 Loan 表和 Account 表. 在 Loan 表中,Loan_ID 是的主码,Account_ID 是外码. 在 Account 表中,Account_ID 是主码. Loan 表和 Account 表可以通过属性 Account_ID 来连接. 元组 ID 传播就是,在 Account 表中增加一列属性 IDset,Account 表中的一个元组在属性 IDset 上的值,就是 Loan 表中可以与这个元组连接的所有元组在属性 Loan_ID 上的值的并集. 也就是说,Loan 表中所有元组的 ID,都传播到 Account 表中可以与之连接的元组的属性 IDset 上.

表 1 元组 ID 传播的一个例子

Loan			
Loan_ID	Account_ID	amount	duration
1	124	1 000	12
2	124	4 000	12
3	108	10 000	24
4	45	12 000	36
5	45	2 000	24

Account			
Account_ID	frequency	date	IDset
124	monthly	960227	1, 2
108	weekly	950923	3
45	monthly	941209	4, 5

4 多关系贝叶斯分类算法 EMBC

本文提出一种有效的多关系贝叶斯分类算法

EMBC. 借助元组 ID 传播的思想,结合传统的朴素贝叶斯分类算法,EMBC 可以直接在多关系上进行分类.

4.1 朴素贝叶斯分类

朴素贝叶斯分类是在单一的关系中进行分类的一种算法,它的基本思路是:对于一个给定的对象 X ,朴素贝叶斯分类通过计算 X 属于一个特定类 c_j 的概率 $P(c_j | X)$,来对 X 进行分类.需要找到可以让 $P(c_j | X)$ 最大化的类 c_j ,将 X 划归为 c_j 所属的类.朴素贝叶斯分类利用统计学中的贝叶斯定理来计算 $P(c_j | X)$,计算公式如下:

$$P(c_j | X) = P(X | c_j)P(c_j)/P(X) \quad (1)$$

式中, $P(c_j | X)$ 为后验概率, $P(X | c_j)$ 为联合概率, $P(c_j)$ 为先验概率.对于所有的类来说, $P(X)$ 均为常数.因此,要最大化 $P(c_j | X)$,只需要最大化 $P(X | c_j)P(c_j)$ 即可.通过训练数据,对于每个类 c_j ,都可以计算出 $P(X | c_j)$ 和 $P(c_j)$ 的值,从而知道可以让 $P(X | c_j)P(c_j)$ 最大化的类 c_j .

朴素贝叶斯分类基于一个简单的假设:类条件独立,它的意思是给定元组的类标号,假定属性值相互独立,即在属性间,不存在依赖关系.朴素贝叶斯分类的详细工作过程在文献[4]给出.

4.2 算法 EMBC 描述

朴素贝叶斯分类在进行分类时,只能处理一张单一的表.在处理多关系时,朴素贝叶斯分类算法在分类之前,必须首先把多关系集成到一个单关系中.可是这种方法存在效率不高的问题,而且集成多张关系容易造成信息丢失和统计偏差,因此分类的准确率也不高.借助元组 ID 传播的思想,EMBC 对朴素贝叶斯分类算法进行拓展,不需要把多关系集成到一个单关系中,可以直接处理多关系分类,即直接对目标关系中类标号未知的每一个元组进行分类.

在多关系数据库中,存在关系之间的链接,以及元组之间的链接.如果关系 R_2 包含关系 R_1 的主码 k ,也就是 R_1 的主码 k 在 R_2 中作为外码,称关系 R_1 和 R_2 之间存在链接;如果 R_1 中的元组 t_1 与 R_2 中的元组 t_2 在属性 k 上具有相同的值,称元组 t_1 与 t_2 之间存在链接.关系之间的链接分为两种类型:直接链接、间接链接.如果关系 R_2 包含关系 R_1 的主码 k ,那么 R_1 与 R_2 之间存在直接链接;如果 R_1 与 R_2 之间不存在直接链接,但它们同时与某一个关系 R_3 存在直接链接,那么 R_1 与 R_2 之间存在间接链接.

多关系贝叶斯分类算法 EMBC 的工作过程如

下描述.

(1)根据分类的任务,在目标关系和各个与目标关系存在直接链接或间接链接的非目标关系中,人工选择与分类任务相关的若干属性,表示成“表.属性”的形式.假设选择出来的属性一共有 m 个,这 m 个属性形成一个属性集,用 $\{A_1, A_2, \dots, A_m\}$ 来表示.

(2)对那些与目标关系存在直接链接的非目标关系,利用元组 ID 传播的方法,将目标关系的每个元组的 ID(包括类标号已知的元组的 ID,以及类标号未知的元组的 ID)和类标号(类标号未知的不需要传播),从目标关系向那些与目标关系存在直接链接的非目标关系进行传播.传播类标号后,针对那些与目标关系存在直接链接的非目标关系中的每个元组,对同属于一个类的标号进行统计计数.按照这种方法,在多关系数据库中,每个与目标关系存在直接链接的非目标关系,都具有目标关系的元组 ID 和类标号.

(3)对那些与目标关系存在间接链接的非目标关系,利用元组 ID 传播的方法,将目标关系的每个元组的 ID 和类标号,从那些与目标关系存在直接链接的非目标关系(这些非目标关系已经具有目标关系的元组 ID 和类标号),传播到那些与目标关系存在间接链接的非目标关系上.按照这种方法,在多关系数据库中,每个与目标关系存在间接链接的非目标关系,都具有目标关系的元组 ID 和类标号.

(4)在目标关系中,对类标号未知的每一个目标元组 t ,获取 t 在前面第(1)步中选择出来的属性 $\{A_1, A_2, \dots, A_m\}$ 上的值.对于选择出来的某个属性,如果该属性在目标关系上,直接从目标关系就可以获取元组 t 在该属性上的值;如果该属性在非目标关系上,可以根据已经传播到这个非目标关系上的目标关系的元组 ID,找到包含元组 t 的 ID 的元组,从而获取 t 在该属性上的值.获取元组 t 在每个选择出来的属性的值后,将这些值表示成一个向量的形式 $X = \{x_1, x_2, \dots, x_m\}$.

(5)为了判断目标元组 t 的类标号,需要找到可以让 $P(c_j | X)$ 最大化的类 c_j .如上一小节所述,由于对于所有的类来说, $P(X)$ 均为常数.因此,要最大化 $P(c_j | X)$,只需要最大化 $P(X | c_j)P(c_j)$ 即可.对每个类 c_j ,需要分别计算出 $P(X | c_j)$ 和 $P(c_j)$ 的值.

(6)根据朴素贝叶斯的假设, $P(X | c_j) = P(x_1 | c_j) \times P(x_2 | c_j) \times \dots \times P(x_m | c_j)$.根据目标关系

上的类标号,以及传播到非目标关系上的类标号,可以分别计算出 $P(x_1 | c_j), P(x_2 | c_j), \dots, P(x_m | c_j)$ 的值,从而得到 $P(X | c_j)$ 的值. $P(c_j)$ 通过下式计算: $P(c_j) = |c_j, D| \div |D|$, 式中, $|c_j, D|$ 是目标关系中训练数据属于类 c_j 的元组个数, $|D|$ 是目标关系中训练数据的元组个数.

(7) 对类标号未知的目标元组 t , 将 t 赋予使 $P(X | c_j)P(c_j)$ 最大的类 c_j .

4.3 EMBC 算法的伪代码

本小节给出 EMBC 算法的伪代码,如下所示.
算法:EMBC,一种有效的多关系贝叶斯分类算法.
输入:一个包含目标关系 R 的多关系数据库 D , 目标关系 R 中包含类标号未知的目标元组.
输出:目标关系 R , 其中每个目标元组的类标号都已知.

方法:
(1) 在目标关系 R , 以及各个与 R 存在链接的非目标关系中, 选择出若干属性 $\{A_1, A_2, \dots, A_m\}$;
(2) 传播 R 的元组 ID 和类标号到非目标关系;
(3) for each R 中类标号未知的目标元组 t
(4) 向量 $\mathbf{X}\{x_1, x_2, \dots, x_m\} \leftarrow t$ 在属性 $\{A_1, A_2, \dots, A_m\}$ 上的值;
(5) for each 类 c_j
(6) 计算 $P(x_1 | c_j), P(x_2 | c_j), \dots, P(x_m | c_j)$, 得出 $P(X | c_j)$;
(7) 计算出 $P(c_j)$, 数组 $B \leftarrow P(X | c_j)P(c_j)$;
(8) 找出数组 B 中最大的值, 并把产生该值的类赋予目标元组 t ;
(9) return R ;

假设目标关系 R_1 , 一个非目标关系 R_2 , 需要对 R_1 中某个目标元组 t 的类标号进行预测. 假设 R_1 的 ID 已经传播到 R_2 上, R_2 的某个属性 A 被选择用来分类. 在 EMBC 算法伪代码的第(4)步, 有可能会遇上两种情况. 第一, 若在 R_2 中的所有元组的 IDset 中, 都找不到 t 的 ID, 那么就不存在 t 在属性 A 上的值. 在这种情况下, 忽略掉属性 A , 也就是说, 仅通过其他选择出的属性, 来对 t 的类标号进行预测. 第二, 在 R_2 中的多个元组的 IDset 中, 同时找到元组 t 的 ID, 那么就会出现 t 在属性 A 上具有多个值的情况. 在这种情况下, 为 t 在属性 A 上的每个值, 都计算一个概率, 然后取平均值, 作为 t 在属性 A 上的概率.

5 实验

5.1 实验方法与参数设置

在实验中, 将本文提出的多关系贝叶斯分类算法 EMBC, 与前人提出的 FOIL 和 CrossMine 进行

比较. 选择分类的准确率, 以及运行时间, 作为两个性能评价指标. 分类准确率的定义为, 分类算法对目标元组的分类结果, 与目标元组实际分类结果的比率; 运行时间的定义为, 分类算法对目标元组分类所花的时间.

使用 PKDD'99 Discovery Challenge 的金融数据库 (<http://lisp.vse.cz/pkdd99/Challenge/chall.htm>) 作为实验数据, 这个金融数据库的描述如表 2 所示.

表 2 金融数据库

关系名称	元组个数	属性个数
Loan	600	6
Account	4 500	4
Client	5 300	4
Disposition	5 300	4
Order	6 400	6
Card	800	4
District	500	16
Transaction	52 000	8

在实验中, 目标关系是 Loan, 将目标关系 Loan 的元组个数从 150 个变化到 600 个. 分类的任务是每笔贷款是否按时还清.

5.2 实验结果与分析

图 1 显示本文提出的 EMBC 算法和其他两种算法关于分类准确率的比较结果. 从图 1 可以看出, CrossMine 的分类准确率最高, EMBC 次之, FOIL 最差. 这是因为 CrossMine 和 EMBC 采用元组 ID 传播的方法, 元组 ID 传播是一种通过把目标元组 ID 附加到非目标关系的元组上、虚拟连接不同关系的方法, 可以避免物理连接, 这种方法只需要少量的数据传递、冗余计算、存储空间, 就可以将元组 ID 在不同关系之间进行传递, 因此这种方法灵活有效, 可以提高分类的准确率. FOIL 采用物理连接方法的代价非常昂贵, 因为物理连接多个关系, 容易造成信息丢失和统计偏差等问题, 所以 FOIL 分类的准确率较低.

图 2 显示本文提出的 EMBC 算法和其他两种算法关于运行时间的比较结果. 从图 2 可以看出, EMBC 的运行时间要优于 CrossMine, 这是因为与基于贝叶斯的 EMBC 相比较, 基于规则的 CrossMine 需要花费较多的时间来构造规则. 此外, 从图 2 还可以看出, FOIL 的运行时间非常长. 这是因为, 当遇上多关系数据库时, FOIL 需要物理连接多个关系, 当规则变长时, 连接的关系就会变广, 可能包含大量的元组, 其中有很多元组是冗余的, 在这

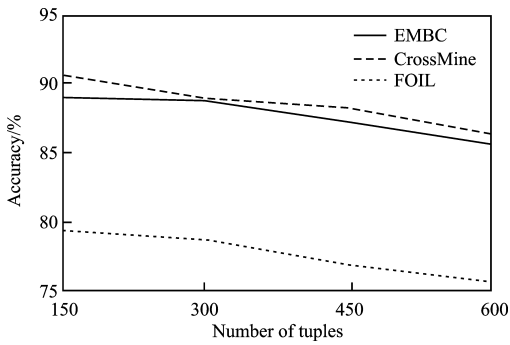


图 1 准确率比较结果

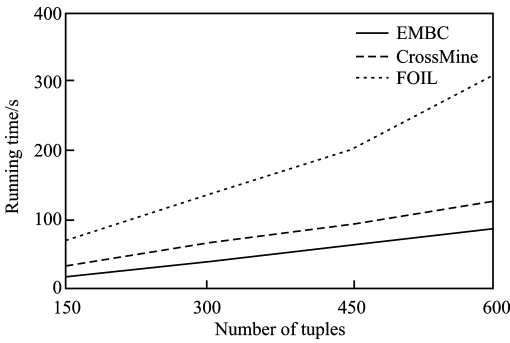


图 2 运行时间比较结果

样的关系上进行分类,相当的耗时,所需要的存储空间也很大。

当元组的个数为 300 时,让 EMBC 算法运行 100 次,对 EMBC 算法每次的运行时间进行比较。图 3 显示不同运行时间之间差异的累积分布函数。从图 3 可以看出,大部分运行时间之间的差异小于 1 s,并且所有运行时间之间的差异都小于 2.3 s。本实验可以说明,EMBC 算法的运行时间是稳定的。

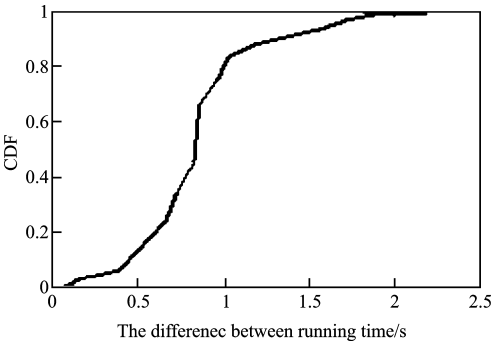


图 3 运行时间之间差异的累积分布函数

基于实验结果,可以做出以下两点总结:(1)虽然本文提出的 EMBC 算法在分类准确率上稍稍差于 CrossMine,但是 EMBC 算法在运行时间上要明显优于 CrossMine;(2)无论是分类准确率,还是运行时间,EMBC 算法都要远远优于 FOIL。综合这两

点总结可以得出,EMBC 算法实现较高的分类准确率,并且极大地减少分类所花的时间,是一种比较有效的多关系分类算法。

6 结束语

传统的朴素贝叶斯分类算法只适用于单关系,在处理多关系时,不得不首先把多关系集成到单关系中,存在分类准确率低和效率低的问题。为了解决这一问题,本文提出一种有效的多关系贝叶斯分类算法 EMBC。借助于元组 ID 传播的思想,结合朴素贝叶斯分类算法,EMBC 可以直接在多关系中对目标关系的目标元组进行分类。基于真实数据的实验表明,EMBC 能实现较高的分类准确率,并且极大地减少运行时间。

参考文献:

[1] 王曙霞,熊曾刚. 海量数据干扰下的危险 Web 数据挖掘技术研究[J]. 微电子学与计算机, 2016, 33(2): 87-91.

[2] 杨秀荣. 并行数据库中异常数据优化分类挖掘方法研究[J]. 微电子学与计算机, 2015, 32(10): 125-128.

[3] VAGHELA V, VANDRA K, MODI N. MR-MNBC: maxrel based feature selection for the multi-relational Naïve Bayesian Classifier[C]// Nirma University International Conference on Engineering. Ahmedabad: NUiCONE, 2013: 1-9.

[4] HAN J, KAMBER M. Data mining: concepts and techniques [M]. FAN M, MENG X, translated. 2nd ed. Beijing: China Machine Press, 2007: 200-206.

[5] YIN X, HAN J, YU P. Cross-relational clustering with user's guidance[C]// Proceedings of the 11th ACM SIGKDD Conference on Knowledge Discovery in Data Mining. Chicago: ACM SIGKDD, 2005: 344-353.

[6] 刘红岩,陈海亮. Graph-NB:一种高效准确的多关系朴素贝叶斯分类算法[J]. 信息系统学报, 2016, 2(1): 1-11.

[7] 徐光美,刘宏哲,张敬尊. 基于特征加权的多关系朴素贝叶斯分类模型[J]. 计算机科学, 2014, 41(10): 283-285.

作者简介:

邓左祥 男,(1983-),博士.研究方向为数据挖掘、自组织网络. E-mail:dengzuoxiang@126.com.

涂 芳 女,(1983-),工程师.研究方向为数据挖掘.