DOI:10.13374/j.issn1001-053x.2008.08.025
第30卷 第8期 北京科技大学学报 Vol.30 No.8
2008年 8月 Journal of University of Science and Technology Beijing Aug. 2008

基于互信息的多关系朴素贝叶斯分类器

徐光美 杨炳儒 秦奕青 张 伟

北京科技大学信息工程学院，北京 100083

摘 要 为进一步提高多关系朴素贝叶斯方法的分类准确率，分析了已有的剪枝方法，并扩展互信息标准到多关系情况下.基于元组号传播方法和面向元组的统计计数方法，给出了基于扩展互信息标准进行属性选择的方法和步骤，并建立了一种基于扩展互信息的多关系朴素贝叶斯分类器.标准数据集上的实验显示，基于扩展互信息标准进行属性选择，可以在不增加算法时间复杂度的前提下，找到与分类属性最相关的属性，并在仅有极少属性参与分类时，得到较高的分类准确率.Mutagenesis数据集上的实验则显示，这种属性选择可以使多关系问题退化为单关系问题，大大降低了分类代价.

关键词 朴素贝叶斯；分类器；多关系数据挖掘；归纳逻辑程序设计；互信息

分类号 TP 391

Multi-relational Naïve Bayesian classifier based on mutual information

XU Guangmei, YANG Bingru, QIN Yiqing, ZHANG Wei

School of Information Engineering, University of Science and Technology Beijing, Beijing, 100083

ABSTRACT To improve the accuracy of multi-relational Naïve Bayesian classifiers, the existing pruning methods were discussed and the attribute filter criterion was upgraded based on mutual information to deal with multi-relational data directly. On the basis of the tuple ID propagation method and counting methods towards tuple, the filter method based on extended mutual information was given, and a multi-relational Naïve Bayesian classifier based on mutual information (MI-MRNBC) was implemented. Experimental results show that, in a multi-relational domain, with the help of the attribute filter based on extended mutual information, the classifier can give a better accuracy without the increase of time complexity. In extraordinary instances, the multi-relational classification degenerates into a single relational one, which extremely decreases the cost of classification.

KEY WORDS Naïve bayesian; classifier; multi-relational data mining (MRDM); inductive logic programming (ILP); mutual information

随着多关系数据挖掘的兴起，各种各样的多关系朴素贝叶斯分类器(multi-relational Naïve Bayesian classifier, MRNBC)被提出，其中大多数方法基于归纳逻辑程序设计(inductive logic programming, ILP)技术形成.这类方法通常都通过用户定义的参数和语言偏置控制规则长度，从而降低问题规模、提高分类准确率.例如：基于ILP—R的多关系朴素贝叶斯分类方法[1]、1BC[2]、1BC2方法[3-4]、Mr—SBC方法[5]和nFOIL方法[6]等；基于CrossMine的元组号传播方法[7-8]；基于关系数据库技术的多关系朴素贝叶斯分类方法——Graph—NB[9].Graph—NB中使用了一种叫做“Cutting off”的剪枝策略，对表进行剪枝，以提高分类准确率.但是这种方法有很多缺点：第一，对于朴素贝叶斯分类而言，参与分类的基本单位是属性而不是表，因此直接删除弱连接的表是明显不合适的.第二，特征选择有两种方法——过滤方法(filter)和打包方法(wrapper)，“Cutting off”剪枝方法是一种打包方法.打包方法是与分类器绑定的剪枝策略，它选择的是与分类器特点相匹配的最好属性，所以理论上能够得到很好的准确率；但打包方法找到的好的属性仅适于当前所用分类器，而且使用打包方法明显需要付出

收稿日期：2007-07-24 **修回日期**：2007-12-17

基金项目：国家自然科学基金资助项目(No.60675030)

作者简介：徐光美(1977—)，女，博士研究生；杨炳儒(1943—)，男，教授，博士生导师，E-mail：bryang_kd@yahoo.com.cn

较高的时间代价[10].

本文扩展互信息标准(它属于过滤方法)到多关系情况下,以提高多关系朴素贝叶斯分类方法的分类性能. 基于元组号传播方法和面向元组的统计计数方法,本文给出了基于扩展互信息标准进行属性选择的方法和步骤,并建立了一种基于扩展互信息的多关系朴素贝叶斯分类器(MI—MRNBC). 标准数据集上的实验显示,基于扩展互信息标准进行属性选择,可以在不增加算法时间复杂度的前提下,找到与分类属性最相关的属性,并在仅有极少属性参与分类时,得到较高的分类准确率.

1 基于互信息的多关系朴素贝叶斯分类器(MI—MRNBC)

1.1 互信息标准的一阶扩展

分类是主要的数据挖掘任务之一,为获得最小的分类错误率,通常需要得到与目标 c 类最大统计依赖的属性集合. 实现最大依赖的最流行的方法是最大相关特征选择:选择与目标类 c 最大相关的属性,相关性则通常根据自相关和互信息来衡量,互信息则是衡量变量依赖广泛使用的测度.

单关系(表)情况下,表内属性与目标属性间的相关性大小可以直接由相应的互信息量大小表示;而多关系情况下,属性分布在不同的关系中,要计算与目标属性间的互信息量需要首先进行关系连接. 为降低关系间连接时的时空代价,本文采用元组号传播方法进行虚拟的连接. 图 1 是多关系情况下,进行元组号传播后的例子.

贷款表

贷款号	账目号	数额	借贷期	已还款金额	类别
1	14	1200	12	100	+
2	14	3000	12	300	+
3	18	9000	24	500	−
4	50	2000	36	200	−
5	50	15000	24	1000	+

账目表

账目号	频率	日期	贷款号	类别
14	每月	95—02—11	1,2	2+,0−
18	每周	96—05—22	3	0+,1−
50	每月	97—01—24	4,5	1+,1−
60	每周	98—08—26	−	0+,0−

图 1 传播后的样例数据库

Fig·1 A sample database of tuple ID propagation

基于图 1,可以很容易地计算每一个属性和目标属性间的互信息量大小. 那么多关系情况下,不同关系中的属性是否应平等对待呢? 文献[9]中实验证明了多关系情况下,每个关系同目标关系间的相关程度是不相同的. 基于上述考虑,为每张表指定一个权值 w,它表示此表与目标表间的相关程度,一般由领域专家指定,w 取值范围为[0,1],其中目标表的值 w 为 1. 因此基于互信息的多关系属性选择步骤如下.

(1) 根据一定知识(如语义关系图[5])进行元组号传播.

(2) 依据传播后的表计算每个属性 x 和目标属性 c 间的互信息值:

$$I(c,x)=\sum_{i,j}p(c_i,x_j)\lg\frac{p(c_i,x_j)}{p(c_i)p(x_j)} \quad (1)$$

其中,$p(c_i,x_j)$为 c_i 和 x_j 的联合概率,$p(c_i)$为 c_i 的概率,$p(x_j)$为 x_j 的概率.

(3) 根据加权互信息值 $wI(c,x)$的大小对所有属性由大到小排序,然后选取前 m(用户指定)个属性参与分类.

1.2 MI—MRNBC 算法

MI—MRNBC 根据传播的类标签的数目和类别进行计数(面向元组的统计计数方法,不考虑元组号重复问题),并用 Laplace 估计方法避免零概率出现,多关系分类公式见文献[9].

输入:数据集,语义关系图 R,交叉验证次数 n,参与分类的属性个数 m.

步骤:

(1) 根据用户要求选择数据集中合适元组,并进行类别属性取值替换等预处理操作,将目标表元组随机分为 n 组(作 n 次交叉验证).

(2) 根据 R 进行元组号传播,此时也要将元组分组标志传播,以判断非目标关系中元组是训练数据还是测试数据.

(3) 从 n 组中选一组数据作为测试数据,其余 $n-1$ 组作为训练数据,并分别进行连续属性离散化.

(4) 对传播后的训练数据进行统计计数,计算所需互信息值和类条件概率值.

(5) 根据加权互信息值为属性排序.

(6) 用多关系朴素贝叶斯分类公式对测试数据进行分类(仅取加权互信息值最大的 m 个属性参与分类).

输出:分类准确率(accuracy).

其中(3)～(6)执行 n 次,最后取 n 次交叉验证

的平均结果.

由于计算互信息值不需要额外的统计计数，因此训练阶段仅需要增加对属性的一次快速排序时间就可以找到所需要的 m 个属性，这不会增加算法运行的时间复杂度(进行属性选择和不进行属性选择算法时间复杂度都为 $O(\max(n_T \cdot n_{T_q}))$，其中 T 为数据集中的任一表，T_q 为表 T 在 R 中的直接前驱，n_T 为 T 中元组数目，n_{T_q} 为 T_q 中元组数目)，而且当这 m 个属性之外的属性存在缺失值时丝毫不会影响分类性能，所以 MI—MRNBC 算法具有高效性和较高的容错性.

2　实验

本文使用多关系领域常用的标准数据集：PKDD CUP 1999 的金融数据集和 Mutagenesis 数据集实验. 本文实验中各表权值 w 都默认为 1.

(1) 金融数据集. 便于比较，根据文献[9]对金融数据集进行修改. 为比较基于互信息的多关系属性选择方法的有效性，本文也实现了一个随机选择属性进行多关系朴素贝叶斯分类的算法 S—MRNBC. 图 2 是 MI—MRNBC 和 S—MRNBC 的分类准确率比较(取 10 次交叉验证的平均结果).

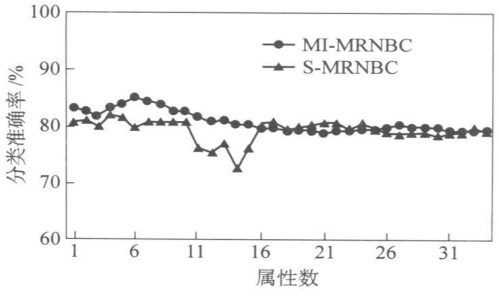

图 2　金融数据集上的分类准确率比较

Fig. 2　Accuracy on financial database

从图 2 可以看出，MI—MRNBC 仅使用极少的的属性(当 m 为 6 时分类准确率最高为 85%)参与分类就可以得到良好的分类效果，并且当属性个数在 1～17 之间变化时都比使用所有的属性参与分类的分类效果更好(所有属性参与分类时，准确率为 79.25%).

(2) Mutagenesis 数据集. 使用回归友好的 Mutagenesis数据集的所有属性参与实验. 图 3 是 MI—MRNBC和S—MRNBC的分类准确率比较.

从图 3 可以看出：MI—MRNBC 仅使用极少的属性就可以得到良好的分类效果，其中当 m 为 3 时分类准确率最高为 88.9%，m 为 1 时分类准确率就可以达到最高为 84.1%；而使用所有属性参与分类时，即 m 为 7 时分类准确率为 78.1%. 而且基于扩展的互信息标准找到的与分类属性最相关的前三个属性是“lumo”、“logp”、“indl”，而这三个属性都是目标表中属性，这意味着不需要进行复杂的多关系连接操作，只用目标表中属性分类就可以得到最好的分类结果，因此多关系分类退化为单关系分类问题，这大大降低了分类的时空代价，一定程度上也提高了分类器的容噪能力.

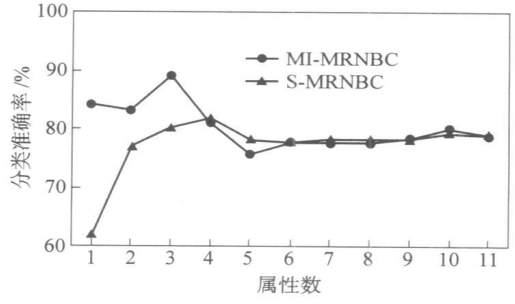

图 3　Mutagenesis 数据集上的分类准确率比较

Fig. 3　Accuracy on Mutagenesis database

3　结论

基于元组号传播方法，本文给出了基于互信息的多关系属性选择方法，并在此基础上形成了基于互信息的多关系朴素贝叶斯分类器(MI—MRNBC). 实验显示使用基于互信息的属性剪枝方法可以在几乎不增加计算时间基础上，找到与分类属性最相关的属性，并在仅用极少的属性参与分类情况下，得到较高的分类准确率.

MI—MRNBC 中根据传播的类标签的数目和类别进行统计计数，计数并没有考虑传播的元组号可能重复的问题，将来的工作可以考虑引入合适的聚合函数，在考虑传播的类标签个数的同时也考虑元组号是否重复. 对于一个表中元组与其他表中多个元组相连接的情况，可以引入合适的概率分布函数定义多个元组的联合概率，改变现在的多关系朴素贝叶斯计算公式.

参　考　文　献

[1] Pompe U, Kononenko I. Naïve Bayesian classifier within ILP-R // *Proceedings of the 5th International Workshop on Inductive Logic Programming*. Belgie: Katholieke Universiteit Leuven, 1995: 417

[2] Peter A, Nicolas L. 1BC: a first-order Bayesian classifier // *Proceedings of the 9th International Workshop on Inductive Logic Programming*. Berlin: Springer-Verlag, 1999, 1634: 92

[3] Nicolas L, Peter A. 1BC2: a true first-order Bayesian classifier //

Proceedings of the 12th International Conference on Inductive Logic Programming. Berlin: Springer-Verlag, 2002: 133

[4] Peter A, Nicolas L. Naïve Bayesian classification of structured data. *Mach Learn*, 2004, 57(3): 233

[5] Ceci M, Appice A, Malerba D. Mr-SBC: a multi-relational Naïve Bayesian classifier // *Lecture Notes in Artificial Intelligence*. Berlin: Springer, 2003: 95

[6] Niels L, Kristian K, Luc D. nFOIL: integrating Naïve Bayes and FOIL // *Proceedings of the 20th National Conference on Artificial Intelligence*. Cambridge, 2005: 795

[7] Yin X X, Han J W, Yang J, et al. Efficient classification across multiple database relations: a CrossMine approach. *IEEE Trans Knowl Data Eng*, 2006, 18 (6): 770

[8] Yin X X, Han J W, Yang J, et al. CrossMine: efficient classification across multiple database relations // *Constraint-Based Mining and Inductive Databases*. Hinterzarten: Springer, 2004: 172

[9] Liu H, Yin X X, Han J W. An efficient multi-relational Naïve Bayesian classifier based on semantic relationship graphs // *Proceedings of ACM-SIGKDD Workshop on Multi-Relational Data Mining*. Chicago, 2005: 39

[10] Peng H C, Long F H, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*, 2005, 27(8): 1226

(上接第 915 页)

使得当磁场垂直于超导薄膜表面时电阻转变展宽现象更为明显. 通过分析超导薄膜的电阻转变数据计算出了 TBCCO/LAO 的各向异性参量, $\gamma = H_{c2}^{ab}(0)/H_{c2}^{c}(0) = 14.58$.

参 考 文 献

[1] Xu X J, Fu L, Wang L B, et al. Dependence of activation energy upon magnetic field and temperature in $YBa_2Cu_3O_{7-x}$ epitaxial thin film. *Phys Rev B*, 1999, 59: 608

[2] Yin D L, Qi Z, Xu H Y, et al. Resistive transition equation of the mixed state of superconductors. *Phys Rev B*, 2003, 67: 092503

[3] Xu S Y, Li Q, Wertz E, et al. High critical current density and vortex pinning of epitaxial MgB_2 thin films. *Phys Rev B*, 2003, 68: 224501

[4] Lu X F, Wang Z, Zhang Y Z, et al. Field and temperature dependence of thermally activated flux flow resistance in $Tl_2Ba_2CaCu_2O_8$ thin films. *Phys C*, 2005, 423: 175

[5] Espinosa-Arronte B, Andersson M. Scaling of vortex-liquid resistivity in high-T_c superconductors. *Phys Rev B*, 2005, 71: 024507

[6] Tinkham M. Resistive transition of high-temperature superconductor. *Phys Rev Lett*, 1988, 61: 1658

[7] Festin O, Svedlindh P. Vortex fluctuation in high-T_c thin films close to the resistive transition. *Phys Rev B*, 2004, 70: 024511

[8] Blatter G, Fegelman M V, Geshkenbein V B, et al. Vortices in high-temperature superconductors. *Rev Mod Phys*, 1994, 66: 1125

[9] Song Y Q, Lee M, Halperin W P, Magnetic-flux-lattice anisotropy of $Tl_2Ba_2CaCu_3O_{10}$ by ^{205}Tl nuclear magnetic resonance. *Phys Rev B*, 1991, 44: 914

[10] Kim D H, Gray K E, Kampwirth R T, et al. Magnetoconductance of $Tl_2Ba_2CaCu_2O_x$ films in fluctuation regime. *Phys Rev B*, 1991, 43: 2910