基于特征加权的多关系朴素贝叶斯分类模型

徐光美 刘宏哲 张敬尊
(北京联合大学信息服务工程重点实验室 北京 100101)

摘 要 为进一步提高多关系朴素贝叶斯方法的分类准确率,分析了已有的特征加权方法,并在将特征加权方法扩展到多关系的情况下结合元组ID传播方法和面向元组的统计计数方法,建立了基于特征加权的多关系朴素贝叶斯分类模型(MRNBC-W)。标准数据集上的实验结果显示,新方法可以在不增加算法时间复杂度的前提下,有效提高金融数据集的分类准确率。文中也给出了结合扩展互信息标准对属性进行过滤后,加权方法和不加权方法的分类比较。

关键词 多关系数据挖掘,朴素贝叶斯,分类,互信息,特征加权
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.10.059

Multi-relational Naïve Bayesian Classifier Using Feature Weighting

XU Guang-mei LIU Hong-zhe ZHANG Jing-zun
(Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China)

Abstract To improve the accuracy of multi-relational naïve Bayesian classifiers, this paper discussed existing feature weighting methods and upgraded the method to deal with multi-relational data directly. Based on the tuple ID propagation method and counting methods towards tuples, a multi-relational naïve Bayesian classifier using feature weighting (MRNBC-W) was given. Experiments on Financial database show that with the help of feature weighting, the classifiers can give better accuracy without increase of time complexity. Furthermore, MRNBC-W based on mutual information (MRNBC-W-MI) was implemented.

Keywords Multi-relational data mining (MRDM), Naïve Bayes, Classification, Mutual information, Feature weighting

1 引言

朴素贝叶斯分类模型(Naïve Bayesian Classifier, NBC)训练简单且具有相当的健壮性和高效性[1],已经成功地应用到分类、聚类及模型选择等数据挖掘的任务中。目前,许多学者致力于研究朴素贝叶斯分类方法的改进模型,两类典型的改进思路分别是:放松特征变量间独立性的限制[2,3],考虑局部相关性,该类方法通常能有效提高分类效果,但会导致计算代价大幅提高;利用打包方法或者过滤方法从不同角度对属性进行加权,以改善分类效果。

多关系朴素贝叶斯分类模型(Multi-relational Naïve Bayesian Classifier, MRNBC)是将统计学习和关系学习相结合来处理分类问题的方法。本文在分析已有研究成果基础上给出了基于特征加权的多关系朴素贝叶斯分类模型(Multi-relational Naïve Bayesian Classifier Using Feature Weighting, MRNBC-W),该方法将属性加权方法扩展到多关系情况下,并基于元组ID传播方法和面向元组的统计计数方法进行概率统计[4,5],以提高多关系朴素贝叶斯分类方法的分类性能。多关系标准数据集上的实验显示,MRNBC-W可以在不增加算法时间复杂度前提下,有效提高金融数据集的分类准确率。为进一步优化分类,本文在加权基础上,给出了基于扩展互信息标准[6]对数据进行属性选择后的分类曲线。

2 相关工作

国内外很多学者致力于研究基于特征加权的朴素贝叶斯分类模型,以期在保证模型的朴素前提下优化模型性能。2007年邓维斌等提出了基于Rough Set的加权朴素贝叶斯分类算法[7]。该文基于粗糙集的重要性理论,提出了基于代数观、信息观以及综合二者的属性权值加权方法,并分析了3种方法的适用情况。2011年邓维斌等基于文献[7]中的思想,从粗糙集方法的信息观点出发,为属性计算加权系数,并对中英文垃圾邮件进行了过滤[8]。王国才等则在2010年提出了另外一种基于粗糙集的加权方法,该方法基于粗糙集理论中的下近似集来计算属性权值[9]。

Harry Zhang[10]等的论文则详细分析了以下几种属性加权方法对分类的影响情况:基于信息增益率计算属性权值、基于爬山法获取属性权值、基于马尔可夫链蒙特卡罗方法获取属性权值,将爬山法和信息增益率方法结合使用来计算属性权值,将马尔可夫链蒙特卡罗法和信息增益率方法相结合来计算属性权值。邓春伟和史焕卿的论文则通过计算属性取值

到稿日期:2013-10-15 返修日期:2014-02-21 本文受国家自然科学基金(61372148),北京市"长城学者"计划项目(CIT&TCD20130320),北京市优秀人才培养(2010D005022000011),北京联合大学校级科研项目(zk201017x)资助。
徐光美(1977—),女,博士,副教授,主要研究方向为数据挖掘等,E-mail:xxtguangmei@buu.edu.cn;刘宏哲(1971—),女,博士,副教授,主要研究方向为语义计算等;张敬尊(1980—),女,硕士,讲师,主要研究方向为数据挖掘等。

对每个类的相关概率和不相关概率来对属性加权[11]。张明卫等于2008年提出了一种基于相关系数的权重求解方法，该方法用相关系数来测量特征属性和分类属性间的线性相关性，并在中医小儿肺炎病例数据集上进行了实验分析和研究[12]。2012年文献[13]对Harry Zhang和张明卫的方法进行了分析，认为基于增益率的加权方法仅仅考虑了每个属性对分类结果的影响程度，而张明卫的方法仅考虑了每个属性对分类属性的影响程度，两种方法都不能充分体现属性应有的权重，因此给出了将两者结合来计算属性权值的方法。

张步良的论文提出基于打包的方法对属性进行加权的方法，即先基于每个属性分别做朴素贝叶斯分类，然后把得到的分类准确率作为该属性的权值[14]。杨敏等则基于偏最小二乘方法建立加权朴素贝叶斯分类模型，通过建立目标属性和非目标属性之间的偏最小二乘回归方程，把回归系数赋给对应的非目标属性，作为相应的权重[15]。文献[16,17]都是基于数据直接获取加权系数，其中，华锐的方法是对每个属性取值计算权值(不考虑所属类别)，而程克非等的方法则是对后验概率计算中的每个条件概率项计算不同的权值(类别不同，权值不同)，以改进朴素贝叶斯的分类能力。

Eibe Frank等在2003年提出了对实例加权的朴素贝叶斯分类方法[18]，该方法通过计算测试样例与训练样例间的距离来为训练样例加权，并通过设置最近邻的大小来限制每个局部分类模型参与的实例数目。实验显示，相对于参数——最近邻的邻域大小，新方法具有很好的鲁棒性，但是实例距离的计算大大增加了算法的时间复杂度。Mark Hall在2007年提出了基于决策树方法对属性加权的方法[19]，具体策略如下：随机抽样部分数据作为训练数据，然后基于抽样数据得到一颗未被剪枝的决策树，如果属性不在这棵树上，则该属性权重置为0，如果属性在决策树上的最小深度为 d，则该属性的权值置为 $1/\sqrt{d}$，将上述工作重复 i 次，每个属性的最终权值取 i 次的平均值。试验显示，该方法的效果好于标准朴素贝叶斯以及基于信息增益和信任度对属性加权的方法。

3 基于特征加权的多关系朴素贝叶斯分类模型

3.1 多关系朴素贝叶斯分类

多关系情况下，为降低关系间连接时的时空代价，本文采用元组ID传播方法[5]，它可以在任何连接路径中传播，而且比实际的物理连接在时间和空间上的代价更小，是一种灵活方便的虚拟连接方法。为了让朴素贝叶斯可以直接处理多表数据，需要修改分类公式。假定目标表(包含分类属性的表)是 T，R 是可以与 T 连接的一个表，表 T 有 m 个属性，表 R 有 n 个属性，对于 T 中的一个元组 $x=(x_1,x_2,\cdots,x_m)$，R 中有 l 个元组可以与 x 连接，这 l 个元组是 $(y_1,y_2,\cdots,y_l)$，每一个元组 y_i 由 n 个属性表示：$y_i=(y_{i1},y_{i2},\cdots,y_{in})$，则元组 x 的分类公式[5]为：

$$C_{MAP}=\arg\max_{c_i\in C} p(x,y_1,y_2,\cdots,y_l|c_i)p(c_i)$$

$$=\arg\max_{c_i\in C}\prod_{j=1}^{m}p(x_j|c_i)\prod_{q=1}^{l}\prod_{o=1}^{n}p(y_{qo}|c_i)p(c_i) \quad (1)$$

3.2 特征权值计算

实际中有很多度量特征属性和分类属性间关联程度的方法，比如信息增益、互信息、相关系数等，研究者也纷纷提出了基于这些标准度量方法计算特征权重的分类模型，但是这些关联度量方法也常常作为属性过滤方法用于提升分类性能。在多关系情况下，属性数目偏多，采用必要的属性过滤手段可以有效改进分类效果，因此本文在基于扩展互信息标准进行属性过滤的前提下[6]，扩展简单的特征加权方法[17]到多关系下，以改进多关系分类效果。具体权值计算策略如下：令数据集合 S 中总的训练样本个数为 $N(S)$，则数据集合中属性 x_i 取值为 x_{ij} 的个数为 $N(x_{ij})$，属性 x_i 的所有的可能取值个数为 n_i，则对于式(1)中的每个条件概率项赋予一个权值 w_i，$w_i=\frac{N(x_{ij})+1}{N(S)+n_i}$。该权值体现了属性的某一个取值在样本中出现的概率，概率越大，说明该属性取值越重要。

因此，基于该权值，我们设计并实现了基于特征加权的多关系朴素贝叶斯分类模型(MRNBC-W)。为进一步改进分类效果，我们也分析了基于扩展互信息标准对MRNBC-W进行属性选择后的分类情况(分类模型简称MRNBC-W-MI)。

4 实验研究

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数据集进行实验。

4.1 金融数据集

为便于比较分类结果，本文对金融数据集作一些修改：随机删除Transaction表中90%的元组，仅保留10%；随机删掉Loan表中的部分正例，使得Loan表中正负例子个数更平衡，并且将连续属性离散化为10个取值。最后整个数据集包含8个表、34个有效属性、75982个元组。8个表分别为：Loan表、Account表、Client表、Dispositon表、Order表、Trans表、Card表和District表。其中目标表为Loan，目标属性为"status"，Loan表中包含324个正元组，76个负元组。表1是不加权的Graph-NB方法和MRNBC-W方法的分类准确率比较(取10次交叉验证的平均结果)。

表1 金融数据集上Graph-NB方法和MRNBC-W方法的分类准确率比较

算法	运行时间(秒)	分类准确率(%)
MRNBC-W	0.4	82.5%
Graph-NB	0.4	79.25%

从表1可以看出，基于特征加权可以有效提高分类准确率，而加权所需计数信息不需要额外计算，加权方法不会牺牲朴素贝叶斯分类的高效性。

图1是基于扩展互信息标准分别对Graph-NB方法(Graph-NB-MI)和MRNBC-W-MI方法进行属性过滤后的分类准确率曲线比较。

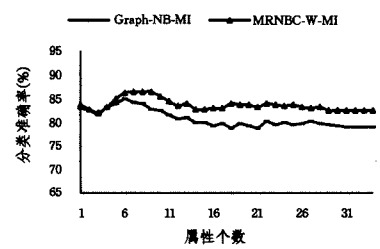

图1 金融数据集上Graph-NB-MI和MRNBC-W-MI的分类准确率比较

从图1可以看出，当参与分类的属性个数 m 从1到34变化时加权方法都要好于不加权方法，而当 m 为7、8、9时

MRNBC-W-MI 方法具有最好的分类准确率(86.5%),当参与分类的属性个数在 5～33 之间变化时,分类准确率都比所有属性参与分类高。

4.2 Mutagenesis 数据集

描述分子结构的 Mutagenesis 数据集被归纳逻辑程序设计(ILP)领域广泛使用。就分类而言,需要将分子表内化合物分为两类:一类是会引起突变的化合物;一类是不会引起突变的化合物。我们选择回归友好的 Mutagenesis 数据集参与实验。通常会使用这个数据集的 3 层背景知识分别做实验,本文所有实验均在背景知识 2 情况下进行(BK2)。它包含 4 个表,共 15218 个元组,4 个表分别为:Mole 表、Atom 表、Bonds 表和 Molatm 表。目标表为 Mole 表,它包含 188 个 Mutagenesis 元组,其中 124 个正元组,64 个负元组。表 2 是不加权的 Graph-NB 方法和 MRNBC-W 方法的分类准确率比较(取 10 次交叉验证的平均结果)。

表 2 Mutagenesis 数据集上 Graph-NB 方法和 MRNBC-W 方法的分类准确率比较

算法	运行时间(秒)	分类准确率(%)
MRNBC-W	0.2	77.8%
Graph-NB	0.2	78.1%

从表 2 可以看出,基于特征加权后,当所有属性参与分类时,分类效果有所下降。

图 2 是 Mutagenesis 数据集上,基于扩展互信息标准分别对 Graph-NB-MI 和 MRNBC-W-MI 方法进行属性过滤后的分类准确率曲线比较。

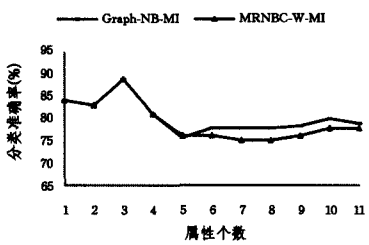

图 2 Mutagenesis 数据集上 Graph-NB-MI 和 MRNBC-W-MI 的分类准确率比较

从图 2 可以看出,当参与分类的属性个数 m 为 3 时,MRNBC-W-MI 和 Graph-NB-MI 方法都具有最好的分类准确率(88.9%)。当 m 为 4 时加权方法好于不加权方法,但当 m 从 5 到 11 变化时加权方法反而降低了分类准确率。一个可能的原因是,该数据集数据量偏小,这种加权使得分类更加倾向于正例,反而在某些情况下降低了分类精度,因此下一步有必要研究更平衡的加权方法,以避免这种偏斜。

结束语 本文基于元组 ID 传播方法和扩展的互信息属性过滤方法,设计并实现了基于特征加权的多关系朴素贝叶斯分类模型 MRNBC-W 和 MRNBC-W-MI,并在标准数据集上进行实验以分析加权对分类的影响程度。

本文在将特征加权方法扩展到一阶情况下,对多关系分类公式中的每个条件概率项赋予一个权值,该权值体现了属性的某一个取值在样本中出现的概率,概率越大,说明该属性取值越重要。虽然该方法有效地提高了金融数据集的分类准确率,但是该权值的计算并没有考虑多关系分类的特点,如何设置符合多关系特点的一阶加权方法是下一步的研究方向。

参 考 文 献

[1] Ratanamahatana C A, Gunopulos D. Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection[C]// Proceedings of Workshop on Data Cleaning and Preprocessing, at IEEE International Conference on Data Mining. Maebashi, Japan, 2002

[2] 石洪波,王志海,黄厚宽,等. 一种限定性的双层贝叶斯分类模型[J]. 软件学报,2004,15(2):193-199

[3] Ong H C, Khoo M Y, Saw S L. An Improvement on the Naïve Bayes Classifier[C]// International Conference on Information and Knowledge Management (2012). Singapore, 2012:190-194

[4] Yin Xiao-xin, Han Jia-wei, Yang Jiong, et al. Efficient classification across multiple database relations: a CrossMine approach [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(6):770-783

[5] Liu Hong-yan, Yin Xiao-xin, Han Jia-wei. An efficient multi-relational naive bayesian classifier based on semantic relationship graphs[C]// Proceedings of the 4th international workshop on Multi-relational mining. Chicago, Illinois, 2005:39-48

[6] 徐光美,杨炳儒,秦奕青,等. 基于互信息的多关系朴素贝叶斯分类器[J]. 北京科技大学学报,2008,30(8):963-966

[7] 邓维斌,王国胤,王燕. 基于 Rough Set 的加权朴素贝叶斯分类算法[J]. 计算机科学,2007,34(2):204-206

[8] 邓维斌,王国胤,洪智勇. 基于粗糙集的加权朴素贝叶斯邮件过滤方法[J]. 计算机科学,2011,38(2):218-221

[9] 王国才,张聪. 一种基于粗糙集的特征加权朴素贝叶斯分类器[J]. 重庆理工大学学报,2010,24(7):86-90

[10] Zhang H, Sheng Li. Learning Weighted Naive Bayes with accurate ranking[C]// The 4th IEEE International Conference on Data Mining (ICDM04). Brighton: IEEE Computer Society, 2004:567-570

[11] 邓春伟,史焕卿. Lucene 的最小风险概率加权朴素贝叶斯算法[J]. 哈尔滨理工大学学报,2012,17(1):63-67

[12] 张明卫,王波,张斌,等. 基于相关系数的加权朴素贝叶斯分类算法[J]. 东北大学学报,2008,29(7):952-955

[13] Guo Bao-en, Liu Hai-tao. Assigning Hybrid-Weight for Feature Attribute in Naïve Bayesian Classifier[C]// 2012 International Proceedings of Computer Science and Information Technology (2012). Hongkong, 2012:86-90

[14] 张步良. 基于分类概率加权的朴素贝叶斯分类方法[J]. 重庆理工大学学报,2012,26(7):81-83

[15] 杨敏,贺兴时,刘平丽,等. 基于属性约简的 PLS 加权朴素贝叶斯分类[J]. 西安工程大学学报,2013,27(1):118-121

[16] 程克非,张聪. 基于特征加权的朴素贝叶斯分类器[J]. 计算机仿真,2006,23(10):92-96

[17] 华锐,梁娜. 特征加权朴素贝叶斯分类器在小样本中的应用[J]. 统计与决策,2012,23:69-71

[18] Frank E, Hall M, Pfahringer B. Locally weighted naive Bayes [C]// The 19th Conference in Uncertainty in Artificial Intelligence(2003). Acapulco, Mexico: Morgan Kaufmann, 2003:249-256

[19] Hall M. A decision tree-based attribute weighting filter for Naive Bayes[C]// Knowledge-Based Systems. 2007:120-126

[20] 张步良. 基于分类概率加权的朴素贝叶斯分类方法[J]. 重庆理工大学学报:自然科学版,2012,26(7):81-83