·人工智能及识别技术· 文章编号:1000-3428(2016)05-0218-06 文献标识码:A 中图分类号:TP18

基于关系选择的多关系朴素贝叶斯分类

毕佳佳,张 晶

(合肥工业大学计算机与信息学院,合肥 230009)

摘 要: 依据多关系数据库中的背景表对分类任务具有的不同大小贡献度,提出一种基于关系选择的多关系朴素贝叶斯分类算法。对关系表进行两轮删减,根据最大信息增益率删掉部分对分类影响较小的关系表,把平均信息增益率作为衡量表对分类的贡献度,根据贡献度选定余下的表用于最终的分类。实验结果表明,该算法能有效提高分类准确率,相比 Graph-NB 算法、Classify_tables 算法及 MRNBC-W 算法分别提高 2.2%,1.1%,0.86%。

关键词: 数据挖掘;多关系;分类;信息增益率;贡献度;关系选择

中文引用格式: 毕佳佳,张 晶. 基于关系选择的多关系朴素贝叶斯分类[J]. 计算机工程,2016,42(5):218-223.

英文引用格式: Bi Jiajia, Zhang Jing. Multi-relational Naive Bayesian Classification Based on Relation Selection[J]. Computer Engineering, 2016, 42(5):218-223.

Multi-relational Naive Bayesian Classification Based on Relation Selection

BI Jiajia, ZHANG Jing

(School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

【Abstract】 Aiming at the problem that background tables have different contribution degree to classification tasks in relation database, this paper proposes a multi-relational naive Bayesian classification based on relation selection. It conducts two rounds of cuts for multi-relational tables, removes part of the relations with little contribution to classification according to their maximum information gain ratio, defines the average information gain ratio as the measure of the contribution of the table, and selects the final relations among the rest relations for classification according to their contribution. Experimental results show that this algorithm can improve classification accuracy effectively. Compared with Graph-NB, Classify_tables and MNBC-W algorithms, the average accuracy rates are improved by 2.2%, 1.1%, 0.86%.

【Key words】 data mining; multi-relation; classification; information gain ratio; contribution degree; relation selection

DOI: 10.3969/j.issn.1000-3428.2016.05.037

1 概述

信息技术的快速发展使得现实社会中的数据规模以爆炸式的速度不断增长。面对如此庞大的数据,商家要想从这些数据中获取有价值的信息来指导后续的行为,就需要利用一些数据挖掘技术分析数据,从而提取出大量隐含的、潜在有用的并且新颖的信息。分类方法是数据挖掘的重要技术之一,如在银行贷款方面,根据用户的收入、信用等信息就可以判断出是否给该用户贷款。目前很多实际应用都是利用数据挖掘中的分类技术来分析这些数据,并提取出较为精确的决策模型,对用户进行预测行为。

在现实生活中,数据基本上都是以二维表的形式存储在关系数据库中,每个数据表中都包含着各种对分类有影响的信息。因此,要想从多关系数据库中提取有价值的信息,就需要利用多关系数据挖掘技术对多关系表中的数据进行分析。多关系分类就是以多个相互联系的关系表为对象对其挖掘分类。根据前人已有的研究方法,多关系分类可以总结为以下2 种:(1)利用传统的分类方法 Propositional Learning 对多关系数据进行处理。该方法需要将多关系数据转化成单个关系。然而,在转化过程中容易造成信息丢失[1]。(2)结合多关系对传统的分类方法扩展更新。这种方法不需要将多个关系表转化成单个关系,便能够直接处理多关系数据[2-3],该方法主要包含 3 个方面:1)基于选择图的多关系分类,如 TILDE[4],MRDTL[5];2)多视图分类[6-7];

基金项目: 国家自然科学基金资助项目(61273292,61305063)。

作者简介: 毕佳佳(1989-),女,硕士研究生,主研方向为数据挖掘;张 晶,副教授。

收稿日期: 2015-03-25 **修回日期:** 2015-05-15 **E-mail:** bijiajia163@163.com

3)基于元组 ID 传播[8]的多关系分类,如 CrossMine[8],Graph-NB[9],RNC[10],FAFS[11]等。这些方法都是在基于元组 ID 传播技术的基础上对背景表中的属性进行评估,也可以对谓词进行评估,极大地提高了多关系分类的效率。

朴素贝叶斯[[12]是一种训练简单、容易理解的分类方法,并且能够达到较好的分类效果。最初它只应用于单表中,后来很多研究者在此基础上加以扩展,可直接处理多关系数据的分类问题。文献[9]提出的 Graph-NB 算法就是基于语义关系图的多关系朴素贝叶斯分类算法,此算法通过深度优先或者宽度优先对语义关系图进行遍历并且优化选择,删除那些与目标表关系较小的表,由此来提高分类精确率。文献[13]提出的是一种利用基于奇异值分解的表的贡献度对语义关系图进行优化剪枝。该方法在保证一定的分类准确率的同时,提高了分类效率,但是该算法在分类准确率上波动比较大,不太稳定。文献[14]对多关系朴素贝叶斯分类器也有诸多研究,如 MRNBC-W 算法,该算法将特征加权的方法扩展到多关系分类中,可以在不增加算法时间复杂度的前提下,有效提高分类准确率,但是该方法中权值的计算没有考虑到多关系分类的特点。本文提出一种基于关系选择的多关系朴素贝叶斯分类(Multi-relational naive Bayesian classification based on Relations Selection,MRS)算法。首先计算出每个背景表的最大信息增益率,并根据最大信息增益率删除部分对分类影响较小的关系表。之后把关系表的平均信息增益率作为衡量表对分类的贡献度,并根据表的贡献度对余下的关系表进行选择,从而选定最终用于分类的关系表,即选择具有较高贡献度的关系表作为最终分类的数据表。

2　相关知识

在现实生活中,数据一般都以表的形式存在数据库中,每个数据表中都包含很多有价值的信息,而表与表之间又存在比较复杂的关系。因此,如何表示这些表之间的关系对后面的处理是很重要的。

2.1　多关系数据库的连接

关系数据库是由多个关系表组成的,它描述了一组实体 $DB=\{E_1,E_2,\cdots,E_n\}$ 以及实体之间的关系。表中的每一行代表着一个元组,而每个元组则代表着由一些不同的属性的值构成的对象。对于一个多关系数据库中的分类任务,通常有一个包含类标签属性的关系表,称其为目标表。除此之外,还有许多与目标表直接或者间接相连的其他表,称为背景表。然而,在背景表中没有类标签,如果想要对背景表中属性的重要性以及背景表本身进行评估,就需要将类标签从目标表中传递到背景表中。而传递的过程则是通过连接属性完成的。数据库中的每个表至少有一个主码属性,其他的则是分类属性或者连接属性。表与表之间皆可以通过主码与外码或者外码与外码之间连接起来。

以 PKDD CUP99 金融数据为例,图 1 是其数据库中各关系的连接结构。在此多关系数据库中,总共有 8 个相互关联的关系表,其中,loan 为目标表;其余的 7 个表则是背景表,背景表根据连接属性直接或者间接地与目标表相连。

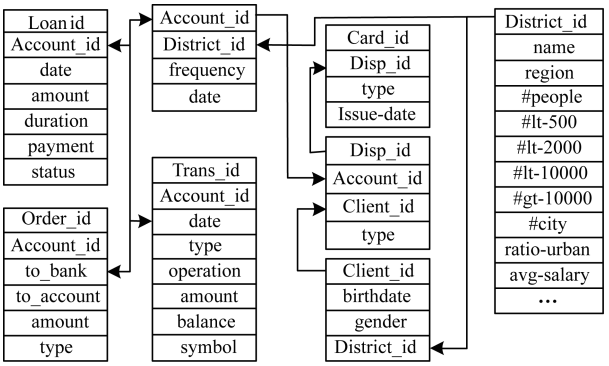

图 1　PKDD CUP 金融数据库各关系的连接结构

2.2　多关系朴素贝叶斯分类

朴素贝叶斯是利用统计学中的贝叶斯定理来对样本进行分类的一种简单概率的分类方法。它的原理简单,易于理解。其思想基础是对于一个待分类的样本,通过计算它属于各个类别的概率值来确定概率值最大的那个类别就是该样本的最终类别。贝叶斯分类应用非常广泛,可应用到各种领域,如对垃圾邮件的过滤[15]、人脸识别[16]等。起初朴素贝叶斯分类只应用于单表数据,后来被扩展到直接处理多个关系表。

对于多关系的朴素贝叶斯,假设 t 是目标表,s 则是与其相连的另外一个关系表,对于表 t 中的一个元组 x:$X=(x_1,x_2,\cdots,x_n)$,s 中有 p 个元组能够与 x 相连接。这些 p 个元组是$(y_{k_1},y_{k_2},\cdots,y_{k_p})$,其中,每个元组 y_{kl} 由 r 个值表示:$y_{kl}=(y_{kl_1},y_{kl_2},\cdots,y_{kl_r})$,元组 x 的类标签的计算如下:

$$\begin{aligned}C_{\mathrm{MAP}} &= \mathrm{argmax}_{c_i\in C}P(c_i\mid X)\\ &= \mathrm{argmax}_{c_i\in C}p(c_i)p(x_1,x_2,\cdots,x_n,y_{k_{11}},\cdots,\\ &\quad y_{k_{1r}},\cdots,y_{k_{p1}},\cdots,y_{k_{pr}}\mid c_i)\\ &= \mathrm{argmax}_{c_i\in C}p(c_i)\cdot\prod_{j=1}^{n}p(x_j\mid c_i)\\ &\quad\cdot\prod_{q=kl}^{kp}\prod_{t=1}^{r}p(y_{qt}\mid c_i)\end{aligned}\tag{1}$$

3 基于关系选择的多关系分类

在多关系分类中,除了可以用于分类的目标表,还有着一些与目标表息息相关的辅助表,即背景表。每个表中都附带着一些对分类有影响的信息。然而,并不是所有背景表中的信息都会对分类结果有所帮助,有的关系表可能没有作用,甚至还会降低分类准确度。因此,就需要想办法删除那些对分类效果没有帮助的关系表。本文提出的基于关系选择的多关系优化选择算法 MRS 对分类准确度有一定的提高。

3.1 多关系表的优化剪枝

在多关系数据库中,每个关系表中都带有对分类有影响的信息,主要体现在分类属性上。为了评价某个属性的分类能力,利用这个属性的信息增益来进行评估[17]。信息增益值的大小代表着这个属性能够给分类系统带来信息量的大小。也就是说,值越大代表着这个属性对分类的帮助就越大。而为了评价一个关系的分类能力,则利用这个表中属性的最大信息增益率来进行评价。对于任意背景表 R 而言,首先利用元组 ID 传播技术完成类标签的传递,把传播后 R 表中的所有分类属性中的最大信息增益率来衡量 R 表的分类能力。

定义 1(信息增益) 假设通过 ID 传播后的表 R 中有 P 个正例元组和 N 个负例元组。A_l 是表 R 的其中一个属性,并且属性 A_l 把表中的元组分成了 k 部分,其中,每部分包含了 P_i 个正例元组和 N_i 个负例元组,即有:

$$gain(A_l) = entropy(P,N) - \sum_{i=1}^{k} \frac{P_i + N_i}{P + N} entropy(P_i,N_i) \tag{2}$$

其中:

$$entropy(P,N) = -\left(\frac{P}{P+N}\mathrm{lb}\frac{P}{P+N} + \frac{N}{P+N}\mathrm{lb}\frac{N}{P+N}\right) \tag{3}$$

当以信息增益来评估属性并对其进行选择时,它会倾向于选择那些取值比较多的属性,而忽略取值少的属性。信息增益率是信息增益的一种扩展,它能够克服信息增益的不足,并且使信息增益达到一定的规范化。

定义 2(信息增益率) 假设表 R 有 P 个正例元组和 N 个负例元组。A_l 是传播后的表 R 中某个属性,则属性 A_l 的信息增益率定义如下:

$$gainratio(A_L) = \frac{gain(A_l)}{-\sum_{i=1}^{k} \frac{P_i + N_i}{P + N}\mathrm{lb}\frac{P_i + N_i}{P + N}} \tag{4}$$

在计算出背景表中每个属性的信息增益率后,则其中的最大值 $\max\limits_{A_l \in R_l} gainratio(A_l)$ 就代表着这个关系表中属性的最大分类能力。而本文的目的是删除分类能力比较小的关系表,也就是说对分类结果影响较小的背景表。因此,给定一个有关信息增益率的阈值 λ,如果背景表的最大信息增益率大于阈值 λ,就保留这些表。反之,则删除那些小于 λ 的关系表。例如以 PKDD CUP99 为例,如果表 order 和表 disp 的最大信息增益率都小于给定的阈值 λ,则说明这 2 个表中的信息对分类没有太大帮助,那么就删除这 2 个表。

3.2 基于贡献度的关系表选择

在根据给定阈值 λ 对关系表进行初步选择后,将对剩余的表进行再次剪枝,选择出最终用于分类的多关系表。一个关系表的最大信息增益率只能片面地代表其最大的分类能力,在根据最大信息增益率对多关系表进行初步筛选后,若仍用最大信息增益率来衡量该表整体的分类能力是不全面的。因此,需要一个能够综合评价关系表对分类的贡献度大小的标准。关系表所有属性的信息增益率的平均值代表了该表的整体分类能力,因此,把通过 ID 元组传播后表 R 中所有分类属性的信息增益率的平均值,定义为这个表对分类的贡献度。

定义 3(表的贡献度) 假设 R 是通过 ID 传播后的关系表,A_i 是 R 中的某一属性。关系表 R 的贡献度则定义为:

$$contribution(R) = \operatorname{avg}_{A_i \in R} gainratio(A_i) \tag{5}$$

在多关系数据中,从目标表开始,其他表与目标表直接或者间接相连,由于多关系数据库复杂的关系,可能会产生多个分支。在经过元组 ID 传播之后,每个背景表都有了 ID 属性,背景表与目标表之间都可以直接互相连接。为了使表与表之间的关系简单明了,并且能够得到最好的分类准确率,本文从目标表开始,将背景表根据其贡献度的大小按照递减的顺序依次线性相连,之后按照这个线性顺序遍历多关系表。

由于本文的最终目标是得到最好的分类准确度,因此使用训练集来帮助训练出最终用于分类的关系表。在根据背景表贡献度的大小线性排列之后,不论背景表是否与目标表直接相连,都可以通过计算这个表中每个分类属性属于每个类别的概率信息,来对目标表中每个训练元组进行分类,并得到一个分类结果。贡献度的大小代表了这个表的分类能力,根据贡献度的大小依次对余下表进行遍历并计

算其概率信息,最终获得使分类结果最好的那些关系表。

以 PKDD CUP99 金融数据为例,在删除表 order 和 disp 之后,对余下的关系表,根据背景表贡献度的大小,可得到下面线性连接关系:loan→account→card→trans→district→client。之后,根据表 loan 中的数据信息完成对目标表中所有训练元组的分类任务,由此获得分类准确度 accuracy1;然后再利用 loan 和 account 这 2 个表中的信息对目标元组再进行分类,得到分类准确度 accuracy2。由此下去,根据上述线性顺序依次加入其余表,并对目标元组进行分类,分别得出每个分类准确度。当加入某一个表后,分类准确度达到最大的时候,再加入其他表,准确度反而变小,就保留这个表以及之前加入的表,而删除其余表。例如,如果在加入 trans 表后,分类准确度达到最大值,则就删除 client 和 district 表。而最终用于分类的关系表就是 loan,account,card 和 trans 表。

4　基于关系选择的多关系分类算法

MRS 算法的具体步骤如下:

输入　n 个关系表,其中 R_t 是目标表;阈值 λ

输出　多个关系表 $T_i, T_{i+1}, \cdots, T_j$

第 1 步　计算目标表的贡献度。

第 2 步　利用元组 ID 传播技术,将类标签和 ID 传递到背景表中。

第 3 步　根据式(4)计算每个背景表的最大信息增益率;该步骤的时间复杂度为 $O(n-1)$。

第 4 步　给定阈值 λ,删除最大信息增益率小于 λ 的背景表。

第 5 步　根据式(5)计算背景表的贡献度;该步骤的时间复杂度为 $O(n-1)$。

根据它们的贡献度,将剩余的 m 个关系表按照贡献度的大小已递减顺序连接,并按照这个顺序进行遍历,结果是 $b[1]b[2]\cdots b[m]$。

第 6 步　给定集合 S,且 $S=b[1]$。

第 7 步　利用 *collectandclassify*(S)函数收集概率信息并对目标表中的元组进行分类,并得到一个分类准确率。

第 8 步　依次加入剩余背景,并对目标元组进行分类,最终得到一个最大的分类准确率,即当加入第 $j(j<m)$ 个背景表时,分类准确率达到最大。

以上就是该算法执行的主要过程。根据最后一步中前 j 个关系表就是最终用于分类的关系表。

在本文中使用基于语义关系图的多关系贝叶斯分类算法[13]对无标签元组进行分类。

5　实验结果与分析

为了对该算法的有效性进行验证,本文在 Windows XP 操作系统上完成其实验过程。CPU 是 Intel(R) Core(TM)2 Duo E75000 2.93 GHz,内存是 1.96 GB。算法是在 eclipse 的 Java 语言上完成的。实验数据是来自 PKDD CUP 99 的金融数据。为了能有效地验证该算法的性能,文中采用了 5 倍交叉验证法。实验分为以下 4 种情况:

(1)数据离散化对比

为了使 PKDD CUP99 金融数据集满足于多关系朴素贝叶斯分类算法的需求,需要对该原始数据集进行离散化,本文在数据预处理阶段对比了 2 种离散化方法,分别是信息熵离散化方法以及 Weka 中有监督的离散化方法。将数据离散化之后,利用 MRS 算法对目标元组进行分类,结果如图 2 所示。

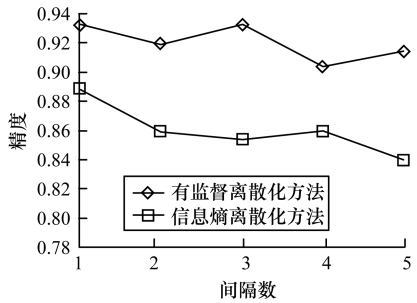

图 2　不同离散化方法的分类结果

由图 2 可知,对于使用不同离散化方法的数据而言,分类精度会在加入不同的关系时达到最大值,同时分类结果还有着很大的差别。使用 Weka 中有监督的离散方法对原始数据预处理,则最终的平均分类精度能够达到 90% 以上;而使用信息熵离散方法时,平均分类精度在 86% 左右。造成这种现象的原因是因为 Weka 中这种有监督的方法本身就对分类有很大的影响,无法有效地验证算法的有效性。因此,本文采用信息熵的离散化方法对实验中数据集进行预处理。

(2)阈值设定

本文在对多关系表进行第一轮删减的时候使用了关于最大信息增益率的阈值 λ,用以删除对分类影响较小的关系表。对于不同的阈值,算法的时间性能和准确率可能会不同,图 3、图 4 是针对不同阈值的对比实验结果。

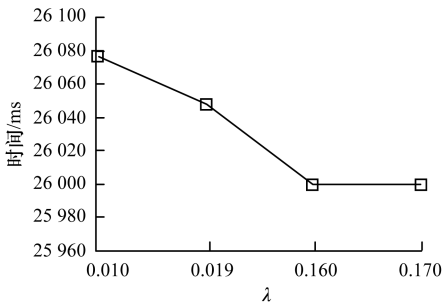

图 3 不同 λ 下的时间性能比较

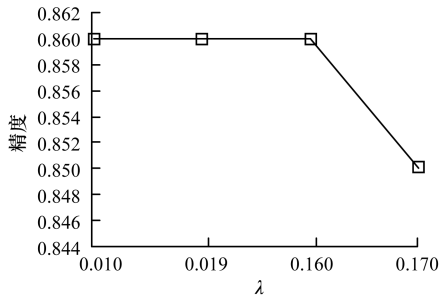

图 4 不同 λ 下的准确率比较

实验结果表明,在 $\lambda=0.160$ 时,时间性能最好,并且准确率最大。因此,本次实验选取 $\lambda=0.160$。

(3)数据规模对比

为了更有效地验证该算法的有效性,本节从数据规模大小的角度对分类效率进行了对比实验。将该数据集平均分为 5 等份,分别以 1 倍规模 ~4 倍规模大小的数据参与分类任务。各规模大小的数据对分类的影响结果如图 5 所示。

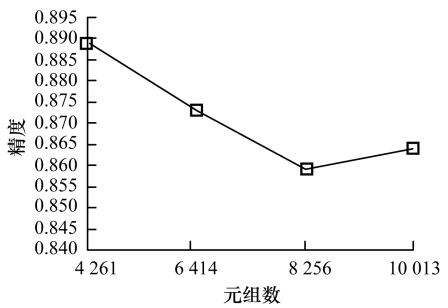

图 5 不同数据规模大小的对比实验结果

该算法的分类精确度随着数据规模的大小先下降后上升,在所有关系元组数目之和为 3 倍规模,即 8 256 个元组时分类准确率较低,但从整体上看,不同规模大小的数据集所进行分类的准确率是比较稳定的。

(4)分类性能比较

将实验结果与 Graph-NB 算法、Classify_tables 算法以及 MRNBC-W 算法进行比较。实验结果如表 1、图 6 所示。由图 6、表 1 可知,MRS 算法在整体上的分类准确度都比这 3 种算法高,平均准确率分别比另外 3 种算法高出了 2.2%,1.1%,0.86%。而且从整体上看,另外 2 种算法的分类准确度有着很大的波动,尤其是 Graph-NB 算法,而 MRS 算法基本上比较稳定,进一步验证了该算法的有效性。

表 1 PKDD CUP99 数据集的分类准确率比较

间隔	Graph-NB	Classify_tables	MRN BC-W	MRS
1	0.840	0.880	0.850	0.888
2	0.880	0.850	0.845	0.859
3	0.790	0.818	0.852	0.854
4	0.804	0.838	0.850	0.860
5	0.878	0.860	0.860	0.840
平均值	0.838	0.849	0.852	0.860

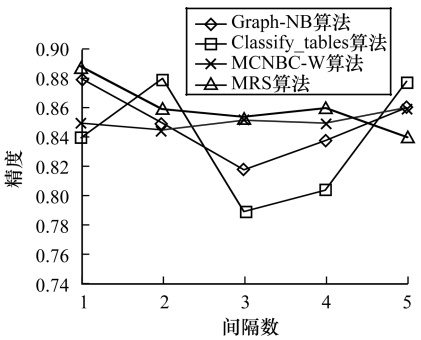

图 6 PKDD CUP 99 数据集的准确率

6 结束语

本文针对多关系数据库中的背景表对分类任务有不同大小的贡献度这一问题,提出一种基于关系选择的多关系朴素贝叶斯分类算法 MRS。本文通过一定的评价标准对多关系数据库进行约减,选择出最终用于分类的关系表,并用多关系朴素贝叶斯进行了测试。实验结果表明,该算法分类准确率有了一定的提高,使得用户下一步的决策更加准确。但是在根据背景表的最大信息增益对多关系表进行剪枝时,如果背景表属性分布不平衡,可能会误删背景表,对分类结果也会造成一定的影响。因此,在今后的研究中,将对关系表的贡献度评价标准做进一步的研究。

参考文献

[1] Kramer S, Lavrac N, Flach P. Propositionalization Approaches to Relational Data Mining[M]. Berlin, Germany: Springer-Verlag, 2001.

[2] Taskar B, Segal E, Koller D. Probabilistic Classification and Clustering in Relational Data[C]//Proceedings of 17th International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan Kaufmann, 2001: 870-876.

[3] Neville J, Jensen D, Gallagher B, et al. Simole Estimators for Relational Bayesian Classifiers [C]// Proceedings of International Conference on Data Mining. Melbourne, Australia: [s. n.], 2003: 609-612.
[4] Blockeel H, de Raedt L, Ramon J. Top-down Induction of Logical Decision Trees[C]//Proceedings of the 15th International Conference on Machine Learning. Madison, USA: [s. n.], 1998: 285-297.
[5] Leiva H A, Gadia S. MRDTL: A Multi-relational Decision Tree Learning Algorithm [C]//Proceedings of the 13th International Conference on Inductive Logic Programming. Berlin, Germany: Springer, 2002: 38-56.
[6] Guo Hongyu, Viktor H L. Multirelational Classification: A Multiple View Approach [J]. Knowledge and Information Systems, 2008, 17(3): 287-312.
[7] 郑利雄,陈　琼,沈勇明. 基于多视图树的多关系分类[J]. 计算机工程,2010,36(19):96-98.
[8] Yin Xiaoxin, Han Jiawei. CrossMine: Efficient Classification Across Multiple Database Relations [C]// Proceedings of 2004 International Conference on Data Engineering. Boston, USA: IEEE Press, 2004: 172-195.
[9] 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,2008,2(1): 1-11.
[10] 郭景峰,苏　哲,刘　强,等. 基于神经网络的多关系数据分类算法[J]. 计算机科学,2008,35(10):94-111.
[11] He Jun, Liu Hongyan, Hu Bo, et al. Selecting Effective Features and Relations for Efficient Multi-relational Classification [J]. Computational Intelligence, 2010, 26(3): 258-280.
[12] Domingos P, Pazzani M. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier[C]//Proceedings of the 13th International Conference on Machine Learning. Madison, USA: [s. n.], 1996: 105-112.
[13] Li Yun, Luan Luan. Multi-relational Classification Based on the Contribution of Tables [C]//Proceedings of International Conference on Artificial Intelligence and Computational Intelligence. Shanghai, China: [s. n.], 2009: 370-374.
[14] 徐广美,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.
[15] 惠　孛,吴　跃. 基于全局的即时垃圾邮件过滤模型的研究[J]. 电子测量与仪器学报,2009,23(5):46-51.
[16] Ouarda W, Trichili H. Combined Local Features Selection for Face Recognition Based on Naive Bayesian Classification [C]//Proceedings of 2013 International Conference on Hybrid Intelligent Systems. Gammarth, Tunisia: [s. n.], 2013: 240-245.
[17] Han Jiawei, Kamber M, Pei J. Data Mining: Concepts and Techniques [M]. San Francisco, USA: Morgan Kaufmann, 2006.

编辑　索书志

(上接第 217 页)

参考文献

[1] Resinick P, Varian H R. Recommender Systems [J]. Communications of the ACM, 1997, 40(3): 56-58.
[2] 王国霞,刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用,2012,48(7):66-76.
[3] 刘枚莲,刘同存,李小龙. 基于用户兴趣特征提取的推荐算法研究[J]. 计算机应用研究,2011,28(5): 1664-1667.
[4] 孙远帅,陈　垚,刘向荣,等. 基于项目层次相似性的推荐算法[J]. 山东大学学报:工学版,2014,44(3): 8-14.
[5] 张建锋,韩伟红,樊　华,等. 基于用户反馈的 top-k 查询修改算法[J]. 计算机研究与发展,2014,51(10): 2206-2215.
[6] 李大学,谢名亮,赵学斌. 基于朴素贝叶斯方法的协同过滤推荐算法[J]. 计算机应用,2010,30(6): 1523-1526.
[7] 杨福萍,王洪国,董树霞,等. 基于记忆效应的协同过滤推荐算法[J]. 计算机工程,2012,38(23):63-66.
[8] 许霄峰,徐炜民. 基于认知复杂度度量的文本推荐模型[J]. 计算机工程与设计,2012,33(10):3990-3994.
[9] 任姚鹏,陈立潮,张英俊,等. 基于潜在语义分析的构件聚类改进方法[J]. 计算机工程,2011,37(4): 67-69.
[10] Somlo G, Howe A. Adaptive Lightweight Text Filtering[C]//Proceedings of the 4th International Symposium on Intelligent Data Analysis. Fort Collins, USA: [s. n.], 2001: 319-329.
[11] Zhang Yi, Callan J, Minka T. Novelty and Redundancy Detection in Adaptive Filtering[C]//Proceedings of the 25th Annual International ACM SIGIR Conference. Pittsburgh, USA: ACM Press, 2002: 81-88.
[12] Robert M G. Entropy and Information Theory [M]. Berlin, Germany: Springer, 2011.
[13] 郭红玉. 基于信息熵理论的特征权重算法研究[J]. 计算机工程与应用,2013,49(10):140-146.
[14] 张玉芳,万斌候,熊忠阳. 文本分类中的特征降维方法研究[J]. 计算机应用研究,2012,29(7):2541-2543.
[15] Salton G. Automatic Text Processing[M]. New York, USA: Addison Wesley Publishers, 1989.
[16] 程　明. 基于文本分类与用户兴趣的个性化搜索与推荐的研究与实现[D]. 广州:中山大学,2006.
[17] Han Jiawei, Kamber M, Pei J. Data Mining: Concepts and Techniques[M]. Burlington, USA: Morgan Kaufmann Publishers, 2011.

编辑　索书志