信息增益区分频繁模式分类方法

陶剑文[1,2],赵杰煜[2],姚奇富[1]
TAO Jian-wen[1,2],ZHAO Jie-yu[2],YAO Qi-fu[1]

1.浙江工商职业技术学院 信息工程系,浙江 宁波 315012
2.宁波大学 信息科学与工程学院,浙江 宁波 315211
1.Department of Information Engineer,Zhejiang Business Technology Institute,Ningbo,Zhejiang 315012,China
2.College of Information Science and Engineering,Ningbo University,Ningbo,Zhejiang 315211,China
E-mail:tjw@zjbti.net.cn

TAO Jian-wen,ZHAO Jie-yu,YAO Qi-fu.Frequent pattern classification method based on information gain.Computer Engineering and Applications,2009,45(7):159-163.

Abstract: The application of frequent patterns in classification appeared in sporadic studies and achieved initial success in the classification of relational data,text documents and graphs.This paper,conducts a systematic exploration of information gain based frequent pattern classification,and provides solid reasons supporting this methodology.By building a connection between pattern frequency and discriminative measures such as information gain,and also develops a strategy to set minimum support in frequent pattern mining for generating useful patterns.Based on this strategy,coupled with a proposed feature selection algorithm,discriminative frequent patterns can be generated for building high quality classifiers.The paper demonstrates that the information gain based frequent pattern classification framework can achieve good scalability and high accuracy in classifying large datasets.
key words: information gain;frequent pattern;classification;discriminative measure

摘 要:基于频繁模式的分类应用研究尚处于初始阶段,但其在关系数据、文本文档与图等方面的分类应用已取得初步成果。系统地研究了基于信息增益区分的频繁模式分类问题,提出了一种基于信息增益区分的频繁模式分类模型(IGFPC),从理论上论证了该模型的可行性。通过建立模式频率与基于信息增益区分度量间的联系,提出了一种在挖掘有用频繁模式上设置最小支持度阀值的方法,基于该方法和提出的特征选择算法(IGPS),生成用以构建高质量模式分类器的区分频繁模式。实验研究显示基于信息增益区分的频繁模式分类框架模型能在分类大数据集上达到较好的扩展性能和较高的分类精度。
关键词:信息增益;频繁模式;分类;区分方法
DOI:10.3778/j.issn.1002-8331(2009)07-0159-05 **文章编号:**1002-8331(2009)07-0159-05 **文献标识码:**A **中图分类号:**TP311

频繁模式挖掘一直以来是数据挖掘研究的热点。目前已产生大量用于不同种类模式挖掘的可扩展的方法,包括项集挖掘[1-3]、序列模式挖掘[4-6]、图模式挖掘[7-8]等。频繁模式挖掘已应用于许多领域,如相关规则挖掘、索引、聚类等[9-11]。频繁模式在分类上的应用也取得一定的成功,如在关系数据、文本数据、图等方面的分类应用[12-18]。

频繁模式反映了项间的强关联性,且其携有潜在的数据语意信息。基于频繁模式的分类思想已在许多不同领域得到应用,包括:(1)相关分类[12-16],其通过产生并分析相关规则来进行分类;(2)图分类[18]、文本分类[17]和蛋白分类[19]。所有这些相关研究与应用在一定程度上显示频繁模式在分类上的可用性。

通过将数据映射到高维空间,特征合并可用于分类,但是,目前基于特征合并的分类方法至少存在两个问题:(1)由于合并特征数是单一特征数的指数级,故枚举所有特征的计算复杂度相当高;(2)稀有特征(不具有表征力的特征)的出现将会降低分类精度(称为过拟合现象)。

通过分析发现,由于在数据集中的有限覆盖率,一个低支持度特征的区分力将受限于某个较低值,从而低支持度的特征在分类上的作用是有限的。利用这个现象,可将频繁模式用于分类。另外,现有的频繁模式挖掘算法能方便地产生模式,从而可以解决面对大数据集时的扩展问题。对于频繁模式挖掘中的最小支持度(min_sup)阀值设置问题,可通过建立一个支持度阀值与区分方法(如信息增益)间的映射,使得由信息增益阀值过滤的特征不会超过相应的 min_sup 阀值。

由于频繁模式的产生仅基于模式频度,而没有考虑模式的预测力,故不经特征过滤的频繁模式的应用仍会导致巨大的特

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.NSFC-60273094);宁波市自然科学基金(the Natural Science Foundation of Ningbo of China under Grant No.2006A610012)。
作者简介:陶剑文(1973-),男,副教授,主要研究方向:模式识别、数据挖掘等;赵杰煜(1965-),男,博士,教授,主要研究方向:人工智能、模式识别;姚奇富(1965-),男,教授,主要研究方向:计算机网络、软件工程。
收稿日期:2008-01-14 **修回日期:**2008-04-14

征空间。这不但会降低模型学习进程,而且会使分类精度恶化。

提出一种基于信息增益区分的频繁模式分类模型(IGFPC),首先通过现有的频繁模式挖掘算法产生模式,然后通过设计的模式选择算法过滤该模式空间,最后通过学习模型对模式进行分类。

相较同类研究,本文的贡献在于:

(1)提出一种基于信息增益的频繁模式分类模型:IGFPC。通过分析模式频度与其预测力间的关系,本文显示频繁模式可提供高质量的分类特征;

(2)IGFPC 利用现有的频繁模式挖掘算法来产生特征,以此达到较好的可扩展性;

(3)在 IGFPC 与基于信息增益特征选择算法间建立了一种形式化的连接关系。该关联性显示 min_sup 阈值与信息增益阈值在过滤低质特征上是相等的,从而提供一个设置 min_sup 阈值的策略;

(4)提出了一个有效的特征选择算法,其选择一个频繁且具有区分力的模式用于分类。

1 相关研究

基于频繁模式的分类与相关分类相关,在后者中,分类器基于高置信度、高支持度的相关规则建立[12-16]。频繁模式与类标签间的相关性用于分类预测。

HARMONY[16]是一个基于规则的分类器,其直接挖掘分类规则。HARMONY 使用实例中心的规则产生方法,对于每个训练实例,具有最高置信的规则被包含在规则集中。HARMONY 较以前的基于规则的分类器具有良好的性能和可扩展性。

本文研究工作有别于相关规则分类的方面包括:(1)利用频繁模式表征一个不同特征空间的数据,其中,任何学习算法可用,而相关规则分类仅采用规则来构建分类模型;(2)在相关规则分类中, 预测过程是发现 top-k 个排序规则用以预测,而本文方法的预测由分类模型完成;(3)通过建立一个与基于信息增益特征选取方法的连接关系,提出一种设置 min_sup 阈值的机制,同时提出一种新颖的特征选取算法。

其它相关研究工作包括:在 NLP 中利用串核(string kernels)[17,19]或字合并分类、利用结构化特征进行图分类[18]。在上述这些研究中,频繁模式被产生且数据被映射到一个较高维特征空间,从而在原始空间非线性可分数据变为在映射空间线性可分。

2 相关概念

设某数据集有 k 个类别属性,某个属性具有一个值集。m 个类集 $C=\{c_1,c_2,\cdots,c_m\}$。每个属性-值对(att,val)被映射为一个唯一的项 $i\in I=\{o_1,o_2,\cdots,o_d\}$。设某个$(att,val)$->$o_i$,这里 att 指属性,val 指一个值。设 x 为一数据点 s 的特征向量,则当 $att(s)=val$ 时,$x_i=1$;否则,$x_i=0$。对于数字属性,连续值需首先进行离散化处理。根据映射,在 B^d 中的数据集表示为 $D=\{x_i,y_i\}_{i=1}^n$,这里 $x_i\in B^d,y_i\in C$。$x_{ij}\in B=\{0,1\}(i\in[1,n],j\in[1,d])$。

定义 1(合并特征) 一个合并特征 $a=\{o_{a1},o_{a2},\cdots,o_{ak}\}$为 I 的一个子集,这里 $o_{ai}\in\{o_1,o_2,\cdots,o_d\}(i\in[1,k])$。$o_i\in I$ 为单一特征。给定一数据集 $D=\{x_i\}$,包含 a 的数据集合表示为 $D_a=\{x_i|x_{iaj}=1,o_{aj}\in a\}$。

定义 2(频繁合并特征) 对于一个数据集 D,如果 $\theta=\frac{|D_\alpha|}{|D|}\geqslant\theta_0$,则合并特征 α 是频繁的,这里 $\theta=\frac{|D_\alpha|}{|D|}$称为 α 在 D 中的相对支持度,θ_0 为最小支持度(min_sup)阈值,且 $0\leqslant\theta_0\leqslant1$。频繁合并特征集记为 $\mathcal{F}$。

给定数据集 $D=\{x_i,y_i\}_{i=1}^n$ 及频繁模式集 $\mathcal{F}$,D 被映射为一个具有 $d'(d'=|I\cup\mathcal{F}|)$个特征的高维特征空间 $B^{d'}$,这里频繁模式集由 θ_0 参数调节。

基于信息增益的频繁模式分类就是在 $I\cup\mathcal{F}$ 特征空间上学习一种分类模型,频繁模式按照最小支持度阈值 θ_0 产生。

设所有项的集合 $I=\{o_1,o_2,\cdots,o_d\}$,I 的非空子集称为项集。一个事务数据集 $D=\{t_1,t_2,\cdots,t_n\}$是项集的集合,这里 $t_i\subseteq I$。对于任意项集 α,则包含 α 的事务集为 $D_\alpha=\{i|\alpha\subseteq t_i$ 且 $t_i\in D\}$。项集 α 的基数$|\alpha|$定义为 α 所包含的项目的数目,即$|\alpha|=\{o_i|o_i\in\alpha\}$。

定义 3(模式的信息增益上界) 设模式 a(由随机变量 X 表征)的信息增益表示为:

$$IG(C|X)=H(C)-H(C|X) \tag{1}$$

式中 $H(C)$指商(entropy),$H(C|X)$指条件商。对于给定的具有固定类分布的数据集,$H(C)$为一常量。模式 a 的信息增益上界 IG_{ub} 定义为:

$$IG_{ub}(C|X)=H(C)-H_{lb}(C|X) \tag{2}$$

式中 $H_{lb}(C|X)$指信息增益的下界。

3 IGFPC 的可行性分析

频繁模式是一种基于单一特征集非线性特征合并形式,由于非线性特征合并的存在,新的特征空间的表征力将增强。另外,由于合并特征携有更多的潜在数据语意,某些频繁模式的区分力将高于单一特征的区分力。从而,得到如下定理:

定理 1 设模式 a 的支持度为 θ,则 a 的信息增益上界 $IG_{ub}(C|X)$与支持度 θ 密切相关,当 θ 小时,$IG_{ub}(C|X)$值低,即非频繁特征具有较低的信息增益上界。

证明 为了简化分析,设 $X\in\{0,1\}$,$C=\{0,1\}$,令 $P(x=1)=\theta$,$P(c=1)=p$,$P(c=1|x=1)=q$。则:

$$\begin{aligned}H(C|X)=&-\sum_{x\in\{0,1\}}P(x)\sum_{c\in\{0,1\}}P(c|x)\log P(c|x)=\\&-\theta q\log q-\theta(1-q)\log(1-q)+(\theta q-p)\log\frac{p-\theta q}{1-\theta}+\\&(\theta(1-q)-(1-p))\log\frac{(1-p)-\theta(1-q)}{1-\theta}\end{aligned} \tag{3}$$

由式(3)知,$H(C|X)$为 p、q 和 θ 的函数。对于给定数据集,p 为一固定值。$H(C|X)$是一凹形函数,当 $q=0$ 或 1 时,如果 $\theta\leqslant p$,$H(C|X)$达到其下界;当 $q=p/\theta$ 或 $1-(1-p)/\theta$ 时,若 $\theta>p$,$H(C|X)$达到其下界。$\theta\leqslant p$ 的情况与 $\theta\geqslant p$ 的情况对称,基于论文空间考虑,本文仅讨论 $\theta\leqslant p$ 的情况。

又由于 $q=0$ 和 $q=1$ 对于 $\theta\leqslant p$ 的情况是对称的,仅讨论 $q=1$ 的情况,下界 $H_{lb}(C|X)$为

$$H_{lb}(C|X)|_{q=1}=(\theta-1)\left(\frac{p-\theta}{1-\theta}\log\frac{p-\theta}{1-\theta}+\frac{1-p}{1-\theta}\log\frac{1-p}{1-\theta}\right) \tag{4}$$

对式(4)求 θ 的偏导数得:

$$\frac{\partial H_{lb}(C|X)|_{q=1}}{\partial\theta}=\log\frac{p-\theta}{1-\theta}-\frac{p-1}{1-\theta}-\frac{1-p}{1-\theta}=\log\frac{p-\theta}{1-\theta}\leqslant\log\ 1=0 \tag{5}$$

以上分析显示信息增益上界 $IG_{ub}(C|X)$是支持度 θ 的函数。$H_{lb}(C|X)|_{q=1}$ 随着 θ 单调递减,即 θ 越小,$H_{lb}(C|X)$越大,而 $IG_{ub}(C|X)$越小,从而可知频率较小的模式的区分力受限于一个较小值。

证毕。

推论 1 根据定理 1,由对称性得出结论:在 $\theta \geqslant p$ 的情况下,较高频率模式的区分力受限于一个较低值(证明从略)。

由上述分析可知,IGFPC 是一种有效的、可扩展的分类方法,其构建一种连接关系,将基于信息增益区分的特征选择算法与基于特征的分类方法很好地联系起来,以达到有效分类的目的。

根据定理 1,总能找到一个 min_sup 阈值 θ^*,其满足:

$$\theta^*=\text{argmax}|\theta(IG_{ub}(\theta)\leqslant IG_0) \quad (6)$$

这里 $IG_{ub}(\theta)$指在支持度 θ 下的信息增益上界,即 θ^* 是最大的支持度阈值,在该点上信息增益上界不再大于 IG_0。

本文特征选择算法过滤所有信息增益小于 IG_0 合并特征,从而,在基于频繁模式的分类方法中,支持度 $\theta \leqslant \theta^*$ 的特征可以安全地过滤掉,因为 $IG(\theta)\leqslant IG_{ub}(\theta)\leqslant IG_{ub}(\theta^*)\leqslant IG_0$。

4 IGFPC 算法设计

采用现有的频繁模式挖掘算法产生候选模式集,然后将设计的基于信息增益区分的模式过滤算法应用于候选模式集,产生新的模式空间,用于训练产生模式分类器。

本文 IGFPC 算法设计主要包括 IGFPC 模型整体实现算法、min_sup 设置算法和特征过滤算法 IGPS,以下将分别描述。

4.1 IGFPC 模型算法

IGFPC 模型主要包括 3 个模块:特征产生、特征过滤、分类模型学习,如图 1 所示。IGFPC 算法实现描述如下。

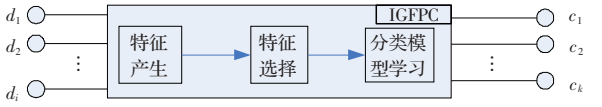

图 1 IGFPC 模型

算法 1 IGFPC 算法

输入:待分类的数据集 $d=\{d_1,d_2,\cdots,d_i\}$;

输出:多个模式分类 $c=\{c_1,c_2,\cdots,c_k\}$;

(1)在特征产生阶段,频繁模式通过一个用户指定的 min_sup 产生,数据 d 根据类标签分区,利用 min_sup,在各分区中产生频繁模式,生成的频繁模式集 $\mathcal{F}$ 构成特征候选集;

(2)在特征过滤阶段,一个特征选择算法 IGPS 应用于 $\mathcal{F}$,特征过滤后产生的选择特征集为 $\mathcal{F}_s$, 从而数据集 D 被转化为 $D'(D'\in B^{\mathcal{F}'})$,此时特征空间 $B^{\mathcal{F}'}$包括单一特征和选择特征集 $\mathcal{F}_s$;

(3)在模型学习阶段,将构建一个基于数据集 D'的特征分类模型;

(4)基于构建的特征分类模型输出模式分类 c。

4.2 min_sup 设置算法

IGFPC 算法实现的关键是如何合理设置 min_sup 参数值。根据定理 1 与推论 1 可知,支持度 θ 太低或太高均导致模式的区分力下降,从而影响分类的精度。提出一种合理设置min_sup 值的算法,具体步骤如下:

算法 2 min_sup 设置算法

(1)在给定类分布 p 的情况下,作为支持度 θ 的函数,计算信息增益的理论上界 $IG_{ub}(\theta)$;

(2)按照文献[20]方法,选择一个信息增益阈值 IG_0 用作特征过滤;

(3)根据式(6),计算 θ^*;

(4)根据 min_sup=θ^* 挖掘频繁模式。

4.3 特征选择算法(IGPS)

尽管频繁模式用于分类是可行的,但并非每个频繁模式的分类效能等同,有必要在用于分类之前,对候选模式集进行一定的模式预处理,即模式过滤,以产生一个具有区分力的特征子集。本文借用文献[21]中的最大边沿相关(Maximal Marginal Relevance,MMR)概念,提出一种用于区分模式选择的算法 IGPS。在实现 IGPS 算法之前,本文先定义分类上下文中的频繁模式相关与冗余的概念:

定义 4(模式相关) 一个相关性度量 S 是一个函数,其映射某个模式 a 到一个实值,使得 $S(a)$是相对于类标签的相关。

定义 5(模式冗余) 一个冗余性度量 R 是一个函数,其映射两个模式 a 与 β 到一个实值,使得 $R(\alpha,\beta)$是 α 与 β 间的冗余。

模式冗余度量两个模式间的相似程度。采用一个改进的 Jaccard 度量[22]去测量两个不同模式间的冗余:

$$R(\alpha,\beta)=\frac{P(\alpha,\beta)}{P(\alpha)+P(\beta)-P(\alpha,\beta)}\times\min(S(\alpha),S(\beta)) \quad (7)$$

根据定义 5, 对于一个闭合模式 α 及其非闭合子模式 β,β 相对于 α 是完全冗余的。故本文算法采用闭合频繁模式[3]作为特征项。

IGPS 算法启发式搜索特征空间, 如果某个特征与给定的类标签相关且其包含与已选特征非常低的冗余,则该特征被选择。初始阶段,具有最高冗余度量的特征被选取,接着,IGPS 算法利用评估的信息增益 g 递增地从 $\mathcal{F}$ 中选择更多模式。如果某个模式在剩下的模式中具有最大的增益,则其被选取。在给定已选模式集 $\mathcal{F}_s$ 下,一个模式 α 的增益计算公式为:

$$g(\alpha)=S(\alpha)-\max_{\beta\in F_s}R(\alpha,\beta) \quad (8)$$

为了确定用于有效分类的频繁模式选择数,本文借用引文[13]的做法,在 IGPS 算法中加入一个数据库覆盖约束因子 δ,设置该参数以确保各训练实例被所选特征至少覆盖 δ 次,这样,在某个用户指定的参数值 δ 下,所选特征数便被自动确定下来。IGPS 算法具体描述如算法 3 所示。

算法 3 特征选择算法 IGPS

输入:频繁模式 $\mathcal{F}$,覆盖阈值 δ,相关性度量 S,冗余性度量 R;

输出:一个选择特征集 $\mathcal{F}_s$。

(1)Let a be the most relevant pattern;

(2)$\mathcal{F}_s=\{a\}$;

(3)while(true)

(4) Find a pattern β such that the gain $g(\beta)$ is the maximum among the set of patterns in $\mathcal{F}-\mathcal{F}_s$;

(5) If β can correctly cover at least one instance

(6) $\mathcal{F}_s=\mathcal{F}_s\cup\{\beta\}$;

(7) $\mathcal{F}=\mathcal{F}-\{\beta\}$;

(8) If all instances are covered δ times or $\mathcal{F}=\phi$

(9) break;

(10) return $\mathcal{F}_s$。

5 实验及分析

5.1 实验设置

实验数据集取自 UCI 机器学习数据库(http://www.ics.uci.edu/~mlearn/MLRepository.html), 对于连续属性进行离散化处理,利用 FPClose[23]算法产生闭合模式,本文提出的 IGPS 算法进行特征选择,选择 Weka[24]中的 LIBSVM[25]和 C4.5 作为分类模型。实验中,将各数据集平均分为 10 个部分,其中一部分作为

测试集,其它部分作为训练集,对各训练集交替地做 10 次交叉验证,选取最好的模型用于测试,10 次测试数据集的分类精度进行平均处理后产生实验结果。

5.2 IGFPC 实验结果

本文主要测试 IGFPC 的分类性能。对于各数据集,产生一个频繁模式集 $\mathcal{F}$,使用特征集 $I\cup\mathcal{F}$ 构建一个分类模型,表示为 Pat_All;将 IGPS 应用于 $\mathcal{F}$ 以产生 $\mathcal{F}_s$,使用特征集 $I\cup\mathcal{F}_s$ 构建一个分类模型,表示为 Pat_FS。为了进行实验比较,本文分别测试了基于单一特征构建的分类模型(表示为 Item_All)和基于选择特征构建的分类模型(表示为 Item_FS)。表 1 和表 2 分别显示了在 SVM 和 C4.5 下的分类结果。在上述 4 种模型种 SVM 均采用线性核模型。

从表 1 清楚地看出,Pat_FS 在大多情况下达到最优的分类精度,其相较于 Item_All 和 Item_FS 模型有较明显的改进,这也进一步说明了两点:

(1)在将数据映射到高维空间的情况下,频繁模式表现出较好的分类性能;

(2)一些频繁模式的区分力高于单一特征。

另外,从表 1 也能看出 Pat_All 模型的性能远比 Pat_FS 模型差,这说明冗余的、非区分的模式常常过拟合模型且恶化模型分类精度。

表 1 基于 SVM 的频繁合并特征与单一特征精度比较

Data	单一特征		频聚合并特征	
	Item_All	Item_FS	Pat_All	Pat_FS
anneal	99.78	99.78	99.33	99.67
austral	85.01	85.50	81.79	91.14
auto	83.25	84.21	74.97	90.79
breast	97.46	97.46	96.83	97.78
cleve	84.81	84.81	78.55	95.04
diabetes	74.41	74.41	77.73	78.31
glass	75.19	75.19	79.91	81.32
heart	84.81	84.81	82.22	88.15
hepatic	84.50	89.04	81.29	96.83
horse	83.70	84.79	82.35	92.39
iono	93.15	94.30	89.17	95.44
iris	94.00	96.00	95.33	96.00
labor	89.99	91.67	94.99	95.00
lymph	81.00	81.62	83.67	96.67
pima	74.56	74.56	76.43	77.16
sonar	82.71	86.55	84.60	90.86
vehicle	70.43	72.93	73.33	76.34
wine	98.33	99.44	98.30	100.00
zoo	97.09	97.09	94.18	99.00

上述结论同样能从表 2 基于决策树的分类模型中得出。

5.3 算法扩展性实验

为了进一步测试本文方法在良好的分类精度下的扩展性能,从 UCI 机器学习数据库中选择了 3 个稠密数据集:Chess、Waveform 和 Letter Recognition(离散集取自 http://www.csc.liv.ac.uk/~frans/KDD/Software/LUCS-KDD-DN/DataSets),对于各数据,*min_sup*=1 用于枚举所有特征合并,将 IGPS 算法应用这些特征。为了进行比较,实验中对支持度阈值进行变化并观测实验结果,如表 3~表 5 所示。

表 3 显示在变化支持度阈值 min_sup 的情况下,对 Chess 数据实验的结果。Chess 数据包含 3 196 个实例、2 个类、73 个项,#Patterns 显示产生的闭合模式数,Time 显示模式挖掘时间与特征选择时间的总和,表中其他 2 列分别显示在 SVM 与 C4.5 模型下的分类精度值。从表 3 看出,在 *min_sup*=1 下,所有模式枚举不能在合理的时间内完成,从而阻止模型的构建。本文提出的模型优势在于在较高的支持度阈值下,能在秒级时间内完成频繁模式挖掘任务且能达到满意的分类精度。表 4、表 5 显示了对另两种数据的相似实验结论。

表 2 基于 C4.5 的频繁合并特征与单一特征精度比较

Data	单一特征		频聚合并特征	
	Item_All	Item_FS	Pat_All	Pat_FS
anneal	98.33	98.33	97.22	98.44
austral	84.53	84.53	84.21	88.24
auto	71.70	77.63	71.14	78.77
breast	95.56	95.56	95.40	96.35
cleve	80.87	80.87	80.84	91.42
diabetes	77.02	77.02	76.00	76.58
glass	75.24	75.24	76.62	79.89
heart	81.85	81.85	80.00	86.30
hepatic	78.79	85.21	80.71	93.04
horse	83.71	83.71	84.50	87.77
iono	92.30	92.30	92.89	94.87
iris	94.00	94.00	93.33	93.33
labor	86.67	86.67	95.00	91.67
lymph	76.95	77.62	74.90	83.67
pima	75.86	75.86	76.28	76.72
sonar	80.83	81.19	83.67	83.67
vehicle	70.70	71.49	74.24	73.06
wine	95.52	93.82	96.63	99.44
zoo	91.18	91.18	95.09	97.09

表 3 Chess 数据集分类精度与时间

min_sup	#Patterns	Time/s	SVM/(%)	C4.5/(%)
1	N/A	N/A	N/A	N/A
2 000	68 967	44.703	92.52	97.59
2 200	28 358	19.938	91.68	97.84
2 500	6 837	2.906	91.68	97.62
2 800	1 031	0.469	91.84	97.37
3 000	136	0.063	91.90	97.06

表 4 Waveform 数据集分类精度与时间

min_sup	#Patterns	Time/s	SVM/(%)	C4.5/(%)
1	9 468 109	N/A	N/A	N/A
80	26 576	176.485	92.40	88.35
100	15 316	90.406	92.19	87.29
150	5 408	23.610	91.53	88.80
200	2 481	8.234	91.22	87.32

表 5 Recognition 数据集分类精度与时间

min_sup	#Patterns	Time/s	SVM/(%)	C4.5/(%)
1	5 147 030	N/A	N/A	N/A
3 000	3 246	200.406	79.86	77.08
3 500	2 078	103.797	80.21	77.28
4 000	1 429	61.047	79.57	77.32
4 500	962	35.235	79.51	77.42

6 结束语

提出一种基于信息增益区分的频繁模式分类框架模型。证明了基于信息增益区分的频繁模式分类的可行性。本研究表明

频繁模式是高质量的特征信息，其具有较好的模型范化能力。提出了一种设置 min_sup 阀值的方法,设计了一种用于选择区分频繁模式的算法 IGPS。实验结果显示 IGFPC 具有较好的分类精度和扩展性能。

参考文献:

[1] Agrawal R,Srikant R.Fast algorithms for mining association rules[C]//Proc of VLDB,1994:487-499.

[2] Han J,Pei J,Yin Y.Mining frequent patterns without candidate generation[C]//Proc of SIGMOD,2000:1-12.

[3] Zaki M J,Hsiao C.CHARM:An efficient algorithm for closed itemset mining[C]//Proc of SDM,2002:457-473.

[4] Agrawal R,Srikant R.Mining sequential patterns[C]//Proc of ICDE,1995:3-14.

[5] Pei J,Han J,Mortazavi-Asl B,et al.PrefixSpan:Mining sequential patterns efficiently by prefix-projected pattern growth[C]//Proc of ICDE,2001:215-226.

[6] Zaki M J.SPADE:An efficient algorithm for mining frequent sequences[J].Machine Learning,2001,42(1/2):31-60.

[7] Kuramochi M,Karypis G.Frequent subgraph discovery[C]//Proc of ICDM,2001:313-320.

[8] Yan X,Han J.gSpan:graph-based substructure pattern mining[C]//Proc of ICDM,2002:721-724.

[9] Agrawal R,Imielinski T.Mining association rules between sets of items in large databases[C]//Proc of SIGMOD,1993:207-216.

[10] Yan X,Yu P S,Han J.Graph indexing:a frequent structure-based approach[C]//Proc of SIGMOD,2004:335-346.

[11] Wang K,Xu C,Liu B.Clustering transactions using large items[C]//Proc of CIKM,1999:483-490.

[12] Liu B,Hsu W,Ma Y.Integrating classification and association rule mining[C]//Proc of KDD,1998:80-86.

[13] Li W,Han J,Pei J.CMAR:accurate and efficient classification based on multiple class-association rules[C]//Proc of ICDM,2001:369-376.

[14] Yin X,Han J.CPAR:classification based on predictive association rules[C]//Proc of SDM,2003:331-335.

[15] Cong G,Tan K,Tung A,et al.Mining top-k covering rule groups for gene expression data[C]//Proc of SIGMOD,2005:670-681.

[16] Wang J,Karypis G.HARMONY:efficiently mining the best rules for classification[C]//Proc of SDM,2005:205-216.

[17] Lodhi H,Saunders C,Shawe-Taylor J,et al.Text classification using string kernels[J].Journal of Machine Learning Research,2002,2:419-444.

[18] Deshpande M,Kuramochi M,Karypis G.Frequent sub-structure-based approaches for classifying chemical compounds[C]//Proc of ICDM,2003:35-42.

[19] Leslie C,Eskin E,Noble W S.The spectrum kernel:A string kernel for svm protein classification[C]//Proc of PSB,2002:564-575.

[20] Yang Y,Pedersen J O.A comparative study on feature selection in text categorization[C]//Proc of ICML,1997:412-420.

[21] Carbonell J,Coldstein J.The use of mmr,diversity-based reranking for reordering documents and producing summaries[C]//Proc of SIGIR,1998:335-336.

[22] Tan P,Kumar V,Srivastava J.Selecting the right interestingness measure for association patterns[C]//Proc of KDD,2002:32-41.

[23] Grahne G,Zhu J.Efficiently using prefix-trees in mining frequent itemsets[C]//ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03),2003.

[24] Witten I H,Frank E.Data mining:practical machine learning tools and techniques[M].2nd ed.[S.l.]:Morgan Kaufmann,2005.

[25] Chang C C,Lin C J.LIBSVM:a library for support vector machines[EB/OL].(2001).http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[26] Duda R,Hart P,Stork D.Pattern classification[M].2nd ed.[S.l.]:Wiley Interscience,2000.

[27] Quinlan J.C4.5:programs for machine learning[M].[S.l.]:Morgan Kaufmann,1993.

(上接 126 页)

[4] 宋晓莉,王劲松,陈源.信息安全风险评估方法研究[J].网络安全技术与应用,2006,12:67-69.

[5] 孙鹏鹏,张玉清,韩臻.信息安全风险评估工具的设计与实现[J].计算机工程与应用,2007,43(9):95-98.

[6] 史亮,庄毅.一种定量的网络安全风险评估系统模型[J].计算机工程与应用,2007,43(18):146-149.

[7] 李继峰.地理信息系统风险评估[J].现代商贸工业,2007,7:189-190.

[8] 梁洪涛,王大萌,黄俊强,等.信息安全风险评估规范在电子政务中的应用[J].信息技术,2007,7:133-135.

[9] 赵冬梅,刘海峰,刘晨光.基于 BP 神经网络的信息安全风险评估[J].计算机工程与应用,2007,43(1):139-141.

[10] 刘宝利,肖晓春,张根度.基于层次分析法的信息系统脆弱性评估方法[J].计算机科学,2006,12:62-64.

[11] 陈其,陈铁,姚林,等.电力系统信息安全风险评估策略研究[J].计算机安全,2007,6:40-42.

(上接 128 页)

[2] Hwang R J,Lee W B,Chang C C.A concept of designing cheater identification methods for secret sharing[J].Journal of Systems and Software,2000,46(1):7-11.

[3] 石润华,黄刘生.一种简单的可验证秘密共享方案[J].计算机应用,2006,26(8):1821-1823.

[4] 唐春明,刘卓军,王明生.一种实用的可验证秘密共享方案[J].计算机工程与应用,2006,42(15):129-133.

[5] 许春香.安全秘密共享及其应用研究[D].西安:西安电子科技大学,2003.

[6] Chien H Y,Jan J K,Tseng Y M.A practical (t,n) multi-secret sharing scheme[J].IEICE Transaction Fundamentals,2000,83(12):2762-2765.

[7] Sun H.On-line multiple secret sharing based on a one-way function[J].Computer Communications,1999,22(8):745-748.

[8] Dehkordi M H,Mashhadi S.An efficient threshold verifiable multi-secret sharing[J].Computer Standards & Interfaces,2008,30:187-190.