特征的支持度与其分类能力的关系研究

尹建芹[1,2],田国会[3],魏 军[1],李金屏[1],林佳本[2]
(1.济南大学信息科学与工程学院 山东省网络环境智能计算技术重点实验室,山东济南 250022;
2.中国科学院太阳活动重点实验室,北京 100012;3.山东大学控制科学与工程学院,山东济南 250061)

摘 要: 频繁模式挖掘在分类问题中得到了广泛的应用,大量的工作利用频繁模式挖掘对分类问题进行特征选择,但对于为什么频繁模式挖掘可以在分类问题中进行有效的特征选择则缺乏系统的研究.为了为频繁模式挖掘在分类问题中的特征选择应用提供理论基础,需要确立特征的支持度与特征分类能力之间的关系,本文以特征的信息增益作为分类能力的评价准则,讨论其与特征支持度之间的联系.首先证明了信息增益是特征支持度的上凸函数;然后,在二类问题和多类问题情况下,分别证明了具有低支持度或高支持度的特征具有有限的信息增益,即具有低支持度或高支持度的特征具有有限的分类能力.最后,通过仿真实验验证了支持度与信息增益之间的关系,为频繁模式挖掘在分类问题中的应用提供了理论基础.

关键词: 频繁模式; 分类; 特征选择; 信息增益
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2015)02-0248-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.02.007

Research on the Relationship of the Support and the Discriminative Ability for Classification of Features

YIN Jian-qin[1,2], TIAN Guo-hui[3], WEI Jun[1], LI Jin-ping[1], LIN Jia-ben[2]
(1. *Shandong Provincial Key Laboratory of Network Based Intelligent Computing, School of Information Science and Engineering, University of Jinan, Jinan, Shandong* 250022, *China*; 2. *Key Laboratory of Solar Activity, National Astronomical Observatories, Chinese Academy of Sciences, Beijing* 100012, *China*; 3. *School of Control Science and Engineering, Shandong University, Jinan, Shandong* 250061, *China*)

Abstract: Frequent pattern mining is used widely in feature selection for classification problem. In order to provide theoretical basis for the application, we established the relationship between the classification discriminative ability and the support of the feature. Information gain was adopted as evaluation criteria, and we discussed the connection between the support of the feature and its discriminative ability. Firstly, we proved the information gain is a concave function about the support of the feature; secondly, we proved the conclusion that the feature with too-high or too-low support has limited discriminative ability under the two classes and multiple classes circumstances separately; Finally, simulation experiments validate our conclusions. And the conclusion provides a theoretical basis for the application of frequent pattern mining in classification problems.

Key words: frequent pattern; classification; feature selection; information gain

1 引言

近年来,频繁模式挖掘越来越多的被用于分类问题,如文本分类[1~3]、网络模体识别[4]以及蛋白质序列分类等[5,6],且取得了良好的效果.尽管在该方面已经有了相当多的研究,但为什么频繁模式挖掘会对分类问题有效这一问题,却研究较少.Cheng 等[7]对这一问题进行了一定的探讨,其针对二分类问题讨论了模式的支持度与分类能力之间的联系,得出了一个非常重要的结论—低支持度特征或高支持度特征具有有限的分类能力.这一结论具有非常重要的意义:由于在大数据集基础上构建分类器的相关算法的可扩展性是一个非常难于解决的问题,而在数据挖掘领域,已经具有了相当多的可扩展性频繁模式挖掘算法[8~10],因此,如果这一结论成立,则可以方便的构造大数据集上的可扩展性分类算法,从而可以有效的解决大数据集上的分类问题.但 Cheng 的工作的不足之处是其较多的采用了说明的过程,缺少必要的证明过程,另外,还有其对于多类问题未

收稿日期:2013-11-05;修回日期:2014-04-21;责任编辑:马兰英
基金项目:国家自然科学基金(No.61075092,No.61203341,No.61173079);山东省高等学校科技发展计划(No.TJY1112);中科院太阳活动重点实验室开放课题(No.TQT1302)

作讨论,作者认为这也是限制该方法在以后的工作中没有得到研究者足够重视的原因.万里等采用了 Cheng 的结论,建立了基于频繁模式的分类框架,并将其应用于时间序列分类,取得了较好的效果[11].Lee 将频繁模式挖掘应用于路网中的轨迹分类问题[12],由于路网轨迹分类是一个典型的多类问题,尽管对于轨迹分类取得了良好的效果,但其在多类问题中用到的模式支持度与信息增益之间的关系是一个未经证明的结论,事实上其对应了某种情况下的特例.本文以信息增益作为分类的评价准则,证明了 Cheng 有关在二分类问题下模式支持度与分类能力之间的联系,并以此为基础,给出了多分类问题下的支持度与分类能力之间的联系,证明了 Lee 所用结论是一种多分类问题下的一种特殊情况,并以实例验证了本文结论.本文工作为频繁模式挖掘在分类问题中的应用提供了理论基础,并为分类问题中可扩展性算法的寻求提供了较好的选择方案.

2　信息增益与特征支持度

在分类问题中,具有有限的类别区分能力特征的加入会增大分类器的复杂度,从而引起过拟合,降低分类器的精度,因此,特征的类别区分能力是分类问题中特征选择的一个重要依据.常用的类别可分性判据有信息增益,KL 散度等[13],本文以特征的信息增益大小作为类别区分能力的衡量准则,以此为基础研究特征的分类能力与其支持度之间的关系.设特征 f 可以表示为随机矢量集 X,则其信息增益可以表示为[13]:

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

由式(1)可以看出,信息增益事实上反应的是特征能够为分类系统带来的信息量的多少,带来的信息量越多,该特征越重要,反之,特征越不重要.对一个分类问题而言,其整个分类系统的信息量记为 $H(C)$,对特征 X,系统包括它和不包括它时信息量将发生变化,而前后信息量的差值就是这个特征给系统带来的信息量.其中,信息量用熵来表示,$H(C)$为系统的熵,其对于特定的分类问题为常量,$H(C\mid X)$ 为条件熵,可以看出条件熵越小,信息增益越大,分类能力越强.在二值特征中,取 $X\in\{0,1\}$,设类别个数为 num_{cla},$P(x)$表示 x 的支持度,其中,x 为一个特定的组合特征,则 $P(\bar{x})$是 x 没出现的支持度,显然有 $P(x)+P(\bar{x})=1$.且条件熵可以表示为式(2):

$$\begin{aligned}&H(C\mid X)\\&=-P(x)\sum_{i=1}^{num_{cla}}P(c_i\mid x)\log_2P(c_i\mid x)\\&-P(\bar{x})\sum_{i=1}^{num_{cla}}P(c_i\mid\bar{x})\log_2P(c_i\mid\bar{x})\end{aligned} \tag{2}$$

根据贝叶斯公式,可得

$$P(c_i\mid\bar{x})=\frac{[1-P(x\mid c_i)]P(c_i)}{1-P(x)} \tag{3}$$

$$P(x\mid c_i)=\frac{P(c_i\mid x)P(x)}{P(c_i)} \tag{4}$$

将式(4)带入式(3),可得式(5)

$$\begin{aligned}P(c_i\mid\bar{x})&=\frac{\left[1-\frac{P(c_i\mid x)P(x)}{P(c_i)}\right]P(c_i)}{1-P(x)}\\&=\frac{P(c_i)-P(c_i\mid x)P(x)}{1-P(x)}\end{aligned} \tag{5}$$

为了表示的方便,设 $P(c_i\mid x)=p_i$,其中 $i=1,2\cdots num_{cla}$,则

$$\begin{aligned}H(C\mid X)&=-P(x)\sum_{i=1}^{num_{cla}}p_i\log_2p_i\\&-\sum_{i=1}^{num_{cla}}[P(c_i)-P(x)p_i]\log_2\left[\frac{P(c_i)-P(x)p_i}{1-P(x)}\right]\end{aligned} \tag{6}$$

同样,为了推理方便,令 $P(x)=\theta$ 为特征 x 的支持度,则有

$$\begin{aligned}H(C\mid X)&=-\theta\sum_{i=1}^{num_{cla}}p_i\log_2p_i\\&-\sum_{i=1}^{num_{cla}}[P(c_i)-\theta p_i]\log_2\left[\frac{P(c_i)-\theta p_i}{1-\theta}\right]\end{aligned} \tag{7}$$

为了得到特征的分类能力与特征支持度的关系,即要求解式(7)中 $H(C\mid X)$与 θ 之间的关系.而可以方便的证明 $H(C\mid X)$是 θ 的上凸函数,下面给出证明.

$$\begin{aligned}\frac{\partial H(C\mid X)}{\partial\theta}&=-\sum_{i=1}^{num_{cla}}p_i\log_2p_i-\sum_{i=1}^{num_{cla}}\left\{-p_i\log_2\frac{P(c_i)-\theta p_i}{1-\theta}\right.\\&\left.+\left[-\frac{p_i}{1-\theta}+\frac{P(c_i)-\theta p_i}{(1-\theta)^2}\right]\frac{(1-\theta)}{\ln2}\right\}\end{aligned} \tag{8}$$

$$\begin{aligned}\frac{\partial^2H(C\mid X)}{\partial\theta^2}&=-\sum_{i=1}^{num_{cla}}\left\{-p_i\left[\frac{-p_i+P(c_i)}{(1-\theta)}\right]\left[\frac{1}{P(c_i)-\theta p_i}\right]\frac{1}{\ln2}\right.\\&\left.+\left[\frac{-p_i}{(1-\theta)}+\frac{[P(c_i)-\theta p_i]}{(1-\theta)^2}\right]\frac{1}{\ln2}\right\}\\&=-\sum_{i=1}^{num_{cla}}\\&\left\{\frac{p_i^2-p_iP(c_i)-\theta p_i^2+\theta p_iP(c_i)+P(c_i)^2-p_iP(c_i)+\theta p_i^2-\theta p_iP(c_i)}{(1-\theta)^2[P(c_i)-\theta p_i]}\right\}\\&\cdot\frac{1}{\ln2}=-\sum_{i=1}^{num_{cla}}\left\{\frac{[p_i-P(c_i)]^2}{(1-\theta)^2[P(c_i)-\theta p_i]}\right\}\frac{1}{\ln2}\leqslant0\end{aligned} \tag{9}$$

由式(8),(9)可以得出,条件熵 $H(C\mid X)$是特征支持度 θ 的上凸函数.为了证明具有高支持度或低支持度的特征具有有限的分类能力,即需要证明具有高支持度或低支持度的特征具有小的信息增益,也就是其信息增益不会超过特定值,因此,只需要证明条件熵关于支持度的下界大于一定值即可,为了分析的方便,本文分两类问题和多类问题进行讨论.

3　特征分类能力与支持度的关系

要求解条件熵的下界,即求解上凸函数式(7)的下界,

如果函数式(7)的定义域为凸集合,则函数的下界会在边界上取得,因此,我们首先考察函数式(7)的定义域.

3.1 二分类问题下特征的分类能力与其支持度之间的关系

在二分类问题下,其定义域 D 为

$$\begin{aligned} &0\leqslant p_i\leqslant 1\\ &p_1+p_2=1\\ &P(c_i)\geqslant\theta p_i \end{aligned} \tag{10}$$

其中,$i=1,2$.

为了化简方便,可以表示为

$$\begin{aligned} &0\leqslant p_i\leqslant 1\\ &p_1+p_2=1\\ &p_i\leqslant\frac{P(c_i)}{\theta} \end{aligned} \tag{11}$$

由式(11),如果$\frac{P(c_i)}{\theta}\geqslant 1$,则最后一个条件退化.即当 $\theta\leqslant P(c_i)$时,则其定义域 D 为

$$\begin{aligned} &0\leqslant p_i\leqslant 1, i=1,2\\ &p_1+p_2=1 \end{aligned} \tag{12}$$

考虑凸集合的定义[14]:

定义 1 若对于任意的 $x_1,x_2\in S$ 以及任意的 $\alpha\in(0,1)$,有 $x_\alpha=\alpha x_1+(1-\alpha)x_2\in S$,则称集合 S 是凸集合[14].

由上凸函数极值的性质,如果其定义域为凸集合,则上凸函数的最小值在边界点上取得[14].

首先根据定义 1,证明二分类问题下的条件熵的定义域为凸集合.

证明 设式(7)的定义域为 D,则

任取 $x_1,x_2\in R$ 以及任意的 $\alpha\in(0,1)$,其中 $x_1=(p_{11},p_{12})$,$x_2=(p_{21},p_{22})$,则有:$0\leqslant p_{11}\leqslant 1,0\leqslant p_{12}\leqslant 1$,$p_{11}+p_{12}=1,0\leqslant p_{21}\leqslant 1,0\leqslant p_{22}\leqslant 1,p_{21}+p_{22}=1$ 以及

$$\begin{aligned} x_\alpha &=\alpha x_1+(1-\alpha)x_2\\ &=(\alpha p_{11}+p_{21}-\alpha p_{21},\alpha p_{12}+p_{22}-\alpha p_{22}) \end{aligned} \tag{13}$$

由于 $0\leqslant\alpha p_{11}+(1-\alpha)p_{21}\leqslant 1,0\leqslant\alpha p_{12}+p_{22}-\alpha p_{22}\leqslant 1$,且有

$$\alpha p_{11}+p_{21}-\alpha p_{21}+\alpha p_{12}+p_{22}-\alpha p_{22}=1 \tag{14}$$

即 $x_\alpha\in D$,因此,集合 D 是凸集合.即该定义域为一个有界闭凸集合,为了形象化起见,将其定义域绘制于图 1(a)中粗线所示;在其他情况下,按照同样方法,可以证明其定义域为有界闭凸集合,如果$\frac{P(c_i)}{\theta}<1$,则其定义域 D 为:

$$\begin{aligned} &0\leqslant p_i\leqslant\frac{P(c_i)}{\theta}, i=1,2\\ &p_1+p_2=1 \end{aligned} \tag{15}$$

如图 1(d)中粗线所示.其他情况事实上是该两种情况的结合,如图 1($b-c$)所示.

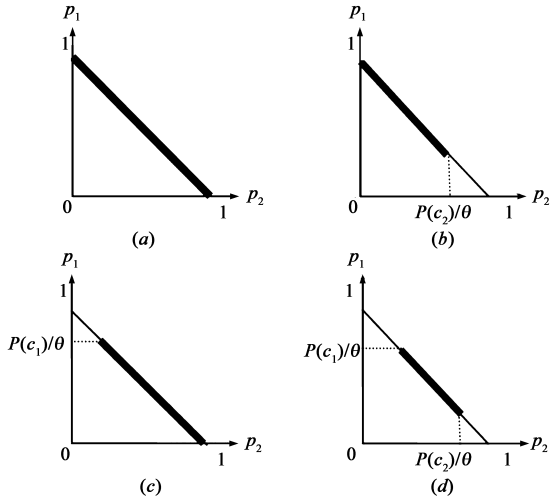

图1 条件熵定义域示意图

由于上凸函数的极值在边界点上取得,因此,条件熵的下界对应了图 1 中的 4 种情况,其极值分别如下:当 $\theta\leqslant\min P(c_i)$时,对应的图 1($a$)中,其下界在边界点上取得,即 $p_1=1,p_2=0$ 或 $p_1=0,p_2=1$ 时取得;当 $\theta\leqslant P(c_1)$,但 $\theta>P(c_2)$时,对应的图 1(b)中,下界在 $p_1=1,p_2=0$ 或 $p_1=1-P(c_2)/\theta,p_2=P(c_2)/\theta$ 时取得;当 $\theta\leqslant P(c_2)$,但 $\theta>P(c_1)$时,对应的图 1(c)中,下界在 $p_1=P(c_1)/\theta,p_2=1-P(c_1)/\theta$ 或 $p_1=0,p_2=1$ 时取得;当 $\theta\geqslant\max P(c_i)$,对应的图 1($d$)中,下界在 $p_1=P(c_1)/\theta,p_2=1-P(c_1)/\theta$ 或 $p_1=1-P(c_2)/\theta,p_2=P(c_2)/\theta$ 时取得.

(1)当 $\theta\leqslant\min P(c_i)$ 时

当 $\theta\leqslant\min P(c_i)$ 时,在 $p_1=1,p_2=0$ 或 $p_1=0,p_2=1$ 时取下界.由于这两种情况是对称的,因此,取第一种情况讨论,将 $p_1=1,p_2=0$ 代入式(7),得到条件熵的下界为:

$$\begin{aligned} H(C|X)_{lb}=&-[P(c_1)-\theta]\log_2\frac{P(c_1)-\theta}{1-\theta}\\ &-[P(c_2)]\log_2\frac{P(c_2)}{1-\theta} \end{aligned} \tag{16}$$

求式(16)关于 θ 的偏导,则得

$$\begin{aligned} &\frac{\partial H(C|X)_{lb}}{\partial\theta}\\ &=\log_2\frac{P(c_1)-\theta}{1-\theta}\\ &\quad+\left[1-\frac{P(c_1)-\theta}{(1-\theta)}-\frac{P(c_2)}{(1-\theta)}\right]\frac{(1-\theta)}{\ln 2}\\ &=\log_2\frac{P(c_1)-\theta}{1-\theta}\leqslant 0 \end{aligned} \tag{17}$$

式(17)说明,当 $\theta\leqslant\min P(c_i)$时,条件熵的下界随着 θ 的增大单调递减,也就是说,θ 越小,条件熵下界越大,从而信息增益上界越小,即支持度较小的特征其信

息增益不大于某个阈值,其值可以根据式(16)计算得到.因此,以信息增益作为类别可分性判据,支持度较小的特征分类能力有限.

(2)当 $\theta \geq \max P(c_i)$ 时

当 $\theta \geq \max P(c_i)$ 时,下界在 $p_1 = P(c_1)/\theta, p_2 = 1 - P(c_1)/\theta$ 或 $p_1 = 1 - P(c_2)/\theta, p_2 = P(c_2)/\theta$ 时取得.同样这两种情况也是对称的,因此,取第一种情况进行证明,将其代入式(7),得到

$$\begin{aligned} & H(C \mid X)_{lb} \\ &= -P(c_1)\log_2 \frac{P(c_1)}{\theta} \\ &\quad -[\theta - P(c_1)]\log_2 \frac{\theta - P(c_1)}{\theta} \end{aligned} \tag{18}$$

求式(18)关于 θ 的偏导,则得

$$\begin{aligned} & \frac{\partial H(C \mid X)_{lb}}{\partial \theta} \\ &= \left\{ \frac{P(c_1)}{\theta} - \log_2 \frac{\theta - P(c_1)}{\theta} \right. \\ &\quad \left. - \theta \left[\frac{1}{\theta} + \frac{P(c_1) - \theta}{\theta^2} \right] \right\} \frac{1}{\ln 2} \\ &= \left\{ \frac{P(c_1)}{\theta} - \log_2 \frac{\theta - P(c_1)}{\theta} - 1 - \frac{P(c_1) - \theta}{\theta} \right\} \frac{1}{\ln 2} \\ &= -\log_2 \frac{\theta - P(c_1)}{\theta} \frac{1}{\ln 2} \geq 0 \end{aligned} \tag{19}$$

式(19)说明,当 $\theta \geq \max P(c_i)$ 时,条件熵的下界随着 θ 的增大单调递增,也就是说,θ 越大,条件熵下界越大,从而信息增益越小,即支持度值较大的模式其信息增益不大于某个阈值,其值可以根据式(18)计算得到,从信息增益的意义来讲,也即说明支持度较大的特征的分类能力有限.其他两种情况是以上两种情况的结合.因此,在二分类问题下,支持度较小或支持度较大的特征具有有限的分类能力.

3.2 多分类问题下特征的分类能力与支持度的关系

在多类问题中,式(7)中条件熵的定义域为

$$\begin{aligned} & 0 \leq p_i \leq 1 \\ & p_1 + p_2 + \cdots + p_{num_{cla}} = 1 \\ & P(c_i) \geq \theta p_i \end{aligned} \tag{20}$$

其中,$i = 1,2,\cdots,p_{num_{cla}}$.

与两类问题一致,为方便起见,将其转化为

$$\begin{aligned} & 0 \leq p_i \leq 1 \\ & p_1 + p_2 + \cdots + p_{num_{cla}} = 1 \\ & p_i \leq \frac{P(c_i)}{\theta} \end{aligned} \tag{21}$$

所以,如果 $P(c_i)/\theta \geq 1$,即 $\theta \leq \min P(c_i)$ 时,最后一个条件退化.同样可以方便证明其定义域为闭凸集合,过程如下:

证明　任取 $x_1, x_2 \in D$ 以及任意的 $\alpha \in (0,1)$,令 $c = num_{cla}$,设 $x_1 = (p_{11}, p_{12}, \cdots p_{1c}), x_2 = (p_{21}, p_{22}, \cdots p_{2c})$,则 $0 \leq p_{1i} \leq 1, 0 \leq p_{2i} \leq 1, i = 1,2,\cdots c$,

$p_{11} + p_{12} + \cdots + p_{1c} = 1$ 以及 $p_{21} + p_{22} + \cdots + p_{2c} = 1$.

令

$$\begin{aligned} x_\alpha &= \alpha x_1 + (1-\alpha)x_2 \\ &= \alpha(p_{11}, p_{12}, \cdots p_{1c}) + (1-\alpha)(p_{21}, p_{22}, \cdots p_{2c}) \\ &= (\alpha p_{11} + p_{21} - \alpha p_{21}, \alpha p_{12} + p_{22} - \alpha p_{22}, \\ &\quad \cdots \alpha p_{1c} + p_{2c} - \alpha p_{2c}) \end{aligned} \tag{22}$$

任取其第 i 项 $\alpha p_{1i} + p_{2i} - \alpha p_{2i}$,则由于 $0 \leq \alpha p_{1i} < \alpha$ 且 $0 \leq p_{2i} - \alpha p_{2i} \leq 1 - \alpha$,因此,$0 \leq \alpha p_{1i} + p_{2i} - \alpha p_{2i} \leq 1 - \alpha + \alpha = 1$;同时,

$$\begin{aligned} & \alpha p_{11} + p_{21} - \alpha p_{21} + \alpha p_{12} + p_{22} - \alpha p_{22} \cdots + \alpha p_{1c} + p_{2c} - \alpha p_{2c} \\ &= \alpha(p_{11} + p_{12} + \cdots + p_{1c}) + (1-\alpha)(p_{21} + p_{22} + \cdots + p_{2c}) \\ &= 1 \end{aligned} \tag{23}$$

即,$x_\alpha \in D$ 因此,集合 D 是凸集合.

同样的,由于上凸函数的下界在边界点上获得.与二分类问题类似,可以得到多种不同的定义域.另外,因为多分类问题对应的定义域属于高维问题,难以进行可视化,且其对应的情况较多,为了表述方便,且考虑我们关心的结论是在较高的支持度和较低的支持度下的情况,因此,取其中两种极限情况讨论,分别在支持度较小和较大时进行.

(1)当 $\theta \leq \min P(c_i)$ 时

当 $\theta \leq \min P(c_i)$ 时,由于 $p_1 + p_2 + \cdots p_c = 1$ 对应了一个超平面,在该种定义域下,边界点为平面的顶点:即 $p_i = 1, p_j = 0$(其中 $j = 1,2,\cdots,c$ 且 $j \neq i$) 时取下界,将其代入式(7),得到条件熵的下界为

$$\begin{aligned} & H(C \mid X)_{lb} = -[P(c_i) - \theta]\log_2 \frac{P(c_i) - \theta}{1 - \theta} \\ & -\sum_{j=1, j\neq i}^{n} [P(c_j)]\log_2 \frac{P(c_j)}{1 - \theta} \end{aligned} \tag{24}$$

求式(24)关于 θ 的偏导,则得

$$\begin{aligned} \frac{\partial H(C \mid X)_{lb}}{\partial \theta} &= -\sum_{i=1}^{num_{cla}} \left[-p_i \log_2 \frac{P(c_i) - \theta p_i}{1 - \theta} \right. \\ &\quad \left. + \left[-\frac{p_i}{1 - \theta} + \frac{P(c_i) - \theta p_i}{(1 - \theta)^2} \right] \frac{(1 - \theta)}{\ln 2} \right] \\ &= \log_2 \frac{P(c_i) - \theta}{1 - \theta} + \left[\frac{1}{1 - \theta} - \frac{P(c_i) - \theta}{(1 - \theta)^2} \right] \\ &\quad \cdot \frac{(1 - \theta)}{\ln 2} - \frac{1 - P(c_i)}{(1 - \theta)^2} \frac{(1 - \theta)}{\ln 2} \\ &= \log_2 \frac{P(c_i) - \theta}{1 - \theta} \leq 0 \end{aligned} \tag{25}$$

式(25)说明,当 $\theta \leq \min P(c_i)$ 时,条件熵的下界随着 θ 的增大单调递减,也就是说,θ 越小,条件熵下界越大,从而信息增益上界越小,即支持度较小的模式其信息增益不大于某个阈值,其值可以根据式(24)计算得

到,从信息增益的意义来讲,也即说明支持度较小的特征分类能力有限.

(2)当 $\theta \geq \max P(c_i)$ 时

当 $\theta \geq \max P(c_i)$,且 $p_j = 1 - \sum_{j=1, j\neq i}^{num_{cla}} p_i = \frac{\theta - 1 + P(c_j)}{\theta} \geq 0$ 时,即 $\theta \geq \sum_{i=1, i\neq j}^{num_{cla}} P(c_i)$ 时,下界在 $p_i = P(c_i)/\theta, i \neq j$,其中,$j = 1,2,\cdots num_{cla}$ 时取得,将其代入式(7),得到

$$H_{lb}(C \mid X) = -\sum_{i=1, i\neq j}^{num_{cla}} P(c_i) \log_2 \frac{P(c_i)}{\theta} - [\theta - 1 + P(c_j)] \log_2 \frac{\theta - 1 + P(c_j)}{\theta} \tag{26}$$

$$\frac{\partial H_{lb}(C \mid x)}{\partial \theta} = -\sum_{i=1, i\neq j}^{num_{cla}} \frac{P(c_i)}{\theta} - \log_2 \frac{P(c_j) + \theta - 1}{\theta} + \theta \cdot \left(\frac{P(c_j) + \theta - 1}{\theta^2} - \frac{1}{\theta}\right) = -\log_2 \frac{P(c_j) + \theta - 1}{\theta} \geqslant 0 \tag{27}$$

式(27)说明,当 $\theta \geq \sum_{i=1, i\neq j}^{num_{cla}} P(c_i)$ 时,条件熵的下界随着 θ 的增大单调递增,也就是说,θ 越大,条件熵下界越大,从而信息增益越小,即支持度值较大的模式其信息增益不大于某个阈值,其值可以根据式(26)计算得到,从信息增益的意义来讲,也即说明频率较大的特征的分类能力有限.

综上,在分类时,支持度较低或支持度较高的序列模式对分类的影响可以限定在一定范围内.

4 实验验证及分析

为了验证以上推理,我们采用 UCI 的不同数据集对本文结论进行了验证,如图 2. 在图 2(a)中给出了在 Breast cancer[15]数据集上的相关结果,Breast cancer 属于二分类问题,其结果分为两类{no-recurrence-events, recurrence-events},且 $P(c_1) = 0.29, P(c_2) = 0.71$. 图 2($a$)中,当 $\theta < P(c_i)$ 时,"$*$"和"o"点分别是 $p_1 = 1, p_2 = 0$ 和 $p_1 = 0, p_2 = 1$ 时的条件熵;当 $\theta > P(c_i)$ 时,"+"和"□"点分别是 $p_1 = P(c_1)/\theta, p_2 = 1 - P(c_1)/\theta$ 和 $p_1 = 1 - P(c_2)/\theta, p_2 = P(c_2)/\theta$ 时的条件熵;实线给出的是整个分类系统的信息增益上界. 由图 2(a)可以看出,当 $\theta < \min P(c_i)$ 时,即 $\theta < P(c_1) = 0.29$ 时,其信息增益的上界在 $p_1 = 1, p2 = 0$ 时取得,当 $\theta > \max P(c_i)$ 时,即 $\theta > P(c_2) = 0.71$ 时,其信息增益的上界在 $p_1 = 1 - P(c_2)/\theta, p_2 = P(c_2)/\theta$ 时取得. 当 $P(c_1) < \theta < P(c_2)$ 时,具体来讲,由图可得,当 $0.29 < \theta < 0.51$ 时,信息增益在 $p_1 = P(c_1)/\theta, p_2 = 1 - P(c_1)/\theta$ 时取上界,当 $0.51 < \theta < 0.71$ 时,信息增益在 $p_1 = 0, p_2 = 1$ 时取上界. 在 $\theta < P(c_1)$ 时,随着支持度的增加信息增益单调递增,当 $\theta > P(c_2)$ 时,信息增益随支持度的增大单调递减. 在确定了有用的信息增益后,则可以由图 2(a)方便的求解支持度,以该支持度为参数,即可以采用不同的频繁模式挖掘算法进行初步的特征选择. 图 2(b)中给出了在 Madelon 数据集[16]上的相关结果,在 Madelon 数据集中,$P(c_1) = 0.5, P(c_2) = 0.5$,"$*$"和"o"点给出的是在 $\theta < P(c_i)$ 时的条件熵,"+"和"□"点是在 $\theta > P(c_i)$ 时的条件熵,实线给出的是信息增益的上界. 由图 2(b)可以看出,当 $\theta < \min P(c_i)$ 时,即 $\theta < P(c_1)$ 时,其信息增益的上界在 $p_1 = 1, p_2 = 0$ 或 $p_1 = 0, p_2 = 1$ 是一致的,因此可以在此时取得信息增益上界,当 $\theta > \max P(c_i)$ 时,即 $\theta > P(c_2)$ 时,其信息增益的上界在 $p_1 = P(c_1)/\theta, p_2 = 1 - P(c_1)/\theta$ 或 $p_1 = 1 - P(c_1)/\theta, p_2 = P(c_1)/\theta$ 同样重合,上界在此时取得. 由图 2(b)可得,当 $\theta < P(c_1) = P(c_2) = 0.5$ 时,信息增益上界随着支持度的增大单调递增,当 $\theta > P(c_1) = P(c_2) = 0.5$ 时,信息增益上界随着支持度的增大单调递减,即高支持度特征和低支持度特征具有有限的分类能力. 图 2(c)中给出了在 Vertebral column 数据集[17]上的相关结果,该数据集分三类,分别为{Normal, Disk Hernia, Spondylolisthesis},$P(c_1) = 0.32, P(c_2) = 0.19, P(c_3) = 0.48$,"×"点表示的是在 $\theta < P(c_2)$ 时,系统在 $p_1 = 0, p_2 = 1, p_3 = 0$ 的条件熵,"o"是在 $\theta < P(c_1)$ 时,系统在 $p_1 = 1, p_2 = 0, p_3 = 0$ 的条件熵,"□"点是在 $\theta < P(c_3)$ 时,系统在 $p_1 = 0, p_2 = 0, p_3 = 1$ 的条件熵,"+"点是在 $P(c_2) < \theta < 1 - P(c_1)$ 时,$p_1 = 0, p_2 = P(c_2)/\theta, p_3 = 1 - P(c_2)/\theta$ 的条件熵,"◇"点是在 $P(c_2) < \theta < 1 - P(c_3)$ 时,$p_1 = 1 - P(c_2)/\theta, p_2 = P(c_2)/\theta, p_3 = 0$ 的条件熵,"△"是在 $P(c_1) < \theta < 1 - P(c_2)$ 时,$p_1 = P(c_1)/\theta, p_2 = 0, p_3 = 1 - P(c_1)/\theta$ 的条件熵,"◆"是在 $P(c_1) < \theta < 1 - P(c_3)$ 时,$p_1 = P(c_1)/\theta, p_2 = 1 - P(c_1)/\theta, p_3 = 0$ 时的条件熵,"$*$"是在 $P(c_3) < \theta < 1 - P(c_1)$ 时,$p_1 = 0, p_2 = 1 - P(c_3)/\theta, p_3 = P(c_3)/\theta$ 的条件熵,"◁"是在 $P(c_3) < \theta < 1 - P(c_2)$ 时,$p_1 = 1 - P(c_3)/\theta, p_2 = 0, p_3 = P(c_3)/\theta$ 的条件熵;"▷"是在 $\theta > \sum_{i=1}^{2} P(c_i)$ 时,在 $p_1 = P(c_1)/\theta, p_2 = P(c_2)/\theta, p_3 = 1 - P(c_1)/\theta - P(c_2)/\theta$ 的条件熵,"▽"是在 $\theta > \sum_{i=2}^{3} P(c_i)$ 时,在 $p_1 = 1 - P(c_2)/\theta - P(c_3)/\theta, p_2 = P(c_2)/\theta, p_3 = P(c_3)/\theta$ 的条件熵,"." 是在 $\theta > \sum_{i=1, i\neq 2}^{3} P(c_i)$ 时,在 $p_1 = P(c_1)/\theta, p_2 = 1 - P(c_1)/\theta - P(c_3)/\theta, p_3 = P(c_3)/\theta$ 的条件熵;实线给出的是信息增益的上界. 由图 2(c)可以看出,当 $\theta < \min P(c_i)$ 时,即 $\theta < P(c_2)$ 时,其信息增益的上界在

$p_1=0, p_2=1, p_3=0$ 时取得,当 $\theta \geq \sum_{i=1, i\neq j}^{3} P(c_i)$ 时,其信息增益的上界在 $p_1 = P(c_1)/\theta$, $p_2 = 1 - P(c_2)/\theta - P(c_3)/\theta$, $p_3 = P(c_3)/\theta$ 时取得.其他情况下的信息增益上界可以从图中方便的得出.同时,由图 2(c)可以看出,当 $\min P(c_i) < \theta < \max P(c_i)$时,其信息增益上界一定不在 $p_i=1, p_j=0$,其中 $j\neq i$ 时取得,而 Lee[12]则采用了该结论,因此,其文中的相关讨论是不严密的.同时,根据本文的结果,则可以方便的确定与信息增益相关的支持度.综上,实验不仅验证了我们关于信息增益与支持度的联系,并且说明了可以方便的根据该联系确定频繁模式挖掘算法所需要的参数,进而进行频繁模式挖掘,以达到特征选择的目的.

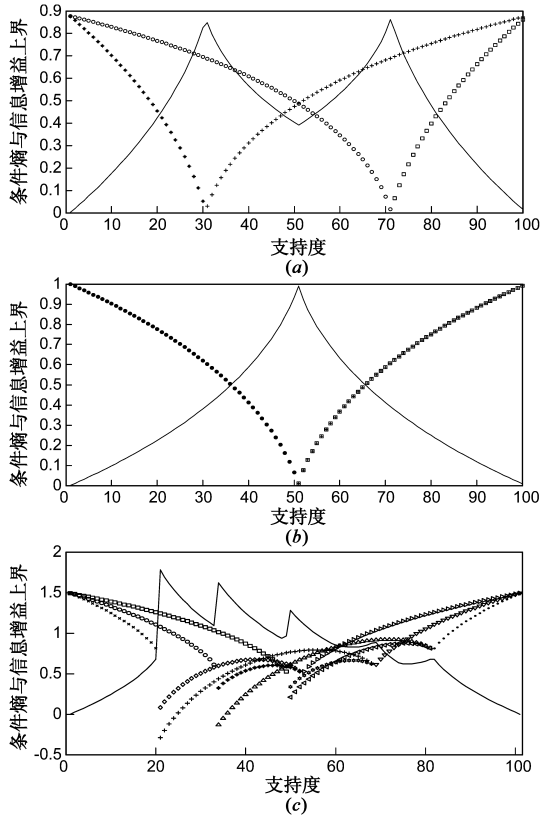

图2　信息熵上界、条件熵与特征支持度之间的关系图

5　结论及展望

本文给出了特征的支持度与条件熵下界或信息增益之间的联系,证明了具有高支持度或低支持度的特征具有有限的分类能力,从而为频繁模式挖掘在分类问题中进行特征选择奠定了理论基础,而由于较多的频繁模式挖掘算法在大数据集上具有可扩展性,因此,本文结论为大数据分类问题的可扩展性算法的寻求提供了一种可行的理论工具.在证明该结论的基础上,通过仿真实验验证了我们的结论.

通过本文工作,为频繁模式挖掘在分类问题中的应用提供了理论基础和理论依据,以该结果为基础,可以方便的确定频繁模式挖掘中所用的支持度参数,从而利用频繁模式挖掘滤除低支持度特征,对于分类器的设计具有重要意义.

参考文献

[1] 陈晓云,陈祎,王雷,李荣陆,胡运发.基于分类规则树的频繁模式文本分类[J].软件学报,2006,17(5):1017 – 1025.
Chen Xiaoyun, Chen Yi, Wang Lei, Li Ronglu, Hu Yunfa. Text categorization based on classification rules tree by frequent patterns[J]. Journal of Software, 2006, 17(5): 1017 – 1025. (in Chinese)

[2] H Lodhi, C Saunders, J Shawe-Taylor, N Cristianini, C Watkins. Text classification using string kernels[J]. Journal of Machine Learning Research, 2002, 2(3): 419 – 444.

[3] Y Li, S M Chung, J D Holt. Text document clustering based on frequent word meaning sequences[J]. Data and Knowledge Engineering, 2008, 64(1): 381 – 404.

[4] 赵建邦,董安国,高琳.一种用于生物网络数据的频繁模式挖掘算法[J].电子学报,2010,38(8):1803 – 1807.
Zhao Jianbang, Dong Anguo, Gao Lin. An algorithm for frequent pattern mining in biological networks[J]. Acta Electronica Sinica, 2010, 38(8): 1803 – 1807. (in Chinese)

[5] Young-Rae Cho, Aidong Zhang. Predicting protein function by frequent functional association pattern mining in protein interaction networks[J]. IEEE Transactions on Information Technology in Biomedicine, 2010, 14(1): 30 – 36.

[6] R Alves, D R Baena, J S A Ruiz. Gene association analysis: a survey of frequent pattern mining from gene expression data [J]. Briefings in Bioinformatics, 2010, 11(2): 210 – 224.

[7] Hong Cheng, Xifeng Yan, Jiawei Han, Chih-Wei Hsu. Discriminative frequent pattern analysis for effective classification[A]. In: IEEE 23rd International Conference on Data Engineering [C]. Istanbul, Turkey, 2007, 716 – 725.

[8] Han Jiawei, Cheng Hong, Xin Dong, Yan Xifeng. Frequent pattern mining: current status and future directions[J]. Journal of Data Mining and Knowledge Discovery, 2007, 15(1): 55 – 86.

[9] H Carl, F John. Sequential pattern mining-approaches and algorithms[J]. ACM Computing Surveys, 2013, 45(2): 1 – 19.

[10] 高琳,覃桂敏,周晓峰.图数据中频繁模式挖掘算法研究综述[J].电子学报,2008,36(8):1603 – 1609.
Lin Gao, Guimin Qin, Xiaofeng Zhou. An overview of algorithms for mining frequent patterns in graph data[J]. Acta Electronica Sinica, 2008, 36(8): 1603 – 1809. (in Chinese)

[11] 万里,廖建新,朱晓民,倪萍.一种基于频繁模式的时间序列分类框架[J].电子与信息学报,2010,32(2):261 –

266.
Li Wan, Jianxin Liao, Xiaomin Zhu, Ping Ni. A frequent pattern based times series classification framework[J]. Journal of Electronics and Information Technology, 2010, 32(2): 261 – 266. (in Chinese)
[12] Lee Jae-Gil, Han Jiawei, Li Xiaolei, Cheng Hong. Mining discriminative patterns for classifying trajectories on road networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(5): 713 – 726.
[13] C M Bishop. Pattern Recognition and Machine Learning[M]. Springer Press, 2006: 55 – 57.
[14] B Stephen, L Vandenberghe. Convex Optimization[M]. Cambridge University Press, England. 2004: 136 – 146.
[15] R S Michalski, I Mozetic, J Hong, N Lavrac. The multi-purpose incremental learning system AQ15 and its testing application to three medical domains[A]. In Proceedings of the Fifth National Conference on Artificial Intelligence[C]. Philadelphia, America. 1986. 1041 – 1045.
[16] I Guyon, R Steve Gunn, A Ben-Hur, G Dror. Result analysis of the NIPS 2003 feature selection challenge[A]. Proceedings on Advances in Neural Information Processing Systems[C]. Vancouver, Canada. 2004. 545 – 552.
[17] A R Rocha, R Sousa, G A Barreto, J S Cardoso. Diagnostic of pathology on the vertebral column with embedded reject option[A]. Proceedings of the 5th Iberian Conference on Pattern Recognition and Image Analysis[C]. Grancanaria, Spain. 2011. 588 – 595,

作者简介

尹建芹 女, 1978 年 11 月出生, 山东潍坊人. 副教授, 博士. 2000 年获山东工业大学工学学士学位, 2002 年和 2013 年获山东大学工学硕士和工学博士学位. 现工作于济南大学信息科学与工程学院, 主要从事模式识别、机器学习及图像处理等相关研究.
E-mail: ise_yinjq@ujn.edu.cn

田国会 男, 1969 年 8 月出生, 河北河间人. 教授、博士生导师. 1990 年、1993 年和 1997 年分别在山东大学、山东工业大学和东北大学获理学学士、工学硕士和工学博士学位. 现为山东大学服务机器人研究室主任, 主要从事服务机器人、智能空间等的研究工作.