计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20180793
56(11): 2410-2423, 2019

基于相似度量的自适应三支垃圾邮件过滤器

谢 秦 张清华 王国胤
(计算智能重庆市重点实验室(重庆邮电大学) 重庆 400065)
(619773142@qq.com)

An Adaptive Three-way Spam Filter with Similarity Measure

Xie Qin, Zhang Qinghua, and Wang Guoyin
(*Chongqing Key Laboratory of Computational Intelligence* (*Chongqing University of Posts and Telecommunications*), *Chongqing* 400065)

Abstract Spam filtering is an important issue in the information age. And, if an important email is wrongly classified, it would lead to an immeasurable cost. Thus, in the field of spam filtering, the ways to improve the accuracy and recall of the filters is the key issue. At present, the binary classification model in machine learning is usually used to deal with spam filtering. However, compared with the three-way decisions, the binary classification model usually leads to a higher cost of misclassification. And, as an important branch of three-way decisions, the three-way decisions with decision-theoretic rough sets can effectively reduce the misclassification cost and further improve the performance of filters. And, it also conforms to human cognition. Nevertheless, few studies consider the effect on classification results induced by the differences among equivalence classes when constructing the loss functions. Therefore, under the framework of the three-way decisions with decision-theoretic rough sets, an adaptive three-way spam filter with similarity measure is proposed. The model calculates the weights of condition attributes according to set variance firstly. Then, a comprehensive evaluation function for describing difference information among equivalence classes based on similarity measure of set is established. Finally, an adaptive model for calculating threshold pairs based on Bayesian decision rules is constructed. Experimental results show that the proposed model performs well in the field of spam filtering.

Key words spam filtering; decision-theoretic rough sets; three-way decisions; similarity measure; thresholds

摘 要 垃圾邮件过滤是信息时代的一个重要研究课题,一封重要邮件被错分会产生不可估量的代价.因此,如何提高过滤器的性能成为垃圾邮件过滤领域中的核心问题.目前,业界通常采用机器学习算法中的二分类模型以处理垃圾邮件过滤问题.然而,较之于三支决策模型,二分类模型会产生较大的错分代价.作为三支决策的一个重要分支,基于决策理论粗糙集的三支决策模型符合人类认知习惯,且能有效地降低错分代价,进而提高过滤器的性能.然而,在构造损失函数时,少有研究考虑由于等价类之间的差异性而对分类结果带来的影响.因此,在基于决策理论粗糙集的三支决策模型的基础上,提出了一种

收稿日期:2018-11-26;**修回日期**:2019-05-05
基金项目:国家自然科学基金项目(61876201,61772096);重庆市研究生科研创新项目(CYS18244)
This work was supported by the National Natural Science Foundation of China (61876201, 61772096) and the Chongqing Postgraduate Scientific Research and Innovation Project (CYS18244).
通信作者:张清华(zhangqh@cqupt.edu.cn)

基于相似度量的自适应三支垃圾邮件分类模型.该模型根据集合方差计算了条件属性的权重,并基于相似度量建立了一种刻画差异信息的综合评价函数,进而根据贝叶斯决策规则构建了一种计算自适应阈值对的方法.实验结果表明所提模型在垃圾邮件过滤领域表现优异.

关键词　垃圾邮件过滤;决策理论粗糙集;三支决策;相似度量;阈值

中图法分类号　TP301.6

互联网时代使人类获取信息的方式更加便捷多样,但也给人类带来了诸如垃圾邮件等有害信息.对于某些人群来说,若邮件被错分则会带来非常大的损失,如风险投资者因一封投资分析报告邮件被错分到了垃圾邮件类,则其投资计划将会因为信息获取不及时而遭受巨大损失.因而提高过滤器的分类准确率及召回率等指标对于垃圾邮件过滤来说至关重要.在业界,垃圾邮件过滤通常被视为一个二分类问题.目前,机器学习领域中的大量二分类模型都可被用来处理垃圾邮件过滤问题,如 k 近邻分析法[1]、朴素贝叶斯(Naive Bayes, NB)分类器[2]、基于集成学习的分类器[3]等.然而,由于二分类模型武断地把邮件归类为垃圾邮件类或合法邮件类,因此会导致决策过程中产生较大的错分代价进而导致分类精度较低.此外,由于二分类模型在分类时过于直接,因此常常无法确定最优阈值.

三支决策模型实质上是在二分类模型的基础上加入了延迟决策域.换言之,三支决策模型的主要优点在于其允许延迟决策的可能性.如果做出延迟决策的代价不高,则三支决策模型是一个有效的选择.此外,从人类认知的角度出发,三支决策模型将不能确定类别的邮件放到延迟决策域,用户可以依据实际需求查看延迟决策域里的邮件,可达到灵活选择有价值的邮件的目的.目前,三支决策研究领域已有不少的成果,如文献[4]将不能合理地判别是垃圾邮件或合法邮件的邮件称为灰色邮件,并提出了 4 种原型检测方法.此外,根据粗糙集[5-6],文献[7-9]提出了决策理论粗糙集模型,并受到了广大学者的关注,如文献[10]基于决策粗糙集提出了一种解决修饰和不修饰情感词语情况下的否定句情感分类模型.进一步地,为给决策粗糙集中的 3 个域赋予一定的语义,文献[11-12]提出了基于决策理论粗糙集的三支决策(three-way decision with decision-theoretic rough sets, 3WD-DTRSs)模型.3WD-DTRSs 模型是一种依据贝叶斯决策规则,以最小化期望代价为目标来确定一对阈值并将待判对象划分为 3 个域(接受域、拒绝域及延迟决策域)的三支决策模型.3WD-DTRSs 模型的研究成果同样丰富,如文献[13]提出了一种基于犹豫模糊决策理论粗糙集的风险决策方法.文献[14]给出了一种用于在给定接受域对象数量的情况下,建立具有给定属性值的三支决策模型.值得提出的是,在 3WD-DTRSs 模型研究中,阈值的计算一直是一个研究热点,研究者已给出一些计算方法[15-19],如文献[20]提出了一种利用相对值确定 3WD-DTRSs 模型的损失函数的方法.垃圾邮件过滤过程从代价的角度可以解释为,将一个实为垃圾邮件的邮件判断为垃圾邮件的代价小于将其判断为延迟决策的代价,而将其判断为延迟决策的代价又小于将其判断为合法邮件的代价.同理,将一个实为合法邮件的邮件判断为合法邮件的代价小于将其判断为延迟决策的代价,而将其判断为延迟决策的代价又小于将其判断为垃圾邮件的代价.而 3WD-DTRSs 模型中的代价逻辑正好与之契合.因此,文献[21]将 3WD-DTRSs 模型用于解决垃圾邮件分类问题,并从代价敏感等角度阐明了模型的优势.

然而,在 3WD-DTRSs 模型研究领域,存在这样一种情况,当 2 个等价类的粗糙隶属度相同时,2 个等价类必定会划分到同一个区域.从模型原理的角度出发,这种现象存在的原因为在 3WD-DTRSs 模型中,粗糙隶属函数反映的是等价类与目标集的交叉程度,而由专家给定的损失函数是将所有对象视作等价的,换言之,对于每一个等价类,只要等价类中的对象与目标集合的交叉度相同,则粗糙隶属度相同,且损失函数都是相同的.然而,当二者所对应的各条件属性值相差比较大的时候,二者在直观上不应该属于同一区域.从条件属性值的角度出发,等价类与目标集合的相似度越大,该等价类在同等条件下更应该被划分到正域,换言之,等价类之间是存在差异的.若仍然按照 3WD-DTRSs 模型的原理,忽略等价类之间的差异性,则在贝叶斯决策准则下推导出来的阈值并不能使得所有等价类达到全局的最小划分代价,进而导致模型精度表现不够理想.然而,目前少有研究在给定或者构造损失函数时考虑

由于等价类之间存在差异而带来的影响.因此,现有的3WD-DTRSs模型在模型性能上还存在一定的研究价值.已有工作[22]考虑等价类之间存在差异性等因素,基于效用理论给出了一种改进的基于决策粗糙集的三支决策模型.然而该模型在基于效用函数构造风险度量函数时,需要主观给定风险偏好参数,则该模型的泛化能力还有待提高.此外,相似度量是一种度量2个对象之间相似度的工具,且相似度量的方法相当丰富[23-24],如相关系数、欧氏距离及集合相似度等.在垃圾邮件过滤领域,文献[25]提出了一种用于垃圾邮件图像聚类分析的非负稀疏相似性度量方法.此外,刘伍颖等人[26]给出了一种基于结构化集成学习的垃圾邮件过滤器.因此,为了解决上述各问题,从集合相似度量角度出发,考虑等价类之间存在差异性等因素,提出了一种基于相似度量的自适应三支垃圾邮件分类器(similarity measure based adaptive three-way spam filter, SBA-3WD-SF).

首先根据集合方差给出了一种计算属性重要度的方法,并基于此提出了一种计算条件属性权重的方法;其次,根据集合相似度给出了一种度量等价类中条件属性值与目标集合中相应属性值的贴近度的方法;然后,进一步给出一种自动确定综合评价函数的方法以度量风险,并根据贝叶斯准则,给出了一种计算自适应阈值的方法.实验结果表明,所提的SBA-3WD-SF在准确率及召回率方面优于二元NB分类器及3WD-DTRSs分类器.换言之,SBA-3WD-SF是合理有效的.

1 基础知识

1.1 基本定义

定义1. 不可分辨关系.[2]给定一个信息系统$S=(U,C\cup D,V,f)$,对于属性子集$\forall B(B\subseteq C\cup D)$,论域$U$上的一个不可分辨关系$IND(B)$可定义为

$$IND(B)=\{(x,y)\mid(x,y)\in U^2,\ \forall b\in B\wedge(b(x)=b(y))\}. \tag{1}$$

定义2. 近似集.[2]给定一个信息系统$S=(U,C\cup D,V,f)$,对于$\forall X(X\subseteq U)$及不可分辨关系$IND(B)$,集合X的上下近似集可分别定义为

$$\overline{apr}(X)=\cup\{E_j\mid E_j\in U/IND(B)\wedge E_j\cap X\neq\varnothing\},\quad \underline{apr}(X)=\cup\{E_j\mid E_j\in U/IND(B)\wedge E_j\subseteq X\}, \tag{2}$$

其中,

$$U/IND(B)=\{X\mid X\subseteq U\wedge\forall x,y\in X\wedge\forall b\in B\wedge(b(x)=b(y))\}$$

表示由等价关系$IND(B)$在论域U上诱导的划分.

从而,非空有限论域可被划分为3个互不相交的区域,即正域($POS(X)$)、负域($NEG(X)$)及边界域($BND(X)$):

$$\begin{aligned}&POS(X)=\underline{apr}(X),\\&NEG(X)=U-\overline{apr}(X),\\&BND(X)=\overline{apr}(X)-\underline{apr}(X).\end{aligned} \tag{3}$$

定义3. 粗糙隶属函数.[27]给定一个信息系统$S=(U,C\cup D,V,f)$,则对于$\forall X(X\subseteq U)$及不可分辨关系$IND(B)$,关于目标集合X的粗糙隶属函数可定义为

$$\mu_X^{IND(B)}(E_j)=\frac{|X\cap E_j|}{|E_j|}, \tag{4}$$

其中,$\mu_X^{IND(B)}(E_j)$表示对象x在属于等价类E_j的条件下,该对象又属于目标集X的概率.显然,作为一种条件概率函数,粗糙隶属函数是一种判别函数.$\mu_X^{IND(B)}(E_j)$简记为$Pr(X\mid E_j)$.

1.2 相关研究

为了突出贡献,本节将分别简要介绍一个二分类及一个三分类垃圾邮件分类器的代表模型,即二元NB分类器与3WD-DTRSs分类器.

1.2.1 二元NB分类器

文献[28]提出了贝叶斯理论,文献[29]提出了特征条件独立性假设,文献[30]将二者结合并提出了NB垃圾邮件过滤器.二元NB分类器是垃圾邮件过滤的一种最常用的二分类方法.假设每一个待判的邮件对象x均可用一个m维的特征向量$\boldsymbol{V}(x)=(v_1(x),v_2(x),\cdots,v_m(x))$表示,$X$表示垃圾邮件类,$\neg X$表示合法邮件类.有别于定义3中的条件概率$Pr(X\mid E_j)$,可根据等价类与目标集的交并而获得,而NB分类器主要是通过将面向单一对象的条件概率$Pr(X\mid\boldsymbol{V}(x))$转换为易于求解的后验条件概率$Pr(\boldsymbol{V}(x)\mid X)$的方式来解决难以直接求解$Pr(X\mid\boldsymbol{V}(x))$的问题.根据贝叶斯定理及全概率公式,给定一个邮件$\boldsymbol{V}(x)=(v_1(x),v_2(x),\cdots,v_m(x))$,其属于垃圾邮件类$X$的后验概率(一种判别函数)为

$$Pr(X\mid\boldsymbol{V}(x))=\frac{Pr(X)Pr(\boldsymbol{V}(x)\mid X)}{Pr(\boldsymbol{V}(x))}, \tag{5}$$

其中,

$$Pr(\boldsymbol{V}(x))=Pr(X)Pr(\boldsymbol{V}(x)\mid X)+Pr(\neg X)Pr(\boldsymbol{V}(x)\mid\neg X),$$

$Pr(X)$为 X 的先验概率,$Pr(\boldsymbol{V}(x)|X)$表示给定邮件 $\boldsymbol{V}(x)$时,$\boldsymbol{V}(x)\in X$ 的似然函数.事实上,似然函数 $Pr(\boldsymbol{V}(x)|X)$是一个联合条件概率 $Pr(v_1(x),v_2(x),\cdots,v_m(x)|X)$,然而,当 m 比较大时,在实践中很难分析 $v_1(x),v_2(x),\cdots,v_m(x)$之间的相互作用.因此,条件独立性假设被应用于贝叶斯分类器,即假设当给定垃圾邮件类 X 时,每一个特征 $v_i(x)(i=1,2,\cdots,m)$与其余特征都是条件独立的,在表达式上体现为

$$Pr(\boldsymbol{V}(x)|X)=Pr(v_1(x),v_2(x),\cdots,v_m(x)|X)=\prod_{i=1}^{m}Pr(v_i(x)\mid X),\tag{6}$$

其中,$Pr(v_i(x)|X)$能很容易地从数据中估计得到.从而式(5)可以表示为

$$Pr(\boldsymbol{V}(x)\mid X)=\frac{\prod_{i=1}^{m}Pr(v_i(x)\mid X)}{Pr(\boldsymbol{V}(x))}.\tag{7}$$

同理,后验概率 $Pr(\neg X|\boldsymbol{V}(x))$可以表示为

$$Pr(\neg X\mid\boldsymbol{V}(x))=\frac{Pr(\neg X)\prod_{i=1}^{m}Pr(v_i(x)\mid\neg X)}{Pr(\boldsymbol{V}(x))}.\tag{8}$$

因此,概率 $Pr(\boldsymbol{V}(x))$可以被消除,即有:

$$\frac{Pr(X|\boldsymbol{V}(x))}{Pr(\neg X|\boldsymbol{V}(x))}=\prod_{i=1}^{m}\frac{Pr(v_i(x)|X)}{Pr(v_i(x)|\neg X)}\frac{Pr(X)}{Pr(\neg X)}.\tag{9}$$

综上所述,对于一个待判邮件 $\boldsymbol{V}(x)$,如果 $\frac{Pr(X|\boldsymbol{V}(x))}{Pr(\neg X|\boldsymbol{V}(x))}$大于阈值 ξ,则 $\boldsymbol{V}(x)$属于垃圾邮件类;反之,$\boldsymbol{V}(x)$属于合法邮件类.

1.2.2　3WD-DTRSs 模型

作为三分类模型的一个代表模型,本节将简要介绍 3WD-DTRSs 模型[11].首先,决策理论粗糙集模型[8]包含 2 个状态 $T=\{X,\neg X\}$及 3 个动作 $A=\{a_P,a_B,a_N\}$.根据 3 个动作,由专家经验给出的损失函数如表 1 所示:

Table 1　Loss Function Given by Expertise

表 1　由专家经验给定的损失函数

T	A		
	a_P	a_B	a_N
X	λ_{PP}	λ_{BP}	λ_{NP}
$\neg X$	λ_{PN}	λ_{BN}	λ_{NN}

由表 1 得知:在垃圾邮件过滤这个背景下,状态 X 及 $\neg X$ 分别表示垃圾邮件类及合法邮件类.λ_{PP},λ_{BP},λ_{NP}分别表示一个本应该属于垃圾邮件类的对象分别采取动作 a_P,a_B,a_N 而产生的损失.同理 λ_{PN},λ_{BN},λ_{NN}分别表示一个本应该属于合法邮件类的对象分别采取动作 a_P,a_B,a_N 而产生的损失.对于任意待判邮件 $x(x\in E_j)$,采取不同动作而产生的期望损失可表示为

$$\begin{aligned}R(a_P|E_j)&=\lambda_{PP}Pr(X|E_j)+\lambda_{PN}Pr(\neg X|E_j),\\R(a_B|E_j)&=\lambda_{BP}Pr(X|E_j)+\lambda_{BN}Pr(\neg X|E_j),\\R(a_N|E_j)&=\lambda_{NP}Pr(X|E_j)+\lambda_{NN}Pr(\neg X|E_j).\end{aligned}\tag{10}$$

根据贝叶斯决策过程中最小化损失原则,则决策规则为

(P)如果 $R(a_P|E_j)\leqslant R(a_B|E_j)$且 $R(a_P|E_j)\leqslant R(a_N|E_j)$,则 $E_j\subseteq POS(X)$;

(B)如果 $R(a_B|E_j)\leqslant R(a_P|E_j)$且 $R(a_B|E_j)\leqslant R(a_N|E_j)$,则 $E_j\subseteq BND(X)$;

(N)如果 $R(a_N|E_j)\leqslant R(a_P|E_j)$且 $R(a_N|E_j)\leqslant R(a_B|E_j)$,则 $E_j\subseteq NEG(X)$.

由于 $Pr(X|E_j)+Pr(\neg X|E_j)=1$,则上述规则只与损失函数及判别函数 $Pr(X|E_j)$有关,一个合理假设为 $\lambda_{PP}<\lambda_{BP}<\lambda_{NP}$,$\lambda_{NN}<\lambda_{BN}<\lambda_{PN}$.且简记为

$$\begin{aligned}\alpha&=\frac{\lambda_{PN}-\lambda_{BN}}{\lambda_{BP}-\lambda_{PP}+\lambda_{PN}-\lambda_{BN}}=\left(1+\frac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}}\right)^{-1},\\\beta&=\frac{\lambda_{BN}-\lambda_{NN}}{\lambda_{NP}-\lambda_{BP}+\lambda_{BN}-\lambda_{NN}}=\left(1+\frac{\lambda_{NP}-\lambda_{BP}}{\lambda_{BN}-\lambda_{NN}}\right)^{-1},\\\gamma&=\frac{\lambda_{PN}-\lambda_{NN}}{\lambda_{NP}-\lambda_{PP}+\lambda_{PN}-\lambda_{NN}}=\left(1+\frac{\lambda_{NP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{NN}}\right)^{-1}.\end{aligned}\tag{11}$$

则决策规则(P)(B)(N)可等价地表示为

(P)如果 $Pr(X|E_j)\geqslant\alpha$,则 $E_j\subseteq POS(X)$;

(B)如果 $\beta<Pr(X|E_j)<\alpha$,则 $E_j\subseteq BND(X)$;

(N)如果 $Pr(X|E_j)\leqslant\beta$,则 $E_j\subseteq NEG(X)$.

综上所述,二分类模型及三支决策模型的差异如图 1 所示.对于二分类模型,如果对象的判别函数 $Pr(X|x)\geqslant\xi$,则该对象被接受;反之,该对象被拒绝.对于三支决策模型,如果对象的判别函数 $Pr(X|x)\geqslant\alpha$,则该对象被接受;如果该对象的判别函数 $Pr(X|x)\leqslant\beta$,则该对象被拒绝;如果该对象的判别函数 $\alpha>Pr(X|x)>\beta$,则该对象被归入延迟决策域.

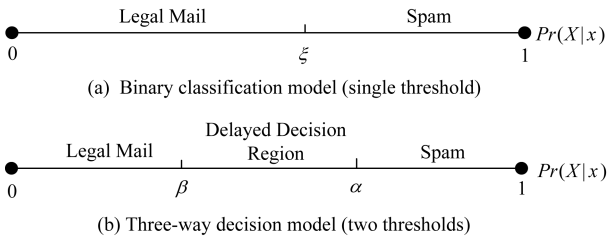


Fig. 1 Binary and three-way classification

图 1 二支及三支分类

2 基于相似度量的三支决策模型

当前,作为一种简单而有效的方法,距离度量被广泛地应用于各个领域.其中,欧氏距离是距离度量中的一种简单且易于理解的方法.在统计领域,欧氏距离[31]常被用来度量 2 个向量的距离.

2 个向量 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_m)$ 和 $\boldsymbol{Z}=(z_1,z_2,\cdots,z_m)$之间的欧氏距离定义为

$$d(\boldsymbol{Y},\boldsymbol{Z})=\sqrt{\sum_{k=1}^{m}(y_k-z_k)^2}, \tag{12}$$

其中,欧氏距离 $d(\boldsymbol{Y},\boldsymbol{Z})$的取值范围为$[0,+\infty)$.

在一个信息系统 $S=(U,C\cup D,V,f)$中,对于任意对象 $x(x\in U)$,各条件属性的取值可以视为一个向量 $\boldsymbol{V}(x)=(v_1(x),v_2(x),\cdots,v_m(x))$,因而欧氏距离不仅可用于度量任意 2 个对象之间的距离而且同样可用于度量 2 个集合之间的距离.基于欧氏距离,用以度量等价类与目标集合间相似度的欧氏距离描述如下.

给定一个信息系统 $S=(U,C\cup D,V,f)$,其中,条件属性集合为 $C=\{c_1,c_2,\cdots,c_m\}$,目标集合为 $X=\{x_1,x_2,\cdots,x_{|X|}\}$, $\forall X\subseteq U$.有限论域 U 上由等价关系 $IND(C)$诱导的划分为 $U/IND(C)=\{E_1,E_2,\cdots,E_s\}$,则对于 $\forall E_i(E_i\in U/IND(C),i=1,2,\cdots,s)$,等价类 E_i 和目标集 X 间的欧氏距离公式表示为

$$d(X,E_i)=\frac{1}{|X|}\sum_{j=1}^{|X|}d(V(x_i),V(x_j)), \tag{13}$$

其中,

$$d(V(x_i),V(x_j))=\sqrt{\sum_{k=1}^{m}(v_k(x_i)-v_k(x_j))^2}$$

表示任意 2 个特征向量 $V(x_i)=(v_1(x_i),v_2(x_i),\cdots,v_m(x_i))$ 和 $\boldsymbol{V}(x_j)=(v_1(x_j),v_2(x_j),\cdots,v_m(x_j))$之间的欧氏距离,且 x_i 表示等价类 E_i 中任意对象,x_j 表示目标集 X 中的对象.

为了便于说明问题,所有例子所用的信息表中的数据都是数值型的,且每个属性的量纲都相同.对于存在非数值型数据的信息表,当前研究中有很多方法可以将字符型数据转换成数值型数据,如 LDA[32]等方法.对于存在不同量纲属性的信息表,采用 min-max 标准化等方法对数据进行统一量纲处理即可.

例 1. 给定一个信息表 $S=(U,C\cup D,V,f)$,其中,有限论域为 $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$,条件属性集合为 $C=\{c_1,c_2,c_3,c_4\}$,目标集合为 $X=\{x_1,x_2,x_7,x_8\}$.例 1 的信息表如表 2 所示:

Table 2 Information Table of Example 1

表 2 例 1 的信息表

U	c_1	c_2	c_3	c_4	d
x_1	3	2	1	1	1
x_2	1	2	1	1	1
x_3	3	2	1	1	2
x_4	1	2	1	1	2
x_5	3	3	1	3	2
x_6	3	3	1	3	2
x_7	3	2	2	1	1
x_8	3	2	2	1	1

由等价关系 $IND(C)$诱导的论域 U 上的划分为 $U/IND(C)=\{E_1,E_2,E_3,E_4\}=\{\{x_1,x_3\},\{x_2,x_4\},\{x_5,x_6\},\{x_7,x_8\}\}$.根据式(12),可以得到目标集合中任意对象 $x_i(x_i\in X,i=1,2,\cdots,|X|)$分别与任意等价类 $E_j(E_j\in U/IND(C),j=1,2,3,4)$中任意对象的距离值(为了行文的简洁,仅给出部分结果):

$$d(V(x_i),V(x_i))=0,i=1,2,\cdots,8,$$
$$d(V(x_1),V(x_2))=2.1603,$$
$$d(V(x_1),V(x_5))=3.5824,$$
$$d(V(x_2),V(x_3))=2.8578,$$
$$d(V(x_2),V(x_5))=4.1834,$$
$$d(V(x_2),V(x_7))=3.0551,$$
$$d(V(x_6),V(x_7))=2.6458.$$

然后,根据式(13),等价类 E_1,E_2,E_3,E_4 与目标集 X 之间的欧氏距离分别为

$$d(X,E_1)=1.6202,d(X,E_2)=2.0676,$$
$$d(X,E_3)=3.6487,d(X,E_4)=1.3039.$$

考虑到不同等价类与目标集之间距离的不同,则可认为等价类之间是存在差异的.因此,每个等价类对应的损失函数也应该是不同的.进一步地,借助

例 2 阐述说明在构建综合评价函数以度量风险时考虑等价类之间的差异性能提高分类器性能的原因.

例 2. 根据例 1,各等价类与目标集合的欧氏距离分别为 $d(X,E_1)=1.620\,2$,$d(X,E_2)=2.067\,6$,$d(X,E_3)=3.648\,7$,$d(X,E_4)=1.303\,9$. 又目标集合为 $X=\{x_1,x_2,x_7,x_8\}$,则对于等价类 $E_1=\{x_1,x_3\}$ 和 $E_2=\{x_2,x_4\}$,其粗糙隶属度均为 $Pr(X|E_1)=Pr(X|E_2)=0.500\,0$. 而由于 $d(X,E_1)=1.620\,2\neq d(X,E_2)=2.067\,6$,所以二者与目标集的距离不同. 且 $d(X,E_2)>d(X,E_1)$,则相比于等价类 E_2,等价类 E_1 与目标集合 X 更相似.换言之,尽管二者的粗糙隶属度相同,若仅有一个等价类能够被划分到正域,考虑到等价类之间的差异性,相比于等价类 E_2,等价类 E_1 更应该被划分到正域.

根据例 1 和例 2,由等价关系诱导的论域划分中的确存在一些与目标集合之间的相似度不相同的等价类,即相对于目标集合来说,等价类之间确实存在差异.因此,在度量风险时,应该对这些等价类有所区分.另外,根据 3WD-DTRSs 模型[11-12],由于根据专家经验给定的损失函数面向整个论域,即对于任意等价类,其损失函数集合均相同,从而在贝叶斯决策准则下推导的阈值并不能使得所有等价类达到全局最优划分.

2.1　基于方差的属性权重度量

在决策过程中,决策者一般通过综合考虑各条件属性而进行决策计划的制定.因此,提出了一种获取信息表中属性权重的方法.给定一个信息表 $S=(U,C\cup D,V,f)$,可以得到论域 U 中所有对象 x 在条件属性 c_i 上的取值,不妨表示为集合 $V_i=\{v_i(x)|x\in U\}$.考虑到集合 V_i 的离散程度越大,则属性 c_i 对论域 U 的划分能力越好,因此,采用方差来度量条件属性的权重,样本方差描述为:

设 $x_1,x_2,\cdots,x_n$ 为取自某总体 X 的样本,则它关于样本均值的样本方差(平均偏差平方和)为

$$var(X)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,\tag{14}$$

其中,$\bar{x}$ 称为该样本的样本均值,且 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$.

定义 4. 基于方差的属性权重. 给定一个信息系统 $S=(U,C\cup D,V,f)$,其中,条件属性集合为 $C=\{c_1,c_2,\cdots,c_m\}$,目标集合为 $X(X\subseteq U)$.则对于 $\forall c_i(c_i\in C,i=1,2,\cdots,m)$,基于方差的属性权重函数可定义为

$$\omega_i=\frac{var(V_i)}{\sum_{k=1}^{m}var(V_k)},\tag{15}$$

其中,$var(V_i)=\frac{1}{n-1}\sum_{j=1}^{n}(v_j-\bar{v}_i)^2$ 表示条件属性 c_i 对应的条件属性值集合 $V_i=\{v_i(x)|\forall x\in U\}$ 的方差,$\bar{v}_i$ 表示集合 V_i 的均值.

例 3. 给定一个信息系统 $S=(U,C\cup D,V,f)$,其中,论域为 $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$,条件属性集合为 $C=\{c_1,c_2,c_3,c_4\}$,目标集合为 $X=\{x_1,x_2,x_7,x_8\}$.例 3 的信息表如表 3 所示:

Table 3　Information Table of Example 3

表 3　例 3 的信息表

U	c_1	c_2	c_3	c_4	d
x_1	3	2	1	1	1
x_2	1	2	1	1	1
x_3	3	2	1	1	2
x_4	1	2	1	1	2
x_5	3	3	1	3	2
x_6	3	3	1	3	2
x_7	3	3	1	3	1
x_8	3	3	2	2	1

根据表 3,条件属性 c_1 对应的条件属性值集合为 $V_1=\{v_1(x)|x\in U\}=\{3,1,3,1,3,3,3,3\}$,则根据式(14),条件属性 c_1 对应的条件属性值集合 V_1 的集合方差为 $var(V_1)=0.857\,1$.同理,条件属性 c_2,c_3,c_4 对应的条件属性值集合 V_2,V_3,V_4 的集合方差分别为 $var(V_2)=0.285\,7$,$var(V_3)=0.125\,0$,$var(V_4)=0.982\,1$.再根据式(15),条件属性 c_1,c_2,c_3,c_4 的权重分别为

$$\omega_1=0.381\,0,$$
$$\omega_2=0.127\,0,$$
$$\omega_3=0.055\,6,$$
$$\omega_4=0.436\,5.$$

2.2　基于相似度量的风险度量

如前所述,在说明等价类与目标集合之间存在差异这个问题时,为便于简单直接地说明问题,所采用的是简单的欧氏距离公式.然而,在使用欧氏距离度量等价类与目标集合的差异时,会存在 2 个等价类与目标集合在距离上相同的情况,而实际上 2 个等价类所对应的属性值集合是不同的情况.因此,为了将等价类之间由于属性值不同而产生的差异作为一个影响因子加入到风险度量的模型中,将基于

集合相似度量方法[33]给出一种考虑等价类与目标集合之间差异性的综合评价函数以度量风险.

定义 5. 集合之间的相似度.[33]假设集合 A 和 B 是非空有限论域 U 上的 2 个子集,$S:U\times U\to[0,1]$是一个映射函数,即$(A,B)\to S(A,B)$.则 $S(A,B)$是集合 A 和 B 之间的相似度函数,当且仅当 $S(A,B)$满足 4 个条件:

1) 对于$\forall A,B\subseteq U$,有 $0\leqslant S(A,B)\leqslant 1$;

2) 对于$\forall A,B\subseteq U$,有 $S(A,B)=S(B,A)$;

3) 对于$\forall A,B\subseteq U$,有 $S(A,A)=1$;

4) 对于$\forall A,B\subseteq U$,当且仅当 $A\cap B=\varnothing$时,有 $S(A,B)=0$.

借用文献[33]中相似度量公式 $S(A,B)=\dfrac{|A\cap B|}{|A\cup B|}$的简洁形式,给出度量目标集合与等价类相似度的定义.

定义 6. 改进的相似度量函数. 给定一个信息系统 $S=(U,C\cup D,V,f)$,其中,条件属性集合为 $C=\{c_1,c_2,\cdots,c_m\}$,目标集合为 $X(X\subseteq U)$. 则对于$\forall c_i(c_i\in C)$,$\forall E_j(E_j\in U/IND(C))$与目标集合 X 之间的改进的相似度量函数可定义为

$$\psi^{c_i}(E_j,X)=\frac{|\{x|(x\in X)\wedge(v_i(E_j)=v_i(x))\}\cup E_j|}{|X\cup E_j|},\quad(16)$$

其中,$|\cdot|$表示集合的势,E_j 表示由等价关系 $IND(C)$在有限论域 U 上诱导的等价类,$v_i(E_j)$表示等价类 E_j 在条件属性 c_i 上的值,$v_i(x)$表示任意对象 x $(x\in U)$在条件属性 c_i 上的取值.

将一个本属于目标集合的对象划分到正域带来的风险要小于将其划分到边界域所带来的风险,且将其划分到边界域带来的风险也小于将其划分到负域带来的风险,同理,对于一个本不属于目标集合的对象,同样有类似的规律.此外,对于目标集来说,任意等价类所对应的条件属性值与之越相似,则将该等价类划分到正域所产生的风险越小.因此,为使得评价函数满足以上规律,不妨借用对数函数的简洁形式,基于集合相似度量的概念,提出对条件属性值打分的评价函数的定义.

定义 7. 基于相似度量的评价函数. 给定一个信息系统 $S=(U,C\cup D,V,f)$,其中,条件属性集合为 $C=\{c_1,c_2,\cdots,c_m\}$,目标集合为 $X(X\subseteq U)$.则在任意划分情况 $\delta(\delta\in H,H=\{\mathrm{PP},\mathrm{BP},\mathrm{NP},\mathrm{NN},\mathrm{BN},\mathrm{PN}\})$下,对于$\forall c_i(c_i\in C,i=1,2,\cdots,m)$,基于相似度量的评价函数可定义为

$$S^{\delta}_{E_j}(c_i)=\ln(\psi^{c_i}(E_j,X)^{-1}\times(\eta_{\delta})^2),\quad(17)$$

其中,常数序列 $\eta_{\mathrm{PP}},\eta_{\mathrm{BP}},\eta_{\mathrm{NP}}$和 $\eta_{\mathrm{PN}},\eta_{\mathrm{BN}},\eta_{\mathrm{NN}}$是基于相似度量的评价函数的系数,且其须满足:

$$2<\eta_{\mathrm{PP}}<\eta_{\mathrm{BP}}<\eta_{\mathrm{NP}},2<\eta_{\mathrm{NN}}<\eta_{\mathrm{BN}}<\eta_{\mathrm{PN}}.\quad(18)$$

例 4. 给定一个信息系统 $S=(U,C\cup D,V,f)$,其中,论域为 $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$,条件属性集合为 $C=\{c_1,c_2,c_3,c_4\}$,目标集合为 $X=\{x_1,x_2,x_7,x_8\}$.信息表的具体信息如例 3 中的表 3 所示.

由等价关系 $IND(C)$诱导的论域 U 上的划分为 $U/IND(C)=\{E_1,E_2,E_3,E_4\}=\{\{x_1,x_3\},\{x_2,x_4\},\{x_5,x_6,x_7\},\{x_8\}\}$.根据定义 6,在属性 c_1 上,等价类 E_1 与目标集合 X 之间的改进的相似度量函数为

$$\psi^{c_1}(E_1,X)=\frac{|\{x_1,x_3,x_7,x_8\}|}{|\{x_1,x_2,x_3,x_7,x_8\}|}=0.800\,0.$$

根据定义 7,不妨将基于相似度量的评价函数的系数取为 $\eta_{\mathrm{PP}}=3.000\,0,\eta_{\mathrm{BP}}=5.000\,0,\eta_{\mathrm{NP}}=9.000\,0,\eta_{\mathrm{NN}}=3.500\,0,\eta_{\mathrm{BN}}=7.000\,0,\eta_{\mathrm{PN}}=12.000\,0$.则根据定义 7,对于条件属性 c_1,在任意的划分情况 $\delta(\delta\in H)$下,等价类 $E_1=\{x_1,x_3\}$与目标集 X 之间基于相似度量的评价函数分别为

$$S^{\mathrm{PP}}_{E_1}(c_1)=2.420\,2,$$
$$S^{\mathrm{BP}}_{E_1}(c_1)=3.442\,0,$$
$$S^{\mathrm{NP}}_{E_1}(c_1)=4.617\,6,$$
$$S^{\mathrm{NN}}_{E_1}(c_1)=2.728\,7,$$
$$S^{\mathrm{BN}}_{E_1}(c_1)=4.115\,0,$$
$$S^{\mathrm{PN}}_{E_1}(c_1)=5.193\,0.$$

同理,对于条件属性 c_2,c_3,c_4,在任意划分情况 $\delta(\delta\in H)$下,等价类 $E_1=\{x_1,x_3\}$与目标集 X 之间的改进的相似度量函数分别为

$$\psi^{c_2}(E_1,X)=\frac{|\{x_1,x_2,x_3\}|}{|\{x_1,x_2,x_3,x_7,x_8\}|}=0.600\,0,$$
$$\psi^{c_3}(E_1,X)=0.800\,0,$$
$$\psi^{c_4}(E_1,X)=0.600\,0.$$

同理,对于等价类 E_2,E_3,E_4,在任意条件属性下及任意划分情况下,基于相似度量的评价函数都可以计算.

性质 1. 给定一个信息表 $S=(U,C\cup D,V,f)$,其中,条件属性集合为 $C=\{c_1,c_2,\cdots,c_m\}$.则给定 $\forall c_i(c_i\in C,i=1,2,\cdots,m)$的条件下,对于$\forall E_j(E_j\in U/IND(C))$,基于相似度量的评价函数满足 $S^{\delta}_{E_j}(c_i)\in(-\infty,+\infty)$.

证明. 因为对于 $\forall c_i(c_i\in C,i=1,2,\cdots,m)$,

$\forall E_j(E_j\in U/IND(C))$及给定的目标集合 $X(\forall X\subseteq U)$,不等式 $0<\psi^{c_i}(E_j,X)\leqslant 1$ 成立. 则当满足条件 $2<\eta_{PP}<\eta_{BP}<\eta_{NP}$,$2<\eta_{NN}<\eta_{BN}<\eta_{PN}$时,$S^{\delta}_{E_j}(c_i)\in(-\infty,+\infty)$成立.　　证毕.

值得指出的是,对于给定的条件属性及在相同的划分动作下,评价值越高的等价类被划分到正域而产生的风险越小.

性质 2. 给定一个信息系统 $S=(U,C\cup D,V,f)$,对于 $\forall E_j(E_j\in U/IND(C))$,$\forall c_i(c_i\in C,i=1,2,\cdots,m)$,如果常数序列满足条件$(\eta_{PN}-\eta_{BN})(\eta_{NP}-\eta_{BP})>(\eta_{BP}-\eta_{PP})(\eta_{BN}-\eta_{NN})$,则基于相似度量的评价函数满足不等式 $(S^{PN}_{E_j}(c_i)-S^{BN}_{E_j}(c_i))(S^{NP}_{E_j}(c_i)-S^{BP}_{E_j}(c_i))>(S^{BP}_{E_j}(c_i)-S^{PP}_{E_j}(c_i))(S^{BN}_{E_j}(c_i)-S^{NN}_{E_j}(c_i))$.

证明. 当 $a\geqslant 1,x>2$ 时,函数 $z=\ln(ax^2)$的斜率大于或等于 1,则有 $x_1-x_2>(\ln(ax_1^2)-\ln(ax_2^2))$. 由定义 6,$S^{\delta}_{E_j}(c_i)=\ln(\psi^{c_i}(E_j,X)^{-1}\times(\eta_{\delta})^2)$,$2<\eta_{PP}<\eta_{BP}<\eta_{NP}$,$2<\eta_{NN}<\eta_{BN}<\eta_{PN}$,且由性质 1 可知 $0<\psi^{c_i}(E_j,X)\leqslant 1$.所以当 $(\eta_{PN}-\eta_{BN})(\eta_{NP}-\eta_{BP})>(\eta_{BP}-\eta_{PP})(\eta_{BN}-\eta_{NN})$成立时,$(S^{PN}_{E_j}(c_i)-S^{BN}_{E_j}(c_i))(S^{NP}_{E_j}(c_i)-S^{BP}_{E_j}(c_i))>(S^{BP}_{E_j}(c_i)-S^{PP}_{E_j}(c_i))(S^{BN}_{E_j}(c_i)-S^{NN}_{E_j}(c_i))$成立.　　证毕.

性质 3. 给定一个信息系统 $S=(U,C\cup D,V,f)$,对于 $\forall E_j(E_j\in U/IND(C))$,$\forall c_i(c_i\in C,i=1,2,\cdots,m)$,如果常数序列满足不等式 $(\eta_{PN}-\eta_{BN})(\eta_{NP}-\eta_{BP})<(\eta_{BP}-\eta_{PP})(\eta_{BN}-\eta_{NN})$,则基于相似度量的评价函数满足不等式 $(S^{PN}_{E_j}(c_i)-S^{BN}_{E_j}(c_i))(S^{NP}_{E_j}(c_i)-S^{BP}_{E_j}(c_i))<(S^{BP}_{E_j}(c_i)-S^{PP}_{E_j}(c_i))(S^{BN}_{E_j}(c_i)-S^{NN}_{E_j}(c_i))$.

证明. 证明过程类似于性质 2.

决策者在做决策时一般是综合考虑每个对象的各条件属性,因此根据定义 7,给出以下综合评价函数,以综合各条件属性的评价值.

定义 8. 综合评价函数.给定一个信息表 $S=(U,C\cup D,V,f)$,其中,条件属性集合为 $C=\{c_1,c_2,\cdots,c_m\}$,目标集合为 $X(X\subseteq U)$. 则对于 $\forall E_j(E_j\in U/IND(C))$,任意划分情况 $\delta(\delta\in H)$下的综合评价函数可以定义为

$$\mu_{\delta}(E_j)=\sum_{i=1}^{m}\omega_i S^{\delta}_{E_j}(c_i),\tag{19}$$

其中,ω_i 表示条件属性 $c_i(c_i\in C,i=1,2,\cdots,m)$的权重,$S^{\delta}_{E_j}(c_i)$表示对于 $\forall c_i(c_i\in C,i=1,2,\cdots,m)$,在任意划分情况 $\delta(\delta\in H)$下的评价函数. 换言之,对于每一个等价类,都可以得到自适应的综合评价函数集合.并且,综合评价值越大,将该对象进行相应划分产生的风险越大.

例 5. 根据例 4,对于任意条件属性 $c_i(c_i\in C)$,已得到在任意划分情况 $\delta(\delta\in H)$下,任意等价类 $E_j(E_j\in U/IND(C))$与目标集合 X 间的评价值.如对于等价类 E_1,相应的评价值为

$$S^{PP}_{E_1}(c_1)=2.4204,S^{BP}_{E_1}(c_1)=3.4420,$$
$$S^{NP}_{E_1}(c_1)=4.6176,S^{NN}_{E_1}(c_1)=2.7287,$$
$$S^{BN}_{E_1}(c_1)=4.1150,S^{PN}_{E_1}(c_1)=5.1930;$$
$$S^{PP}_{E_1}(c_2)=2.7081,S^{BP}_{E_1}(c_2)=3.7297,$$
$$S^{NP}_{E_1}(c_2)=4.9053,S^{NN}_{E_1}(c_2)=3.0164,$$
$$S^{BN}_{E_1}(c_2)=4.4026,S^{PN}_{E_1}(c_2)=5.4806;$$
$$S^{PP}_{E_1}(c_3)=2.4204,S^{BP}_{E_1}(c_3)=3.4420,$$
$$S^{NP}_{E_1}(c_3)=4.6176,S^{NN}_{E_1}(c_3)=2.7287,$$
$$S^{BN}_{E_1}(c_3)=4.1150,S^{PN}_{E_1}(c_3)=5.1930;$$
$$S^{PP}_{E_1}(c_4)=2.7081,S^{BP}_{E_1}(c_4)=3.7297,$$
$$S^{NP}_{E_1}(c_4)=4.9053,S^{NN}_{E_1}(c_4)=3.0164,$$
$$S^{BN}_{E_1}(c_4)=4.4026,S^{PN}_{E_1}(c_4)=5.4806.$$

根据例 3,属性 c_1,c_2,c_3,c_4 的权重分别为 $\omega_1=0.3810,\omega_2=0.1270,\omega_3=0.0556,\omega_4=0.4365$.因此,对于 $E_1=\{x_1,x_3\}$,在不同的划分情况 $\delta(\delta\in H)$下的综合评价值分别为

$$\mu_{PP}(E_1)=2.5824,\mu_{BP}(E_1)=3.6041,$$
$$\mu_{NP}(E_1)=4.7797,\mu_{NN}(E_1)=2.8908,$$
$$\mu_{BN}(E_1)=4.2770,\mu_{PN}(E_1)=5.3550.$$

性质 4. 给定一个信息系统 $S=(U,C\cup D,V,f)$,对于 $\forall E_j(E_j\in U/IND(C))$及 $\forall c_i(c_i\in C,i=1,2,\cdots,m)$,若 $(S^{PN}_{E_j}(c_i)-S^{BN}_{E_j}(c_i))(S^{NP}_{E_j}(c_i)-S^{BP}_{E_j}(c_i))>(S^{BP}_{E_j}(c_i)-S^{PP}_{E_j}(c_i))(S^{BN}_{E_j}(c_i)-S^{NN}_{E_j}(c_i))$成立,则 $(\mu_{PN}(E_j)-\mu_{BN}(E_j))\times(\mu_{NP}(E_j)-\mu_{BP}(E_j))-(\mu_{BP}(E_j)-\mu_{PP}(E_j))(\mu_{BN}(E_j)-\mu_{NN}(E_j))$成立.

证明. 因为

$$(\mu_{PN}(E_j)-\mu_{BN}(E_j))(\mu_{NP}(E_j)-\mu_{BP}(E_j))-(\mu_{BP}(E_j)-\mu_{PP}(E_j))\times(\mu_{BN}(E_j)-\mu_{NN}(E_j))=$$
$$\sum_{i=1}^{m}\omega_i(S^{PN}_{E_j}(c_i)-S^{BN}_{E_j}(c_i))\times\sum_{i=1}^{m}\omega_i(S^{NP}_{E_j}(c_i)-S^{BP}_{E_j}(c_i))-\sum_{i=1}^{m}\omega_i(S^{BP}_{E_j}(c_i)-S^{PP}_{E_j}(c_i))\times\sum_{i=1}^{m}\omega_i(S^{BN}_{E_j}(c_i)-S^{NN}_{E_j}(c_i)),$$

而由已知条件 $(S^{PN}_{E_j}(c_i)-S^{BN}_{E_j}(c_i))(S^{NP}_{E_j}(c_i)-S^{BP}_{E_j}(c_i))>(S^{BP}_{E_j}(c_i)-S^{PP}_{E_j}(c_i))(S^{BN}_{E_j}(c_i)-S^{NN}_{E_j}(c_i))$,则:

$$\sum_{i=1}^{m}\omega_i(S_{E_j}^{\mathrm{PN}}(c_i)-S_{E_j}^{\mathrm{BN}}(c_i))\times\sum_{i=1}^{m}\omega_i(S_{E_j}^{\mathrm{NP}}(c_i)-S_{E_j}^{\mathrm{BP}}(c_i))-\sum_{i=1}^{m}\omega_i(S_{E_j}^{\mathrm{BP}}(c_i)-S_{E_j}^{\mathrm{PP}}(c_i))\times\sum_{i=1}^{m}\omega_i(S_{E_j}^{\mathrm{BN}}(c_i)-S_{E_j}^{\mathrm{NN}}(c_i))>0,$$

从而 $(\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{BN}}(E_j))(\mu_{\mathrm{NP}}(E_j)-\mu_{\mathrm{BP}}(E_j))>(\mu_{\mathrm{BP}}(E_j)-\mu_{\mathrm{PP}}(E_j))(\mu_{\mathrm{BN}}(E_j)-\mu_{\mathrm{NN}}(E_j))$成立.

证毕.

2.3 SBA-3WD-SF 模型阈值推导

根据决策粗糙集模型[11-12],SBA-3WD-SF 模型同样有 2 个状态 $T=\{X,\neg X\}$及 3 个动作 $A=\{a_{\mathrm{P}},a_{\mathrm{B}},a_{\mathrm{N}}\}$,则相应的综合评价函数可由表 4 表示:

Table 4 Comprehensive Evaluation Function

表 4 综合评价函数

T	A		
	a_{P}	a_{B}	a_{N}
X	μ_{PP}	μ_{BP}	μ_{NP}
$\neg X$	μ_{PN}	μ_{BN}	μ_{NN}

根据贝叶斯决策过程[11],采取不同动作 $a_{\mathrm{P}},a_{\mathrm{B}},a_{\mathrm{N}}$ 时,对于 $\forall E_j(E_j\in U/IND(C))$,基于相似度量的期望风险分别为

$$\begin{aligned}R(a_{\mathrm{P}}|E_j)&=\mu_{\mathrm{PP}}Pr(X|E_j)+\mu_{\mathrm{PN}}Pr(\neg X|E_j),\\R(a_{\mathrm{B}}|E_j)&=\mu_{\mathrm{BP}}Pr(X|E_j)+\mu_{\mathrm{BN}}Pr(\neg X|E_j),\\R(a_{\mathrm{N}}|E_j)&=\mu_{\mathrm{NP}}Pr(X|E_j)+\mu_{\mathrm{NN}}Pr(\neg X|E_j).\end{aligned}\tag{20}$$

根据贝叶斯决策准则,使得期望风险最小的一套动作,便是最优划分计划.从而,可得到 3 条决策规则:

(P)如果 $R(a_{\mathrm{P}}|E_j)\leqslant R(a_{\mathrm{B}}|E_j)$且 $R(a_{\mathrm{P}}|E_j)\leqslant R(a_{\mathrm{N}}|E_j)$,则 $E_j\subseteq POS(X)$;

(B)如果 $R(a_{\mathrm{B}}|E_j)\leqslant R(a_{\mathrm{P}}|E_j)$且 $R(a_{\mathrm{B}}|E_j)\leqslant R(a_{\mathrm{N}}|E_j)$,则 $E_j\subseteq BND(X)$;

(N)如果 $R(a_{\mathrm{N}}|E_j)\leqslant R(a_{\mathrm{P}}|E_j)$且 $R(a_{\mathrm{N}}|E_j)\leqslant R(a_{\mathrm{B}}|E_j)$,则 $E_j\subseteq NEG(X)$.

这 3 条决策规则只与判别函数 $Pr(X|E_j)$及各综合评价函数有关.根据定义 6 及性质 2,各综合评价函数满足条件 $\mu_{\mathrm{PP}}(E_j)<\mu_{\mathrm{BP}}(E_j)<\mu_{\mathrm{NP}}(E_j)$,$\mu_{\mathrm{NN}}(E_j)<\mu_{\mathrm{BN}}(E_j)<\mu_{\mathrm{PN}}(E_j)$.因此,在此大小关系的前提下,使得决策风险最小的决策规则表示为

$$(\mathrm{P})\begin{cases}R(a_{\mathrm{P}}|E_j))\leqslant R(a_{\mathrm{N}}|E_j))\Leftrightarrow Pr(X|E_j)\geqslant\dfrac{\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{NN}}(E_j)}{\mu_{\mathrm{NP}}(E_j)-\mu_{\mathrm{PP}}(E_j)+\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{NN}}(E_j)},\\R(a_{\mathrm{P}}|E_j))\leqslant R(a_{\mathrm{B}}|E_j))\Leftrightarrow Pr(X|E_j)\geqslant\dfrac{\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{BN}}(E_j)}{\mu_{\mathrm{BP}}(E_j)-\mu_{\mathrm{PP}}(E_j)+\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{BN}}(E_j)}.\end{cases}$$

$$(\mathrm{B})\begin{cases}R(a_{\mathrm{B}}|E_j))\leqslant R(a_{\mathrm{N}}|E_j))\Leftrightarrow Pr(X|E_j)\geqslant\dfrac{\mu_{\mathrm{BN}}(E_j)-\mu_{\mathrm{NN}}(E_j)}{\mu_{\mathrm{NP}}(E_j)-\mu_{\mathrm{BP}}(E_j)+\mu_{\mathrm{BN}}(E_j)-\mu_{\mathrm{NN}}(E_j)},\\R(a_{\mathrm{B}}|E_j))\leqslant R(a_{\mathrm{P}}|E_j))\Leftrightarrow Pr(X|E_j)\leqslant\dfrac{\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{BN}}(E_j)}{\mu_{\mathrm{BP}}(E_j)-\mu_{\mathrm{PP}}(E_j)+\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{BN}}(E_j)}.\end{cases}$$

$$(\mathrm{N})\begin{cases}R(a_{\mathrm{N}}|E_j))\geqslant R(a_{\mathrm{B}}|E_j))\Leftrightarrow Pr(X|E_j)\leqslant\dfrac{\mu_{\mathrm{BN}}(E_j)-\mu_{\mathrm{NN}}(E_j)}{\mu_{\mathrm{NP}}(E_j)-\mu_{\mathrm{BP}}(E_j)+\mu_{\mathrm{BN}}(E_j)-\mu_{\mathrm{NN}}(E_j)},\\R(a_{\mathrm{N}}|E_j)\geqslant R(a_{\mathrm{P}}|E_j)\Leftrightarrow Pr(X|E_j)\leqslant\dfrac{\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{NN}}(E_j)}{\mu_{\mathrm{NP}}(E_j)-\mu_{\mathrm{PP}}(E_j)+\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{NN}}(E_j)}.\end{cases}$$

进一步,为了得到决策规则的简约形式,简记处理为

$$\begin{aligned}\alpha_j&=\frac{\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{BN}}(E_j)}{\mu_{\mathrm{BP}}(E_j)-\mu_{\mathrm{PP}}(E_j)+\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{BN}}(E_j)},\\\beta_j&=\frac{\mu_{\mathrm{BN}}(E_j)-\mu_{\mathrm{NN}}(E_j)}{\mu_{\mathrm{NP}}(E_j)-\mu_{\mathrm{BP}}(E_j)+\mu_{\mathrm{BN}}(E_j)-\mu_{\mathrm{NN}}(E_j)},\\\gamma_j&=\frac{\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{NN}}(E_j)}{\mu_{\mathrm{NP}}(E_j)-\mu_{\mathrm{PP}}(E_j)+\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{NN}}(E_j)}.\end{aligned}\tag{21}$$

根据性质 4,则有不等式:

$$\frac{\mu_{\mathrm{BP}}(E_j)-\mu_{\mathrm{PP}}(E_j)}{\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{BN}}(E_j)}<\frac{\mu_{\mathrm{NP}}(E_j)-\mu_{\mathrm{BP}}(E_j)}{\mu_{\mathrm{BN}}(E_j)-\mu_{\mathrm{NN}}(E_j)},$$

而因为不等式:

$$\frac{b}{a}>\frac{d}{c}\Rightarrow\frac{b}{a}>\frac{b+d}{a+c}>\frac{d}{c}(a,b,c,d>0)$$

成立,则有不等式:

$$\frac{\mu_{\mathrm{BP}}(E_j)-\mu_{\mathrm{PP}}(E_j)}{\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{BN}}(E_j)}<\frac{\mu_{\mathrm{NP}}(E_j)-\mu_{\mathrm{PP}}(E_j)}{\mu_{\mathrm{PN}}(E_j)-\mu_{\mathrm{NN}}(E_j)}<\frac{\mu_{\mathrm{NP}}(E_j)-\mu_{\mathrm{BP}}(E_j)}{\mu_{\mathrm{BN}}(E_j)-\mu_{\mathrm{NN}}(E_j)}$$

成立,即 $0\leqslant\beta_j<\gamma_j<\alpha_j\leqslant1$ 成立.因此,规则(P)(B)(N)可以等价地表示为

(P)如果 $Pr(X|E_j)\geqslant\alpha_j$,则 $E_j\subseteq POS(X)$;

(B)如果 $\beta_j<Pr(X|E_j)<\alpha_j$,则 $E_j\subseteq BND(X)$;

(N)如果 $Pr(X|E_j)\leqslant\beta_j$,则 $E_j\subseteq NEG(X)$.

综上所述,根据决策规则,给出了一个阈值自适应的三支决策分类器.

2.4　算法步骤

根据 2.1～2.3 节,所提模型的基本步骤可由算法 1,2,3 表示.在算法 1,2,3 中,m 表示条件属性的个数,$\omega_i(i=1,2,\cdots,m)$表示每个条件属性的权重.$\delta(\delta\in H)$表示任意划分情况.E_j 表示由等价关系 $IND(C)$在论域 U 上诱导的划分 $U/IND(C)$中的任意等价类,s 表示 $U/IND(C)$中的元素个数.pos,bnd,neg 分别表示正域、边界域及负域.

算法 1. 计算每个条件属性的权重.

输入:$S=(U,C\cup D,V,f)$;

输出:ω_i.

① for $i=1$ to m

$$var(V_i)=\frac{1}{n}\sum_{j=1}^{n}(v_j-\overline{v}_i)^2;$$

$$Temp+=var(V_i);$$

② end for

③ for $i=1$ to m

$$\omega_i\leftarrow\frac{var(V_i)}{Temp};$$

④ end for

算法 2. 构造综合评价函数.

输入:$S=(U,C\cup D,V,f),X,\omega_i,\eta_\delta$;

输出:$\mu_\delta(E_j)$.

① for each $c_i\in C$

for each $E_j\in U/IND(C)$

$$\psi^{c_i}(E_j,X)=\frac{|\{x|(x\in X)\wedge(v_i(E_j)=v_i(x))\}\cup E_j|}{|X\cup E_j|};$$

② end for

③ end for

④ for $i=1$ to m

while $2<\eta_{PP}<\eta_{BP}<\eta_{NP}$ && $2<\eta_{NN}<\eta_{BN}<\eta_{PN}$ && $(\eta_{PN}-\eta_{BN})(\eta_{NP}-\eta_{BP})>(\eta_{BP}-\eta_{PP})(\eta_{BN}-\eta_{NN})$ do

⑤ $\mu_\delta(E_j)\leftarrow\sum_{i=1}^{m}\omega_i\ln(\psi^{c_i}(E_j,X)^{-1}\times(\eta_\delta)^2)$;

⑥ end while

⑦ end for

算法 3. 计算模型的分类准确率(acc)及召回率($recall$).

输入:$S=(U,C\cup D,V,f),X,\mu_\delta(E_j)$;

输出:$acc,recall$.

① 初始化 $pos=\varnothing,bnd=\varnothing,neg=\varnothing$.

② for $j=1$ to s

$$\alpha_j\leftarrow\frac{\mu_{PN}(E_j)-\mu_{BN}(E_j)}{\mu_{BP}(E_j)-\mu_{PP}(E_j)+\mu_{PN}(E_j)-\mu_{BN}(E_j)}$$

and

$$\beta_j\leftarrow\frac{\mu_{BN}(E_j)-\mu_{NN}(E_j)}{\mu_{NP}(E_j)-\mu_{BP}(E_j)+\mu_{BN}(E_j)-\mu_{NN}(E_j)};$$

$$Pr(X|E_j)\leftarrow\frac{|X\cap E_j|}{|E_j|};$$

③ if $Pr(X|E_j)\geqslant\alpha_j$ then

④ $pos\leftarrow pos\cup E_j$;

⑤ else if $\beta_j<Pr(X|E_j)<\alpha_j$ then

⑥ $bnd\leftarrow bnd\cup E_j$;

⑦ else $neg\leftarrow neg\cup E_j$;

⑧ end if

⑨ end for

⑩ $$acc\leftarrow\frac{|X\cap pos|+|\neg X\cap neg|}{|U|},$$

$$recall\leftarrow\frac{|X\cap pos|}{|X|};$$

⑪ return $acc,recall$.

通过分析算法 1,2,3,所提算法的时间复杂度为 $T(n)=O(n^2)$.因此,所提算法在现实生活中是有效可行的.

3　对比实验

3.1　实验指标

在垃圾邮件过滤领域,用户通常关心的是邮件正确分类数量占总体邮件的比例.当前评价垃圾邮件过滤器性能的方法中,常用的评价指标为准确率(acc)及召回率($recall$)等[34].

3.1.1　准确率

对于垃圾邮件过滤器,垃圾邮件被分到垃圾邮件类及合法邮件被分到合法邮件类的数量占邮件总量比例越高,则该过滤器的性能越好.因此,不妨采用准确率来衡量过滤器的性能.

$$acc=\frac{n_{s\to s}+n_{l\to l}}{|U|},\tag{22}$$

其中,$n_{s\to s}$表示实为垃圾邮件的对象被分为垃圾邮件类的数量,$n_{l\to l}$表示实为合法邮件的对象被分为合法邮件的数量,$|U|$表示论域中的所有邮件的数量.

3.1.2　召回率

对于垃圾邮件过滤器,垃圾邮件被划分到垃圾邮件类的占实际垃圾邮件比例越高,则说明该过滤器对垃圾邮件的识别能力越好,即过滤器的性能也越好.因此,不妨采用召回率来衡量过滤器的性能.

$$recall=\frac{n_{s\to s}}{N_s}, \tag{23}$$

其中,N_s 表示实为垃圾邮件的对象个数.

3.2 数据预处理与参数设置

相较于传统的二分类模型,由于三支决策模型增加了延迟决策域,则其能有效降低错误分类代价,从而有效提高分类的准确率.特别地,相较于 3WD-DTRSs 模型[11],所提的 SBA-3WD-SF 模型能提高准确率及召回率.为了验证这一点,使用当前垃圾邮件过滤研究领域的 2 个常用数据集,其一是来自 UCI 机器学习数据资料库的 Spambase 数据集,其二是来自文献[35]提供的 PU1 Corpus 数据集,二者的基本信息如表 5 所示:

Table 5 Basic Information of Datasets

表 5 数据集基本信息

Datasets	Sources	Sample Size	Spam	Legal Mail
Spambase	UCI	4 601	1 813	2 788
PU1 Corpus	Androutsopoulos	1 099	481	618

对于 Spambase 数据集,实验随机选取了 3 000 个样本作为训练数据集,剩下的 1 601 个样本作为测试数据集.首先,采用 Entropy-MDL 方法[36]对训练数据集及测试数据集进行离散化处理,然后选用条件互信息最大化法[37]对数据集进行特征选择,从 57 个条件属性中选取了 20 个条件属性.对于 PU1 Corpus 数据集,由于该语料库本身已被分为 10 个部分,因此实验随机选取了其中的 8 个部分作为训练数据集,剩下的 2 个部分作为测试数据集.进一步采用基于互信息的单变量特征提取方法(KBest feature selection)对数据集进行特征选择,并依据互信息最大化原则,选择了前 200 个条件属性.

为测试模型的性能,针对 3WD-DTRSs 模型[11]及所提 SBA-3WD-SF 模型,实验分别随机给出了 5 组参数如表 6 及表 7 所示.根据 1.2 节中二元 NB 分类器[30]的原理,其模型的阈值相应地可扩展表示为 $\xi=\lambda_{NP}/\lambda_{PN}$,如表 8 所示.

Table 6 Loss Functions of 3WD-DTRSs Model[11]

表 6 3WD-DTRSs 模型[11]的损失函数

Case	λ_{NP}	λ_{BP}	λ_{PP}	λ_{PN}	λ_{BN}	λ_{NN}
Case1	9.000 0	8.000 0	1.000 0	9.000 0	7.000 0	3.000 0
Case2	7.500 0	6.000 0	1.500 0	8.500 0	6.500 0	2.500 0
Case3	8.300 0	7.500 0	2.200 0	7.900 0	7.100 0	1.800 0
Case4	6.000 0	5.000 0	1.000 0	5.000 0	4.000 0	1.000 0
Case5	8.500 0	7.500 0	0.600 0	2.400 0	2.000 0	0.800 0

Table 7 Coefficients of Evaluation Functions Based on Similarity Measure

表 7 基于相似度量的评价函数的系数

Case	η_{NP}	η_{BP}	η_{PP}	η_{PN}	η_{BN}	η_{NN}
Case1	9.000 0	8.000 0	3.000 0	9.000 0	7.000 0	3.000 0
Case2	7.500 0	6.000 0	3.500 0	8.500 0	6.500 0	2.500 0
Case3	8.300 0	7.500 0	2.200 0	7.900 0	7.100 0	2.800 0
Case4	6.500 0	5.000 0	2.300 0	5.000 0	4.000 0	2.400 0
Case5	8.500 0	7.500 0	2.600 0	5.400 0	3.000 0	2.800 0

Table 8 Threshold of Binary NB Classifier[30]

表 8 二元 NB 分类器[30]的阈值

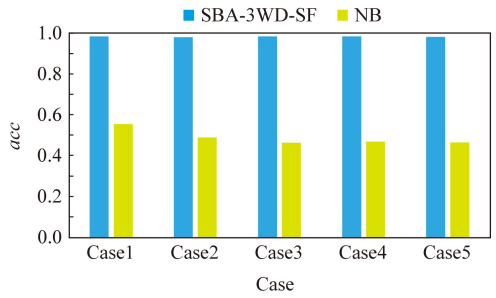
Case	ξ
Case1	1.000 0
Case2	1.680 0
Case3	2.045 0
Case4	2.912 0
Case5	3.808 0

3.3 实验结果与分析

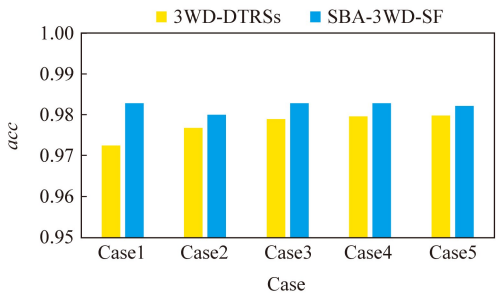
首先,针对分类器的准确率,SBA-3WD-SF 模型与二元 NB 分类器[30]及 3WD-DTRSs 模型[11]在数据集 Spambase 上的表现分别如图 2(a)(b)所示.从图 2(a)(b)中可以发现在 Spambase 数据集上,3WD-DTRSs 模型及 SBA-3WD-SF 模型的分类准确率均优于二元 NB 模型,且 SBA-3WD-SF 模型要优于 3WD-DTRSs 模型.另外,如图 3(a)(b)所示,可以发现 3WD-DTRSs 模型及 SBA-3WD-SF 模型的分类准确率均优于二元 NB 模型,且 SBA-3WD-SF 模型的分类准确率优于 3WD-DTRSs 模型.

其次,针对分类器的召回率,SBA-3WD-SF 模型与二元 NB 分类器及 3WD-DTRSs 模型在数据集 PU1 Corpus 上的表现分别如图 3(c)及图 3(d)所示.在 PU1 Corpus 数据集上,3WD-DTRSs 模型及 SBA-3WD-SF 模型的召回率均优于二元 NB 模型,且 SBA-3WD-SF 模型的召回率要优于 3WD-DTRSs 模型.在 Spambase 数据集上,如图 2(c)(d)所示,3WD-DTRSs 模型及 SBA-3WD-SF 模型的召回率均优于二元 NB 模型,且 SBA-3WD-SF 模型的召回率要优于 3WD-DTRSs 模型.

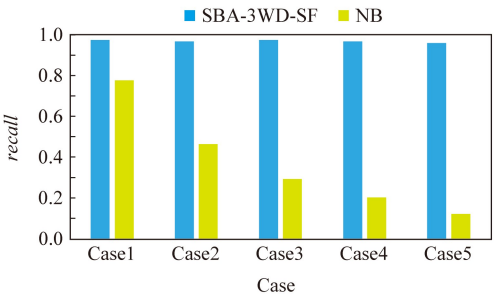
在表 6~8 中的任意参数条件下,3WD-DTRSs 模型[11]及 SBA-3WD-SF 模型在 Spambase 或 PU1 Corpus 数据集上的分类准确率及召回率都要优于二元 NB 模型[30],且 SBA-3WD-SF 模型无论在分类



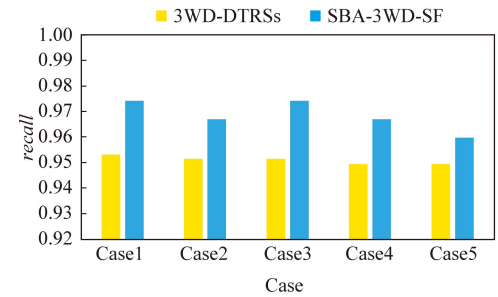
(a) Accuracy of SBA-3WD-SF and NB on Spambase



(b) Accuracy of SBA-3WD-SF and 3WD-DTRSs on Spambase



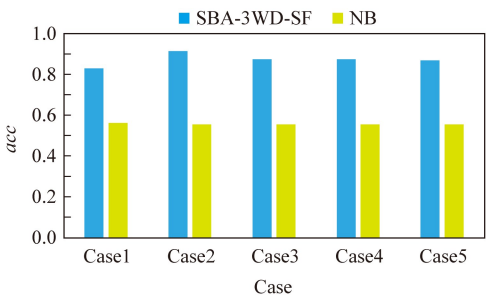
(c) Recall of SBA-3WD-SF and NB on Spambase



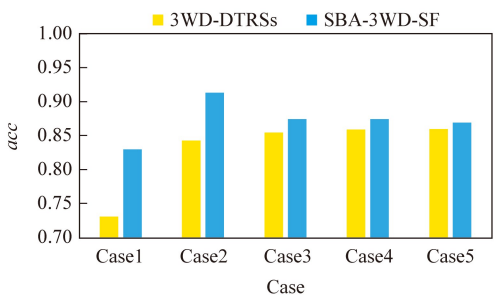
(d) Recall of SBA-3WD-SF and 3WD-DTRSs on Spambase

Fig. 2　Results of comparative experiments on Spambase dataset

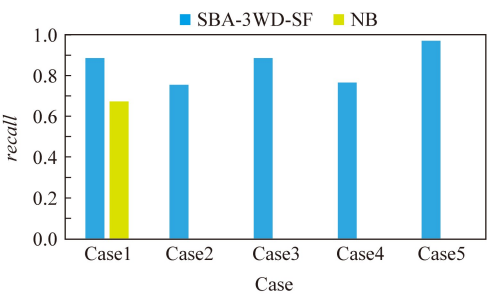
图 2　Spambase 数据集上的对比结果



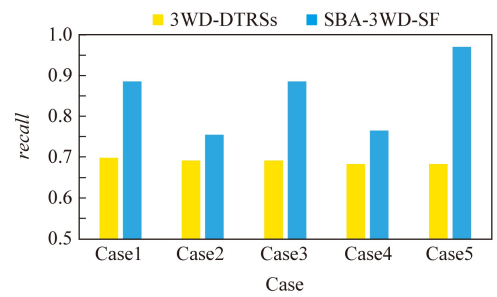
(a) Accuracy of SBA-3WD-SF and NB on PU1 Corpus



(b) Accuracy of SBA-3WD-SF and 3WD-DTRSs on PU1 Corpus



(c) Recall of SBA-3WD-SF and NB on PU1 Corpus



(d) Recall of SBA-3WD-SF and 3WD-DTRSs on PU1 Corpus

Fig. 3　Results of comparative experiments on PU1 Corpus

图 3　PU1 Corpus 数据上的对比结果

准确率还是召回率上也同样明显优于 3WD-DTRSs 模型.而表 6～8 中的参数都是随机给定的,因此根据极大似然法原理,3WD-DTRSs 模型及 SBA-3WD-SF 模型的分类准确率及召回率优于二元 NB 模型,且 SBA-3WD-SF 模型无论在分类准确率还是召回率上也同样明显优于 3WD-DTRSs 模型.

特别地,由于 NB 分类器为了求解后验概率而采用了特征条件独立性假设这一严格的条件,因此该模型牺牲了一定的分类准确率.值得指出的是,如图 3(c)所示,NB 模型的鲁棒性较差,而 3WD-DTRSs

及 SBA-3WD-SF 模型则具有良好的鲁棒性.此外,SBA-3WD-SF 模型在 3WD-DTRSs 模型的基础上,在进行风险度量时考虑了等价类之间差异带来的影响,因此,在理论上,SBA-3WD-SF 模型的分类准确率及召回率都优于 3WD-DTRSs 模型.而对于二分类及三支决策来说,从理论上,相较于二分类模型,三支决策模型由于增加了延迟决策域而能有效降低分类风险代价,从而提高模型的分类准确率及召回率等.从实验结果上来看,由图 2 和图 3 可知在 2 个垃圾邮件分类的数据集上,SBA-3WD-SF 模型及 3WD-DTRSs 模型的分类准确率及召回率的确都优于二元 NB 模型,SBA-3WD-SF 模型也要明显优于 3WD-DTRSs 模型.此外,由图 2(b)(d)、图 3(b)(d)可以发现,在 Spambase 或者 PU1 Corpus 数据集上,所提模型 SBA-3WD-SF 相较于 3WD-DTRSs 模型来说,分类准确率及召回率提高的幅度没有特别大,而在垃圾邮件过滤领域,即使只有一封邮件被错误分类都会产生不可估量的代价.因此,在垃圾邮件过滤领域,SBA-3WD-SF 模型具有显著的价值.

4 结 语

垃圾邮件分类中分类准确率及召回率等是用户最为关注的问题.相较于二分类模型,增加了延迟决策域的三支决策模型能有效降低错分代价,从而提高分类准确率及召回率.通过度量等价类与目标集的距离,发现相对于目标集合来说,等价类之间的确存在差异.因此,在 3WD-DTRSs 模型的基础上提出了一种基于相似度量的阈值自适应三支垃圾邮件过滤模型.首先根据集合方差提出了一种计算条件属性权重的方法,然后考虑等价类之间的差异性带来的影响,基于等价类之间的相似度重新构建了综合评价函数,然后依据贝叶斯决策准则推导出自适应阈值.在垃圾邮件过滤领域,过滤器的性能至关重要.对比实验表明所提 SBA-3WD-SF 模型在分类准确率及召回率等指标上都优于 3WD-DTRSs 模型[11]及二元 NB 模型[30],即验证了 SBA-3WD-SF 模型的合理性及有效性,同时也说明了 SBA-3WD-SF 模型在垃圾邮件过滤领域的价值.值得提出的是,除了损失函数,影响垃圾邮件过滤器的性能的因素有很多,如判别函数及参数的求解方式等.因此,今后的研究工作将重点关注判定函数的合理构建及更优的参数求解方式.

参 考 文 献

[1] Androutsopoulos I, Paliouras G, Karkaletsis V. Learning to filter spam e-mail: A comparison of a Naive Bayesian and a memory-based approach [J]. Computer Science, 2012, 97(2): 1-13

[2] Sahami M, Dumais S, Heckerman D, et al. A Bayesian approach to filtering junk e-mail [C] //Proc of Learning for Text Categorization. Menlo Park, CA: AAAI, 1998: 98-105

[3] Schapire R E, Singer Y. Boostexter: A boosting-based system for text categorization [J]. Machine Learning, 2000, 39(2/3): 135-168

[4] Yih W, Mccann R, Kolcz A. Improving spam filtering by detecting gray mail [C] //Proc of the 4th Conf on Email and Anti-Spam. Redmond: Microsoft, 2007 [2018-09-25]. https://www.researchgate.net/publication/221650806_Improve_Spam_Filtering_by_Detecting_Gray_Mail

[5] Wang Guoyin, Yao Yiyu, Yu Hong. A survey on rough set theory and applications [J]. Chinese Journal of Computers, 2009, 32(7): 1229-1246 (in Chinese)
(王国胤, 姚一豫, 于洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009, 32(7): 1229-1246)

[6] Pawlak Z, Polkowski L, Skowron A. Rough sets [J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356

[7] Yao Yiyu. Decision-theoretic rough set models [C] //Proc of Int Conf on Rough Sets and Knowledge Technology. Berlin: Springer, 2007: 14-16

[8] Yao Yiyu, Wong S K M. A decision theoretic framework for approximating concepts [J]. International Journal of Man-Machine Studies, 1992, 37(6): 793-809

[9] Deng Xiaofei, Yao Yiyu. Decision-theoretic three-way approximations of fuzzy sets [J]. Information Sciences, 2014, 279: 702-715

[10] Zhang Zhifei, Miao Duoqian, Nie Jianyun, et al. Sentiment uncertainty measure and classification of negative sentences [J]. Journal of Computer Research and Development, 2015, 52(8): 1806-1816 (in Chinese)
(张志飞, 苗夺谦, 聂建云, 等. 否定句的情感不确定性度量及分类[J]. 计算机研究与发展, 2015, 52(8): 1806-1816)

[11] Yao Yiyu. Three-way decision: An interpretation of rules in rough set theory [C] //Proc of Int Conf on Rough Sets and Knowledge Technology. Berlin: Springer, 2009: 642-649

[12] Yao Yiyu. Three-way decisions with probabilistic rough sets [J]. Information Sciences, 2010, 180: 341-353

[13] Liang Decui, Liu Dun. A novel risk decision making based on decision-theoretic rough sets under hesitant fuzzy information [J]. IEEE Transactions on Fuzzy Systems, 2015, 23(2): 237-247

[14] Zhang Qinghua, Lü Gongxun, Chen Yuhong, et al. A dynamic three-way decision model based on the updating of attribute values [J]. Knowledge-Based Systems, 2018, 142: 71-84

[15] Li Wen, Miao Duoqian, Wang Weili, et al. Hierarchical rough decision theoretic framework for text classification [C] //Proc of the 9th IEEE Int Conf on Cognitive Informatics. Piscataway, NJ: IEEE, 2010: 484-489

[16] Liu Dun, Yao Yiyu, Li Tianrui. Three-way investment decisions with decision-theoretic rough sets [J]. International Journal of Computational Intelligence Systems, 2011, 4(1): 66-74

[17] Liang Decui, Liu Dun. Deriving three-way decisions from intuitionistic fuzzy decision-theoretic rough sets [J]. Information Sciences, 2015, 300: 28-48

[18] Li Huaxiong, Zhou Xianzhong. Risk Decision making based on decision-theoretic rough set: A three-way view decision model [J]. International Journal of Computational Intelligence Systems, 2011, 4(1): 1-11

[19] Liu Dun, Liang Decui, Wang Changchun. A novel three-way decision model based on incomplete information system [J]. Knowledge-Based Systems, 2016, 91: 32-45

[20] Liang Decui, Pedrycz W, Liu Dun. Determining three-way decisions with decision-theoretic rough sets using a relative value approach [J]. IEEE Transactions on Systems Man & Cybernetics Systems, 2017, 47(8): 1785-1799

[21] Zhou Bing, Yao Yiyu, Luo Jigang. Cost-sensitive three-way email spam filtering [J]. Journal of Intelligent Information Systems, 2014, 42(1): 19-45

[22] Zhang Qinghua, Xie Qin, Wang Guoyin. A novel three-way decision model with decision-theoretic rough sets using utility theory [J]. Knowledge-Based Systems, 2018, 159: 321-335

[23] Li Fan, Xu Zhangyan. Similarity measure between vague sets [J]. Journal of Software, 2001, 12(6): 922-927 (in Chinese)
(李凡, 徐章艳. Vague 集之间的相似度量[J]. 软件学报, 2001, 12(6): 922-927)

[24] Shi Shengfei, Zhang Wei, Li Jianzhong. A complex event detection algorithm based on correlation analysis [J]. Journal of Computer Research and Development, 2014, 51(8): 1871-1879 (in Chinese)
(石胜飞, 张伟, 李建中. 一种基于模式匹配与相关性分析的事件检测算法[J]. 计算机研究与发展, 2014, 51(8): 1871-1879)

[25] Gao Yan, Choudhary A, Hua Gang. A nonnegative sparsity induced similarity measure with application to cluster analysis of spam images [C] //Proc of 2010 IEEE Int Conf on Acoustics Speech & Signal Processing. Piscataway, NJ: IEEE, 2010: 14-19

[26] Liu Wuying, Wang Ting. Structured ensemble learning for email spam filtering [J]. Journal of Computer Research and Development, 2012, 49(3): 628-635 (in Chinese)
(刘伍颖, 王挺. 结构化集成学习垃圾邮件过滤[J]. 计算机研究与发展, 2012, 49(3): 628-635)

[27] Pawlak Z, Skowron A. Rough membership functions: A tool for reasoning with uncertainty [J]. Algebraic Methods in Logic and Computer Science, 1993, 28: 135-150

[28] Good I. The estimation of probabilities: An essay on modern Bayesian methods [J]. Biometrics, 1966, 23(1): 158-161

[29] Mitchell T. Machine Learning [M]. New York: McGraw Hill, 1997

[30] Pantel P. SpamCop-a spam classification & organization program [C] //Proc of Learning for Text Categorization. Menlo Park, CA: AAAI, 1998: 95-98

[31] Basic B D. Distance Measures [M]. Berlin: Springer, 2011

[32] Blei D, Ng A, Jordan M. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2012, 3(4/5): 993-1022

[33] Zhang Qinghua, Wang Guoyin, Xiao Yu. Approximate sets of rough sets [J]. Journal of Software, 2012, 23(7): 1745-1759 (in Chinese)
(张清华, 王国胤, 肖雨. 粗糙集的近似集[J]. 软件学报, 2012, 23(7): 1745-1759)

[34] Zhou Zhihua. Machine Learning [M]. Beijing: Tsinghua University Press, 2016 (in Chinese)
(周志华. 机器学习[M]. 北京: 清华大学出版社, 2016)

[35] Androutsopoulos I, Paliouras G. Learning to filter unsolicited commercial e-mail [J]. National Center for Scientific Research, 2004, 10(1): 4324-4330

[36] Fayyad U. Multi-interval discretization of continuous-val-ued attributes for classification learning [C] //Proc of Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 1993: 1022-1027

[37] Bridle J. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1989: 211-217

Xie Qin, born in 1993. Master candidate. Her main research interests include three-way decisions, rough sets, machine learning and uncertain information processing.

Zhang Qinghua, born in 1974. PhD, professor, PhD supervisor. His main research interests include rough sets, fuzzy sets, granular computing and uncertain information processing.

Wang Guoyin, born in 1970. PhD, professor, PhD supervisor. Senior member of the IEEE. His main research interests include rough set, granular computing, knowledge technology, data mining, neural network, and cognitive computing.