计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20131552
52(5):1005-1013,2015

基于集成学习的中文文本欺骗检测研究

张 虎 谭红叶 钱宇华 李 茹 陈 千
(山西大学计算机与信息技术学院 太原 030006)
(zhanghu@sxu.edu.cn)

Chinese Text Deception Detection Based on Ensemble Learning

Zhang Hu, Tan Hongye, Qian Yuhua, Li Ru, and Chen Qian
(*School of Computer & Information Technology, Shanxi University, Taiyuan* 030006)

Abstract Deception detection is important in the field of information security. Existing researches show that one third of the interpersonal communication involves the potential deceptions, and there are large amounts of deceptive messages in the more and more Web information. If the deception is potentially dangerous to people's life, the survival of enterprise and the stability of the country, then the negligence of deception may lead to incalculable loss. In the massive amounts of information the scale of the non-deceptive texts is much larger than the scale of the deceptive texts, so people remain unsuccessful and inefficient in detecting those deceptive messages by the existing methods, and it is desirable to create an automated method which could help people flag the possible deceptive messages. In this paper, we built a deception detection model based on ensemble learning to solve the imbalance of the existing data sets. Firstly a novel bisecting k-means method is proposed to cut the training sample set, and the separate classifiers are trained by using each pair of positive and negative samples, and then each test sample category value is calculated by the classifiers, and finally a novel min-max modular approach is used to integrate each category result. Experimental results verify the effectiveness of this method.

Key words deception; deception detection; ensemble learning; cutting samples; min-max modular support vector machine (M3-SVM)

摘 要 欺骗信息检测是信息安全领域中的重要研究内容.现有的研究表明,三分之一的人际交往中会涉及到潜在的欺骗,大量的欺骗信息充斥在各种各样的通信媒介中,在海量的网络信息中欺骗性数据的规模通常远小于非欺骗性数据的规模,已有方法还不能很好地适应于准确高效地欺骗检测,迫切期望提出一种能高效地检测欺骗信息的方法.针对具有非平衡性的海量网络信息,提出了一种基于集成学习的欺骗行为检测方法.通过改进的二分 k-means 划分方法对训练样本集进行分解,分别在每对正负样本集上学习各自独立的分类器,然后利用每个独立分类器分别计算待测样本的类别输出值,并采用结合个体分类器分类正确率的最小最大模块化方法集成每个判别结果.实验结果验证了该方法的有效性.

关键词 欺骗;欺骗检测;集成学习;样本划分;最小最大模块化支持向量机

中图法分类号 TP391

收稿日期:2013-11-29;**修回日期**:2014-06-23

基金项目:国家自然科学基金项目(61005053,61100138,61373082,61322211);国家"八六三"高技术研究发展计划基金项目(2015AA015407);新世纪优秀人才支持计划基金项目(20121401110013);山西省回国留学人员科研资助项目(2013-022);山西省高等学校科技创新项目(2015104);中国民航大学信息安全评测中心开放课题基金项目(CAAC-ISECCA-201402)

欺骗已在社会科学的很多学科中被广泛研究,它是指信息发送者故意传递错误信息并导致信息接收者得出错误的结论[1].迄今为止,国外对欺骗检测的研究主要集中在3个方面:欺骗理论研究、欺骗检测实验研究和欺骗检测数据集研究.欺骗理论研究开展相对较早,已成为欺骗检测研究的理论基础,并用于构建实验假设的理论主要有:媒介丰富性理论(media richness theory, MRT)[2]、社会存在理论(social presence theory, SPT)[3]、媒介扩展理论(channel expansion theory, CET)[4]和人与人之间的欺骗理论(interpersonal deception theory, IDT)[5].欺骗检测实验研究从不同角度分析了欺骗检测中的特定表现、特殊问题、具体假设、相关概念和所用的方法等.George等人[6-9]的4项研究调查了不同通信方式对欺骗发生和检测的影响;Qin等人[10-12]的3项研究着重分析了有效的欺骗线索;Zhou等人[13-14]的2项研究着重分析了有效的欺骗检测模型.近年来,Blair等人[15]将情景、状态、位置等背景知识加入到欺骗检测中;Evans等人[16]将语言线索提升到心理认知层次.针对中文文本的欺骗检测研究始于2007年,目前还处于探索阶段,Zhang等人[17-18]的2项研究着重探索了中文文本欺骗检测语料库构建、欺骗性线索获取和欺骗检测模型构建.

从欺骗检测数据集来看,目前大多数研究主要集中在面对面交流和语音通话等丰富性媒介通道中[19],而对通过计算机网络媒介进行人与人交流的欺骗检测研究还相对较少.以计算机网络为介质的通信可以采用基于音频、视频、文本或三者相结合等形式进行信息传输,当前网络上绝大多数的信息是通过计算机以文本的形式进行传递,因此,面向文本的欺骗检测研究具有重要意义.目前,针对中文文本数据集探索欺骗检测方法主要遇到了3个挑战:1)目前欺骗性数据集仅靠人工构建,不易采集和判别[17-18];2)欺骗性数据集的规模远小于非欺骗性数据集的规模;3)信息增长的速度远远超过了现有方法的处理能力.如何应对这3个挑战成为当前欺骗检测研究迫切需要解决的问题,本文着重研究后2个挑战的处理策略,试图发现具有非平衡性的海量网络信息中的欺骗信息,期望提出一种能高效地检测欺骗信息的方法.

1 欺骗检测模型

处理非平衡数据集时,传统分类方法通常关注样本个数较多的类别样本尽量分类准确时,倾向于忽视样本个数较少的样本类别,欺骗检测关注欺骗性数据的检测结果,即少数样本类样本的标注结果,因此,传统分类方法在非平衡数据集上的欺骗检测无法达到预期的效果.目前,解决不平衡分类问题的策略大致可分为2类:1)从训练集入手,通过改变训练集样本分布,降低不平衡程度;2)从学习算法入手,根据算法在解决不平衡问题时的缺陷,适当地修改算法使之适应不平衡分类问题.

信息过载对信息存储、信息处理等方面提出挑战,传统的处理方法无法高效地解决数据的大规模、高维度等问题.一种典型的解决办法是采用集成学习的思路:1)对任务或数据按某种方式分割;2)同时处理若干个可并行的任务或数据;3)按一定策略对其结果集成.Bagging技术和boosting技术是目前效果较好的2种集成学习方法,他们都通过将 T 次学习得到的分类法 $C_1, C_2, \cdots, C_T$ 组合起来,从而形成一个改进的分类法 C_*,但实验表明他们在非平衡数据集上表现一般.Lu等人[20]在1999年提出一种基于分布式系统的最小最大模块化网络(min-max modular, M3)的集成学习方法,该集成学习方法在训练阶段把所要解决的大规模的二类问题按照一定规则分解为一系列更小的、相互独立的二类子问题,最后通过一定的集成规则将各个二类子问题的结果集成,得到原问题的结果.2004年,Lu等人[21]将M3网络和支持向量机(support vector machine, SVM)相结合,将SVM作为基本分类器,提出了最小最大模块化支持向量机(M3-SVM),该方法针对大规模数据集中的二类问题取得了较好的效果.近年来,最小最大模块化网络方法采用不同的基分类器已经应用到许多模式分类领域中,如语音标注、工业图像的故障判断、文本分类、人脸属性分类和性别分类等[22-25].具体来说,集成学习是使用多个学习器来解决同一问题,它通过调用简单的分类算法以获得多个不同的基学习器,然后采用某种方式将这些基学习器组合成一个集成学习器[26].该方法的关键是基分类器的生成和基分类器结果的组合,它能显著提高一个学习系统的泛化能力,并能提高学习器的分类精度[27].

基于以上分析,本文提出了样本集划分和个体分类器集成相结合的中文文本欺骗检测模型,主要包括3项研究内容:1)欺骗线索选择;2)样本集划分与个体分类器训练;3)个体分类器集成.本文在Zhang等人[17]提出的欺骗线索的基础上着重探索了后2个问题的解决办法.针对样本集划分,提出了

一种改进的二分 k-means 的划分方法,基于此对大规模数据集进行分解,然后分别利用每对正负样本训练一个 SVM 分类器;对于个体分类器集成,利用每个独立 SVM 分类器分别计算待测样本的类别输出值,然后结合个体分类器的分类正确率,通过最小最大模块化方法集成各个 SVM 分类器的输出结果,基于此确定待测样本的类别. 欺骗检测的流程如图 1 所示:

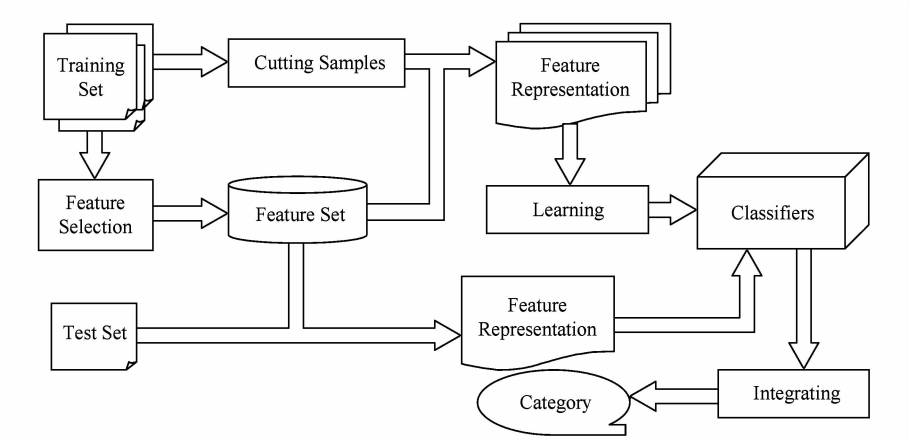


Fig. 1　Deception detection model.

图 1　欺骗检测模型

2　任务分解与集成策略

2.1　基于聚类的样本集划分

如何找到一种有效且复杂度较低的训练集划分方法,从而得到相对平衡的划分子集对集成学习模型非常重要. 传统常用的划分方法是随机划分,该方法容易实施,但忽视了样本集内部不同样本的共性和差异,尤其对于样本集中的小概率事件,更难以获取它们的潜在特征. 针对此方法的不足,本文基于 k-means 方法提出了一种改进的二分 k-means 划分方法.

1) k-means 聚类算法

典型的 k-means 算法中,初始点簇中心的选择是随机的,算法在选取随机的初始聚类中心时,可能会得到一个次最优聚类,难以获得较高的聚类质量. 针对 k-means 类型的聚类算法所遇到的问题,国内外许多学者对其进行了研究. Yu 就 k-means 的泛化[28]、模糊 k-means 的模糊因子选择[29]等问题展开了深入地研究并给出许多重要的结论. Huang 等人针对特征非等效问题,提出了自动加权的 k-means 聚类算法[30]. Xiong 等人研究了数据的异常分布对 k-means 聚类算法的影响,提出了"均匀效应"现象,并给出了理论分析[31].

2) 改进的二分 k-means 算法

二分 k-means 算法是在初始状态将整个训练集划分成 2 个簇,而后选择 1 个簇再次进行二分聚类,依次进行下去. 二分 k-means 在每次二分聚类时,都选择用 k-means 算法进行聚类,每次聚类就相当于对样本集进行一次划分. 二分 k-means 具有不受初始质心的影响,但是由于每次二分中的 k-means 算法依然是随机选取的质心,所以得到的解仍然是局部的. 为了弥补这些不足,本文从 2 个方面改进了二分 k-means 算法:

① 基于最大最小距离算法思想,通过计算簇内各样本点间的最大距离确定 2 个最大距离点,将其分别作为 2 类的初始聚类中心.

② 融合类内和类间距离,构造新的准则函数为

$$J=\frac{J_{\mathrm{W}}+\lambda}{J_{\mathrm{B}}}, \tag{1}$$

其中,J_{W} 为误差平方和,表示类内距离准则;J_{B} 为类间距离准则;λ 为平滑因子,令 $\lambda=0.05$.

$$J_{\mathrm{W}}=\sum_{j=1}^{k}\sum_{\boldsymbol{\chi}^{(i)}\in\omega_j}\left\|\boldsymbol{\chi}^{(i)}-\boldsymbol{\mu}_j\right\|^2, \tag{2}$$

$$J_{\mathrm{B}}=\sum_{j=1}^{k}(\boldsymbol{\mu}_j-\boldsymbol{\mu})'(\boldsymbol{\mu}_j-\boldsymbol{\mu}), \tag{3}$$

其中,ω_j 表示第 j 类,$\boldsymbol{\chi}^{(i)}$ 为 ω_j 类中第 i 个样本的矢量,$\boldsymbol{\mu}_j$ 为 ω_j 类的平均矢量,$\boldsymbol{\mu}$ 为总的平均矢量 $\boldsymbol{\mu}=\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\chi}^{(i)}$.

上述改进的二分 k-means 算法的实施步骤如下:

① 输入划分样本子集数目 k;

② 基于最大距离选取 2 个样本作为初始簇中心,用式(1)所列的改进准则函数对样本集进行二分 k-means 聚类,得到 2 个簇;

③ 选出一个样本个数最多的簇重新进行步骤②,再次进行二分聚类;

④ 反复迭代以上步骤②③,直到聚类中簇的个数等于 k.

理论上,好的初始类中心的选择能够极大地减少算法所需时间,减少聚类结果的误差总和;融合类内、类间距离的准则函数能够同时保证类内距离较小和类间距离较大,提高聚类质量.

2.2 结合分类器分类正确率的最小最大模块化模型

欺骗检测本质上是将待测样本标注为正类或负类,是一种典型的二类问题的任务分解,其任务分解和个体分类器集成的形式化表示如下:

1) 二类问题的任务分解[20-21]

一个分类问题可以分解成一组子问题,由于每个二类问题都含有较大数量的样本,所以需要进一步划分成一系列小的相对平衡的二类子问题. 具体过程:

① 给定一个二类分类问题 $S=X^+ \cup X^-$,其中 $X^+=\{(X_i,+1)\}_{i=1}^{N^+}$ 表示正类样本集,$X^-=\{(X_i,-1)\}_{i=1}^{N^-}$ 表示负类样本集. 其中,X_i 表示第 i 个训练样本;N^+ 和 N^- 分别表示正类和负类样本的数目,则训练样本总数为 $N=N^++N^-$.

② 训练阶段. 根据事先确定的分解常数 K^+ 和 K^-,按照某种训练集分解方法,将原训练集 X^+ 和 X^- 分别分解为 K^+ 和 K^- 个包含样本数量大致相等且互不相交的子集:

$$X_i^+ = \{(X_m,+1)\}_{m=1}^{N_i^+},$$
$$X^+=\bigcup_{i=1}^{K^+} X_i^+,\ \bigcap_{i=1}^{K^+} X_i^+ = \varnothing, \tag{4}$$
$$i=1,2,\cdots,K^+;$$
$$X_j^- = \{(X_n,-1)\}_{n=1}^{N_j^-},$$
$$X^-=\bigcup_{j=1}^{K^-} X_j^-,\ \bigcap_{j=1}^{K^-} X_j^- = \varnothing, \tag{5}$$
$$j=1,2,\cdots,K^-,$$

其中,N_i^+ 和 N_j^- 分别满足 $\sum_{i=1}^{K^+} N_i^+ = N^+$ 和 $\sum_{j=1}^{K^-} N_j^- = N^-$,$\varnothing$ 表示空集. 于是原二类分类问题 S 被分解成 $K^+\times K^-$ 个规模较小的二类分类子问题,即

$$S_{i,j} = X_i^+ \cup X_j^-,$$
$$i=1,2,\cdots,K^+,$$
$$j=1,2,\cdots,K^-. \tag{6}$$

由于 $K^+\times K^-$ 这个子问题在处理时不需要相互通信,可将得到的这 $K^+\times K^-$ 个子问题 $S_{i,j}$($i=1,2,\cdots,K^+$;$j=1,2,\cdots,K^-$)使用通常的分类器并行或串行训练,得到 $K^+\times K^-$ 个子分类器,表示为

$$M_{i,j}(\chi)$$
$$i=1,2,\cdots,K^+, \tag{7}$$
$$j=1,2,\cdots,K^-.$$

2) 个体分类器集成

集成学习的另一个任务是如何把这些个体分类器组合起来,M3 的组合策略依赖于 2 条集成规则,即 min 规则和 max 规则[19-20]. 本文基于这 2 条规则,将个体分类器的分类正确率引入到分类器集成过程,形成本文所用的集成策略.

① min 规则. 对拥有相同正类训练样本集和不同负类训练样本集的分类结果取最小值;

② max 规则. 对拥有相同负类训练样本集和不同正类训练样本集的分类结果取最大值.

$K^+\times K^-$ 个二类子问题训练得到的 $K^+\times K^-$ 个个体分类器可以通过 K^+ 个 min 规则和一个 max 规则按下列方式集成在一起:

$$M_i(\chi) = \min_{1\leqslant j\leqslant K^-} M_{i,j}(\chi)C_{i,j}(\chi),$$
$$i=1,2,\cdots,K^+, \tag{8}$$

其中,$M_{i,j}(\chi)$ 代表子问题训练后得到的个体分类器的输出函数,本文 $M_{i,j}(\chi)$ 为 SVM 分类器输出的 $f(\chi)$,未采用 $\mathrm{sgn}(f(\chi))$;$C_{i,j}(\chi)$ 代表个体分类器的权值,表示为

$$C_{i,j}(\chi) = \begin{cases} 1/P_{i,j}(\chi), & M_{i,j}(\chi)\geqslant 0, \\ P_{i,j}(\chi), & M_{i,j}(\chi)<0, \end{cases} \tag{9}$$

其中,$P_{i,j}(\chi)$ 表示个体分类器的分类正确率,$P_{i,j}(\chi)=\sqrt{P^+\times P^-}$,$P^+$ 和 P^- 分别表示该分类器对正类样本和负类样本的分类正确率.

$$M(\chi) = \max_{1\leqslant i\leqslant K^+} M_i(\chi), \tag{10}$$

其中,$M_i(\chi)$ 代表由具有相同正类训练样本的 K^- 个二类支持向量机通过 min 规则集成后得到的传递函数;而 $M(\chi)$ 就是利用 max 规则把 K^+ 个 $M_i(\chi)$ 进一步集成得到,也就是问题最终的解,$M(\chi)$ 可进一步表示为

$$M(\chi) = \max_{1\leqslant i\leqslant K^+}\ \min_{1\leqslant j\leqslant K^-} M_{i,j}(\chi)C_{i,j}(\chi), \tag{11}$$

通过计算 $M(\chi)$ 可以判断待测样本的类别.

3 实验结果与分析

实验分别采用随机划分(radom method, RM)、传统 k-means 划分(KM)和改进的二分 k-means 划

分(novel bisecting k-means, NBKM)对样本集进行划分,然后采用结合个体分类器分类正确率的 M3-SVM 模型(classification-precision and M3-SVM, CM3-SVM)在同样的欺骗数据集上进行了同类实验.

3.1　评价指标

传统的以准确率为评测指标的分类器倾向于降低稀有类的分类效果,不重视稀有类对分类性能的评价.因此以准确率作为分类器评测指标不再适合非平衡数据集情况,这就需要采用新的评价指标来描述非平衡数据集的分类性能.

对于非平衡数据集的每一个测试样本,二类分类器有 4 种可能的标注结果,如表 1 所示:

Table 1　Decision Results of Sample

表 1　样本标注结果

Sample Classification	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

表 1 中,TP 是本属于正类且被判别为正类的样本个数,FP 是本属于负类且被判别为正类的样本个数,FN 是本属于正类且被判别为负类的样本个数,TN 是本属于负类且被判别为负类的样本个数.

基于以上 4 种判决结果,本文采用 F-measure (F_m)和相对灵敏度(relative sensitivity, RS)作为模型评价测度. F-measure 方法是召回率(recall, R)和正确率(precison, P)的组合,主要要求在召回率和正确率平衡的前提下尽可能将其最大化,定义为

$$F_m = \sqrt{R \times P}, \tag{12}$$

其中,计算正类的和负类的 F_m 时,R 和 P 的值的计算公式分别为

$$\begin{cases} R^+ = \dfrac{TP}{(TP+FN)}, \\ P^+ = \dfrac{TP}{(TP+FP)}, \\ R^- = \dfrac{TN}{(TN+FP)}, \\ P^- = \dfrac{TN}{(TN+FN)}. \end{cases} \tag{13}$$

RS 可以用来判断分类器是否对正类和负类有相同的分类能力,如果 RS 的值偏离 1 很多,那么说明分类器是有偏的,定义为

$$RS = \frac{M_{\text{Sensitivity}}}{M_{\text{Specificity}}}, \tag{14}$$

其中,$M_{\text{Sensitivity}} = TP/(TP+FN)$, $M_{\text{Specificity}} = TN/(TN+FP)$.

3.2　实验结果和分析

1) 实验 1. 确定本文的实验基线.

Zhang 等人[17]的实验中选择相对平衡的数据集,随机选取 700 篇欺骗性文本和 1 000 篇非欺骗性文本作为训练集,同时选择 793 篇欺骗性文本和 1 191 篇非欺骗性文本作为测试集,实验结果如表 2 所示:

Table 2　Experimental Results for the Balanced Data Sets

表 2　平衡数据集上不同分类器的实验结果

Classification Model	Non-deceptive			Deceptive			RS
	P^+	R^+	F_m^+	P^-	R^-	F_m^-	
Bayes	0.851	0.761	0.80	0.690	0.799	0.74	0.95
KNN	0.828	0.797	0.81	0.711	0.752	0.73	1.06
SVM	0.885	0.908	0.90	0.857	0.822	0.84	1.1

实验结果表明,利用 Sigmoid 核函数的 SVM 模型在表 2 的三组欺骗检测实验中得到最好的实验结果,但该实验的一个重要问题是所选择的数据集中 2 类文本的比例不太符合真实分布,因此,为了使实验数据比例更符合真实的分布,本文改变了数据集中欺骗性文本和非欺骗性文本的比例,基于此按照下边模式进行了实验.实验选取 900 篇欺骗性文本和 9 000 篇非欺骗性文本作为数据集;同时,为了得到较为可靠的实验结果,实验采用了 3 折交叉验证法,将数据集等分成 3 组,保证每一组中正负类的样本数比例均与整体数据集一致(各组包括 300 篇欺骗性文本和 3 000 篇非欺骗性文本),将其中的 2 组作为训练数据,1 组作为测试数据进行实验,并将这个过程重复 3 次(Test1, Test2, Test3),最后取这 3 次实验结果的平均值作为基线.实验结果如表 3 所示:

Table 3　Experimental Results for the Imbalanced Data Sets without Cutting Samples

表 3　无样本划分的非平衡数据集的实验结果

Experiments	Non-deceptive			Deceptive			RS
	P^+	R^+	F_m^+	P^-	R^-	F_m^-	
Test1	0.939	0.924	0.93	0.342	0.397	0.37	2.33
Test2	0.943	0.919	0.93	0.355	0.443	0.40	2.07
Test3	0.937	0.913	0.93	0.310	0.390	0.35	2.34
Average	0.940	0.919	0.93	0.335	0.410	0.37	2.25

表 3 中的评价指标 F_m^- 较低,这表明 Zhang 等人[17]提出的方法对非平衡数据集中样本较少的负

类样本标注结果较差,同时,RS 指标也表明所选择的分类器是有偏的.

2) 实验 2. 基于数据集划分的欺骗检测实验.

基于表 3 的 3 组实验数据集,利用随机划分、k-means 划分和改进的二分 k-means 划分方法,并采用结合个体分类器分类正确率的 CM3-SVM 模型分别进行了 3 次实验,对于所有划分都是把正类样本划分为 3～10 份,即 k 取值为 3～10. 实验结果如表 4 至表 6 所示,其中黑体数据行为 TOP 值.

Table 4 Experimental Results for Random Methods

表 4 随机划分方法的实验结果

k	R^+	P^+	R^-	P^-	F_m^+	F_m^-	RS
3	0.917	0.949	0.510	0.380	0.933	0.440	1.797
4	0.914	0.952	0.537	0.385	0.933	0.455	1.704
5	0.913	0.955	0.573	0.397	0.934	0.477	1.592
6	0.903	0.956	0.583	0.375	0.929	0.468	1.547
7	0.889	0.961	0.640	0.366	0.925	0.484	1.390
8	0.880	0.964	0.673	0.360	0.921	0.492	1.307
9	**0.879**	**0.965**	**0.683**	**0.362**	**0.921**	**0.497**	**1.287**
10	0.871	0.963	0.663	0.340	0.916	0.475	1.313

Table5 Experimental Results for k-means Methods

表 5 k-means 划分方法的实验结果

k	R^+	P^+	R^-	P^-	F_m^+	F_m^-	RS
3	0.917	0.955	0.567	0.406	0.936	0.479	1.618
4	0.913	0.958	0.597	0.406	0.935	0.492	1.530
5	0.906	0.958	0.607	0.392	0.932	0.488	1.493
6	0.903	0.964	0.663	0.405	0.933	0.519	1.361
7	0.897	0.968	0.707	0.408	0.932	0.537	1.270
8	**0.895**	**0.975**	**0.767**	**0.422**	**0.934**	**0.569**	**1.167**
9	0.881	0.970	0.730	0.381	0.925	0.527	1.207
10	0.876	0.977	0.797	0.392	0.925	0.559	1.100

Table 6 Experimental Results for Novel Bisecting k-means Methods

表 6 改进的二分 k-means 方法的实验结果

k	R^+	P^+	R^-	P^-	F_m^+	F_m^-	RS
3	0.915	0.955	0.573	0.403	0.935	0.481	1.596
4	0.909	0.960	0.617	0.403	0.934	0.499	1.474
5	0.906	0.963	0.657	0.411	0.934	0.520	1.380
6	0.903	0.969	0.707	0.421	0.935	0.546	1.278
7	0.893	0.975	0.770	0.418	0.933	0.567	1.159
8	**0.886**	**0.984**	**0.860**	**0.430**	**0.934**	**0.608**	**1.030**
9	0.878	0.981	0.830	0.406	0.928	0.580	1.058
10	0.873	0.983	0.847	0.401	0.926	0.582	1.031

实验表明,当正负样本的比例为 10∶1时,结合 CM3-SVM 模型和 KM 的方法(CM3-SVM-KM)、结合 CM3-SVM 模型和 NBKM 的方法(CM3-SVM-NKKM)对正类划分为 8 份($k=8$)时会得到最好的结果;而结合 CM3-SVM 模型和 RD 的方法(CM3-SVM-RD)是将正类划分为 9 份($k=9$)时得到最好的结果.

3) 实验结果分析.

将实验 2 中 3 种方法的结果与实验 1 的结果比较,评价指标 F_m^+,F_m^- 和 RS 的变化趋势如图 2 至图 4 所示:

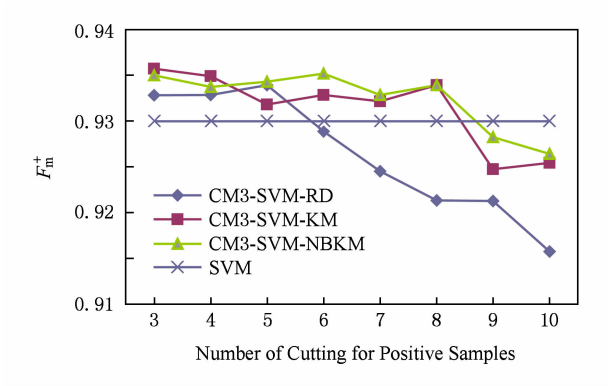


Fig. 2 The trend of F_m^+ for four methods.

图 2 不同方法 F_m^+ 的变化趋势

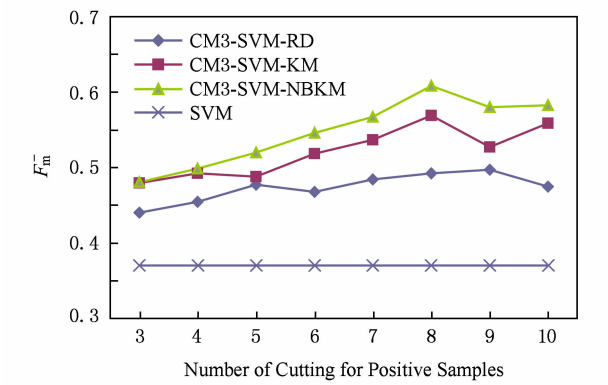


Fig. 3 The trend of F_m^- for four methods.

图 3 不同方法 F_m^- 的变化趋势

显然,3 种划分方法在正类的 F_m^+ 测度上都取得较为稳定的结果,但在负类的 F_m^- 测度和 RS 测度上则表现出较大差别. 随机划分虽然算法复杂度较低,但其在划分过程中破坏了数据集原本的结构属性,不能得到相对较好的聚类结果,因此对于样本较少的负类表现出较差的性能,而且分类器对正负类的偏差也相对较大. k-means 和改进的二分 k-means 复杂度大致相同,但后者容易聚类出相对均衡的簇,因此,后者的 F_m^- 比前者的略高,且后者的

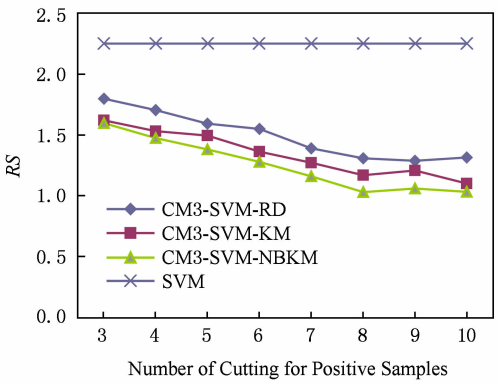


Fig. 4　The trend of *RS* for four methods.

图 4　不同方法 *RS* 的变化趋势

RS 值也更加趋近于 1. 同时,实验还有 3 点需要说明:

① 所有方法的 P^- 都比较小,这是因为正类样本数远大于负类样本数,比例很小的正类样本判定为负类样本也会对负类的正确率造成极大的影响. 如本实验中的 3 000 个正类样本中有 342 判定为负类,而 300 个负类样本中有 42 个判定为正类,这样计算 P^- 仅为 0.43,显然,实际的欺骗检查中我们允许少数的正类样本判定为负类,但要尽量使少的负类样本判定为正类,因为负类的错判比正类的错判代价要大很多.

② 本次实验未采用并行机制对多个个体分类器进行并行训练,所以对模型耗时没有准确的统计. 理论上,当训练集划分数越大,即可并行训练的个体分类器越多,总耗时(多个并行分类器执行时间之和)会越大,但在不考虑并行设备代价时,单位时间内可训练的数据量会随着 k 的增大而增长.

③ 本文分别尝试了将负类训练样本划分为 2 份和 3 份、正类样本划分为对应等规模份数的实验,实验结果与上述结果相差不大,得出的结论是:对于现有数据集,负类有共同的特征,并且当正类划分的类别太多时,部分类别之间的相关性会很强.

4　结论与展望

论文提出的改进二分 k-means 和 CM3-SVM 相结合的学习方法能同时兼顾欺骗检测数据集的不平衡性和大规模性,在样本集划分过程中对正负训练样本集选择不同的划分数,可以保证在组成 2 类子问题时他们的正负训练样本集规模相差不大,进而训练出较好的分类器. 采用的集成学习机制除了能提高分类精度外,还可以提高模型的训练效率. 同时,为了进一步加强该方法的泛化能力,改善自动欺骗检测的效果,后续仍需进行以下相关研究:

1) 本文提出的方法统一采用 SVM 子分类器和 M3 集成策略,所用的子分类器和集成策略都较单一. 因此为了进一步改进该模型,后续需研究不同子分类器的集成策略.

2) 实验所用的数据规模不够大,离海量数据级更是相差甚远,后续研究需探索人工构建欺骗性数据集的方法、基于半监督学习或无监督学习的样本集构建方法. 同时,为了进一步验证本文提出的方法,也可以将该方法应用到其他具有非平衡特性的海量数据分类问题中.

参　考　文　献

[1] Buller D, Burgoon J. Strategic Interpersonal Communication [M]. Mahwah, NJ: Lawrence Erlbaum Associates Publishers, 1994: 191-223

[2] Daft R, Lengel R. Organizational information requirements, media richness, and structural design [J]. Management Science, 1986, 32(5): 554-570

[3] Short J, Williams E, Christie B. The Social Psychology of Telecommunications [M]. New York: Wiley Publisher, 1976

[4] Carlson J, Zmud R. Channel expansion theory and the experiential nature of media richness perceptions [J]. Academy of Management Journal, 1999, 42 (2): 153-170

[5] Buller D, Burgoon J. Interpersonal deception theory [J]. Communication Theory, 1996, 6 (3): 203-242

[6] Blair J, Burgoon J, Strom R. Heuristics and modalities in determining truth versus deception [C] //Proc of the 38th Annual Hawaii Int Conf on System Sciences. Los Alamitos, CA: IEEE Computer Society, 2005: 19-25

[7] George J, Marett K, Tilley P. Deception detection under varying electronic media and warning conditions [C] //Proc of the 37th Annual Hawaii Int Conf on System Sciences. Los Alamitos, CA: IEEE Computer Society, 2004: 327-336

[8] George J, Marett K. Inhibiting detection and its detection [C] //Proc of the 37th Annual Hawaii Int Conf on System Sciences. Los Alamitos, CA: IEEE Computer Society, 2004: 337-346

[9] Zhou L, Sung Y. Cues to deception in online Chinese groups [C] //Proc of the 41st Annual Hawaii Int Conf on System Sciences. Los Alamitos, CA: IEEE Computer Society, 2008: 146-153

[10] Qin T, Burgoon J, Nunamaker J, et al. An exploratory study on promising cues in deception detection and application of decision tree [C] //Proc of the 37th Annual Hawaii Int Conf on System Sciences. Los Alamitos, CA: IEEE Computer Society, 2004: 357-366

[11] Zhou L, Zhang D. Can online behavior unveil deceivers? [C] //Proc of the 37th Annual Hawaii Int Conf on System Sciences. Los Alamitos, CA: IEEE Computer Society, 2004: 317-326

[12] Hancock J, Curry L, Goorha S, et al. Automated linguistic analysis of deceptive and truthful synchronous computer-mediated communication [C] //Proc of the 37th Annual Hawaii Int Conf on System Sciences. Los Alamitos, CA: IEEE Computer Society, 2004: 22-31

[13] Zhou L, Burgoon J, Twitchell D, et al. A comparison of classification methods for predicting deception in computer-mediated communication [J]. Journal of Management Information Systems, 2004, 20(4): 139-165

[14] Zhou L, Shi Y, Zhang D. A statistical language modeling approach to online deception detection [J]. IEEE Trans on Knowledge and Data Engineering, 2008, 20(8): 1077-1080

[15] Blair J, Levine T, Shaw A. Content in context improves deception detection accuracy [J]. Human Communication Research, 2010, 36(3): 423-442

[16] Evans J, Michael S, Meissner C, et al. Validating a new assessment method for deception detection: Introducing a psychologically based credibility assessment tool [J]. Journal of Applied Research in Memory and Cognition, 2013, 2(1): 33-41

[17] Zhang H, Fan Z, Zheng J. An improving deception detection method in computer-mediated communication [J]. Journal of Networks, 2012, 7(11): 1811-1816

[18] Zheng Jiaheng, Zhang Hu. A study deception detection method for Chinese text [J]. Journal of Shanxi University, 2009, 32(4): 541-545 (in Chinese)
(郑家恒, 张虎. 面向中文文本的欺骗行为检测方法研究[J]. 山西大学学报, 2009, 32(4): 541-545)

[19] Qin T. Identification of reliable cues for an automatic deception detection system [D]. Tucson, AZ: University of Arizona, 2007

[20] Lu B, Ito M. Task decomposition and module combination based on class relations: A modular neural network for pattern classification [J]. IEEE Trans on Neural Networks, 1999, 10(5): 1244-1256

[21] Lu B, Wang K, Utiyama M, et al. A part-versus-part method for massively parallel training of support vector machines [C] //Proc of IJCNN'04. Los Alamitos, CA: IEEE Computer Society, 2004: 735-740

[22] Ma Q, Lu B. Part of speech tagging with min-max modular neural networks [J]. Systems and Computers in Japan, 2002, 33(7): 30-39

[23] Huang B, Lu B. Fault diagnosis for industrial images using a min-max modular neural network [G]// LNCS 3316: Proc of the 11th Int Conf on Neural Information Processing. Berlin: Springer, 2004: 843-847

[24] Lian Huicheng. A study on min-max modular networks and facial attribution classification [D]. Shanghai: Shanghai Jiao Tong University, 2007 (in Chinese)
(连惠城. 最小最大模块化网络及人脸属性分类研究[D]. 上海: 上海交通大学, 2007)

[25] Zheng J, Lu B. A support vector machine classifier with automatic confidence and its application to gender classification [J]. Neurocomputing, 2011, 74(11): 1926-1935

[26] Zhang Chunxia, Zhang Jiangshe. A survey of selective ensemble learning algorithms [J]. Chinese Journal of Computers, 2011, 34(8): 1399-1410 (in Chinese)
(张春霞, 张讲社. 选择性集成算法综述[J]. 计算机学报, 2011, 34(8): 1399-1410)

[27] 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)

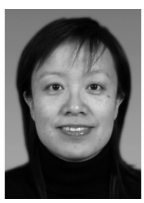
[28] Yu Jian. General c-means clustering model [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1197-211

[29] Yu Jian. On the fuzziness index of the FCM algorithms [J]. Chinese Journal of Computers, 2003, 26(8): 968-973 (in Chinese)
(于剑. 论模糊 C 均值算法的模糊指标[J]. 计算机学报, 2003, 26(8): 968-973)

[30] Huang Z, Ng M, Rong H, et al. Automated variable weighting in k-means type clustering [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668

[31] Xiong H, Wu J, Chen J. k-means clustering versus validation measures: A data-distribution perspective [J]. IEEE Trans on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(2): 318-331

Zhang Hu, born in 1979. PhD. Lecturer in Shanxi University. Member of China Computer Federation. His main research interests include natural language processing and social computing.



Tan Hongye, born in 1971. Associate professor in Shanxi University. Member of China Computer Federation. Her main research interests include natural language processing, complex networks and social computing(tanhongye@sxu.edu.cn).

Qian Yuhua, born in 1976. Professor and PhD supervisor in Shanxi University. Senior member of China Computer Federation. His main research interests include granudar computing and machine learning for big data(jinchengqyh@126.com).

Li Ru, born in 1963. Professor and PhD supervisor in Shanxi University. Senior member of China Computer Federation. Her main research interests include natural language processing and data mining (liru@sxu.edu.cn).

Chen Qian, born in 1983. Lecturer in Shanxi University. Member of China Computer Federation. His main research interests include text mining and topic modeling (chenqian@sxu.edu.cn).

2015 年《计算机研究与发展》网络与信息安全专辑(正刊)征文通知

——“网络安全与隐私保护研究进展”

近年来,随着计算机网络的不断发展,数据量的井喷式增长催生了大量新型网络应用,随之产生的网络安全问题也更加突出.网络安全除了网络密码安全、网络系统安全外,由于云计算、大数据等新型应用模式需求,产生了大量的安全与隐私保护问题,这些问题的研究不仅关乎国家安全、经济发展和社会稳定,也关乎网络安全学科本身的进步.为推动我国学者在这方面的研究、及时报道我国学者的最新研究成果,《计算机研究与发展》将于 2015 年 10 月出版网络与信息安全专辑——网络安全与隐私保护研究进展,欢迎相关领域的专家学者和科研人员踊跃投稿.

征文内容

本专辑包括(但不限于)下列网络安全与隐私保护的相关主题:

1) 网络密码安全(面向网络环境的密码算法、协议及应用密码技术解决网络安全问题);

2) 网络系统安全(网络架构安全、网络空间安全、网络设备安全);

3) 新型应用模式下的安全与隐私保护(云计算、大数据等新型应用模式下的安全与隐私保护).

投稿要求

1) 论文应属于作者的科研成果,数据真实可靠,具有重要的学术价值与推广应用价值,未在国内外公开发行的刊物或会议上发表或宣读过,不存在一稿多投问题.作者在投稿时,需向编辑部提交投稿声明.

2) 论文应包括题目、作者信息、摘要、关键词、正文和参考文献,论文一律用 word 排版,论文格式体例格式请参考《计算机研究与发展》近期文章.

3) 论文需附通信作者的联系地址、电话或手机及 E-mail 地址.

4) 论文请通过期刊网站(http://crad.ict.ac.cn)进行投稿,并在留言中注明“网络与信息安全 2015 专题”.(否则按自由来搞处理)

重要日期

征文截止日期: 2015 年 6 月 15 日　　录用通知日期: 2015 年 7 月 25 日

作者修改稿提交日期:2015 年 8 月 25 日　　出版日期:2015 年 10 月

特邀编委

徐秋亮　教授　山东大学　xql@sdu.edu.cn

张玉清　教授　中国科学院大学、国家计算机网络入侵防范中心　zhangyq@ucas.ac.cn

董晓蕾　教授　华东师范大学　dongxiaolei@sei.ecnu.edu.cn

联系方式

联 系 人:徐秋亮　xql@sdu.edu.cn　13969133581

编辑部　crad@ict.ac.cn　010-62620696,010-62600350

通信地址: 北京 2704 信箱《计算机研究与发展》编辑部

邮政编码: 100190