文章编号:2095-476X(2014)01-0094-04

一种具有分类细化功能的垃圾语言识别方法

李小娇, 马吉明, 张向梅

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450001)

摘要:为了筛选出散播垃圾语音的用户,建立了一种采用基于加权k-means和支持向量机的垃圾语言识别方法.该方法依据用户的历史通信活动建立通信行为网络模型,用加权的k-means算法对用户进行半监督聚类,然后从每个类中均匀选取部分用户数据作为训练集,采用支持向量机获得训练模型用以预测剩余用户数据.实验结果表明,该方法的用户分类更细化,并具备预测功能,有一定的机器学习能力,可用于大客户发现及关联客户发现和业务推荐等.

关键词:数据挖掘;k-means;支持向量机;垃圾语音

中图分类号:TP399 **文献标志码**:A DOI:10.3969/j.issn.2095-476X.2014.01.020

A SPIT recognition method with refined classification

LI Xiao-jiao, MA Ji-ming, ZHANG Xiang-mei

(*College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450001, *China*)

Abstract: In order to screen the spreading spam over Internet telephony (SPIT) user, a recognition method was built based on weighted k-means and support vector machine (SVM). This method built a communication network model according to historical communication activities of customers, and clustered semi-supervised by weighted k-means algorithm. Then it equally selected part of customers data from each classified cluster as the training set and finally processed the rest data by using SVM method. Experimental data showed that this method could make the classification more refined and had forecast function and certain ability of machine learning. It can be used for the discovery of important customers, relevant customers and service recommendation, etc.

Key words: data mining; k-means; support vector machine (SVM); spam over Internet telephony (SPIT)

0 引言

随着"三网融合"进程的加快,网络电话在实际中已得到广泛应用.网络电话将模拟的声音信号压缩与封包后,以数据封包的形式在互联网上进行语音信号的传输.与传统通信业务相比,网络电话的优势是广泛地运用互联网,服务更好而且价格相对便宜.

垃圾语音SPIT(spam over Internet telephony)是利用网络电话传输的消息,它是一种非预期的语音发送行为,大量的垃圾呼叫可能会导致网络超载和呼叫业务失常.此外,垃圾语音以音频作为内容承载手段,使得较为成熟的基于内容过滤的垃圾邮件检测方法难以实施.如何识别并屏蔽网络电话的垃圾语音,成为网络电话发展需要克服的问题.

2009年,何光宇等[1]利用用户的反馈信息建立评判模型,用加权的朴素贝叶斯分类算法获得评判结果来识别垃圾语音,准确率较高,但需要用户的

收稿日期:2012-11-29

作者简介:李小娇(1987—),女,河南省长葛市人,郑州轻工业学院硕士研究生,主要研究方向为数据挖掘.

参与,增加了维护难度. 2012 年,何光宇等[2]通过要求主叫在发送语音会话之前消耗自身大量系统资源破解谜题,实现对垃圾语音攻击的防范,这种方法在预防垃圾语音方面有优势,不足之处是占用的资源较多. 王菲等[3]提出基于用户信誉的分布式 VoIP 垃圾语音过滤模型,该信誉模型的主观性较强,也需要用户的参与. 张卫兵等[4]提出一种结合 Boyer Moore 模式匹配算法和布鲁姆过滤器的垃圾语音检测方法,用该方法进行精确匹配需要更多的时间,而且算法中分组元素的个数和数据分片大小的设定过于主观.

针对上述问题,本文以实际通话数据为基础,提取出用户的关键属性,用加权 k-means 的方法将属性相似的用户分至同一类,从每一类中挑选出部分用户作为支持向量机的训练集,经训练获得训练模型,用得到的训练模型预测其余用户的分类.

1　加权 k-means 算法

1.1　传统 k-means 算法

聚类分析是数据挖掘方法的一种,根据数据对象及其关系的信息将数据对象分组. 其目标是组内的对象相似而组间的对象不同. 组内的相似性越大,组间差别越大,聚类就越好[5]. 判断相似性的准则是对象间的距离,对象间的距离越小表示对象的相似度越大,越有可能被分为一组[6].

传统 k-means 算法是一种简单但重要的聚类分析技术:设有 n 个数据对象的数据集 $X=\{X_1,X_2,\cdots,X_n\}$,从中随机选择 k 个样本点作为初始中心点集合 $A=\{A_1,A_2,\cdots,A_k\}$,其余样本点被逐个分到与自身距离最近的中心点,被分到一个中心点的点集为一个簇;然后依据样本点到中心点欧氏距离误差平方和最小的准则重新分配中心点,直到聚类中心不发生变化[7].

1.2　加权 k-means 算法

传统的 k-means 算法不考虑对象中每个变量在聚类过程中体现作用的不同,而是统一看待,即每个属性的特征权重是等值的[8]. 要解决的问题中样本点的每个属性对聚类结果的贡献是不同的,因此需要在 k-means 算法计算样本点之间距离的公式中加入权重因子,将原来的距离公式

$$d(X_i,X_j)=\sqrt{\sum_{h=1}^{d}(X_{ih}-X_{jh})^2}$$

转换为

$$d'(X_i,X_j)=\sqrt{\sum_{h=1}^{d}W_h\times(X_{ih}-X_{jh})^2}$$

其中,$W=(W_1,W_2,\cdots,W_d)$,W 是各个属性权重值集合.

2　支持向量机算法

支持向量机 SVM(support vector machine)是一种基于统计学习理论的模式识别方法,在解决小样本、非线性及高维模式识别中表现出许多特有的优势,并能够推广应用到函数拟合等其他机器学习问题中. 支持向量机算法自提出以后,以其良好的分类性能,在社会生活中的各领域获得了广泛的应用[9].

SVM 本质上是一种分类算法,待分类的样本点可以是任意 n 维空间数据点,通过这些点找到 1 个 $n-1$ 维的超平面将样本点分开,这个超平面通常被称为分类器. 如果能找到使得属于不同类的数据点间隔最大的那个面,那么这个分类器就称为最大间隔分类器. 在分割超平面的两边建有 2 个互相平行的超平面,分割超平面使 2 个平行超平面的距离最大化,假定平行超平面间的距离或间距越大,分类器的总误差越小.

SVM 擅长应付样本数据线性不可分的情况,主要通过松弛变量和核函数技术来实现. 松弛变量是用来处理影响分类的离散点的一种方法,而常用的核函数有线性核函数、多项式核函数、径向基核函数、Sigmoid 核函数. 采用核函数方法向高维空间映射时并不增加计算的复杂度,且有效地克服了维数灾难问题[10]. 核函数不同对分类性能有影响,核函数相同但参数不同也影响分类性能[11].

在 Matlab 环境下进行 SVM 的相关运算,采用的工具包是 LIBSVM.

LIBSVM 的使用步骤是:

1)按照 LIBSVM 软件包所要求的格式预处理数据集;

2)对数据进行归一化处理;

3)一般选用径向基核函数;

4)采用交叉验证选择最佳参数 c 与 g;

5)采用最佳参数 c 与 g 对整个训练集进行训练,获取 SVM 模型;

6)利用获取的模型进行测试与预测.

3 SPIT 识别方法

3.1 提取通信用户属性

首先建立用户通信行为网络模型,用有向图表示. 每个用户为一个节点,用户 A 主叫用户 B,则 A 和 B 之间用由 A 指向 B 的单向箭头连接. 每个通信用户用 5 个属性描述,各属性的定义如下.

1)局部集聚系数 $C_{(i)}$:一个顶点 $V_{(i)}$ 的局部集聚系数 $C_{(i)}$ 等于所有相邻顶点之间所连边的数量除以相邻顶点之间可以连出的最大边数[12]. 如图 1 所示,A,B,D 这 3 个节点的局部集聚系数依次为 2/3,1/3,1/3. 假如 A,B,D 节点代表 3 个不同的人,则说明 A 的朋友之间相互联系更紧密. 因此,局部集聚系数反映了用户间联系是否紧密,垃圾用户的局部集聚系数总小于正常用户.

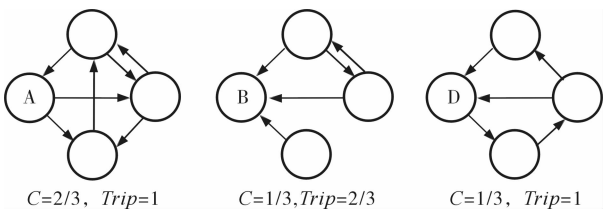

图 1 局部集聚系数和三角形节点率示例

2)三角形节点率 $Trip_{(i)}$:用户邻域节点中具有邻域间通信的节点的比例. 正常用户在生活中总有固定的朋友圈,朋友之间也一般会产生通话,而垃圾用户往往向不认识的人拨打电话,没有固定的交流圈,因此垃圾用户的三角形节点率也低于正常用户. 另外,在局部集聚系数相同的情况下,三角形节点率可以进一步判断用户群联系的紧密程度. 如图 1 所示,节点 B 和节点 D 具有相同的局部集聚系数,节点 B 的三角形节点率小于 D,说明节点 D 的朋友间相互联系更紧密.

3)重复呼叫率:与节点 i 有重复呼叫的邻域数和节点 i 的总邻域数之比,称为重复呼叫率. 垃圾用户往往不会重复呼叫某一个人,而是随机拨打电话,因此垃圾用户的重复呼叫率更低.

4)拨打/接听率:节点 i 拨打电话的次数与节点 i 被拨打的次数之比为拨打/接听率. 垃圾用户拨打出去的电话数量要远大于接听到的电话数量,因为很少有人会给垃圾用户打电话,所以垃圾用户的拨打/接听率的值往往更大.

5)平均通话时长:节点 i 作为主叫的所有呼叫的通话时长的算术平均数为平均通话时长. 实际生活中如果接听到来自垃圾用户的电话,通常会挂掉电话或拒接,因此垃圾用户作为主叫的通话时长总是比较短.

3.2 聚类

第 1 步:数据分析. 为了使聚类结果不受非重要属性的干扰,需要对所有的属性按照重要性分配权重,再进行聚类. 在对聚类的贡献方面,从大到小依次为属性 1(局部集聚系数),属性 2(三角形节点率),属性 3(重复呼叫率),属性 4(拨打/接听率),属性 5(平均通话时长),其权重分配为 0.35,0.25,0.2,0.1 和 0.1.

由于属性 4 数值分布在[10,51],属性 5 的数值分布在[38.13,156.79],其余属性均小于或等于 1,而 k-means 聚类使用样本点之间的欧式距离判断节点之间相似性,数值较大的属性会影响距离判断的准确性,因此需要进行归一化. 经实验,所有属性归一化到区间[0,1]最佳.

第 2 步:对待聚类数据集进行加入特征权重的 k-means 算法,筛选出可疑垃圾用户组. 为了深入观察通话用户的行为,不能简单地把聚类结果分为垃圾用户和非垃圾用户 2 类,需要有更精细的分类,经多次聚类设定聚类簇数为 10 较为合理. 从分得的 10 组数据中将属性 1,属性 2,属性 3,属性 5 的数值相对较小而属性 4 数值相对较大的分组设为可疑垃圾用户组.

第 3 步:对可疑垃圾用户组再进行加入特征权重的 k-means 算法,筛选出垃圾用户组.

3.3 建立训练集

采用 SVM 的目的是利用已分类数据获得训练模型,然后用训练模型预测新数据. 从 3.2 节中得到的 10 个分类中选取训练集. 为了使训练集更丰富,随机取每个类的 2/3 用户放入训练集. 经训练得到训练模型,然后用训练模型预测测试集分类的精确度.

采用径向基函数作为 SVM 的核函数,训练模型中参数 c 和 g 的选取,使用交叉验证的方法. 参数 c 和 g 在一定的范围内取值,选择能够达到最高验证分类准确率中参数 c 最小的那组作为最佳参数. 这样做是因为过高的 c 会导致过度拟合的情况发生. 在实际操作中,c 和 g 的取值一般为 $[2^{-10},2^{10}]$. 经过反复试验,为了缩短算法执行时间,又能保证选

到最优参数,设定算法在[2^{-4},2^{5}]的范围内选择最佳的 c 和 g.

4　实验结果与分析

实验的计算机配置为:机型为 HP Pro 3348 MT,CPU 为 Intel Core i5—3470,主频为 3.20 GHz,内存为 4 GB,内存频率 1 600 MHz.

实验数据来源于 3 000 名用户的通话数据,共有 2.024×10^5 条通话记录,所有用户属性的平均值如下:局部集聚系数为 0.622 8,三角形节点率为 0.971 3,重复呼叫率为 0.541 2,拨打/接听率为 1.466 2,平均通话时长为 97.340 7 s. 处理数据的软件为 Matlab R2009a,工具包为 LIBSVM-mat-2.89-3.

4.1　聚类结果与分析

对所有用户进行加权 k-means 计算,得到 10 个分类,每一类中用户数分别为 75,157,330,330,242,111,515,455,245,540. 每个类各个属性归一化后的平均值如图 2 所示.

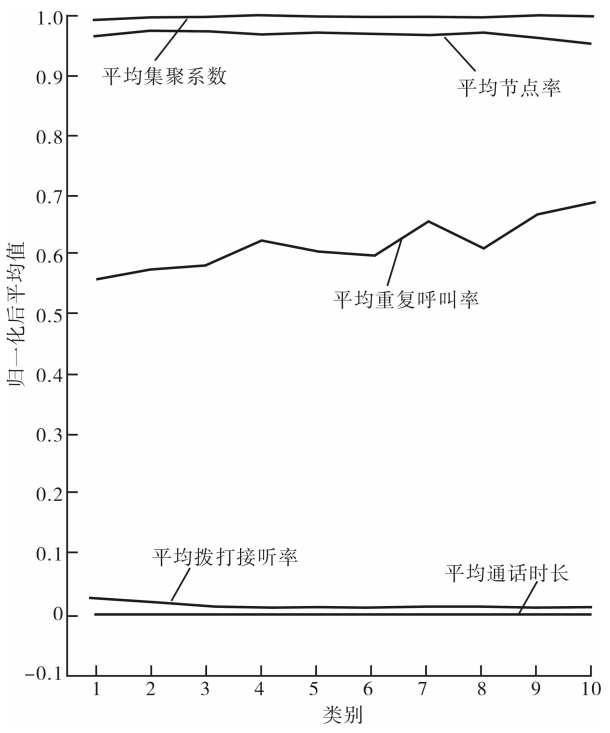

图 2　10 个用户组的属性特征

从图 2 可以看出,类别 1 的平均集聚系数、平均节点率、平均通话时长最短,平均拨打接听率最大,因此该类用户是垃圾用户的可能性最大,其次是类别 2,类别 3. 而类别 9 和类别 10 的平均集聚系数最大,平均拨打接听率最小,因此它们为正常用户的可能性最大.

类别 1 中有 75 项数据,为进一步观察数据之间的关系,将类别 1 用加权聚类的方法再分为 a,b,c 3 个小组,聚类结果显示每一小组中用户人数分别为 3,57 和 15. 各小组平均属性值如表 1 所示,其中小组 a 平均聚集系数、平均三角形节点率最小,平均通话时长最短,平均拨打/接听率最大,是最可疑垃圾用户组,其次是小组 b 和 c. 最可疑小组 a 的平均重复呼叫率并不是最小,但是与小组 b 和 c 差别很小. 小组 a 和 b 共 60 人,占总数的 2%.

表 1　可疑垃圾用户组属性平均值

类别	平均集聚系数	平均三角形节点率	平均重复呼叫率	平均拨打/接听率	平均通话时长/s
a	0.535 6	0.954 9	0.455 9	2.429 7	51.500 7
b	0.568 6	0.970 4	0.449 7	2.053 3	63.601 2
c	0.583 6	0.978 0	0.432 1	1.492 9	66.627 4

4.2　SVM 结果与分析

按照目前常用的交叉验证的方法选择出最优 SVM 参数 $c = 0.25$,$g = 0.062\ 5$,对应的训练模型的准确率为 97.551 2%.

按照训练集得到的模型预测测试集,以 k-means 方法标记的类标号为参照,分类准确率为 97.097 1%. 由于事先并不知道哪些用户是垃圾用户,被标记为垃圾用户的依据是 k-means 的聚类结果,因此分类准确率只是个相对值,但是分类结果在一定程度上可以作为判断垃圾用户的依据. 比如,可以依照此模型对 3 000 名用户以外的无标签通信用户进行可疑度判断,而不需要先经过加权 k-means 算法获得类标签.

5　结语

本文建立了一种基于加权 k-means 和 SVM 的识别 SPIT 方法,依据用户的历史通信活动建立通信行为网络模型,用加权的 k-means 算法对用户进行半监督聚类,然后从每个类中均匀选取部分用户数据作为训练集,采用支持向量机获得训练模型用以预测剩余用户数据. 该方法不同于只将用户分为垃圾用户和非垃圾用户 2 类的方法,不但能有效筛选出垃圾用户,用户分类更为细化,而且有预测功能,具备一定的机器学习能力. 另外,该方法使用范围广,通过改变用户属性权值,可得到特定类型的垃圾用户分类和训练集. 如果从数据集中提取出用户

(下转第 108 页)

位、快速处理,并控制影响范围. 近 3 a 来,企业共发生辅料质量问题 10 次,每次追查隔离有相同质量问题的辅料库存需耗时 2 h,追查范围涉及到多个业务部门. 系统上线后,在系统中 10 min 之内查找到相关的问题小件及分布情况,并为成品追溯提供有效数据支持.

3)避免生产机台投料差错. 通过物料批次有效识别物料小件信息,在物流各环节均进行扫码校验. 在物料移动时,对处于质检状态、有质量问题或者过期的物料,用户扫码时手持终端自动提示,从而有效避免了发错料、用错料的情况. 近 3 a 来,企业共发生辅料用料错误共 14 次,每次造成 2 件成品报废. 系统上线 1 a 以来,纳入批次管理的物料均无用料错误的情况发生,从而避免了用料错误带来的经济损失.

参考文献:

[1] 李民,秦现生,李盘靖,等. 流程产品质量可追溯性[J]. 西北工业大学学报,2002,20(3):506.

[2] 严祥辉. 浅谈基于条码技术的卷烟辅料批次质量追溯系统建设[J]. 海峡科学,2012(11):26.

[3] 陈国清. 卷烟质量跟踪与追溯条码系统的开发应用[J]. 福建质量信息,2008(10):8.

[4] 王青亮. 基于批次管理的产品追踪溯源的研究[D]. 哈尔滨:哈尔滨工业大学,2006.

[5] 韩云辉,韩磊,范黎,等. 烟用材料编码及应用[J]. 郑州轻工业学院学报:自然科学版,2012,27(4):87.

[6] 李歆. 基于 J2EE 设计模式的 Web 应用模型研究与实现[D]. 武汉:武汉大学,2005.

[7] 杨少波. J2EE Web 核心技术:Web 组件与框架开发技术[M]. 北京:清华大学出版社,2011.

(上接第 97 页)

其他属性,比如有效通信率、节点的邻域数量和双向通信率等,则可以用在大客户发现方面,也可以用于关联客户发现和业务推荐等. 但是由于事先并不知道垃圾用户有哪些,算法可能会漏报垃圾用户,以后在实际使用中需要对算法再进行有针对性的改进.

参考文献:

[1] 何光宇,闻英友,赵宏. 基于反馈评判的 SPIT 检测与防范方法[J]. 东北大学学报:自然科学版,2009,30(4):526.

[2] 何光宇,闻英友,赵宏. 固定移动融合网络中基于资源挑战的垃圾语音防范方法[J]. 计算机学报,2012,35(1):38.

[3] 王菲,莫益军,黄本雄. 基于信誉的 P2P-VoIP 垃圾语音过滤模型[J]. 华中科技大学学报:自然科学版,2008,36(8):62.

[4] 张卫兵,魏更宇,黄玮,等. 一种基于布鲁姆过滤器的网络垃圾语音检测方法[J]. 信息工程大学学报,2010,11(5):557.

[5] Tan P N,Steinbach M,Kumar V. 数据挖掘导论[M]. 范明,范宏建,译. 北京:人民邮电出版社,2011.

[6] 夏惠芬,董卫民. 基于关联规则的 Web 挖掘技术研究[J]. 现代电子技术,2011(16):101.

[7] 张雪凤,张桂珍,刘鹏. 基于聚类准则函数的改进 k-means 算法[J]. 计算机工程与应用,2011,47(11):123.

[8] 李健森,白万民. 一种改进的距离度量的聚类算法[J]. 电子设计工程,2012,20(22):86.

[9] 王立梅,李金凤,岳琪. 基于 k 均值聚类的直推式支持向量机学习算法[J]. 计算机工程与应用,2013,49(14):144.

[10] 徐红,彭力,陈容. 基于优化支持向量机的人脸表情分类[J]. 计算机应用研究,2013,30(8):2541.

[11] 奉国和. SVM 分类核函数及参数选择比较[J]. 计算机工程与应用,2011,47(3):123.

[12] Watts D J,Strogatz S H. Collective dynamics of 'small-world' networks[J]. Nature,1998,393(6684):440.