doi:10.11835/j.issn.1000-582X.2018.08.006

一种多分类的微博垃圾用户检测方法

杨　云[1],徐光侠[2],雷　娟[3]

(1.国网重庆市电力公司信息通信分公司,重庆 400014;
2.重庆大学博士后流动站,重庆 400044;3.国网重庆市电力公司电力科学研究院,重庆 401123)

摘要:针对微博多类垃圾用户的检测问题,设计了一种基于模糊多类支持向量机的垃圾用户检测方法。首先,采用一对多 SVM(support vector machines)的构造思想来构造多分类器,并针对每类用户的分类器重新选择训练集;然后,利用构造好的训练集来训练多分类器,经过反复调整参数,得到 5 个用户分类器;最后,针对多分类器的不可分样本,采用模糊聚类来进行模糊处理,即在垂直于 SVM 的最优分类面上定义一个改进的隶属度函数,选择最大隶属度对样本进行再分类。实验结果表明,该方法在保证垃圾用户检测效果的前提下,可以解决多分类中存在的混分和漏分问题。

关键词:微博垃圾用户检测;多分类;模糊处理;隶属度函数

中图分类号:TP393　　**文献标志码**:A　　**文章编号**:1000-582X(2018)08-044-12

A multi-classification method for detecting microblog spam users

YANG Yun[1], *XU Guangxia*[2], *LEI Juan*[3]

(1. State Grid Chongqing Information & Telecommunication Company, Chongqing 400014, P.R.China;
2. Postdoctoral Research Station of Chongqing University, Chongqing 400044, P.R.China; 3. State Grid Chongqing Electric Power Co. Electric Power Research Institute, Chongqing 401123, P.R.China)

Abstract: Based on fuzzy multi-class support vector machine, a method for detecting microblog spammers is designed. Firstly, a multi-class SVM(support vector machines) is used to construct multi-classifiers, and a training set is re-selected for each type of user's classifier. Then, the constructed training set is used to train the multi-classifier, and five user classifiers are obtained after repeated remediation. Finally, for the non-separable samples of multiple classifiers, fuzzy clustering is used to perform the fuzzy processing. An improved membership function is defined on the optimal classification plane perpendicular to the SVM, and the maximum membership degree is used to reclassify the samples. Experimental results show that this

收稿日期:2018-04-02

基金项目:国家自然科学基金项目(61772099);中国博士后基金(2014M562282);重庆市博士后项目(XM2014039);重庆市人工智能技术创新重大主题专项(cstc2017rgzn-zdyf0140);重庆市高校优秀成果转化资助项目(KJZH17116)。

Supported the National Natural Science Foundation of China (61772099), China Postdoctoral Fund (2014M562282), the Project Postdoctoral Supported in Chongqing(XM2014039)), the Artificial Intelligence Technology Innovation Important Subject Projects of Chongqing(cstc2017rgzn-zdyf0140), and the University Outstanding Achievements Transformation Funding Project of Chongqing(KJZH17116).

作者简介:杨云(1964—)男,高级工程师,主要从事网络信息安全、大数据安全及智能电网等研究,(E-mail)story_kb24@163.com。

method can solve the problems of mixing and missing points in multi-classification under the premise of ensuring the detection effect of spammers.

Keywords: microblog spammer detection; multi-classification; fuzzy processing; degree of membership function

微博垃圾用户检测一直是微博反垃圾研究的工作重点,从检测方法上可以分为两类:基于监督学习算法的检测方法和基于无监督学习算法的检测方法[1]。在监督学习检测方面,文献[2]使用 Twitter 提供的 API (application programming interface)抓取了活跃的 Twitter 用户,讨论了一些基于用户和基于内容的垃圾用户和合法用户之间不同的特征,利用这些特征来帮助检测垃圾用户。文献[3]开发了一个数据采集器,收集了两万多个微博用户发布的微博,分析了垃圾用户和正常用户之间不同的行为特征和文本特征,整合了 SVM、RF(random forest)和 NB(naive bayes)3 种监督学习算法并实现了自动垃圾信息检测框架。文献[4]从爬取的新浪微博数据集中研究一组与消息内容和用户行为相关的最重要的一套特征,并将其应用于基于 SVM 的垃圾用户检测模型,用于检测垃圾信息的发送者。

在无监督学习检测算法方面比较具有代表性的是 Tan 等[5]提出一种无监督的社交网络垃圾检测方法,该方法通过观察垃圾用户行为模式,发现垃圾用户通过不断改变行为来逃避检测。文献[6]考虑了每个用户在用户关系中的活动序列,提取社交网络特征,并使用马尔可夫模型混合标记垃圾用户,最后利用基于马尔可夫随机场的统计关系模型检测垃圾用户群。文献[7]提出一种在线社交网络垃圾用户检测方案,该方案从 Facebook 网页中抓取用户数据,然后按照用户的兴趣进行分组,每个组内建立图结构,根据节点间的相似度区分用户。

综上所述,现有的微博垃圾用户研究方案中,基本上都是对全局的垃圾用户特征进行分析,将问题转化为二分类问题,即分出垃圾用户和正常用户,缺乏针对某类垃圾用户的特征分析,这样会使垃圾用户逃避系统的检测。笔者在模糊多类支持向量机算法[8]基础上,设计了一种垃圾用户检测方法。该方法在进行垃圾用户检测时,利用 CDF(cumulative distribution function)曲线寻找每类垃圾用户区分度高的特征,并利用这些特征构造多类分类器。这种多分类器基于一对多 SVM 算法,然后针对一对多 SVM 产生的混分样本和漏分样本进行模糊处理,最后得到基于 FMSVM(fuzzy multi-class support vector machine)的一种垃圾用户检测方法。

1　相关定义

1.1　多分类思想

定义 1　多分类,给定一组训练实例$(\boldsymbol{X}_1,\boldsymbol{Y}_1),(\boldsymbol{X}_2,\boldsymbol{Y}_2),\cdots,(\boldsymbol{X}_n,\boldsymbol{Y}_n)$,典型地,每个实例 $\boldsymbol{X}_i(i=1,2,\cdots,n)$是一个 m 维向量,$\boldsymbol{Y}_i$ 是一个有 $K(K\geqslant 1)$个类别的向量,分类的任务是从训练实例中学习一个模型 $f:\boldsymbol{X}\rightarrow\boldsymbol{Y}$,从而对新的实例给出一个值得信赖的类别预测。

多类分类(multi-class classification)学习的分类器需要对一个样本指定唯一的类别,常见的多类分类器的构造策略有 2 种:第 1 种是基于距离或者后验概率一次性地给出所有类别的度量,然后判定度量结果,选择度量值最大的类别作为预测类别;第 2 种的思想是将多类分类问题拆解为多个二值分类问题,然后结合全部二值分类的结果,得到组合分类器。

第 1 种的指导思想看起来简单,但由于该类最优化问题的设计方法和求解过程相对来说过于复杂,还要同时计算出所有多类分类决策函数,计算量太大,实现难度大、可行性比较弱,未被广泛应用。因此,笔者选择第 2 种策略来构造多分类器。

1.2　多分类 SVM

SVM 是统计机器学习理论(statistical learning theory, SLT)的核心内容,它基于 VC 维(vapnik-chervonenkis dimension)理论和结构风险最小化原理[9]。现如今,SVM 作为分类算法已经得到了相当广泛的应用,常应用在遥感图像识别[10]、人脸检测[11]以及入侵检测[12]等方面,是目前最常用、效果最好的分类器之一。它能够在有限的样本数据中权衡模型的学习能力与复杂性,并自主寻找多分类有比较好的区分能力

的支持向量，达到不同样本类之间间隔最大化的目的，此外，SVM 其本身的优化目标是结构化风险最小，容易抓住数据和特征之间的非线性关系，在小样本训练集上能够得到比其他算法好很多的结果，因此，笔者将运用 SVM 构造多分类器。

SVM 最初的设计目的是将样本分为正类或者负类，但是本研究的目的是要解决多分类问题，因此，选择间接的一对多 SVM 来构造多类分类器。

定义 2 一对多 SVM，该算法是最早出现也是目前应用最为广泛的多类分类方法之一[13]。与一对一 SVM 构建分类器的思想不同，对于 K 类分类问题，其步骤是用一类和剩下其他所有类判别分类，即构造 K 个二值分类器。对于第 $i(1\leqslant i\leqslant K)$ 个分类器，将 i 类中训练样本作为正类，然后将其他类的训练样本作为负类，待分类样本通过 K 个分类器分类，找出属于正类的一个，得到最后的分类结果。

1.3 隶属度函数

SVM 算法对数据中存在的孤立点和干扰数据是非常敏感的，在一对多 SVM 分类器中会存在决策函数出现混分或者漏分的情况，如图 1 所示，此时无法判断待测样本类的类别，因此，需要对决策函数进行处理，或者对决策函数的分类结果作进一步判断。

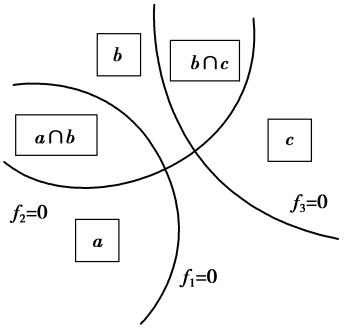


图 1 决策函数出现误差

Fig.1 The error of the decision function

Zadeh[14]提出“隶属度函数”来反映了隶属度问题。在给出的模糊集中，要对论域 U 上面的每个元素给出隶属度，隶属度表示一个元素属于一个集合的程度，隶属度越高就表示属于该集合的程度越大，隶属度越低就表示属于该集合的程度越小。

常用的隶属度函数有 4 种，分别是三角形隶属函数、钟型隶属函数、高斯隶属函数和两边型高斯隶属函数。其中，高斯隶属函数具有良好的抗干扰能力，性能良好，计算方便，模糊化结果更接近于人的认知特点，因此，笔者在垃圾用户分类中选择高斯隶属函数。

定义 3 高斯隶属函数(gaussian membership function)[15]，它是典型的模糊控制算法之一，由 2 个参数 $\{c,\sigma\}$ 来描述：

$$g(x,c,\sigma)=e^{-\frac{(x-c)^2}{2\sigma^2}}。\tag{1}$$

公式(1)中，x 为自变量，c 为某类聚类中心，σ 决定了高斯函数的陡度。通过公式(1)可以得出：$0<g(x,c,\sigma)\leqslant 1$，当 x 距离聚类中心 c 越近，那么其函数值就越大，也就是隶属于该类的可能性就越高。

在多类垃圾用户分类的过程中，由于垃圾用户的类别是确定的且其识别范围属于个体的识别，因此，笔者在对垃圾用户的归类问题中，使用模糊识别的直接方法，而直接方法的判别标准是最大隶属度原则，就是将样本归入到隶属度最大的类别中[15]，具体定义如下：

定义 4 设 $A_i\in F(U)(i=1,2,\cdots,n)$ 对 $u_0\in U$，若存在 i，使得：

$$A_i(u_0)=\max\{A_1(u_0),A_2(u_0),\cdots,A_n(u_0)\},\tag{2}$$

式中：$F(U)$ 为论域 U 上的模糊集的全体成员；$A_i(u_0)$ 为隶属度函数中的最大值，那么认为 u_0 相对地隶属于 A_i。

2 基于 FMSVM 的垃圾用户检测模型框架

FMSVM 框架采用的网页爬虫方法是 UID(user IDentification)遍历爬取策略。UID 就是新浪微博提供给每一个用户的 ID 号，唯一对应且不会改变。UID 遍历全网爬虫的算法是根据指定的 UID 段去爬取微博用户数据，该算法从 UID 的初始字段 $\boldsymbol{X}$ 爬取到结束字段 $\boldsymbol{Y}$。该 UID 爬虫程序包括 3 个模块：模拟登录、网页爬虫和网页内容解析。首先根据新浪微博的特点，实现程序对微博网页的模拟登录；接着通过 HTTP 协议使用 GET 方法采集网页数据并对该数据进行解析。这种方法通过模拟正常用户使用浏览器客户端浏览微博的过程，不依赖于微博平台开放 API，可以根据自己的需求灵活改变爬取数据字段。

针对新浪微博多个用户的特征数据 $\boldsymbol{N}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\}$，其中 $\boldsymbol{X}_i(i=1,2,\cdots,n)$ 为用户 i 的特征向量，n 表示用户数，这里的特征为微博用户的统计特征。从上述特征数据中选取部分微博用户作为训练样本，并

对其进行标注。假设 $\boldsymbol{L}=\{(\boldsymbol{X}_1,\boldsymbol{Y}_1),(\boldsymbol{X}_2,\boldsymbol{Y}_2),\cdots,(\boldsymbol{X}_l,\boldsymbol{Y}_l)\}$表示有类标签的样本集,其中,$l$ 表示被标记的用户数,$\boldsymbol{Y}_i\in\{0,1,2,3,4\}$表示用户 i 的类标签,0 表示正常用户类标签,1 表示营销广告型垃圾用户类标签,2 表示重复发送型垃圾用户类标签,3 表示过度关注型垃圾用户类标签,4 表示主动骚扰型垃圾用户类标签。基于 FMSVM 的垃圾用户检测模型框架如图 2 所示。

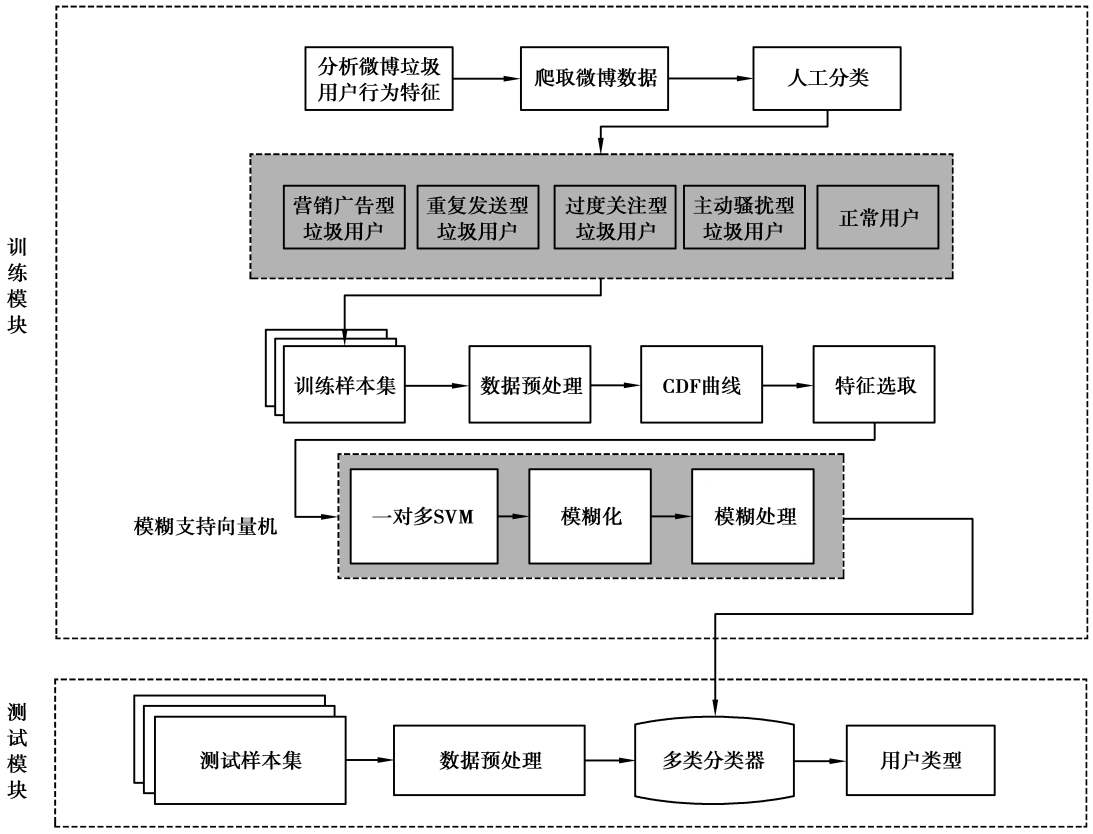


图 2　基于 FMSVM 的垃圾用户检测模型框图

Fig.2　Framework of Spammer Detection Model Based on FMSVM

基于 FMSVM 的垃圾用户检测模型大体上可以分为两部分:训练模块和测试模块。训练模块利用训练样本构建多分类垃圾用户分类器,测试模块主要是利用测试样本集对分类器进行测试,检测分类效果。

第 1 步,分析微博现状,根据用户的行为模式将微博用户分为 5 类:营销广告型垃圾用户 $Spammer1$、重复发送型垃圾用户 $Spammer2$、过度关注型垃圾用户 $Spammer3$、主动骚扰型垃圾用户 $Spammer4$ 和正常用户 $User$。利用爬虫程序获取用户原始数据集 $\boldsymbol{D}=\{\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_n\}$,其中 $\boldsymbol{M}_i(i=1,2,\cdots,n)$是用户 i 的原始数据,包括粉丝数、微博数等直接从网页爬取的基本信息,然后利用原始数据集 D 通过预处理得到微博用户的特征值向量 $\boldsymbol{N}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\}$。

第 2 步,对 5 类用户在统计分析的基础上进行人工标记处理,获得有类标签样本集 $\boldsymbol{L}=\{(\boldsymbol{X}_1,\boldsymbol{Y}_1),(\boldsymbol{X}_2,\boldsymbol{Y}_2),\cdots,(\boldsymbol{X}_l,\boldsymbol{Y}_l)\}$,其中 $Spammer1$ 的标签样本集为 $\boldsymbol{L}_1$、$Spammer2$ 的标签样本集为 $\boldsymbol{L}_2$、$Spammer3$ 的标签样本集为 $\boldsymbol{L}_3$、$Spammer4$ 标签样本集为 $\boldsymbol{L}_4$,$User$ 的标签样本集为 $\boldsymbol{L}_0$。

第 3 步,构造一对多 SVM 分类器。具体步骤是:针对 $Spammer1$ 的分类器 $Classifier1$,将 $Spammer1$ 标为正类,其他 4 类用户标为负类,得到新的标签样本集 $\boldsymbol{LS}_1$,利用 CDF 曲线寻找 $Spammer1$ 与其他 4 类用户区分度高的特征值 $\boldsymbol{Feature}1=\{\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_{k_1}\}$,其中 k_1 是被选出来的区分度高的特征个数总数,在 $\boldsymbol{LS}_1$ 中只保留下 $Feature1$ 的特征,利用 SVM 构建 $Spammer1$ 的分类器。以此类推,完成其他 4 类用户的分类器,分别为 $Spammer2$ 的分类器 $Classifier2$、$Spammer3$ 的分类器 $Classifier3$、$Spammer4$ 的分类器 $Classifier4$、$User$ 的分类器 $Classifier0$。

第4步,利用这五类分类器构造一对多SVM分类器,先判断多分类有无不可分样本,针对不可分样本引入高斯隶属度函数 $g\text{aussian}(x,c,\sigma)$进行模糊处理,最后得到FMSVM分类器。

第5步,利用待测用户测试FMSVM分类器的各项指标性能。

3 基于FMSVM的垃圾用户检测算法

FMSVM算法是在一对多SVM算法的基础上引入模糊隶属度函数,并在对不可分样本进行处理的时候,与原算法进行融合。因此,可以得出该算法的具体实施步骤,并对该步骤进行代码实现,其伪代码如表1所示。

表1 基于FMSVM的垃圾用户检测算法

Table 1 Spammer user detection algorithm based on FMSVM

Algorithm A Microblog spammer detection model based on FMSVM

Input: Data set $\boldsymbol{D}$

Output: Classification Model F(classifier)

Procedure begin

1. /* init Dataset */
2. Use set $\boldsymbol{D}$ to calculate eigenvalues $\boldsymbol{N}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\}$
3. Use CDF to find eigenvalues with the high degree of differentiation
4. Get eigenvalues of each user $Feature_i(0\leqslant i\leqslant 4)$
5. Use $Feature_i(0\leqslant i\leqslant 4)$ to construct training data set $\boldsymbol{D}_i(0\leqslant i\leqslant 4)$
6. /* one-against-all SVM construction stage */
7. **for each** training data set $\boldsymbol{D}_i(0\leqslant i\leqslant 4)$
8. Initialize the penalty parameters C_i
9. According to the SMO algorithm to calculate each optimal solution $\alpha^{(i)*}$
10. According to the optimal solution $\alpha^{(i)*}$, calculate $\omega^{(i)*}$ and $b^{(i)*}$
11. **end for**
12. Get the $Classifier_i(0\leqslant i\leqslant 4)$: $sign(\omega^{(i)*}*x+b^{(i)*})$
13. Input test set to Classifier0,Classifier1,Classifier2,Classifier3,Classifier4
14. Get the classification results for each samples
15. /* Fuzzy processing */
16. **if**(non separable samples existed)
17. Fuzzy processing for non separable samples
18. Get sample category
19. **return** $F(classifier)$

Procedure end

针对上述伪代码,构建每一类的SVM分类器,需要初始化惩罚参数 C,然后根据SMO(Sequential Minimal Optimization)算法计算每个最优解 $\alpha^{(i)*}$,最后根据最优解 $\alpha^{(i)*}$,计算 $\omega^{(i)*}$ 和 $b^{(i)*}$。

表1中模糊处理过程为:对于不可分样本 x,先计算第 i 类用户能与其他类区分开的决策函数:$D_i(x)=w_i^Tx+b_i$,当 $D_i(x)=0$ 时,超平面形成最优分类面,那么通过超平面的分割,属于第 i 类用户的 $D_i(x)=1$,属于其他四类用户的 $D_i(x)=-1$,也就是对于不可分样本 x,当:

$$D_i(x)>0。\tag{3}$$

如果公式(3)只有一个 i 满足,那么样本 x 就属于第 i 类用户。如果公式(3)中有多个 i 满足,那么就存在混分样本,为了解决这类不可分情况,对满足公式(3)的样本点引入高斯隶属度函数,下面具体阐述模糊处理过程。

首先,对于用户类别 i 在第 j 类用户的最优分类面 $D_j(x)=0$ 的垂直方向上定义一个隶属度函数

$m_{i,j}(x)$,那么当 $i=j$ 时:

$$m_{i,j}(x)=\begin{cases}1, & for\ D_i(x)\geqslant 1\\ D_i(x), & otherwise\end{cases};\tag{4}$$

当 $i\neq j$ 时:

$$m_{i,j}(x)=\begin{cases}1, & for\ D_j(x)\leqslant -1\\ -D_j(x), & otherwise\end{cases}。\tag{5}$$

当 $i=j$ 时,如果 $D_i(x)\geqslant 1$,则表示只存在 i 类的训练样本数据,那么假设此时 $m_{i,j}(x)$的值是 1,否则 $m_{i,j}(x)$的值就是 $D_i(x)$。当 $i\neq j$ 时,如果 $D_j(x)\leqslant -1$,则表示类别 i 的样本在最优分类面 $D_j(x)=0$ 的负值区域,那么假设此时 $m_{i,j}(x)$的值是 -1,否则 $m_{i,j}(x)$的值为 $-D_j(x)$。

接下来,求 $m_{i,j}(x)(j=1,\cdots,n)$的最小值来定义用户类别 i 的隶属度函数:

$$m_i(x)=\min_{j=1,\cdots,n} m_{i,j}(x)。\tag{6}$$

根据公式(6)将样本 x 归入类别:

$$\arg\max_{i=1,\cdots,n} m_i(x)。\tag{7}$$

如果从公式(4)和公式(5),x 满足:

$$D_k\begin{cases}>0,\ k=i;\\ \leqslant 0,\ k\neq i,k=1,\cdots,n。\end{cases}\tag{8}$$

且有 $m_i(x)>0$ 和 $m_j(x)\leqslant 0(j\neq i,j=1,\cdots,n)$,将向量 x 归入到类别 i 中。混分情况得到解决,再看多分类中的漏分样本,即

$$D_i(x)<0。\tag{9}$$

对于满足高斯隶属度函数的样本 $i_1,\cdots,i_l(l>1)$,从公式(4)到公式(6),对 $m_k(x)$定义如公式(10)所示:

$$m_k(x)=\min_{j=i_1,\cdots,i_l} -D_j(x)k\in i_1,\cdots i_l。\tag{10}$$

在公式(9)中根据最大隶属度原则,选择最大的 $m_k(x)(k\in i_1,\cdots i_l)$赋给 $D_k(x)$,即

$$m_k(x)=D_k(x)。\tag{11}$$

至此,解决了一对多 SVM 中漏分的问题。

4　实验与结果分析

4.1　实验数据

通过网页爬虫的方式来遍历爬取了新浪微博中 20 000 多名用户的个人信息及其发表的 149 090 条微博信息,其中,选取 4 928 名用户作为训练数据集,这些数据被保存在 MySQL 数据库中。

4.2　FMSVM 分类结果分析

在构造一对多 SVM 分类器之前,需要先构造 5 个分类器的训练集。针对每一类样本的数量,在完整的负类样本集中再抽取相应的数量,例如,针对正常用户,将正常用户标记为 0,将其余 4 类垃圾用户标记为 -1。5 个分类器的训练集样本个数如表 2 所示。

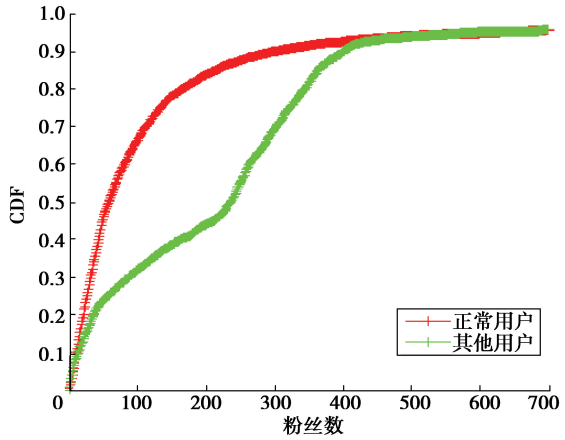
表 2　构造 5 类用户分类器的训练数据集

Table 2　Five user classifier training data sets

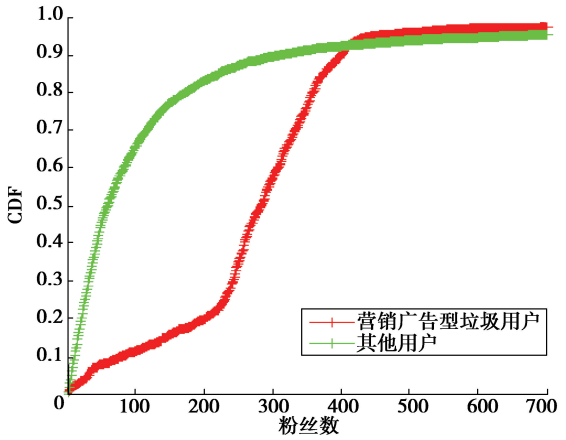
训练数据集	该类用户数量	其他用户数量
主动骚扰型垃圾用户	105	120
过度关注型垃圾用户	103	120
重复发送型垃圾用户	338	350
营销广告型垃圾用户	892	900
正常用户	1 000	1 000

对于每一类用户的分类器，训练集样本的特征选择是十分重要的步骤。为了区分开正类与负类，特征的选择要尽量选出区分度较高的特征，实验通过绘制特征的CDF曲线来寻找每一类用户与其余4类用户具有较大区分度的特征，利用这些特征作为每一类用户分类器的样本特征。

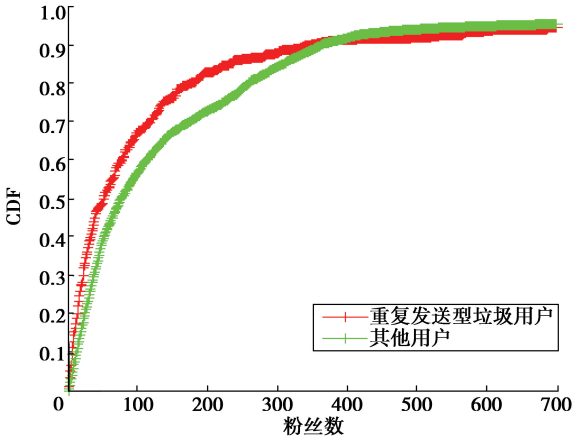
以表2中的粉丝数特征为例，绘制出来该特征在5类样本集中的CDF曲线。如图3所示，从图3(a)和图3(b)可以看出，正常用户样本和营销广告垃圾用户样本的粉丝数特征与其他用户的区分度较大；从图3(c)、(d)、(e)可以看出，其他3类用户的粉丝数特征区分度较小，因此，粉丝数是正常用户和营销广告垃圾用户要选择的特征之一。



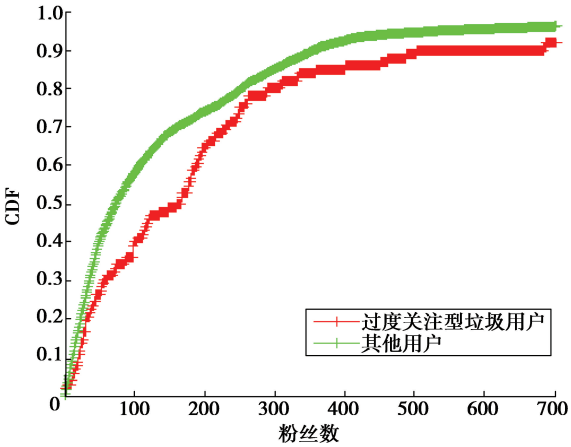
(a)正常用户的粉丝数CDF曲线



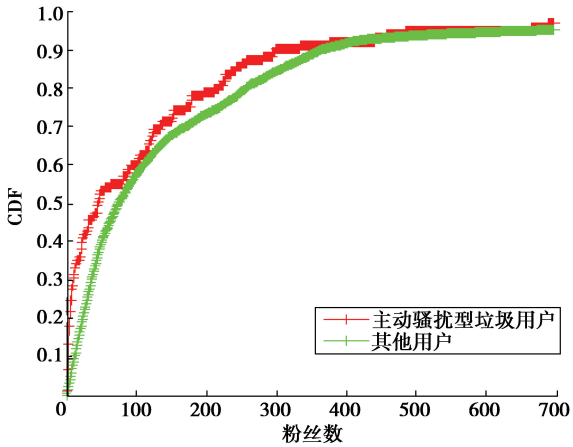
(b)营销广告型垃圾用户的粉丝数CDF曲线



(c)重复发送型垃圾用户的粉丝数CDF曲线



(b)过度管制型垃圾用户的粉丝数CDF曲线



(e)主动骚扰型垃圾用户的粉丝数CDF曲线

图3 粉丝数在五类样本集中的CDF曲线

Fig.3 Number of followers in the CDF curve of five sample groups

最后,针对每一类样本选择了如表 3 所示的各类分类器的训练集特征集。

表 3　各类分类器的特征集

Table 3　Feature sets of various classifiers

训练集名称	特征名称
营销广告型垃圾用户	粉丝数,关注数,总微博数,日发微博数,话题平均数,URL 平均数,图片平均数,原创微博话题平均数
重复发送型垃圾用户	总微博数,日发微博数,话题平均数,平均文本相似性,URL 平均数,原创微博数,原创微博率
过度关注型垃圾用户	关注数,关注/粉丝比,用户名誉度,URL 平均数,转发平均数,话题平均数,原创微博数
主动骚扰型垃圾用户	日发微博数,提及平均数,URL 平均数,图片平均数,平均文本相似性,原创微博提及平均数
正常用户	粉丝数,总微博数,日发微博数,评论平均数,URL 平均数,图片平均数

为了得到一对多 SVM 分类器,利用表 2 中的训练样本,使用第 2 章 FMSVM 框架中步骤 3 来训练每一个用户分类器,得到 5 类用户的 SVM 分类器。为了检测一对多 SVM 分类器的效果,选择训练集样本作为本次的测试样本,将这 5 类测试样本输入到一对多 SVM 分类器中,然后查看每类样本的测试情况。以营销广告垃圾用户为例,将营销广告型垃圾用户测试样本送入到这 5 个分类器中,先统计只分类成正常用户、营销广告型垃圾用户、重复发送型垃圾用户、过度关注型垃圾用户和主动骚扰型垃圾用户的数量,最后统计不可分样本的数量(即混分样本和漏分样本),最后结果如表 4 所示。

表 4　一对多 SVM 分类结果

Table 4　One-to-many SVM classification results

测试样本	C_1	C_2	C_3	C_4	C_5	不可分
正常用户(1 000)	938	22	12	6	0	22
营销广告型(892)	56	750	21	13	1	51
重复发送型(338)	18	25	263	5	2	25
过度关注型(103)	2	7	6	76	0	12
主动骚扰型(105)	8	5	7	3	63	19

从表 4 中可以得出,正常用户的分类器能够较好地将正常用户分类出来,测试样本中的 1 000 个正常用户中有 938 个样本被正确地分出,比其他 4 个垃圾用户的正确分类比例都要高。在垃圾用户检测方面,要在尽量保证正常用户不会误判的前提下,提高垃圾用户的检测效果,正常用户分类器的分类结果达到了要求。再看其他 4 类垃圾用户的分类情况,从分类结果可以看出,营销广告型垃圾用户和重复发送型垃圾用户由于多是真人操作的垃圾用户,其行为会模仿正常用户,因此,相较于其他主动策略型垃圾用户会较多地被误判

为正常用户。营销广告型垃圾用户和重复发送型垃圾用户之间也存在相似行为,分类结果也表明这两类垃圾用户之间被误判为对方垃圾用户的概率要比剩下的主动型垃圾用户和过度关注型用户大。过度关注型垃圾用户分类器的不可分样本最少。主动骚扰型垃圾用户分类器的分类效果最差,105个主动骚扰型垃圾用户只有63个被正确分类,但是其他用户被误判为主动骚扰型垃圾用户的情况也是最少的。此外,从结果可以看出每类用户都存在不可分情况,因此,引入模糊处理十分有必要。

通过分析一对多SVM多类分类器的结果,发现一对多SVM存在较多的不可分样本,针对不可分样本,利用模糊处理方法继续对不可分样本进行分类。分类结果如表5所示。

表5 模糊处理结果

Table 5 Fuzzy processing result

分类结果	0类不可分样本(22)	1类不可分样本(51)	2类不可分样本(25)	3类不可分样本(12)	4类不可分样本(19)
0类	15	5	1	0	0
1类	4	31	4	2	1
2类	3	10	18	2	3
3类	0	3	0	8	0
4类	0	2	2	0	15

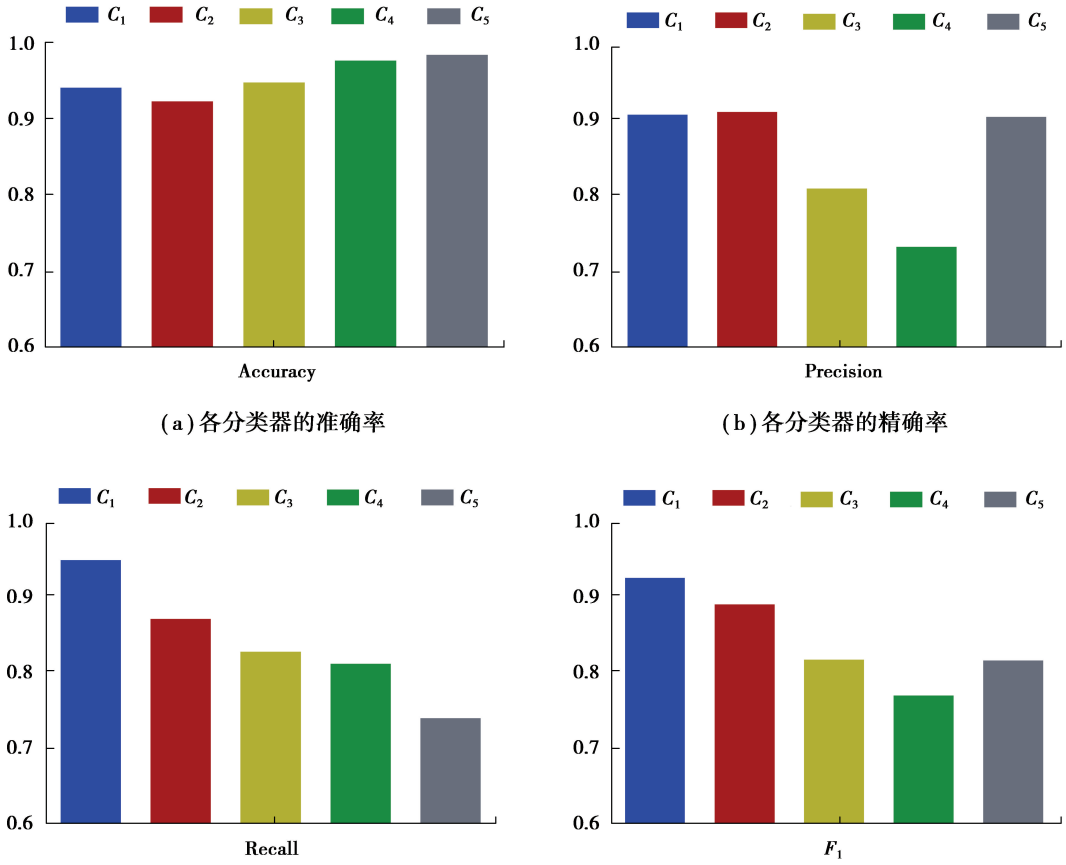
通过模糊处理,不可分样本得到了新的分类。从分类结果来看,每类用户的不可分样本中大部分都被正确地分类到相对应的类别中,少部分被错分为其他类别。以营销广告型垃圾用户的不可分样本为例,51个不可分样本中,有31个被正确地分到营销广告型垃圾用户、5个被分到正常用户、10个被分到重发送型垃圾用户、3个被分到过度关注型垃圾用户和2个被分到主动骚扰型垃圾用户,除了本身31个被正确分类之外,重复发送型垃圾用户由于与营销广告型垃圾用户的行为较为相似,不可分样本也较多地被分到重复发送型垃圾用户中。结合上文的一对多SVM分类结果与本节的模糊处理结果,可以得到FMSVM的最后结果,分类结果如表6所示。

表6 FMSVM分类器最后结果

Table 6 FMSVM classifier final result

测试样本	C_1	C_2	C_3	C_4	C_5
0类(1 000)	953	26	15	6	0
1类(892)	61	781	31	16	3
2类(338)	19	29	281	5	4
3类(103)	2	9	8	84	0
4类(105)	8	6	10	3	78

从表6的最后结果来看,通过模糊处理,不可分样本得到了较好的处理。单一的评价指标在垃圾识别方面是比较片面的,因此,还需要评估每一个分类器的各项指标,在计算每一类用户分类器指标的时候,将该类作为正类,其他4类作为负类,将问题转化为二值分类,评价指标分别为:准确率、精确率、召回率和F_1值,结果如图4所示。



(a)各分类器的准确率　(b)各分类器的精确率

(c)各分类器的召回率　(d)各分类器的F_1值

图 4　各分类器指标

Fig.4　Various classifier indicators

从图 4 可知，五类用户的分类器中，正常用户分类器的各项指标都比较优秀，这说明正常用户能够被很好地分类出来，并且较少地被误判为其他用户。四类垃圾用户分类器中，过度关注型垃圾用户和主动骚扰型垃圾用户在准确率方面都比较高，这是由于样本数量较少而负类样本数量较多，即提高了公式中 TN 的值，因此，这两类样本的准确率上升，其他两类垃圾用户分类器的准确率也能保持在 90%以上。

营销广告型分类器和主动骚扰型分类器的精确率较高，说明被检测出的这两类垃圾用户中真实的用户占比较大，重复发送型分类器的准确率在 80%以上，过度关注型分类器的准确率最低，但也在 70%以上。四类垃圾用户分类器中，过度关注型垃圾用户和主动骚扰型垃圾用户分类器的召回率较低，其原因是这两类用户本身的样本数偏少，而且存在较多样本被错分为其他类别；营销广告型垃圾用户和主动骚扰型垃圾用户的样本数较多，正确分类的样本数也较多，因此，这两类垃圾用户的召回率较高。最后，分析综合评价指标 F_1 值，四类垃圾用户分类器中，除了过度关注型分类器的 F_1 值较低之外，其余三类垃圾用户分类器的 F_1 值均在 80%以上，重复转发型分类器和主动骚扰型分类器的 F_1 值基本相同，营销广告垃圾用户的 F_1 值最高。

总体来说，4 类垃圾用户的各项指标令人满意，尤其是垃圾用户数量最多的广告营销型垃圾用户，该类垃圾用户分类器的各项指标都比较高，能够起到检测广告营销型垃圾用户的作用。

4.3　多分类器对比分析

随后，在多分类器的对比上，选择 MLP(multiLayer perceptron)[16]、MCC(multi-class classifier)与 FMSVM 进行对比。实验样本选择表 2 中的样本集，多分类器的评价指标依然选择准确率、精确率、召回率和 F_1 值，各项指标的计算方法同 FMSVM，最后得到图 5 的各项指标结果。

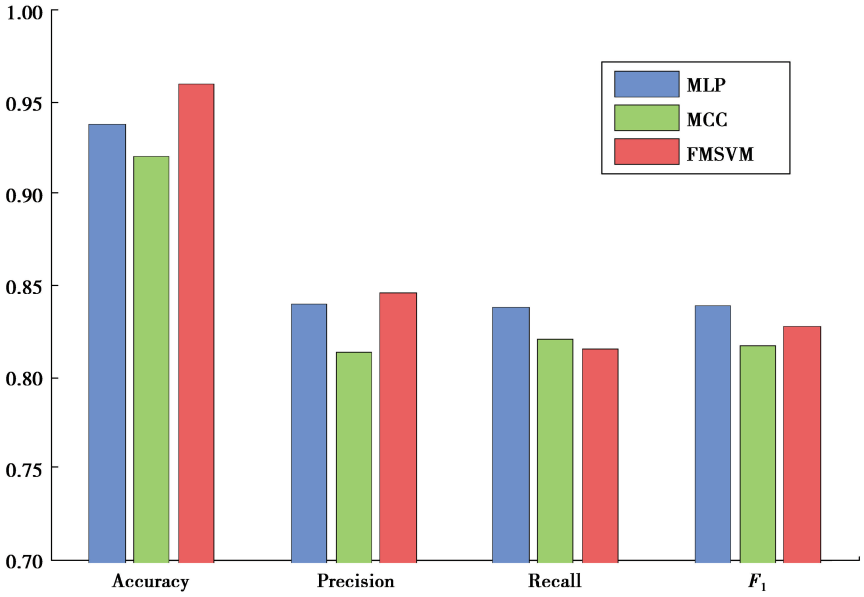


图5 多分类器与FMSVM指标对比

Fig.5 Comparison between multi-class classifiers and FMSVM

如图5所示,通过柱状图的形式展示各项评价指标实验结果。从图5可以看出,FMSVM在准确率上高于MLP和MCC,说明了FMSVM中加入模糊处理能够提高正确分类的样本数。在精确率方面,FMSVM略高于MLP,明显高于MCC,精确率得到保证意味着检测出来的垃圾用户中真实的垃圾用户占比高,也说明正常用户被误判的比例较低,保证精确率是有效检测垃圾用户的前提条件。在召回率上,MLP要略高于FMSVM;相比于MCC,FMSVM比MCC低0.6%,基本持平。最后分析综合评价指标 F_1 值,3种多分类器中,FMSVM处在中等水平,相比MLP低1.1%,相比MCC高1.1%。总之,FMSVM算法的垃圾用户检测能力还是较好的。

5 结 语

设计了一种基于FMSVM的垃圾用户检测方法。该检测方法在一对多SVM分类器的基础上,首先,在构造每一类用户分类器的时候,需要选择不同的训练集,每一类用户的训练集要将该类用户标记为正类,其余用户标记为负类;然后,通过特征选择选取区分度高的特征组合成训练集;最后,通过训练得到5类用户的5个分类器。针对多分类的混分样本和漏分样本,该检测方法引入模糊处理,在垂直于SVM的最优分类面上定义一个隶属度函数,解决多分类的不可分情况。实验对比了多分类算法,对比结果表明,设计的基于FMSVM的垃圾用户检测方法在多分类情况下,能够达到较好的检测效果。

参考文献:

[1] 张玉清, 吕少卿, 范丹. 在线社交网络中异常帐号检测方法研究[J]. 计算机学报, 2015, 38(10): 2011-2027.
ZHANG Yuqing, LV Shaoqing, FAN Dan. Anomaly detection in online social networks [J]. Chinese Journal of Computers. 2015, 38(10): 2011-2027.(in Chinese)

[2] Mccord M, Chuah M. Spam detection on twitter using traditional classifiers[C]// Autonomic & Trusted Computing-International Conference, Banff: ACM Press, 2011: 175-186.

[3] Ma Y, Niu Y, Ren Y, et al. Detecting spam on sina weibo[J]. International Workshop on Cloud Computing and Information Security, 2013, 3(2): 93-96.

[4] Zheng X, Zeng Z, Chen Z, et al. Detecting spammers on social networks[J]. Neurocomputing, 2015, 159(1): 27-34.

[5] Tan E, Guo L, Chen S, et al. UNIK: unsupervised social network spam detection[C]// Acm International Conference on Conference on Information & Knowledge Management. San Francisco: ACM Press, 2013: 479-488.

[6] Fakhraei S, Foulds J, Shashanka M, et al. Collective spammer detection in evolving multi-relational social networks[C]// Acm Sigkdd Conference on Knowledge Discovery & Data Mining. Sydney: ACM Press, 2015: 1769-1778.

[7] Ahmed F, Abulaish M. Identification of sybil communities generating context-aware spam on online social networks[M]. Berlin: Springer, 2013: 268-279.

[8] Abe S. Fuzzy support vector machines for multilabel classification[J]. Pattern Recognition, 2015, 48(6): 2110-2117.

[9] Loosli G, Canu S. Comments on the "core vector machines: fast SVM training on very large data sets"[J]. Journal of Machine Learning Research, 2007, 8(2): 291-301.

[10] Zhou J H, Qin J, Gao K, et al. SVM-based soft classification of urban tree species using very high-spatial resolution remote-sensing imagery[J]. International Journal of Remote Sensing, 2016, 37(11): 2541-2559.

[11] Cui P, Yan T. A SVM-based feature extraction for face recognition[J]. Communication in Computer and Information Science, 2016(623): 120-126.

[12] Wang W, Liu J, Pitsilis G, et al. Abstracting massive data for lightweight intrusion detection in computer networks[J]. Information Sciences, 2018(433/434): 417-430.

[13] Zhu W, Zhong P. A new one-class SVM based on hidden information[J]. Knowledge-Based Systems, 2014, 60(2): 35-43.

[14] Gao C, Ge Q, Jian L. Rule extraction from fuzzy-based blast furnace SVM multiclassifier for decision-making[J]. IEEE Transactions on Fuzzy Systems, 2014, 22(3): 586-596.

[15] 杨纶标, 高英仪, 凌卫新. 模糊数学原理及应用[M].5 版. 广州:华南理工大学出版社, 2011.
YANG Lunbiao, GAO Yingyi, LING Weixin. Principle and application of fuzzy mathematics[M].5th ed. Guangzhou: South China University of Technology Press, 2011.

[16] Lunghi P, Ciarambino M, Lavagna M. A multilayer perceptron hazard detector for vision-based autonomous planetary landing[J]. Advances in Space Research, 2016, 58(1):131-144.

(编辑　王维朗)