一种新的中文文本分类算法
—One Class SVM-KNN算法

刘 文,吴 陈
(江苏科技大学 智能信息处理实验室,江苏 镇江 212003)

摘 要:中文文本分类在数据库及搜索引擎中得到广泛的应用,K-近邻(KNN)算法是常用于中文文本分类中的分类方法,但K-近邻在分类过程中需要存储所有的训练样本,并且直到待测样本需要分类时才建立分类,而且还存在类倾斜现象以及存储和计算的开销大等缺陷。单类SVM对只有一类的分类问题具有很好的效果,但不适用于多类分类问题,因此针对KNN存在的缺陷及单类SVM的特点提出One Class SVM-KNN算法,并给出了算法的定义及详细分析。通过实验证明此方法很好地克服了KNN算法的缺陷,并且查全率、查准率明显优于K-近邻算法。

关键词:中文文本分类;支持向量机;K-近邻;One Class SVM-KNN
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2012)05-0083-04

A New Text Classification Algorithm—One Class SVM-KNN

LIU Wen, WU Chen
(The Opening Laboratory of Intelligent Computing, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract: Text classification is widely used in database and search engine. KNN is widely used in Chinese text categorization, however, KNN has many defects in the application of text classification. The deficiency of KNN classification algorithm is that all the training samples are kept until the samples are classified. When the size of samples is very large, the storage and computation will be costly, which will result in classification deviation. One class SVM is a simple and effective classification algorithm in one class. To solve KNN problems, a new algorithm based on harmonic one-class-SVM and KNN was proposed, which will achieve better classification effect. The experiment result is shown that the recall computed using the proposed method is obviously more highly than the KNN method.

Key words: Chinese text classification; support vector machine; K-nearest neighbour; One Class SVM-KNN

0 引 言

随着信息技术的发展,特别是Internet的应用和普及,互联网上的信息以及各种科技文献成爆炸式地出现在人们面前,然而大部分信息仍是以文本形式存储,如何从这些海量数据中快速、准确、全面地提取人们需要的信息是信息检索重要的研究课题之一,然而对于这种半结构或无结构化的数据[1],获取特定信息的手段却较弱,导致搜索困难和信息利用率低下。由此数据挖掘中的文本挖掘及搜索引擎中的信息检索等研究领域出现了前所未有的高潮[2-4]。

文本分类是一种确定文档所属类别的情报分析方法,随着文本挖掘技术的发展[5],目前常用的分类方法有Naive Bayes、Decision tree、K-近邻、SVM算法等。其中分类效果相对较好、用的比较多的方法为K最近邻算法和支持向量机算法,但K最近邻算法存在类倾斜现象[6],并且存在存储和计算开销比较大等缺陷。文中在单类分类支持向量机和K最近邻算法的基础上,提出了One Class SVM-KNN算法,可以很好地解决K-近邻算法存在的缺陷,实验证明该方法具有更好的分类效果。

收稿日期:2011-09-28;**修回日期**:2011-12-29
作者简介:刘 文(1985-),男,山东临沂人,硕士研究生,研究方向为数据挖掘;吴 陈,教授,研究方向为数据挖掘、模式识别等。

1 文本预处理中的关键技术

1.1 文本表示模型

文本表示就是将文本转化为在机器上可以描述和计算的形式的过程[7],是文本分类中的重要步骤之一。现有文本表示模型有布尔模型、向量空间模型、概率模型、图模型等,向量空间模型是常用的文本表示模型。

向量空间模型[5](Vector Space Model, VSM)是20世纪60年代末由GerardSalton等人提出的,它采用简洁的特征向量来表示文本,即一篇文本D可以表示为:

$D=D(t_1,w_1,\cdots,t_i,w_i,\cdots,t_n,w_n)$, $i=1\cdots n$;其中 t_i 表示文本中的字、词或短语成为项,w_i 为项 t_i 的权重表示 t_i 在文本中的重要程度。为了简化分析,t_i 在文本中的先后顺序可以暂时不考虑并要求 t_i 互异,此时可以把 $t_1,t_2,\cdots,t_n$ 看成一个 n 维的坐标系,而 $w_1,w_2,\cdots,w_n$ 为相应的坐标值,所以 $D(w_1,w_2,\cdots,w_n)$ 可以被看成是 n 维空间中的一个向量,则 $D(w_1,w_2,\cdots,w_n)$ 称为文本 D 的向量表示。

1.2 文本的特征选择

一篇文本有成千上万个字词组成,如果把每一个字词都做为文本的特征来进行分类,显然维数过大影响分类的效果,而且文本中的某些字和词对文本的类别判别无作用,因此提高分类准确率可以去除区分度较小的噪音特征项[8],加快运行速度可以去除重要性较低的低频特征,所以在分类之前,对文本进行特征选择是必要的,常用的特征选择方法有文档频次、互信息、信息增益、χ^2统计等。文献[9]中对这些方法做了描述,并进行了改进。

2 算法描述

2.1 单类 SVM 算法

只有一种类别的样本作为训练样本集的分类问题叫做单类分类问题,通常划分为类的样本叫做正常类[10],而其它数据叫做异常类或负类。单类支持向量机是指解决单一类别分类问题的支持向量机,这种单类分类支持向量机应用十分广泛,可应用于图像抽取、人脸检测、异常检测、文本检测等领域。

基于密度的方法和基于边界的方法是单类问题的典型方法。基于密度和基于边界的方法在文献[10]中做了详细介绍,文中主要采用基于边界的方法。

算法简要描述如下:

给定包含 n 个样本点的训练集合 $X=\{x_1,\cdots,x_n\}$,寻找一个超球体,使其在包含样本 数尽可能多的情况下,而使超球体的半径尽可能的小。设球心为 O,半径为 R,通过样本点到球心 O 的距离是否大于半径 R,来判断样本点是否属于该类,要使超球体的半径 R 尽可能的小,并且能够包含类中的所有或大部分样本点。由于敏感的样本点往往离球心比较远,因此允许一些样本点在球体的外面。

在此引入松弛变量 ξ_i,得到下面的限制条件:

$$(x_i-o)\ (x_i-o)^{\mathrm{T}}\leqslant R^2+\xi_i \tag{1}$$

其中,$\xi_i\geqslant 0$,要使球体半径 R 和松弛变量 ξ_i 这两项最小化:

$$f(R,o,\xi_i)=R^2+c\sum_i\xi_i \tag{2}$$

其中,为了平衡不被包含的样本的数目和球体的体积引入常数 c。在加入了限制条件(1)后构造 Lagrange 函数:

$$L(R,o,\alpha_i,\xi_i)=R^2+c\sum_i\xi_i-\sum_i\alpha_i\{R^2+\xi_i-(x_i-2ox_i+o^2)\}-\sum_i\beta_i\xi_i \tag{3}$$

其中 $\alpha_i\geqslant 0,\beta_i\geqslant 0$,则可将问题(2)转化为其对偶问题:

$$\min_o\sum_{i,j=1}^{n}o_io_j(x_i\cdot x_j)-\sum_{i=1}^{n}o_i(x_i\cdot x_j)$$
$$\text{s.t.}\ \sum_{i=1}^{n}o_i=1,\alpha=\sum_{i=1}^{n}o_ix_i\quad 0\leqslant o_i\leqslant c,i=1,\cdots,n \tag{4}$$

通常通过比较一个样本点到圆心的距离和半径的大小,从而来判断样本点是正常类,还是非正常类,当一个测试样本点 y 的距离小于半径时,样本点 y 为正类,否则为非正常类。

$$(y-o)\ (y-o)^{\mathrm{T}}=(y\cdot y)-2\sum_{i=1}^{n}(y\cdot x_i)+\sum_{i,j=1}^{n}o_io_j(x_i\cdot x_j)\leqslant R^2 \tag{5}$$

能够使上面等式成立,并且满足 $O_i\neq 0$ 的向量叫做支持向量,少量球面上的支持向量决定了球体的半径。

2.2 KNN 算法

K-近邻算法是一种简单、有效、懒散、无参数的分类方法,主要思想是通过计算待测样本与训练集样本之间的距离,寻找与测试样本距离最小的 K 个数据样本,然后根据这 K 个样本的类别属性判断测试样本属于哪个类别。

虽然 KNN 分类算法在文本分类、图像分类、网页分类中得到广泛应用,但它还存在一些缺陷:

①它是懒散的无参数化的分类方法,需要存储所有的训练样本,每次分类都需要计算待测样本与训练集中每个样本之间的距离,所以存储和计算开销量非常大[11]。

②对于 K-近邻分类器,由于训练样本分布的不均匀性在类边缘处可能会造成测试样本类别的误判。如图 1 中的类 1 和类 2 所示:由于类 1 的样本数量明显比类 2 多,所以 x 样本很容易被误分为类 1 的样本。

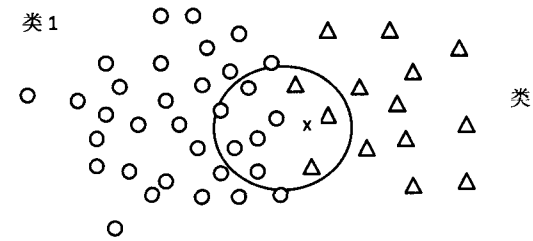

图 1 KNN 分类示意图

另外判断分类样本间是否存在类倾斜现象采用文献[8]中提出的方法,即采用类密度和标准差来判定倾斜现象,其中类密度定义公式为:

$$c_i = \frac{\sum_{j=1}^{n} c_{ij}}{n}, \quad i = 1,\cdots,m; \quad j = 1,\cdots,n$$

其中 c_{ij} 表示类别 L_i 中文本 a_j 与类内最近邻的相似度,m 为文本类别数,n 为 L_i 类中的文本个数。

标准差公式定义为:

$$\delta = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})}$$

其中 N 为文本类别数,x_i 表示第 i 类文本密度,$\bar{x}$ 表示文本类的平均密度。

由上面类密度和标准差的定义可以将类的倾斜定义为:给定一个 T 为文本训练集,共有 t 类数据,每类的训练样本密度分别为 $C_1,\cdots,C_t$,假设它的标准差为 $\delta, X_0 = \min(C_1,\cdots,C_t)$。当 $X_0 < \delta$ 时,称文本训练集 T 是类倾斜的;如果 $X_0 > \delta$,那么称文本训练集 T 不是类倾斜的。

2.3 One Class SVM—KNN 的算法描述

单类 SVM 在解决单一类别问题上有很好的效果,但不适用于多类分类问题,KNN 算法虽然有较好的应用但还存在一些缺陷,所以吸取两者的优点组合出新的分类算法,One Class SVM-KNN 算法。

根据 K-近邻算法的近邻规则,当测试样本不在某类的边界区域时,不管类中心区域有没有分类的训练样本,只要在类边界区域有适当的训练样本,就可以根据这些在边界区域的训练样本,实现对测试样本的正确分类;当测试样本在某类的边界区域时,那么它的最近邻也主要集中在那些处于边界区域的训练样本中,综上可以得出类中心区域的大多数训练样本在分类过程中不会起到太多的作用。所以采用 One Class SVM 对训练样本进行适当的裁剪,同时减小因类倾斜现象带来的误差。

分类过程中先用单类 SVM 将训练集中的每一类生成分类器,然后将整个训练集作为测试样本集进行测试,将不属于该类而被误判为此类的样本,以及属于此类而没有被正确分类的样本和各类的支持向量机组成新的训练集,然后用 KNN 算法以新的训练集训练分类器进行二次修正,分类器训练总体流程如图 2 所示。

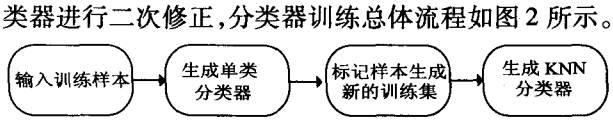

图 2　分类器总体流程

新的训练集中的样本基本上是处于各类的边缘或交叉区域,是容易被误分的样本群,将这些特殊样本作为新的训练集,当待测样本具有与这些样本相似的性质时便减少了被误分的概率,从而提高分类效果。

如图 3 所示,类 1 和类 2 经过单类支持向量机分类器后,形成的训练集样本数量大大缩减,数据分布相对均匀,从而在一定程度上解决了由于类的分布不均导致的类倾斜问题。

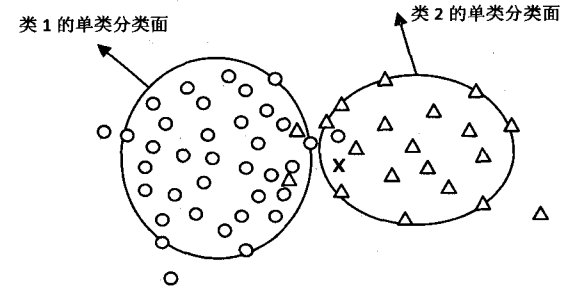

图 3　单类分类面

算法具体过程如下:

假定给定训练样本集 $T = \{(x_1,y_1),\cdots,(x_l,y_l)\} \in (R^n \times Y)^l, x_i \in R^n, y_i \in Y = \{1,2,\cdots,N\}, i = 1\cdots l$。

第一步,首先通过单类 SVM 算法,对第 y_i 类样本进行分类,从而得到分类器 f_i 标记支持向量。

第二步,用训练样本集作为测试样本,通过上步得到的分类器,标记被误分到该类的样本和属于该类而没分到该类的样本。

第三步,循环第一、第二步将训练集中的每个类生成一个单类分类器,并将每类中标记的误分到每类的样本和属于该类而没分到该类里的样本,以及每类的支持向量组成新的训练样本集 T_2,T_2 作为 KNN 算法的训练集,生成 KNN 分类器。

第四步,对未知类别的样本分类时首先经过每类的分类器 f_i,然后通过以 T_2 为训练集的 KNN 算法进行修正,样本的类别为 f_i 和 KNN 结果一致的类别。如果经过 f_i 分类后待测样本不属于任何类,则以 KNN 的分类结果为该类的类别。如果 f_i 和 KNN 的分类结果不一致,采用 KNN 算法的分类结果为该样本的类别。

3　实验及结果分析

实验数据采用中科院计算所提供的中文文本语料库 TanCorp1.0,此语料库共收集 12 类文本,共 14150 篇文章,本实验在其中抽取了财经、教育、艺术、电脑、科技、人才等六类文本,每类抽取 300 篇文章共 1800 篇作为实验数据。

实验环境为 matlab 环境,工具箱主要采用台湾大学林智仁副教授提供的 LIBSVM 工具箱,中文文本分类性能的评估,采用通用的国际评估指标,即 Recall、Precision、F1 评估值等三项主要指标。对应的公式为:

①查准率(Precision,简记为 P)$P = cp_i/p_i$;

②查全率(Recall 简记为 R)$R = cp_i/c$;

③F1 评估值，$F1 = 2RP/(R+P)$，其中 c 表示实际属于类 c_i 的样本数，p_i 表示分类器预测为类 c_i 的样本数，cp_i 表示正确分到类 c_i 的样本数。

实验取每类中的 200 篇文章作为训练样本，其余作为测试样本，由 2.2 节可以计算出训练样本各类的类密度为{2,5,7,3,4,15}，$x_0 = 2$，标准差 $\delta = 4.32$，$x_0 < \delta$，通过进行两两计算可以得出财经和电脑、财经和人才存在倾斜现象，分别采用 KNN 算法和 One Class SVM-KNN 算法进行测试，同时设置分类计时器 t 用于验证分类速度有没有提高，分类结果数据如表 1 所示。

表 1 KNN 与 One Class SVM-KNN 实验结果

	KNN 算法			One Class SVM-KNN		
类别	查全率	查准率	F1	查全率	查准率	F1
财经	0.7000	0.7303	0.7150	0.7800	0.8654	0.8205
教育	0.8100	0.7712	0.7912	0.8200	0.8826	0.8501
艺术	0.6900	0.7145	0.7020	0.8000	0.8504	0.8246
电脑	0.8200	0.8028	0.8113	0.8700	0.9371	0.9023
科技	0.7900	0.7625	0.7760	0.8100	0.9074	0.8559
人才	0.7100	0.6834	0.6964	0.7700	0.8529	0.8093

由表 1 可以看出由于财经和电脑、财经和人才存在倾斜现象，采用传统的 KNN 算法进行分类效果不太理想，而经过改进的算法 One Class SVM-KNN 进行测试，整体的分类效果明显提高，而且财经、艺术、人才的分类效果提高幅度更明显。

采用 KNN 算法分类所用时间 $t = 95s$，采用 One Class SVM-KNN 分类的时间 $t = 45s$，由此可以看出改进的分类算法在分类速度上也明显提高。

4 结束语

KNN 算法是数据挖掘领域的重要分类方法之一[12]，KNN 算法实现简单被广泛应用于数据挖掘、模式识别、图像处理等领域。文中通过将 One Class SVM 和 KNN 结合组成新的分类算法：One Class SVM—KNN，通过上面实验分析可知，One Class SVM-KNN 方法可以很好地解决传统的 KNN 方法存在的类倾斜及存储与计算开销大等缺陷，分类效果明显提高，是一种可行的方法。

参考文献：

[1] 唐 菁，沈记全，杨炳儒. 基于 Web 的文本挖掘系统的研究与实现[J]. 计算机科学，2003，30(1)：60-63.

[2] 姜 鹤，陈丽亚. SVM 文本分类中一种新的特征提取方法[J]. 计算机科学与发展，2010，20(3)：17-19.

[3] 马忠宝，刘冠蓉. 基于支持向量机的中文文本分类模型研究[J]. 计算机技术与发展，2006，16(11)：70-72.

[4] 李荣陆. 文本分类及其相关技术研究[D]. 上海：复旦大学，2005.

[5] Lewis D D，Yang Y，Rose T，et al. Rcv1：A New Benchmark Collection for Text Categorization Research[J]. Journal of Machine Learning Research，2004(5)：361-397.

[6] 许高建. 基于 web 的文本挖掘技术[J]. 计算机技术与发展，2007，17(6)：187-190.

[7] Salton G，Wong A，Yang C S. A Vetor Space Model for Automatic Indexing[J]. Communications of ACM，1975，18(11)：613-620.

[8] 闫 晨. KNN 文本分类研究[D]. 秦皇岛：燕山大学，2010.

[9] Manevitz L，Yousef M. One class SVMs for document classification[J]. Journal of Machine Learning Research，2002(2)：139-154.

[10] 冯爱民，薛 晖，刘学军，等. 增强型单类支持向量机[J]. 计算机研究与发展，2008，45(11)：1845-1864.

[11] 曹建芳，王鸿斌. 一种新的基于 SVM-KNN 的 web 文本分类算法[J]. 计算机与数字工程，2010(4)：59-61.

[12] Fernandez J，Montances E，Diaz I，et al. Text categorization by a machine learning based term selection[C]//Proc. of the Database and Expert Systems Applications Conference. Zaragoza：[s. n.]，2005：253-262.

(上接第 82 页)

[5] 齐 越，沈旭昆，段米毅，等. 基于 Perlin 噪音绘制云的方法[J]. 系统仿真学报，2002(9)：1204-1207.

[6] Max N. Computer Animation of Cloud[C]//Proc of Computer Animation. [s. l.]：IEEE，1994：167-174.

[7] 龚 琳，王善斌，顾大权. 基于计算机视觉和粒子系统的三维云模拟[J]. 中国体视学与图像分析，2008，13(1)：59-62.

[8] 石教英，蔡文立. 科学计算可视化算法与系统[M]. 北京：科学出版社，1996.

[9] 吴 晶，徐晓刚，陈新来，等. 云的模拟技术综述[J]. 计算机应用研究，2009，26(4)：1205-1209.

[10] 胡 香，游 雄，武玉国. 体云的光照模型及其算法实现[J]. 测绘科学技术学报，2006，23(3)：205-207.

[11] 李玲娟，张 敏. 云计算环境下关联规则挖掘算法的研究[J]. 计算机技术与发展，2011，21(2)：43-50.

[12] 徐江斌，赵 健，杨 超，等. 真实感云的快速建模[J]. 小型微型计算机系统，2010，8(8)：1590-1594.

[13] 李占德，张政保，文家福，等. 用于图像认证的小波域双重脆弱水印算法研究[J]. 计算机技术与发展，2011，21(2)：181-184.