第26卷 第3期
2011年6月

郑州轻工业学院学报(自然科学版)
JOURNAL OF ZHENGZHOU UNIVERSITY OF LIGHT INDUSTRY(Natural Science)

Vol.26 No.3
Jun. 2011

文章编号:1004-1478(2011)03-0001-04

基于KSVM的网络评论情感分类研究

张素智, 孙培锋

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450002)

摘要:结合机器学习方法中的SVM算法和KNN算法各自的优势,提出一种KSVM分类算法,采用具有语义倾向的词并综合其词性作为特征项,将一些网络评论进行情感分类,以判断一篇评论是正面还是反面. 实验表明,运用该算法对网上的一些评论进行分类,可以达到较高的准确率.

关键词:语义倾向度;情感文本分类;情感特征选择;KSVM

中图分类号:TP391 **文献标志码**:A

Sentiment classification of network comments based on KSVM

ZHANG Su-zhi, SUN Pei-feng

(*College of Comp. and Com. Eng., Zhengzhou Univ. of Light Ind., Zhengzhou* 450002, *China*)

Abstract: A KSVM classification algorithm by combining the advantages of SVM algorithm and KNN algorithm in machine learning is proposed. Some with semantics tendency and combined the parts of speech is chosen as the characteristic items, and the proposed algorithm is applied in sentiment classification of network comments to judge one comment positive or negative. Experimental results showed that the proposed algorithm can classify some comments online with a higher accuracy.

Key words: semantics tendency; sentiment classification; sentiment feature selection; KSVM

0 引言

随着互联网的快速发展,网络已经成为一个便捷的信息交流平台,越来越多的人开始利用网络发布信息或发表自己的观点,并利用网络查看相关评论信息,以此来辅助用户的最终决策. 然而,这些评论没有经过分类整理,正面与反面的评论往往混在一起,不便于查看. 由此引出了文本分类领域一个新的研究方向——基于情感的文本分类. 对文本所表达的情感等主观内容进行分类,可以分别查看不同倾向的评论,并对评论的总体情感倾向迅速做出判断.

目前可用于评论分类的技术主要分为2类:一是机器学习技术,包括在普通文本分类中成熟应用的Naive Bayes算法、最大熵算法、SVM算法、KNN算法等;二是基于语义分析的技术,主要有SO-PMI分类法、SO-LAS分类法等. 本文拟结合机器学习分类技术中KNN算法和SVM算法各自的优点,对网络评论分类提出一种新的KSVM算法.

1 文本分类框架

文本分类过程主要包括文本预处理、特征选

收稿日期:2011-03-24

基金项目:河南省重点科技攻关项目(082102210054);河南省自然科学基金资助项目(0411010500)

作者简介:张素智(1965—),男,河南省孟州市人,郑州轻工业学院教授,博士,主要研究方向为Web数据库、分布式计算和异构系统集成.

择、文档表示及特征权重计算.

1.1 文本预处理

文本预处理是指在做中文分词之前进行的去噪处理,即将一些与文档内容不相干且出现频率较高的词语从文本中去掉,这些词语往往未携带有用信息,对特征的提取无辅助作用. 本文采用中国科学院计算技术研究所研制的汉语词法分析系统(ICTCLAS),对原始语料进行停用词过滤,留下需要的形容词、副词、名词、动词等.

1.2 特征选择

本文研究的重点是情感分类,即重在分析那些具有感情色彩的词,故只选择那些能表达语义倾向的词来进行分析. 采用基于情感词典的统计方法可以判定一个词语是倾向于正面还是负面[1]. 本文采用基于 How Net 计算出现在文本中的每个词的语义倾向度来计算语义倾向值的方法:选用具有强烈褒贬倾向的40个正面基准词(如:优秀)和40个负面基准词(如:恶劣),通过计算词汇与基准词之间关联的紧密程度来判断语义倾向度. 其计算公式如下:

$$Se\text{-}tendency(\omega) = \sum Similarity(Pkey,\omega) - \sum (Nkey,\omega)$$

$$Similarity(p_1,p_2) = \alpha/(d+\alpha)$$

其中,$Similarity(key,\omega)$ 为2个单词义源相似度的最大值,$Pkey$ 为正面基准词,$Nkey$ 为负面基准词;p_1,p_2 为词汇的义源;d 为 p_1,p_2 在义源层次体系中的路径长度,是一个常数;α 为一个可调节的参数[2]. 计算结果大于0时为正面评论,小于0时为负面评论. 本文选取正(负)面形容词,正(负)面副词,正(负)面名词,正(负)面动词作为特征项.

1.3 文档表示及特征权重计算

文档表示是指将计算机无法理解的自然语言表达转化为计算机能理解的结构化数据的过程. 本文采用常见的向量空间模型(VSM),将训练和测试样本表示成由特征项构成的向量空间,即 $D(\omega_1,\omega_2,\cdots,\omega_i)$. 其中,$D$ 表示文档,ω_i 表示第 i 个特征词元的权重. 本文采用 *tf-idf* 对其进行计算,计算公式如下:

$$tf\text{-}idf(t_k,d_j) = tf(t_k,d_j) \times \log\frac{N}{n(t_k)}$$

其中,特征频率 *tf* 表示特征词元在网络评论中出现的次数,文档频率 *df* 则是基于全体语料训练获得的恒定值,$tf\text{-}idf(t_k,d_j)$ 表示特征 t_k 在文档 d_j 中的 *tf-idf* 值;$tf(t_k,d_j)$ 表示特征 t_k 在文档 d_j 中出现的次数;N 表示总文本数;$n(t_k)$ 表示出现特征 t_k 的文档数. 对于文档长度不一致的问题,一般还需要对 *tf-idf* 值进行归一化处理,从而得到特征 t_k 在文档 d_j 中的权值 w_{kj},计算公式为

$$w_{kj} = \frac{tf\text{-}idf(t_s,d_j)}{\sqrt{\sum_{s=1}^{T}(tf\text{-}idf(t_s,d_j))^2}}$$

其中,T 表示一篇文本中的词数.

2 文本分类算法

2.1 支持向量机算法(SVM)

SVM(support vector machines)是 Vapnik 于1995年提出的一种新的统计学习方法[3-4],它建立在 VC 理论及结构风险最小化理论的基础上,主要基于以下3个方面:

1)基于结构风险最小化原则,利用 VC 维来降低机器学习的风险,从而提高其推广能力;

2)基于对有限样本信息的模型复杂性(即其对特定训练样本的学习精度)与学习能力(即准确地识别任意样本的能力)的考虑,并寻求这两者的最佳初衷,从而提高其泛化能力;

3)基于泛函中的 Mercer 定理,定义合适的内积函数(即核函数),通过非线性变换将样本空间映射到高维特征空间,并在其中寻求最优分类超平面.

2.2 K 近邻算法(KNN)

近邻法(简称 NN),是模式识别非参数法中的重要方法之一[5],它是一种典型的延迟学习方法[6]. NN 最大的特点是认为所有类中全部样本点都是代表点,故而在分类时需要计算所有训练样本与待识别样本 x 之间的距离,与 x 距离最近的训练样本所属类别即为分类结果.

KNN 是 NN 的推广算法,在分类时选出 x 的 K 个最近邻,其 K 个近邻中的多数所属类别就为待识别样本 x 的类别. 它是在已知类别的训练样本条件下,按最近距离原则对待分类数据进行分类. 该算法的基本思路是:在给定新文本后,考虑在训练文本集中与该新文本距离最近(最相似)的 K 篇文本,以这 K 篇文本所属的类别来判定新文本类别.

2.3 SVM-KNN分类算法

通过对SVM分类时错分样本的分布的分析得知：SVM分类器的出错样本点与其他的分类器一样，都分布在分界面附近[7-8]，欲提高其分类性能，就应该充分利用分界面附近的样本所蕴含的信息. 同时由SVM基本理论可知：分界面附近的样本基本是支持向量. 因此本文结合SVM和KNN，针对空间中样本的不同分布，对其采用不同的分类算法.

将SVM和KNN分类器进行结合，其出发点是可以将SVM看成是每类只取1个代表点的1NN分类器[9]，即SVM对每类支持向量只取1个样本作为代表，但因其分布不规律，所取的代表点有时并不能较好地代表该类，此时考虑到NN是将每类所有支持向量作为代表点，将SVM与KNN相结合，从而提高其分类准确率. SVM-KNN算法如图1所示.

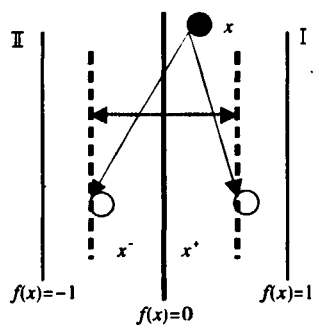

图1　SVM-KNN算法

具体过程为：对于待识别样本 x，首先计算其与2类支持向量代表点 x^+ 和 x^- 的距离差，并对其距离差做如下考虑：

1）当距离差小于给定的阈值，如图1中的区域Ⅰ，即 x 离分界面较近，此时若采用SVM算法，只计算 x 与1个代表点的距离，则容易出现错分现象；

2）若距离差大于给定的阈值，如图1中区域Ⅱ，即 x 离分界面较远，此时采用SVM算法一般不会出错.

因此，对于上述情况1）中当SVM算法错分时，采用KNN算法，将各类中的每个支持向量作为代表点，计算其与待识别样本的距离并对其做出判断. SVM-KNN分类算法具体步骤如下（首先采用任何一种SVM算法求出相应的支持向量和系数以及常数 b）：

步骤1：if $T \neq \varnothing$ take $x \in T$

else break；

步骤2：take $g(x) = \sum_i y_i \alpha_i k(x_i, x) + b$

步骤3：if $|g(x)| > \varepsilon$ take $f(x) = \mathrm{sgn}(g(x))$ as output；

else if $|g(x)| < \varepsilon$ adopt KNN algorithm as output；

步骤4：$T \leftarrow T - \{x\}$，go to 步骤1.

其中，设 T 为测试集，k 为KNN的个数，步骤3中的KNN算法采用与通常的KNN分类算法相同的流程，即将支持向量集作为分类算法的代表点集合. 本文在计算测试样本与每个支持向量的距离时并非采用常规的欧式距离公式，而是采用下式进行计算：

$$d(x_i, x_j) = \frac{\sum_{k=1}^{n} x_{ik} \times x_{jk}}{\sqrt{\left(\sum_{k=1}^{n} x_{ik}^2\right)\left(\sum_{k=1}^{n} x_{jk}^2\right)}}$$

该式为余弦公式，表示2个样本之间的夹角，值越小，表示2个样本的相似性越大. 算法中 ε 表示分类阈值，通常设为1左右，当 $\varepsilon = 0$ 时，该算法就为SVM算法.

3 实验结果与分析

3.1 实验数据来源

实验原始语料是从网上下载的书评，共538篇评论，其中正面评论273条，负面评论265条. 将2类文本随机分成4份，3份作训练集，1份作测试集. 对训练中的文本进行手工的正反面标注. 本实验的运行环境是Windows XP操作系统，所用的仿真软件是Matlab.

3.2 结果分析

文档分类中普遍使用的性能评估指标有召回率（Recall，简记为 r）、准确率（Precision，简记为 p）. 召回率和准确率的定义分别为 $r = \frac{a}{a+c}$ 和 $p = \frac{a}{a+b}$. 其中，a 为被正确分到该类的文本数，b 为被错误分到该类的文本数，c 为属于该类但被错误分到其他类别的文本数. 它们反映了分类质量的2个不同方面，故需对其综合考虑. 因此，引入一种新的评估指标：F_1-*measure*，其数学公式为 $F_1 = \frac{2 \times r \times p}{r + p}$.

用上述方法对语料进行处理得到文本的向量

空间表示.然后将其载入Matlab中,分别利用KNN,SVM及本文所提出的KSVM算法对其进行分类,实验结果如表1所示.

通过对实验结果分析可知,相对于KNN,SVM等传统的文本分类算法,该算法是一种性能较好的分类算法,能够明显提高分类的准确率和召回率,F_1-*measune* 最高可比传统算法提高0.04以上.

表1 3种算法分类结果

算法	评论	总数	判断为正面	判断为负面	召回率	准确率	F_1-*measure*
KNN算法	正面	273	247	26	0.905	0.915	0.910
	负面	265	23	242	0.913	0.903	0.908
SVM算法	正面	273	249	24	0.912	0.905	0.908
	负面	265	26	239	0.902	0.909	0.905
KSVM算法	正面	273	261	12	0.956	0.919	0.937
	负面	265	23	242	0.913	0.953	0.928

4 结论

人类情感作为心理行为的重要组成部分,是一个很复杂的心理现象,它包含各种各样的内容.本文通过文本预处理将情感分类转化为汉语的计算机理解,结合在文本分类领域中有着成熟应用的SVM算法和KNN算法,提出了一种基于KSVM的情感分类方法.

本研究还处在初步阶段,尚存在一些不足.针对本系统的错误分类现象,今后应该对语料进行进一步的丰富和检验,扩展特征,完善基准词.

参考文献:

[1] 朱嫣岚,闵锦,周雅倩,等.基于How-net的词汇语义倾向计算[J].中文信息学报,2006,20(1):14.

[2] 刘群,李素建.基于《知网》的词汇语义相似度计算[C]//第三届汉语词汇语义学研讨会论文集,台北:[s.n.],2002:59-76.

[3] 丁琼.基于向量空间模型的文本自动分类系统的研究与实现[D].上海:同济大学,2007.

[4] 李蓉,叶世伟,史忠植.SVM-KNN分类器——一种提高SVM分类精度的新方法[J].电子学报,2002,30(5):745.

[5] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000,26(1):32.

[6] Vapnik V N. The Nature of Statistical Learning Theory [M]. New York:Springer Verlag,1995.

[7] Li Rong,Ye Shiwei,Shi Zhongzhi. A effective classified algorithm of support vector machine with multi-representative points based on nearest neighbor principle[C]//Proc of Int Conf on Into-tech and Info-net, Beijing:[s. n.], 2001:113-119.

[8] Chin K K. Support Vector Machines applied to Speech Pattern Classification[D]. Cambridge:Cambridge University,1998.

[9] Vapnik V N. Estimation of dependencies based on empirical data[M]. Berlin:Springer Verlag,1982.