一种基于关联分析的 KNN 文本分类方法

范恒亮,成卫青
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:KNN 算法在数据挖掘的分支-文本分类中有重要的应用。在分析了传统 KNN 方法不足的基础上,提出了一种基于关联分析的 KNN 改进算法。该方法首先针对不同类别的训练文本提取每个类别的频繁特征集及其关联的文本,然后基于对各个类别文本的关联分析结果,为未知类别文本确定适当的近邻数 k,并在已知类别的训练文本中快速选取 k 个近邻,进而根据近邻的类别确定未知文本的类别。相比于基于传统 KNN 的文本分类方法,改进方法能够较好地确定 k 值,并能降低时间复杂度。实验结果表明,文中提出的基于改进 KNN 的文本分类方法提高了文本分类的效率和准确率。

关键词:数据挖掘;文本分类;KNN;关联分析
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2014)06-0071-04
doi:10.3969/j.issn.1673-629X.2014.06.018

An Improved KNN Approach of Text Classification Based on Association Analysis

FAN Heng-liang, CHENG Wei-qing
(School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: The KNN algorithm is largely applied in text classification, one branch of data mining. On the basis of analyzing the deficiencies of the traditional KNN method, an improved KNN algorithm based on association analysis is proposed in this paper. In this method, frequent feature sets for each class of training documents and associated documents should be extracted in advance. When a document with unknown class is to be classified, by the use of the results of association analysis, the number of nearest neighbors, k can be decided, k nearest neighbors can be found quickly from all classes of training documents, and the class of the document can be decided by the classes of its neighbors. Compared with the traditional KNN algorithm, this method has greatly improved in the selection of the best number of nearest neighbors. Moreover, it can also reduce the time complexity of the algorithm. The experimental results show that the proposed algorithm has better efficiency and accuracy in text classification.

Key words: data mining; text classification; KNN; association analysis

0 引 言

随着网络信息技术的飞速发展,Internet 的信息资源呈现指数级的增长趋势,而文本作为最基本的信息载体,其分类技术已经成为现代信息处理的一大热点。目前比较常用的文本分类算法有:朴素贝叶斯[1]、支持向量机[2]、神经网络[3]、决策数[4]、K-最近邻(K-Nearest Neighbor)[4]等方法。其中,基于经典 KNN 的文本分类方法简单有效,是分类效果最好的方法之一,但也有一些明显的缺点:

第一,确定待分类文本的类别时,需要计算其与训练样本集合中全部样本的相似度,之后从中选出与其相似度最高的前 k 个样本,一般情况,文本分类时的训练样本常常规模很大,由于在成千上万训练样本上计算与待分类文本的相似度,随着训练样本数目的增加,分类性能就会很快下降;

第二,它是一种懒惰的文本分类学习方法,在对测试样本分类时计算量大,消耗的时间较多,随着训练样本规模的增加,分类耗时急剧上升,造成分类时间是非线性的;

第三,KNN 算法必须指定 k 值,而如何确定待分类文本的近邻数目,尚缺乏较好且广泛适应的方法,k 的选取对类别判定起到很重要的作用,k 取得过大或过小都会降低文本分类的准确性。

文中将针对传统 KNN 文本分类方法存在时间复

收稿日期:2013-08-31 **修回日期**:2013-12-06 **网络出版时间**:2014-02-24
基金项目:国家自然科学基金资助项目(61170322,71171117);江苏省自然科学基金资助项目(BK2010524)
作者简介:范恒亮(1987-),男,安徽安庆人,硕士,研究方向为数据挖掘;成卫青,副教授,通讯作者,研究方向为网络测量和模式识别。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140224.0922.054.html

杂度高和 k 值难以确定等问题，提出一种基于关联统计分析的 KNN 改进算法，以提高文本分类的效率和准确率。

1 基于 KNN 的文本分类方法

传统 KNN 分类算法原理如图 1 所示，已知有三角形和矩形两个类别，待分类的对象是图 1 中的圆形，判别该对象是三角形还是矩形的过程为：首先计算被测对象与训练集合中的每个对象的相似度，降序排列从而得到被测对象的近邻；其次，如果是 $k=3$，由于三角形所占比例为 2/3，所以圆形对象被判定为三角形类别，如果 $k=5$，由于矩形比例为 3/5，因此圆形对象被赋予矩形类别。可见，传统 KNN 算法在执行过程中必须计算被测对象与每一个训练对象的相似性距离，且近邻数目 k 无法确定，从而影响 KNN 算法分类的准确性。

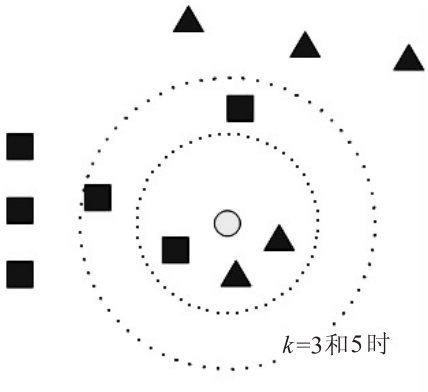

图 1 传统 KNN 算法

基于传统 KNN 方法的文本分类主要包括如下过程：

(1)文本的预处理[5]主要包括去除停用词、分词等。与英文不同，中文各个词之间没有固定的分隔符，因此对于中文文本而言，分词就成为一个必要的环节，且分词的准确性直接影响到文本分类的结果。

(2)经过预处理后，训练文本中所提取特征词数量仍然较大，且其中包含很多对分类效果没有影响的词。特征选取是从预处理选取的特征词中进行筛选，选取更为重要的特征词。主要的方法有：互信息、交叉熵、信息增益、χ^2统计方法(CHI-Square)、文本证据权等[6]。一般采用基于各个类别选取特征的方法。

(3)扫描并统计每一篇训练文本在特征空间中的向量，确定向量中每一维的权重，计算权重一般采用 TF-IDF[7]的方法。

(4)对于一篇待分类的文本样本，提取特征词并按照同样的方式计算其文本向量的各维权重，之后与每一篇训练文本计算相似度，一般采用余弦距离[8]的方法：

$$\mathrm{sim}(d_i,d_j)=\frac{\sum_{k=1}^{M} w_{ik} * w_{jk}}{\sqrt{\left(\sum_{k=1}^{M} w_{ik}^{\ 2}\right) * \left(\sum_{k=1}^{M} w_{jk}^{\ 2}\right)}} \tag{1}$$

其中，w_{ik}为文本向量 d_i的第 k 维属性权重；M 是文本特征向量的维度。

(5)按照文本相似度降序排列，选出与测试文本最相邻的 k 个训练文本。

(6)基于测试文本与其 k 个近邻的相似度以及 k 个近邻的类别，计算测试文本属于每一个类别的权重[9]：

$$\mu_j(X)=\sum_{i=1}^{k}\mu_j(X_i)\,\mathrm{sim}(X,X_i) \tag{2}$$

其中，$\mu_j(X_i)\in\{0,1\}$含义为文本 X_i是否属于 C_j；$\mathrm{sim}(X,X_i)$表示测试文本与训练文本的相似度。决策方法为：如果 $\mu_l(X)=\max\mu_j(X)$，则决策 $X\in C_l$，即类别权重最大的作为测试文本的所属类别。

2 基于关联分析改进的 KNN 文本分类算法

2.1 基本思路

文中提出利用关联分析[10]对基于 KNN 的文本分类方法进行改进。基本思路举例如图 2 所示，说明如下：设共有两个类别 A 和 B，4 个特征词 a_1,a_2,b_1,b_2；易知 4 个特征词可以产生 $C_4^1+C_4^2+C_4^3+C_4^4=15$ 种非空集合$\{a_1\},\{a_2\},\{b_1\},\{b_2\},\{a_1,a_2\},\cdots,\{a_1,a_2,b_1,b_2\}$，采用关联分析算法 Apriori 分别记录每种情况下包含该情况特征集合的文本对象；比如包含特征 a_1的对象具体有 5 个，包含特征词 a_1,a_2,b_1的对象有 3 个，含有特征词 a_1,a_2,b_1,b_2的对象有 2 个，这样可以建立特征词频繁项集；当需要对待分类文本(a_1,a_2,b_1)进行分类时，可直接查找出特征词频繁项集$\{a_1,a_2,b_1\}$所对应的文本，作为待分类文本的初始近邻，并确定待分类文本最终近邻数 k，再进行相似度的计算，根据前 k 个近邻的类别决定待分类文本的类别，例如图 2 中的圆形对象判定为矩形类别。

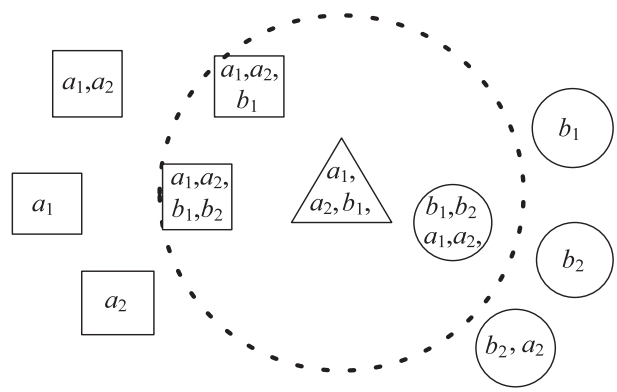

图 2 改进 KNN 方法的原理图

可见，改进算法相对于传统算法在执行分类时大

大降低了时间复杂度,且能够较好提高分类的准确性。不过在前期需要建立特征词频繁项集和关联的训练文本对象集合之间的映射,这需要一定的时空消耗。

2.2　算法描述

文中提出一种基于关联分析改进 KNN 的文本分类方法,该方法通过利用关联分析结果选取近邻,避免了传统 KNN 的不足。相对于传统方法,该方法在时间复杂度和 K 值选取方面有较好的改进。改进方法大致分为两个阶段:

(1) 基于关联分析提取频繁特征词集合及其关联的训练文本。

Step1:设文本类别总数为 m,类别为 $c_1,c_2,\cdots,c_m$,每个类别的训练样本数记为 $N_1,N_2,\cdots,N_m$;对训练集中的文本进行预处理,利用 χ^2 统计方法,对训练集中各类别文本分别选取一定数量,记为 N_f 的特征词(例如每个类别取 50 个特征);

Step2:扫描所有训练文本,将每个文本表示为由所有类别的特征词构成的 $m \cdot N_f$ 维文本向量,利用 TF-IDF 和基于 χ^2 统计方法的特征评价函数[7]计算特征权重,将权重设为:TF-IDF * 基于 χ^2 的特征评价值;

Step3:提取每个类别的频繁特征集及其关联的文本;本步仅考虑每个训练文本所属类别的特征,其余的暂且忽略;对每个类别分别处理,包括如下步骤:

Step3.1:将该类别的每个文本看作是单个事务(transaction),将其包含的该类别的特征看作是事务的数据项,项集也就是该类别的特征词集合,设置最小支持度,利用 Apriori 算法[11-13]得到文中该类别满足最小支持度阈值的所有项集,即产生所有文中该类别的频繁项集;

Step3.2:对每一个频繁项集保存其关联的训练文本,包含某频繁项集中所有特征的训练文本即为该频繁项集关联的训练文本。

(2)利用关联分析结果,确定待分类文本的初始近邻并确定最终的近邻数 k,再基于近邻类别进行文本分类。

Step1:对于待分类文本,先进行预处理,再利用已提取出的各类别的特征词表示该文本,得到 $m \cdot N_f$ 维文本向量,再利用 TF-IDF 和基于 χ^2 统计方法的特征评价函数计算特征权重,将权重设为:TF-IDF * 基于 χ^2 的特征评价值;

Step2:对待分类文本的文本向量中属于各个类别的特征词的权重分别求和并降序排列,选取排列在前 3 的类别,记为 c_x,c_y,c_z,及特征;

Step3:根据 Step2 获取的待分类文本的文本向量中属于前 3 个类别的特征词,分别在其对应的类别中查找最大频繁项集,并获取相关联的训练文本,这些训练文本都作为待分类文本的初始近邻;设相关联的训练文本集合分别为 I_x,I_y,I_z,文本数目分别为 n_x,n_y,n_z,设定 $k=\min(2.5*n_x,n_x+n_y+n_z)$;

Step4:计算待分类文本与每个初始近邻文本的余弦相似度;

Step5:将相似度降序排列,选取前 k 个训练文本,统计属于 3 个类别的文档数目,分类别累加相似度,进而得到待分类文本与每个类别近邻文本相似度的平均值,平均值最大的类别判定为待分类文本的类别。

传统 KNN 方法的权重计算一般采用 TF-IDF 方法,改进算法采用 TF-IDF * 基于 χ^2 的评价函数方法,增加了特征词权重因素的考虑,能够更加准确地给出文本向量每一维的权重。

3　实验验证与结果分析

3.1　实验设计

为对文中提出的改进算法进行有效性实验验证,实验中采用中科院计算技术研究所研制的汉语词法分词系统 ICTCLAS 进行分词;利用 Apriori 算法产生频繁项集与相关数据的关联信息,文本分类器使用 Visual C++开发;训练集和测试集采用复旦大学李荣陆博士收集的中文语料库,其类别和文本数量情况如表 1 所示,该训练样本库和测试样本库主要来自 BBS 论坛、门户网站新闻版面、相关的杂志和期刊;此外,中间一些数据的处理等工作使用 WEKA 等工具。

表 1　训练文本集和测试文本集分布情况

	医药	政治	教育	环境	经济
训练文本	135	330	140	130	218
测试文本	90	90	90	90	90

分类效果评估采用典型信息检索评价标准:查全率(recall)、查准率(precision)和 F_1 测试值[14],各指标描述如下:

$$\text{查全率}=\frac{\text{分类正确的文本数}}{\text{实际分类文本数}}$$

$$\text{查准率}=\frac{\text{分类正确的文本数}}{\text{应有文本数}}$$

$$F_1=\frac{2\times\text{查全率}\times\text{查准率}}{\text{查全率}+\text{查准率}}$$

3.2　结果分析

传统 KNN 算法和文中提出的改进算法的文本分类实验结果如图 3 所示。

从实验结果看,改进 KNN 算法的分类准确率总体要优于传统的 KNN 算法,并且在查全率方面也有一定的改善。观察传统 KNN 算法的实验结果可以发现医药和经济两个类别的查全率较低,政治的查准率较低,结合上面的数据信息可知,由于政治类别的训练文本

数量较其他类别的文本多,且医药和经济的类别特征与政治的类别特征有部分是相同的,从而造成将医药和经济类别的文本误分到了政治中。而改进的 KNN 算法由于对各个类别的文本向量进行了关联性分析,并且使用待分类文本与各类别近邻样本的相似度的平均值进行类别判别,从而有效地减少了文本类别误判情况的发生。

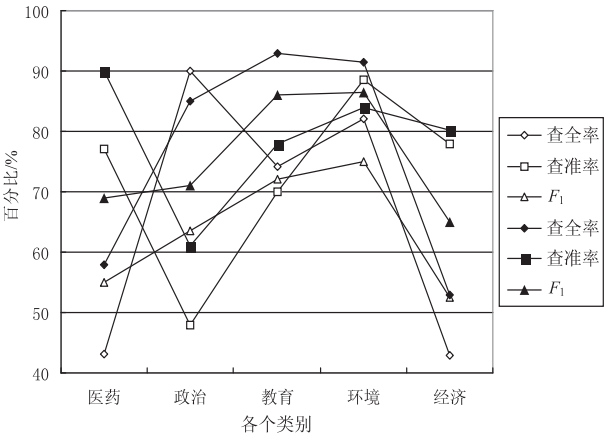

图 3 传统 KNN 方法和改进 KNN 方法的实验效果比较

上述实验表明,根据关联分析产生的特征词和训练文本之间的相关信息来选定测试文本的近邻是可行的;相对于传统的 KNN 方法在近邻的选择上能够大幅度减少计算量,一般为传统的方法时间复杂度的 1/3;同时通过频繁特征词集合查找测试文本在各类别训练样本中的近邻,不同类别中的近邻数对于确定测试文本的最近邻个数 k 具有重要参考价值。

实验结果显示,文中的近邻选择方法能够减少相似度不高的训练文本参与测试文本的类别判别,同时由于利用 Apriori 算法产生频繁项集设置了最小支持度,因此,文中的方法也不是一种对样本的局部特征非常敏感的方法,所以能够在一定程度上提高分类的准确率。

4 结束语

文中提出了一种基于关联分析的改进 KNN 文本分类方法,对近邻数量 k 的确定有较好的改进,同时能大大减少分类的时间复杂度。实验表明,相对于传统的方法,改进 KNN 的文本分类方法时间复杂度较低且分类准确率较高。

文中方法也存在不足:首先利用 Apriori 算法对各个类别的文本向量进行关联分析在时间和空间上都是不小的消耗。其次对于某个测试文本通过查找满足最小支持度的频繁项集的统计信息,不能完全准确地查找到与其邻近的所有文本向量,实验中一般只能找到测试文本 80% 左右的近邻。此外,如果某个类别的特征不是特别明显,会导致不能或只能提取较少的符合最小支持度的关联信息。下一步考虑利用 Apriori 的改进算法来提取关联信息,同时在保证较低时间复杂度的情况下寻找更好的方法来提高测试文本近邻的查全率。

参考文献:

[1] McCallum A, Nigam K. A comparison of event models for Naive Bayes text classification[C]//Proc of AAAI workshop on learning for text categorization. [s. l.]: AAAI Press, 1998: 41-48.

[2] 叶志刚. SVM 在文本分类中的应用[D]. 哈尔滨:哈尔滨工程大学,2006.

[3] Han J, Kamber M. 数据挖掘概念与技术[M]. 孟小峰,译. 北京:机械工业出版社,2005.

[4] 王 煜. 基于决策树和 K 最近邻算法的文本分类研究[D]. 天津:天津大学,2006.

[5] 薛翠方,郭炳炎. 汉语文本特征词的抽取方法[J]. 情报学报,2000,19(3):242-247.

[6] 王建会. 中文信息处理中若干关键问题的研究[D]. 上海:复旦大学,2004.

[7] 程 军. 基于统计的文本分类技术研究[D]. 北京:中国科学院,2003.

[8] 宋 玲,马 军,连 莉. 文档相似度综合计算研究[J]. 计算机工程与应用,2006,42(30):160-163.

[9] 吕震宇,赵 爽,林永民. KNN 在中文文本分类中的应用研究[J]. 计算机与现代化,2008(11):69-72.

[10] 刘城霞. 基于 MS 关联规则数据挖掘模型的应用与探讨[J]. 计算机技术与发展,2013,23(1):25-28.

[11] Han Jiawei, Pei Jian, Yin Yiwen. Mining frequent patterns without candidate generation[C]//Proceedings of the 2000 ACM SIGMOD international conference on management of data. New York, NY, USA: ACM, 2000: 1-12.

[12] Agrawal R, Imieliski T, Swami A. Mining association rules between sets of items in large database[C]//Proceedings of the ACM SIGMOD conference on management of data. New York, NY, USA: ACM, 1993: 207-216.

[13] Lu Hongjun, Feng Ling, Han Jiawei. Beyond intratransaction association analysis: mining multidimensional intertransaction association rules[J]. ACM Transactions on Information Systems, 2000, 18(4): 423-454.

[14] Yang Yiming, Pedersen J O. A comparative study on feature selection in text categorization[C]//Proc of 14th international conference on machine learning. San Francisco, CA, USA: Morgan Kaufmann Publisher Inc, 1997: 412-420.

一种基于关联分析的KNN文本分类方法

作者： 范恒亮， 成卫青， FAN Heng-liang， CHENG Wei-qing
作者单位： 南京邮电大学 计算机学院,江苏 南京,210003
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(6)