文章编号：2095-2163(2019)03-0269-04 中图分类号：TP311 文献标志码：A

K 近邻算法在政府采购数据挖掘中的研究与应用

王 宏，门 博，雷 娜
（西安石油大学 计算机学院，西安 710065）

摘 要：随着政府采购信息化水平的不断提升，政府部门在履行职责过程中沉淀了大量的政府采购数据。本文分析了政府采购中标信息的要素，在研究有监督机器学习方法和文本数据处理流程的基础上，选取 K 近邻分类算法将其应用于文本分类中，形成政府采购项目领域模型之后，再对各中标公司在各领域的出现情况进行分析，并研究在取不同 K 值情况下分类的准确率。

关键词：政府采购数据；文本数据处理；有监督的机器学习；K 近邻分类算法；领域模型

Research and application of K Nearest Neighbor algorithm in government procurement data

WANG Hong, MEN Bo, LEI Na
(School of Computer Science, Xi´an Shiyou University, Xi´an 710065, China)

【Abstract】 With the continuous improvement of the level of government procurement information, government departments have deposited a large amount of government procurement data in the course of performing their duties. This paper analyzes the elements of government procurement bid-winning information. On the basis of studying the supervised machine learning method and text data processing flow, the K-nearest neighbor classification algorithm is selected and applied to the text classification to form the government procurement project domain model. The successful bidders analyzes the occurrence of each field and studies the accuracy of classification under different K values.

【Key words】 government procurement data; text data processing; supervised machine learning; K-nearest neighbor classification algorithm; domain model

0 引 言

近几年，国家将“政府信息公开”提升为“政府数据开放”。数据的开放使得政府积累的数据可以更好地被利用和分析，也意味着，这些数据可以公开获得并可以进行研究。政府采购就是指国家各级政府为从事日常的政务活动或为了满足公共服务的目的，利用国家财政性资金和政府借款购买货物、工程和服务的行为，政府部门在长期的采购过程中沉淀了大量的各类数据，这些数据涉及到政府采购的各个方面，包括采购人、招标机构、中标人和采购项目信息等，这其中隐藏着各方面之间的千丝万缕的关联关系，只有通过挖掘才能得到有价值的信息[1]。

在数据挖掘中，分类是一种很重要的工具。本文将采用 K 近邻分类算法对政府采购的中标信息进行分析，以获取一些关联信息。

1 分类介绍

分类是一种有监督的学习方法，包括了对文本的分类。文本分类能根据预先定义的主题类别，按照一定规则将文档集合中未知类别的文本自动分为一个或几个类别的过程[2]。或者说能够根据已被分类的训练文本集，通过特征选择、特征提取等方法得到特征项，也可以通过训练得到文本分类器，然后以此分类器对待分类文本集进行文本分类[3]。目前，经典的分类算法有：贝叶斯分类法、决策树、K 近邻分类算法、支持向量机等[4]。对此可做研究阐述如下。

1.1 K 近邻分类算法

K 近邻分类算法是一个理论上比较成熟的分类算法，其核心思想是：如果一个样本在特征空间中的 K 个最相似的样本中的大部分属于某一类，则该样本也属于这一类，K 近邻算法的分类决策规则依据少数服从多数的思想。在类的决策上，将只依靠一个或几个近邻的点来判定待分类点的类别。

1.2 算法描述

(1)选定训练数据，选定的数据不能太多，也不宜过少。

作者简介：王 宏(1968-)，男，副教授，硕士生导师，中国计算机学会会员，主要研究方向：服务计算、系统集成、移动云计算。

收稿日期：2019-02-26

(2)计算待分类数据与各个训练数据之间的距离(采用欧式距离)。

(3)选取与该数据点距离最小的 K 个点, K 一般为奇数。

(4)统计这 K 个点所出现类别的频率。

(5)选取 K 个点中出现频率最高的类作为待分类数据的类。

K 近邻算法流程如图 1 所示。

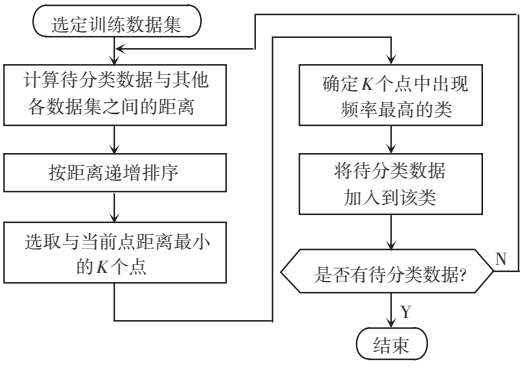

图 1 K 近邻算法流程图

Fig. 1 K-nearest algorithm flow chart

1.3 K 近邻算法的优缺点

K 近邻算法优点如下:

(1)算法易实现。

(2)对数据噪音有较强的忍耐能力。

(3)分类时只依赖于最相邻的 K 个点,因此只需要选择合适的 K 值。

K 近邻近算法缺点如下:

(1)K 近邻分类算法使用惰性学习方法,没有主动的学习。

(2)不同的 K 值会导致 K 近邻算法的精度有差距[5]。

2 政府采购数据处理

政府部门在履行职责过程中沉淀了大量的数据[6],通过网络爬虫已经可以在各省的政府采购网站上获取到公开的采购数据,包括中标公告(结果公告)等。如图 2 所示。图 2 中,描述的是已经获得的中标结果公告表的部分信息。其中,第一列代表着中标项目,第二列代表着中标公司。

program_name	hitted_orgaization_name
陕西省安康市8个县土地承包登记正射影像图制作项目	国家测绘地理信息局第一地形测量队◇
陕西省卫生宣传教育中心健康公益宣传片拍摄制作项目	恒圣（北京）文化传媒有限公司◇
公开选聘2015年-2017年陕西省政府债券信用评级机构采购	中债资信评估有限责任公司◇

图 2 中标公告表

Fig. 2 Winning bidding form

对中标信息分析的总体思路是:首先按领域对政府采购数据中标项目进行分类,分析出每个项目属于哪一个领域,得到分类结果,因为每条记录中包含了中标项目与中标公司,一个中标公司对应一个或多个项目,项目属于哪个领域则相当于该中标公司出现于哪个领域中。也就是通过对项目的领域属性的分类,以期获得各中标公司关注点主要放在那些领域的结论。

这里,将对中标项目进行分类,但 K 近邻算法只能对数值型数据进行处理,所以需要把文本数据转换成数值型数据。对中标项目名称进行文本处理,将文本数据转化为数值型数据,使其可以使用 K 近邻算法进行分类。具体的文本数据处理流程如图 3 所示。

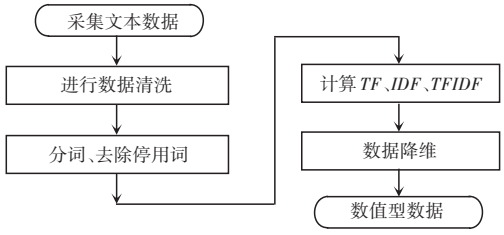

图 3 文本数据处理流程图

Fig. 3 Text data processing flow chart

对文本数据进行数据清理,主要是达到格式标准化与重复数据清除的目标。

接下来,对政府采购数据项目文本数据进行分词并移除停用词。采用 Python 语言编写程序,使用 jieba 分词对文档进行分词,针对实际的数据设置需要移除的停用词。如:“陕西省”、“咸阳市”、“西安市”等地区名,这些词在大部分数据中都会出现,并不能代表该文本。实际操作中对“陕西省安康市 8 个县土地承包登记正射影像图制作项目”与“陕西省卫生宣传教育中心健康公益宣传片拍摄制作项目”进行分词并移除停用词得到结果:[’土地 承包 登记 正 射影 像图 制作’],[’卫生 宣传教育 中心 健康 公益 宣传片 拍摄 制作’]。

对得到的分词结果分别按顺序计算 TF、IDF,最后得到 $TF-IDF$。其中,TF、IDF 以及 $TF-IDF$ 的含义可分述如下。

(1)词频 TF:在一份给定的文档里,词频指的是某一个给定的词语在该文档中出现的次数。这个属性是对词数的归一化,以防止其偏向长的文档[7]。对于在某一个文档里的词语来说,其重要性可用如下公式来表示:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \quad (1)$$

其中，$n_{i,j}$ 分子是该词在文档中的出现次数，而 $\sum_k n_{k,j}$ 则是在文档中所有字词的出现次数之和。

(2)逆向文档频率 *IDF*：逆向文档频率是一个词语普遍重要性的度量。某一特定词语的 *IDF*，可以由总文档数目除以包含该词语的文档数目，再将得到的商取对数，由此得到如下公式：

$$idf_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|}, \quad (2)$$

其中，$|D|$ 表示语料库中的文档总数，$|\{j:t_i \in d_j\}|$ 表示包含词语的文档数目，如果该词语不在语料库中，就会导致分母为零，因此一般情况下使用 $1 + |\{j:t_i \in d_j\}|$ 作为分母。

(3) *TF* - *IDF*[8] 权值：*TF* 与 *IDF* 的乘积。*TF* - *IDF* 是一种用于信息检索与数据挖掘的常用加权技术，是一种统计方法，用来确定一个词对于一个文本的重要程度。词的重要性随着该词在文本中的出现正比例增加，而随着其在语料库中的出现而下降。对此可做出如下公式表示：

$$w_{i,j} = tf_{i,j} * idf_i. \quad (3)$$

得到 *TFIDF* 之后，已经将文本数据转化成数值型数据，此时的数值数据有着高维度的特性，不利于计算和可视化展示。比如 K 近邻算法中存在着“维度灾难”的问题，即随着维度的增加，看似相近的 2 个点之间的距离却越来越大，将多维数据转化成二维或三维数据更利于数据的使用。主要的降维方法有：主成分分析法、线性判别分析法等降维方法[9]，本文选用主成分分析法进行数据降维。主成分分析(PDA)最初是在二十世纪初由 Karl Pearson 对非随机变量引入，又称简单 K-L 变换的理论与方法，这是一种对数据进行分析与处理的统计学算法，旨在利用降维的思想，将多指标转化成几个综合指标，其中每个综合指标都能反映原始指标的大部分信息，且所含信息互不重复[10]。

使用 Python 语言有一个好处就是会有很多可以方便使用的工具包，比如此处进行数据降维可以使用 sklearn 包下的 PCA(主成分分析法)工具模块，使用 PCA 对得到的 *TF* - *IDF* 数据进行降维，得到数据如：

[0.117 543 596 857 529,　0.010 702 118 710 152]，[0.176 951 967 884 651 2,0.008 137 108 014 717]。此时第一个数据代表着“陕西省卫生宣传教育中心健康公益宣传片拍摄制作项目”文本数据，第二个则代表“陕西省卫生宣传教育中心健康公益宣传片拍摄制作项目”。

3　政府采购数据应用分析

领域是对所属行业的一个划分，根据中国的行业分类并结合所获得政府采购数据实际的情况，本次研究将所有项目划分为 5 个领域、即：专业服务、电子电工、医药卫生、信息产业以及建筑建材。

3.1　领域分类

对降维后得到的数据，使用 K 近邻算法进行分类。首先 K 近邻分类算法需要一些训练数据集，从降维后的数据中选取训练数据集，添加上领域分类标签，对其可表述如下：

[0.207 055 796 653 806 83,0.058 263 573 135 836 28]专业服务，

[0.169 101 780 998 629 58,0.008 371 312 576 148 626]电子电工，

[0.133 684 676 453 191 18,0.024 987 790 298 132 31]医药卫生，

[0.259 447 879 333 501 83,0.010 690 718 265 212 179]信息产业，……

训练集中数值部分代表已分类的文本数据，后面则是按领域分类的标签。按照算法描述步骤计算待分类数据与各个训练数据之间的欧式距离，选取与该数据点距离最小的 *K* 个点；按少数服从多数原则，这 *K* 个点大部分属于信息工程领域类，则将该数据分类为信息工程领域，依次重复，直到数据分类结束。其中，*K* 的取值不同，分类效果也有差异。当取不同 *K* 值时，研究得到的分类准确率结果见表 1。

表 1　取不同 *K* 值分类准确率比较

Tab. 1　Comparison of classification accuracy of different *K* values

K 值	准确率/%
3	85
4	86
5	91
6	84

可以看出，选取 *K* 值为 5 准确率最高，并绘制出散点图，如图 4 所示。图 4 中，“+”点代表建筑建材领域，五角星代表医药卫生领域，三角形代表专业服务领域，六边形代表电子电工领域，奔驰形状代表信息产业领域。

3.2 中标公司在各领域出现情况分析

通过检索数据库,查询每一个中标公司所对应的项目,通过该项目的领域分类来判断中标公司所参与的领域。经由分析得出:医院只出现于医药卫生领域,电子科技类公司多数出现于信息产业领域,少部分则出现在电子电工领域。中标公司所出现的领域与公司自身的性质相同,比如保洁服务公司只出现于专业服务领域。但有些公司可能出现于 2 个领域,比如电子科技类公司。

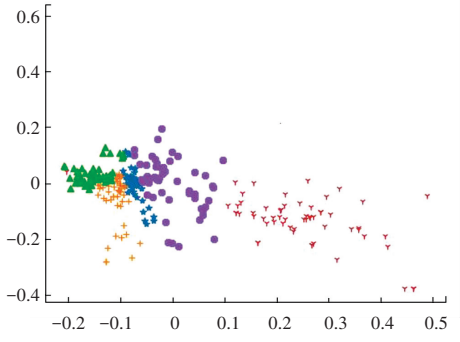

图 4 分类图

Fig. 4 Classification map

4 结束语

本文使用K近邻分类算法对政府采购数据进行项目领域的研究分类,通过不断地适配 K 值来得出较好分类结果,讨论了变量在算法调优中的重要作用。并且为今后针对政府采购数据各要素关联性分析中应用 K 近邻算法提供了途径。

参考文献

[1] 全姣. 政府采购资金使用数据挖掘研究[D]. 重庆:重庆理工大学,2011.
[2] 陆旭. 文本挖掘中若干关键问题研究[M]. 合肥:中国科学技术大学出版社, 2008.
[3] 王仁武. Python 与数据科学[M]. 上海:华东师范大学出版社, 2016.
[4] 李荣陆,胡运发. 基于密度的 kNN 文本分类器训练样本裁剪方法[J]. 计算机研究与发展,2004,41(4):539-545.
[5] 皮亚宸. K 近邻分类算法的应用研究[J]. 通讯世界,2019(1):286-287.
[6] 万如意. 大数据分析在政府采购领域中的应用:数据、技术与案例[J]. 中国政府采购,2015(12):52-56.
[7] 徐戈,王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报,2011,34(8):1423-1436.
[8] 朱晓霞,宋嘉欣,孟建芳. 基于主题—情感挖掘模型的微博评论情感分类研究[J/OL]. 情报理论与实践:1-11[2018-12-21]. https://kns.cnki.net/kcms/detail/11.1762.G3.20181219.1124.006.html.
[9] 吕皓,周晓纪. 基于主题模型的技术预见文本分析[J]. 情报探索,2018(10):52-59.
[10] 叶凌箭,王朗,马修水,等. 基于主元分析的最优状态检测技术[J]. 计算机与应用化学,2014,31(1):15-18.

(上接第 268 页)

大数据聚类算法。采用分段线性拟合方法进行用户行为特征大数据线性规划处理,提取用户行为特征大数据的互信息特征量,结合联合关联规则检测方法进行用户行为特征多维度文本数据的统计分析,构建大数据分布的关联属性样本集,采用联合半监督学习分类器进行数据分类,结合多传感量化跟踪识别方法进行聚类中心自动搜索,提高聚类收敛性。研究得知,采用本文方法进行用户行为特征多维度文本数据聚类处理的信息融合性能较好,数据聚类中心的自动搜索能力较强,提高了大数据分类检索能力,具有很好的应用价值。

参考文献

[1] 毕安琪, 董爱美, 王士同. 基于概率和代表点的数据流动态聚类算法[J]. 计算机研究与发展, 2016, 53(5): 1029-1042.
[2] 蒋芸,陈娜,明利特,等. 基于 Bagging 的概率神经网络集成分类算法[J]. 计算机科学,2013,40(5): 242-246.
[3] 孙力娟, 陈小东,韩崇,等. 一种新的数据流模糊聚类方法[J]. 电子与信息学报, 2015, 37(7): 1620-1625.
[4] 张红蕊,张永,于静雯. 云计算环境下基于朴素贝叶斯的数据分类[J]. 计算机应用与软件,2015,32(3):27-30.
[5] 梁聪刚,王鸿章. 微分进化算法的优化研究及其在聚类分析中的应用[J]. 现代电子技术,2016,39(13):103-107.
[6] 李昆仑, 关立伟, 郭昌隆. 基于聚类和改进共生演算法的云任务调度策略[J]. 计算机应用, 2018, 38(3): 707-714.
[7] 文政颖,李运娣. 语义指向性特征聚类的图像检索算法研究[J]. 计算机技术与发展,2017,27(4):83-88.
[8] 林楠,史苇杭. 基于多层空间模糊减法聚类算法的 Web 数据库安全索引[J]. 计算机科学,2014,41(10):216-219.
[9] 廖大强. 面向多目标的云计算资源调度算法[J]. 计算机系统应用, 2016, 25(2):180-189.
[10] 徐建. 用遗传算法评价部分股市常用技术指标的探索[J]. 智能计算机与应用,2018,8(5):158-160.