

一种聚类模式下基于密度的改进 KNN 算法

刘海峰, 庞秀梅, 张学仁

(解放军理工大学 理学院, 江苏 南京 210007)

摘 要: KNN 是基于实例的算法, 训练样本的数量影响 KNN 的分类性能. 合理的样本剪裁可以提高分类器的效率. 提出了一种聚类条件下基于密度的 KNN 改进模型. 首先使用聚类方法对训练集进行基于类别的选择, 裁剪边缘样本以减少噪音; 再基于类别密度对样本进行加权, 改善 k 近邻选择时大类别、高密度训练样本的占优现象. 试验结果表明, 本文提出的改进 KNN 分类算法提高了 KNN 的分类效率.

关键词: k-最近邻; 文本聚类; 样本密度; 文本分类

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2011)07-0125-03

An Improved Density—Based KNN Algorithm under Clustering

LIU Hai-feng, PANG Xiu-mei, ZHANG Xue-ren

(Institute of Sciences, PLA University of Science and Technology, Nanjing 210007, China)

Abstract: KNN is one of the arithmetic which based on the instance. The number of training samples influence on the classification performance of KNN. Reasonable sample cut can improve the efficiency of classification. This paper proposes an improved density—based KNN model under the clustering conditions. Firstly, basing on the types, we used the clustering method to choice the samples in order to reduce the noise samples. Then we weighted samples basing on sort density and overcame the defect that the big class, high density of training samples dominated in the KNN. The result of test shows that the improved KNN classification algorithm improves the efficiency of KNN classification.

Key words: KNN; text clustering; sample density; text categorization

1 引言

Web 是最大的信息资源库, 文本是最主要的信息载体. 信息检索的有效性和实时性对文本信息处理技术提出了越来越高的要求. 作为文本挖掘领域的重要研究内容, 文本自动分类技术在提高信息检索效率方面具有重要的现实意义和巨大的应用前景. 文本分类(text categorization, TC)是指在给定的类别体系下对未知类别文本根据其内容将其自动划归到一类或多类别的过程. 文本分类器的计算开销和泛化能力是评价其分类性能的重要指标. 作为文本分类中最常用的分类算法之一, KNN 分类器在准确率和召回率方面表现出众. 但是 KNN 的分

类模式同时决定了该算法的精度会受到样本分布状况的影响. 针对这一问题, 文中提出了一种对训练样本的两步骤处理方法, 首先对训练样本进行筛选, 使得清洗后的训练样本对相关类别更具有代表性; 然后引入权重因子对训练样本进行加权, 改善了 k 近邻算法对大类别、高密度样本倾斜现象, 提高了 KNN 算法的分类性能.

2 KNN 算法原理、问题及解决途径

作为一种基于实例的文本分类方法, KNN 算法被认为是向量空间模型(Vector Space Model)下最好的分类算法之一^[1]. 利用 KNN 算法对待分类文本进行分类的过程是:^[2]

收稿日期: 2010-10-25; **修回日期:** 2010-11-21

基金项目: 国家自然科学基金项目(71071161)

对于待分类文本 $d_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T$, 分类器在训练集里查找与 d_j 最相似的 k 个文本, 使用这 k 个文本的类别作为 d_j 的候选类别; 将 k 个文本和 d_j 的相似度作为 k 近邻文本所在类的类别权重, 把 d_j 归入相似度最大的类别中. 计算公式为:

$$p(d_j, c_i) = \sum_{d_s \in d_{jk\text{-round}}} \text{sim}(d_j, d_s) \delta(d_j, c_i) \quad (1)$$

其中, $p(d_j, c_i)$ 表示 d_j 与类别 c_i 的相似度, $i = 1, 2,$

$\dots, r$; $\delta(d_s, c_i) = \begin{cases} 1, & d_s \in c_i \\ 0, & d_s \notin c_i \end{cases}$ 为示性函数, $d_{jk\text{-round}}$ 表

示 d_j 的 k 近邻样本集, $\text{sim}(d_j, d_s)$ 为文本 d_j 与 d_s 的相似度, 使用常用的向量夹角余弦计算:

$$\text{sim}(d_j, d_s) = \frac{\sum_{i=1}^n x_{ij} x_{is}}{\sqrt{\sum_{i=1}^n x_{ij}^2 \sum_{i=1}^n x_{is}^2}} \quad (2)$$

类别决策函数为:

$$C_{d_j \text{ fall into}} = \arg \max_{c_i} (p(d_j, c_i)) \quad (3)$$

作为一种有监督机器学习非参数方法, KNN 分类器在基于统计的模式识别方面非常有效, 具有鲁棒性、概念清晰等优点, 对于未知和非正态分布数据能够得到较高的正确率, 在文本分类方面已经是最常用的分类器之一. 但是该算法具有一些明显的弱点, 主要体现在下述三个方面:

(1) KNN 是一种懒惰的学习方法, 计算过程均是在分类时才开始进行, 因此体现出训练时快但用于分类时慢的特点, 随着训练样本数的增加分类用时急剧上升;

(2) KNN 本质上是一种基于实例的算法, 在分类时需要计算待分类文本 d_j 与所有训练文本的相似度, 时间复杂度与样本数量成正比. 而文本分类时的训练样本常常是大样本, 随着训练样本数的增加分类性能下降很快^[3];

(3) 如果定义类 c_i 内所含有的训练样本数量占整个训练集样本数的比例 c_i 为内样本密度的话, 那么不论是基于距离还是基于向量夹角余弦计算文本之间的相似度, 样本密度大的类别在 k 近邻的选择时更易于被选中. 也就是说 KNN 分类器的分类效率受到训练样本分布情况的影响很大^[4], 但是在实际分类时数据分布倾斜却是普遍的现象. 即使是著名的英文文本分类语料 Reuters-21578, 最大类含有 2877 个文本, 最小的类仅含有几个文本, 而 82% 的类里文本数低于 100 个. 在这种训练条件下在进行文本分类训练时, 大类别样本占有密度上的优势从而更容易被选中, 势必影响分类的准确率.

解决前两个问题的关键是减少训练文本的数量. 但是训练样本太少必然会丢失太多的分类信息, 而训练样本太多又会加大计算开销, 因此选取一定数量的、具有最大类别表现能力的训练文本成为提高 KNN 分类能力的关键所在. 文中采取的对策是使用聚类方法对每个类别的训练文本进行删减, 去掉每个簇的边缘样本, 仅保留部分簇中心区域的样本作为训练样本.

解决第三个问题的方法在于引进权重系数. 加大密度较小类别训练样本权重, 提高它们与待分类文本之间的相似度, 同时减小密度较大类别的样本权重, 在 k 近邻选择时对大类别样本的作用加以抑制, 改善大类别样本引起的密度占优现象.

3 聚类模式下一一种基于密度的改进 KNN 分类算法

3.1 基于聚类的训练样本的类内删减

从判定待分类文本类别的作用方面讲, 对于类别 c_i 内的训练样本来说, 接近类中心的样本的作用当然优于类边缘的样本. 从这个角度看, 以类内样本的几何平均作为类中心, 删除在位置上处于类别边缘的部分样本即可达到样本裁剪的目的. 然而进一步考虑会发现, 即使是同一个类别的文本, 由于所表达的主题上的差异, 它们在这个类别区域内也应该属于不同的小区域, 通常情况下大类别下面也会有若干个小类别. 因此, 简单以几何位置上的类边缘作为标准进行噪音删除, 有可能错误地删除部分小类别的大部分甚至全部样本. 这是因为一些小类在位置上可能整体处于大类别的边缘位置. 这种方式会造成较大的信息丢失, 不能从整体角度反映出类内样本的分布上的差异. 因此, 应该选择出能代表 c_i 的不同主题、不同子类的最具有代表性的样本作为训练样本, 才能提高它们对 c_i 类的标引能力.

基于上述考虑, 文中的策略是对训练样本按照不同类别, 首先在各个类别内进行样本聚类, 使得属于不同主题、不同小类的样本聚集为不同的簇, 即对类内样本进行“簇”的分割, 使之反映出类内样本的总体分布状况; 然后在各个簇内删除与簇中心相似度小于给定阈值 ϵ 的边缘样本, 将各个簇的剩余样本作为 c_i 类的新的训练样本, 最后将各个类别剩余样本组成 KNN 算法的训练样本集, 这种处理方式可以在有效地去除分类噪音的同时, 最大限度地保留训练样本的分类信息.

3.2 基于类别密度的训练样本的 KNN 加权

KNN 算法受到训练样本的分布状况影响较大. 一般说来,不同类别的样本,类内样本数较多的类别的样本在 KNN 选择时容易被选中^[5]. 针对这个问题,考虑引入权重系数对高密度类别样本的重要性进行抑制,同时加大小类别样本的权重.

为此考虑在式(1)引入修正因子:

$$\lambda_i = \frac{\log(1 + \frac{N}{n_i})}{\sqrt{\sum_{i=1}^r \log^2(1 + \frac{N}{n_i})}} \quad (4)$$

这里 N 表示训练集样本数, n_i 表示类 c_i 内训练样本数. 显然 n_i 越大,权重 λ_i 越小. 这种处理方式可以使样本点分布较密集区域的样本权重减小,改善由于训练样本分布不均而对分类算法的影响.

此时式(1)改进为:

$$p(d_j, c_i) = \sum_{d_s \in d_{jk-\text{round}}} \lambda_i \times \text{sim}(d_j, d_s) \delta(d_j, c_i) \quad (5)$$

4 聚类模式下一一种基于类别加权的改进 KNN 分类算法

综上分析,改进的 KNN 文本分类算法如下:

(1) 使用向量空间模型将文本表示成向量形式,使用常用的 TF-IDF 模型^[6]进行特征赋权;

(2) 使用经典的 k-means 聚类方法^[7]对训练样

本进行聚类,文中实验取类簇数 $k = 5$;

(3) 对于各个簇内样本使用算术平均方法计算簇中心,使用式(2)计算簇内样本与簇中心相似度,将相似度小于阈值 ϵ (实验中取 $\epsilon = 0.15$) 的训练样本删除,剩余样本组成训练样本集;

(4) 使用式(2)计算训练文本与待分类文本 d_j 的相似度,使用式(5)计算 d_j 与类别 c_i 的相似度,其中取 KNN 算法中参数 $k = 25$;

(5) 根据式(3)确定待分类文本 d_j 的所属类别.

5 实验及其结果分析

文中对上述方法的分类效果进行了实验. 实验数据来自复旦大学计算机信息与技术系国际数据库中心. 从中选取经济、农业、体育、环境、政治、军事 6 类共计 2 318 篇文本. 试验时采用 4 折交叉验证方法,将这 2 318 篇文本分为 4 份,3 份为训练集,1 份为测试集;每份轮流作为测试集循环测试共 4 次,取平均值为测试结果. 分类效果评估指标使用常用的查准率、查全率与 F_1 测试值:

查准率 = 分类的正确文本数 / 实际分类文本数;
查全率 = 分类的正确文本数 / 应有文本数;

$$F_1 = (\text{查准率} \times \text{查全率}) \times 2 / (\text{查准率} + \text{查全率})$$

为了与经典的 KNN 分类算法的分类效果进行比较,在数据下面分别以 old、new 标注. 数据统计如表 1 所示.

表 1 文本分类测试结果比较

类 别	经济	农业	体育	环境	政治	军事	平均
查准率 _{new}	0.824 6	0.887 2	0.901 5	0.802 4	0.810 3	0.864 8	0.848 5
查准率 _{old}	0.786 1	0.821 9	0.832 7	0.781 3	0.802 7	0.825 1	0.808 3
查全率 _{new}	0.841 6	0.813 2	0.853 1	0.764 9	0.842 6	0.832 7	0.824 7
查全率 _{old}	0.817 2	0.778 2	0.801 1	0.743 8	0.825 1	0.816 1	0.796 9
F_1 值 _{new}	0.833 0	0.848 6	0.876 6	0.783 2	0.826 1	0.848 5	0.836 0
F_1 值 _{old}	0.801 3	0.799 5	0.816 6	0.762 1	0.813 7	0.820 6	0.802 3

可以看出,改进的 KNN 分类算法在农业、体育、军事类别的文本分类效果较好,最高 F_1 值为 0.876 6,环境与政治两类 F_1 值略低一点,但是最低也达到 0.783 2. 改进后 F_1 测试值平均提高约 3.5%.

6 结束语

文中在基于聚类的基础上对训练样本进行裁剪,去除了部分的噪音样本;在此基础上针对训练样本分布差异提出了一种基于密度的文本加权方法,

改善了 KNN 分类器的大类别样本占优现象.

文本分类的另一个瓶颈是文本特征的高维性问题. 特征降维技术是文本信息处理所必须面对的又一个主要问题. 目前在文本分类方面常用的分类器除了 KNN 方法以外,还有决策树、朴素贝叶斯分类器、支持向量机分类器等. 探讨在合理的特征降维模式下提高这些分类器的性能,使之能够适合 Web 文本的实时在线分类需求,还要进一步研究.

(下转第 131 页)

地地址绑定到这个端口上,而是直接使用 connect() 向指定的服务器发送连接请求,如果请求被接收,下一步就可以进行数据流通信了。TCP 通信是服务器端和客户机端通信的函数应用过程如图 4 所示。

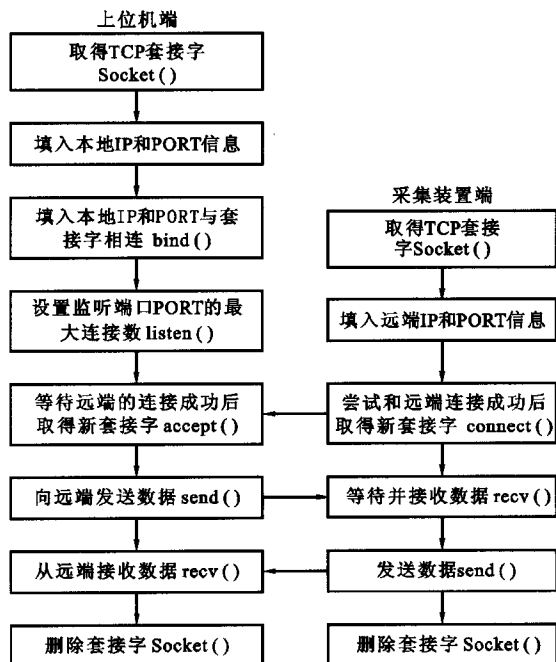


图4 上下位机通讯

上位机主要负责对所有转速的监测,通过 TCP 协议和采集装置进行网络通信,以保证网络数据的快速、可靠和准确。Qt 提供了 QTcp Socket 类和 QTcpServer 类用于编写 TCP 客户端和服务端应用程序。QTcpSocket 类采用异步工作方式,提供了 TCP 协议的通信接口;QTcpServer 类在服务器端处理外来的 TCP 客户端连接。

5 结束语

文中提出的测频法,保证在全段运行时即快速

又准确的测量出离心机转速。该测试装置样机达到了设计技术指标,满足离心机转速测试的要求。与其它方法相比具有较强的实用性和安全性。可广泛应用于核工业生产厂,具有极强的推广价值和广阔的应用前景。

参考文献:

- [1] samsung. S3C2440A 32-BIT RISC MICROPROCESSOR USER'S MANUAL. [EB/OL]. [2010-09-08]. <http://www.datasheetsite.com/datasheet/s3c2410x>.
- [2] DAVICOM. DM9000 Datasheet. [EB/OL]. [2010-09-08]. <http://www.datasheetarchive.com/om9000-datasheet.html>.
- [3] 孙琼. 嵌入式 Linux 应用程序开发详解[M]. 北京:人民邮电出版社, 2006: 145-51.
- [4] 孙天泽,袁元菊,张海峰. 嵌入式 Linux 驱动开发指南—基于 ARM9 处理器[M]. 北京:电子工业出版社, 2005.
- [5] 何轩,夏应清,李祥,等. 2005. 12 基于以太网的嵌入式 WEB 服务器的设计与实现. 微电子学与计算机.
- [6] S3C2440 DM9000 驱动移植[EB/OL]. [2010-09-08]. <http://blog.csdn.net/baozhongchao/archive/2009/09/10/4538182.aspx>.
- [7] Trolltech. Qt reference documentation [EB/OL]. [2010-03-13]. <http://qt.nokia.com/doc/4.5/index.html>.
- [8] 吴骏,肖志涛. 基于 S3C2410A 的嵌入式网络通信系统设计与实现[J]. 计算机应用研究, 2008, 25(6): 1897-1900.

作者简介:

丁莉女, (1972-), 硕士, 讲师. 研究方向为嵌入式系统开发和研究。

(上接第 127 页)

参考文献:

- [1] 苏金树,张博锋,徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报, 2006, 17(9): 1848-1859.
- [2] 一种基于数据偏斜的改进 KNN 文本分类[J]. 微电子学与计算机, 2010, 27(3): 51-53.
- [3] 印鉴,谭焕云. 基于统计量的 KNN 文本分类算法[J]. 小型微型计算机系统, 2007, 28(6): 1094-1097.
- [4] 张孝飞,黄河燕. 一种采用聚类技术改进的 KNN 文本分类方法[J]. 模式识别与人工智能, 2009(22) 6: 936-940.
- [5] 沈志斌,白清源. 基于加权修正的 KNN 文本分类算法

[J]. 计算机科学, 2008, 35(10A): 123-126.

- [6] Salton G, Buckley C. Term-weighting approaches in automatic retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523.
- [7] 范明,孟小峰. 数据挖掘概念与技术[M]. 北京:机械工业出版社, 2004.

作者简介:

刘海峰男, (1962-) 博士, 副教授, 计算机学会会员. 研究方向为模式识别、文本挖掘。

庞秀梅女, (1969-) 副教授. 研究方向为数据挖掘。

张学仁男, (1955-) 副教授. 研究方向数据挖掘。