

一种基于数据偏斜的改进 KNN 文本分类

刘海峰, 陈 琦, 刘守生, 苏 展
(解放军理工大学 理学院, 江苏 南京 210007)

摘 要: KNN 是一种简单、有效、非参数的分类算法. 针对样本分布偏斜的分类环境, 首先提出了一种改进的特征选择方法进行特征降维, 在此基础上进一步提出了一种基于分布的改进 KNN 方法用于文本分类, 降低了分布偏斜问题对决策函数的影响. 试验表明, 所提出的改进 KNN 文本分类方法具有较好的分类性能.

关键词: 特征选择; 文本分类; 改进 KNN; 相似度

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2010)03-0051-03

An Improved KNN Text Categorization Method Based on Data Uneven

LIU Hai-feng, CHEN Qi, LIU Shou-sheng, SU Zhan

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

Abstract: KNN is a simple, valid and non-parameter method often applied in categorization. Under the condition that the samples distribution is uneven, we first put forward an improved weighting way in feature reduction; then we improve the KNN basing on density in automatic text categorization. This way reduces the impact from the uneven distribution. we have a test about text categorization. The result shows that these methods have a better precision than the common KNN.

Key words: feature selection; text classification; improved KNN; similarity

1 引言

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利用的有效性和准确性上具有重要的现实意义和巨大的应用前景. 在向量模型下, 特征降维的效率以及分类器的性能是文本自动分类的两个核心技术. 作为一种简单、有效、非参数的文本分类算法, KNN 分类器在准确率和召回率方面表现出众, 成为文本分类中常用的分类器. 但是 KNN 的决策模式决定了该算法的精度会受到样本分布状况的影响. 文中针对这一问题, 在使用一种改进的互信息方法进行特征降维的基础上, 提出了一种基于分布的改进 KNN 方法用于文本分类并取得了较好的效果^[1-2].

2 一种改进的互信息特征选择方法

2.1 基于互信息的特征选择方法及特点

使用向量模型进行文本的结构化表示是目前比较成熟和有效的文本表示的主流方法. 向量模型下进行文本自动分类的最大困难是特征空间的高维性问题. 高维的特征向量带来的直接问题是计算机需要大量的存储空间和很高的计算能力, 并且容易引起过拟合现象. 合理降维不仅能大量降低处理开销, 而且在很多情况下可以改善分类器的效果. 因而寻找有效的特征降维方法降低特征空间的维数, 抑制大量的“噪音特征”对分类效率的影响, 成为基于统计的文本分类算法的核心技术, 各种有效的特征降维方法的研究成为基于机器学习的文本分类研究热点而得到了学者的持续关注.

目前, 在中文文本分类中, 常用的特征选择方法

主要有基于文本频率(document frequency, DF)方法;基于特征之间、特征与文本之间相关信息度量的互信息(multi-information, MI)、信息增益(information gain, IG)、 χ^2 统计(Chi-square)、期望交叉熵(expected cross entropy, ECE)等模型以及它们的变种.研究表明^[3]:DF 和 χ^2 分类性能较好,IG 对不同的语料分类波动较大,MI 效果最差;在特征项维数较小时,DF、 χ^2 性能优势效果明显.

互信息(MI)是统计学和信息论相结合的一个评估方法.它体现了特征项和文本相互关联程度的度量.互信息量越大,说明两者的关联程度越高,特征项对确定文本类属的作用越大.特征项 t_i 与文本类别 c_k 之间的互信息计算公式如下:

$$MI(t_i, c_k) = \log \frac{p(t_i, c_k)}{p(t_i)p(c_k)} = \log p(t_i | c_k) - \log p(t_i) \quad (1)$$

而特征 t_i 在文本集里相应的互信息为

$$MI(t_i) = \sum_{k=1}^l \log \frac{p(t_i, c_k)}{p(t_i)p(c_k)} \quad (2)$$

式中, $p(t_i, c_k)$ 为类别 c_k 中出现特征项 t_i 的文本数与训练集里的文本数之比, $p(t_i)$ 为训练集里出现 t_i 的文本数与训练集里的文本数之比; $p(t_i | c_k)$ 为类别 c_k 中出现 t_i 的文本数的比率.

该模型的显著优点是考虑了低频词所带有的信息量,但是该方法容易引起过度学习,这是由于该模型对低频词的词频过于敏感造成的.

2.2 基于互信息的特征选择方法问题及改进

互信息方法的概率依据是基于文本集里各个类别的文本数量的分布是大致相等这一前提的.但是当类别 c_i 和类别 c_j 所含的文本数差别很大,比如前者在数量上远远大于后者;此时即使特征项 t_r 与 c_i 的关联程度远远小于其与 c_j 的关联程度,但是通过式(2)的计算却反映出 t_r 与 c_i 的互信息优于其与 c_j 的互信息.这种不合理的现象是由于文本在文本集里不同类别之间的分布不均衡而造成的.

一种常见的现象是训练样本分布不均,从频数上看大类别的特征主要出现在该类别内而在其他类别出现的比例却很小.但是式(2)却是基于整个训练集的特征频数的度量,使得低频词的作用被放大,而类别特征的作用被抑制.为此,不仅应该考虑训练样本的类别分布状况,将类别因素考虑进去,而且应该考虑特征出现时训练样本的分布状况,为此,将式(2)改进为

$$MI(t_i) = \sum_{j=1}^l p(c_j) p(c_j | t_i) \log \frac{p(t_i, c_j)}{p(t_i)p(c_j)} \quad (3)$$

上式改善了由于训练样本的类别分布不均对互信息的影响.

3 数据偏斜条件下一种改进的KNN分类方法

KNN方法是一种简单有效的非参数方法,它是一种预测性的分类算法(有监督学习),并不需要产生额外的数据来描述规则,它的规则本身就是数据.在准确率和召回率方面表现出众成为文本分类中常用的分类器.其指导原则是根据待分类文本的 k 个最近邻样本来预测待分类文本的类别.

3.1 经典KNN算法特点与不足

作为一种基于实例的机器学习算法,KNN算法被认为是向量模型(Vector Space Model, VSM)理论下最好的分类算法之一^[4].在文本分类中KNN算法的具体步骤如下:

(1) 将待分类文本 d 表示成和训练文本构成的维数一致的特征向量;

(2) 根据距离函数计算待分类文本向量 d 和训练文本向量的相似度,可以使用两向量之间欧氏距离 $|d - d_i|$ 计算;选择与 d 相似度最大(距离最小)的 k 个文本作为 d 的 k 个最近邻,也可以使用向量夹角余弦计算;

(3) 根据 d 的 k 个最近邻依次计算文本类别 $c_1, c_2, \dots, c_r$ 相应的权重,计算公式为

$$p(d, c_j) = \sum_{i=1}^k \text{sim}(d_i, d) \delta(d_i, d) \quad (4)$$

式中, $\text{sim}(d_i, d)$ 表示文本 d_i 与文本 d 之间相似度,常用的向量夹角余弦计算;

$$\delta(d_i, d) = \begin{cases} 1, & d_i \in c_j \\ 0, & d_i \notin c_j \end{cases} \quad \text{为二值类属函数.}$$

(4) 将待分类文本 d 归入权重最大的类别.

作为一种有监督的机器学习的非参数方法,KNN分类器在文本分类方面效率高,使用简便.但是该算法的一个明显的弱点是其分类效率受到训练样本分布情况的影响较大,当数据分布倾斜现象严重时,其分类性能可能变得很差.相应的改进研究得到一些学者的关注^[5-6].

3.2 一种基于样本分布密度的改进KNN方法

实验发现KNN分类器对样本的分布较为敏感,特别是在分布不均匀的样本集上,KNN分类器分类

效果差异较大.这是因为在样本点分布稀疏的区域, k 个近邻点所组成的局部邻域显然要比在样本点分布比较密集的区域大,而密集区域文本之间的相似度要大于分布稀疏的区域里文本之间的相似度,这样对于待分类文本而言,由于其所处的训练样本区域密度上的差异,在为其选择 k 个最近邻文本时,分布密集的文本簇相比较分布稀疏的文本更容易被选中,从而使得密集区域文本在被选中的机会上占有优势.因此降低数据密集区域文本的相似度,加大数据稀疏区域文本之间的相似度,以此来消除由于训练文本在类别分布上的不均对文本之间相似性度量的影响可以改善 KNN 分类器的分类性能^[7].

由于距离小的两个训练样本之间的相似度大,反之距离大的训练样本之间的相似度较小,所以用下式作为修正因子:

$$\rho(c_j)=\frac{1-\text{sim}(\boldsymbol{d},\boldsymbol{d}_i))^2}{\frac{1}{|c_j|}\sum_{\boldsymbol{d}_j\in c_j}\frac{1}{|\boldsymbol{d}-\boldsymbol{d}_j|}}$$

对式(4)进行修正:

$$\hat{p}(\boldsymbol{d},c_j)=\sum_{i=1}^k\text{sim}(\boldsymbol{d}_i,\boldsymbol{d})\delta(\boldsymbol{d}_i,\boldsymbol{d})\times\rho(c_j)\tag{5}$$

在式(5)下,大类别训练样本与测试样本的相似度被降低,小类别训练样本与测试样本的相似度被放大.

测试文本相应的类别决策函数为

$$C_{d\text{ fall into}}=\arg\max\hat{p}(\boldsymbol{d},c_j)\tag{6}$$

这种新的类属计算方式可以使处于样本点分布

较密集区域的样本点之间的相似性降低,而使处于样本点分布较稀疏的区域的样本点之间的相似性增大,从而降低由样本点分布不均对样本之间相似性度量的影响.

4 实验及其结果分析

文中对上述方法的分类效果进行了试验.试验数据为网上随机下载的 4 195 篇文本,其中分为经济(520 篇)、房地产(416 篇)、体育(673 篇)、计算机(918 篇)、文艺(910 篇)以及军事(758 篇).试验时采用 4 分交叉试验法,将 4 195 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,取平均值为测试结果.使用东北大学自然语言处理实验室提供的分词软件进行分词,使用禁用词表过滤停用词、人工删除冷僻的特低频词、并剔除了虚词、助词、人称代词、特高频词等预处理后建立候选特征项集合得到特征个数 6 145 个,先用式(3)进行特征选择,选择权值最大的 275 个特征构造文本向量,使用我们提出的改进 KNN 分类器进行分类,决策函数为式(6).分类效果评估指标使用常用的查准率、查全率以及测试值:

其中查准率=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;

$$F_1=\frac{\text{查准率}\times\text{查全率}\times 2}{\text{查准率}+\text{查全率}}\tag{7}$$

试验数据结果统计如表 1 所示.

表 1 分类试验的 F_1 值结果统计

类 别	经济	房地产	体育	计算机	文艺	军事	平均值
传统 KNN F_1 值	0.774	0.796	0.821	0.839	0.798	0.784	0.802
改进 KNN F_1 值	0.837	0.819	0.887	0.872	0.857	0.846	0.853

从表 1 可以看出,改进的 KNN 分类算法在体育、计算机、文艺与军事类别的文本分类效果较好,经济与房地产两类准确率略低,可能是这两类共用词组相对多一些.另外从计算开销上两种算法在速度上差别很小,但是改进的算法在文本分类的效率方面其 F_1 值提高了 6.4%,还是较普通的算法有显著的改善.

5 结束语

特征降维的合理性和分类器的性能是文本分类方面两个核心问题,文中在特征选择和分类器改造两个方面进行了研究,提出了一种特征降维模式下的基于密度的 KNN 文本分类算法.特征降维问题

是文本处理所必须面对的主要问题之一,是制约提高文本分类效率的瓶颈;根据训练集训练文本的不同分布状况构造合适的分类器成为影响分类效果的因素,提高分类器在大样本集、样本分布不均等环境下的健壮性是文本分类领域需要研究的关键问题,探讨两者之间的合理结合方法是我们今后要进一步研究的工作.

参考文献:

[1] Kim H, Howland P, Park H. Dimension reduction in text classification with support vector machines[J]. Journal of Machine Learning Research, 2005, 6(1):37-53.

RoTS算法和 TSNM 算法^[5]分别在 0.65s 和 0.35s 左右得到了多媒体系统 MMS 的映射后的 NoC 通讯功耗 130.06 mW. 从实验结果可以看出, 文中提出的算法在 3s 左右下获得的 NoC 通讯功耗的最优值明显优于 RoTS 和 TSNM, 降低系统功耗接近 10% 左右.

(2) 算法搜索效率的评估. EPAM 算法, RoTS 算法和 TSNM 算法具有较快的收敛速度, 在算法执行 1s 以内获得最优解, 后两种算法随机产生一个新的初始解, 在解决组合优化问题的初期 NoC 通讯功耗值急剧下降, 收敛速度较快, 但降到某个值后, 容易陷入局部最优从而导致搜索停滞. 由于本文中预处理方法的有效性, 以及对于未映射节点通讯功耗的合理估计, 在 3s 左右获得问题的最优解, 时间复杂度在实验可接受的范围之内, 大大提高了对映射空间的搜索效率.

6 结束语

低功耗优先的 NoC 映射问题已逐渐成为 NoC 设计阶段研究的热点, 文中通过分析通讯功耗与通讯流量之间的数学关系模型, 利用分支限界算法降低了通讯流量, 该算法首先采用高效的贪心策略对映射过程做预处理, 将中间过程划分成几个候选图, 利用预处理得到的结果对映射空间做进一步深度优先搜索. 实验结果表明文中方法有效地降低了 NoC 的通讯流量和通讯功耗, 保证了片上服务质量, 为其他拓扑结构以及不规则 IP 的映射方法的拓展研究打下了基础, 我们下一步工作将针对此类情况展开研究.

参考文献:

- [1] Benini L, Micheli G D. Networks on chips: a new soc paradigm [J]. IEEE Computer, 2002, 35(1):70-78.
- [2] Hu J, Marculescu R. Energy-aware mapping for tile-based NoC architectures under performance constraints [C]// Proceedings of Asia South Pacific Design Automation Conference. Kitakyushu, 2003: 233-239.
- [3] Lei T, Kumar S. A two-step genetic algorithm for mapping task graphs to a network on chip architecture [C]// Proceedings of the Euromicro Symposium on Digital System Design, Belek. Antalya, 2003: 180-187.
- [4] Zhou Ganmin, Yin Yongsheng, Hu Yonghua, et al. NoC mapping based on ant colony optimization algorithm [J]. Computer Engineering and Applications, 2005, 41(18): 7-11.
- [5] Chang Zhengwei, Xie Xiaona, Sang Nan, et al. An improved tabu search algorithm for network-on-chip mapping [J]. Journal of Computer-Aided Design and Computer Graphics, 2008(2): 155-160.
- [6] 魏建军, 康继昌, 雷艳静. NoC 的平衡设计 [J]. 微电子学与计算机, 2007, 24(5): 54-57.
- [7] 常政威, 谢晓娜, 熊光泽. 求解二次分配问题的改进禁忌搜索算法 [J]. 微电子学与计算机, 2008, 25(2): 21-24.

作者简介:

欧阳一鸣 男, (1963-), 副教授. 研究方向为嵌入式系统综合与测试、人工智能和模式识别等.

张 岚 男, (1984-), 硕士研究生. 研究方向为嵌入式系统综合与测试.

梁华国 男, (1959-), 教授, 博士生导师, 中国计算机学会容错计算专业委员会委员. 研究方向为嵌入式系统综合与测试、数字系统设计自动化.

(上接第 53 页)

- [2] 刘海峰, 王元元, 姚泽清, 等. 文本分类中一种混合型特征降维方法 [J]. 计算机工程, 2009, 35(2): 194-196.
- [3] 牛强, 王志晓, 陈岱, 等. 基于支持向量机的 Web 文本分类算法 [J]. 微电子学与计算机, 2006, 23(9): 102-104.
- [4] 苏金树, 张博峰, 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.
- [5] 王煜, 王正欧, 白石. 用于文本分类的改进 KNN 算法 [J]. 中文信息学报, 2007, 21(3): 76-81.
- [6] 王煜, 白石, 王正欧. 用于 Web 文本分类的快速 KNN 算

法 [J]. 情报学报, 2007, 26(1): 60-64.

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、文本分类.

陈 琦 女, (1979-), 博士研究生, 讲师. 研究方向为决策分析.

刘守生 男, (1965-), 博士, 副教授. 研究方向为遗传算法、神经网络.

苏 展 女, (1982-), 硕士, 助教. 研究方向为机器学习.