ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.7, No.35, December 2011.

E-mail: jslt@cccc.net.cn
http://www.dnzs.net.cn
Tel:+86-551-5690963 5690964

基于聚类的KNN算法改进

樊东辉[1,2],王治和[1],陈建华[1],许虎寅[1]

(1.西北师范大学 数学与信息科学学院,甘肃 兰州 730070;2.驻马店职业技术学院,河南 驻马店 463000)

摘要:通过研究KNN算法,提出了一种利用训练集文本聚类结果改进KNN算法的方法,首先将训练集文本采用DBSCAN算法聚进行聚类,将训练集文本分为若干个簇,然后采用KNN算法对测试文档进行测试,最后用距离最近的n个簇中的若干训练集文本使用KNN算法对测试文本进行分类。实验表明,改进后的算法降低了计算量,提高了效率,同时对聚类结果有了一定的改进。

关键词:KNN算法;DBSCAN算法;训练集

中图分类号:TP391 文献标识码:A 文章编号:1009-3044(2011)35-9033-02

An Improved KNN Algorithm Based on Clustering

FAN Dong-hui[1,2], WANG Zhi-he[1], CHEN Jian-hua[1], XU Hu-yin[1]

(1.College of Mathematics and Information Science, Northwest Normal University, Lanzhou 730070, China; 2.Zhumadian Vocational and Technical College, Zhumadian 463000, China)

Abstract: By studying the KNN algorithm, I proposed the method a training set of text clustering results using KNN algorithm to improve, first the training set using text DBSCAN algorithm to cluster together, the text is divided into a number of training set clusters, then using KNN algorithm test document for testing, and finally with the nearest cluster of n number of training set in the text text using KNN algorithm to classify the test. Experiments show that the improved algorithm reduces the amount of computation, improve efficiency, while a certain improvement of clustering results.

Key words: feature selection; document frequency; word frequency

1 概述

文本自动分类是指对未知类别的文档进行自动处理,判断它所属类别。随着各种形式的文本文档以指数级的速度增长,有效的信息检索、内容管理等应用变得愈加重要和困难。文本自动分类作为一个有效的解决办法,已成为一项具有实用价值的关键技术。现如今已有诸多分类技术和方法被提出来,例如KNN算法(K-Nearest Neighbour)[1]、贝叶斯网络(Bayes Network)[2]、支持向量机(SVM)[3]等。

其中KNN算法简单、有效,计算时间和空间线性于训练集的规模被广泛采用。但由KNN算法具体步骤可以知道:KNN是非积极学习方法,基本上不学习;再者每个训练集样本需要与训练集中样本进行计算,计算量非常大;还有因为要与单个训练集样本进行计算,易受单个样本的影响[4]。针对其局限性,我们提出改进的KNN算法就是在训练集样本中先进行聚类,然后利用KNN算法计算测试集样本与训练集样本簇之间的距离,选用较近的n个簇,用这n个簇中的训练集样本和测试集样本再采用KNN算法来确定测试集样本的分类。

2 基于传统算法的改进

2.1 传统的kNN算法

对于测试集中每一个测试文本,都需要计算它与训练集中每个文本的距离,然后把距离排序找到离该测试文本最近的k个文本,根据测试文本与训练文本的距离来给该测试文档的候选类别按公式(1)评分。如果有属于同一个类别的,就将该类别中的文本的打分求和作为该类别的得分。最后,将得分排序,测试文本将被分配给得分最高的那个类别。

$$SCORE(c|x)=\sum_{d\in knn} sim(x,d)I(d,c) \quad (1)$$

x是一个测试集文本,c是训练集的类别,d是距离x最近的k个文本之一;

sim(x,d) 是文本x与文本d的相似度,这里指的是距离;

I(d,c) 是表示d是否属于类c,如果属于类c则为1,否则为0。

2.2 改进的IKNN算法

首先对训练集文本进行聚类,采用DBSCAN算法

算法过程如下:

第一步:如果文本对象P未被归入某个簇或标记为噪声,就检查它的指定半径邻域r,如果指定半径邻域内包含的对象数目大于

收稿日期:2011-11-06

作者简介:樊东辉(1977-),男,硕士研究生,主要研究方向为数据挖掘,聚类。

等于给定的值m,就建立新簇C,将p的指定半径领域r中所有点加入该簇C;

第二步:对C中所有尚未被处理(归入某个簇或标记为噪声)的对象q,检查它的指定半径邻域,如果该邻域内包含对象数目大于等于给定的值m,将该邻域中没有归入任何一个簇的对象加入C;

第三步:重复第二步,继续检查C中未被处理对象,直到没有新的对象加入当前簇C:

第四步:重复以上步骤,直到所有对象都被处理。

其中关键参数为作为密度计算的距离表示的半径,密集点所必需的在指定半径内拥有的最少的其他点的数目。通过这两个参数我们就可以计算在任何点周围的密度值。

这样,训练集中文本就聚为若干个类了。每个簇的类别由簇中多数文本类别而定。

然后结合KNN算法,计算测试集文本与训练集文本簇之间的距离,这样可以减少计算量和个别孤立点对测试集文本的影响。

具体算法:

第一步:对于任一个给定的测试集文本,计算与训练集中各个簇的距离,采用(2)式为测试集文本评分

$$SCORE(c|x)=\sum_{d\in knn} sim(x,t)I(t,c) \quad (2)$$

其中x是一个测试集文本,c是训练集的类别,t是距离x最近的k个文本簇之一。

sim(x,t) 是文本x与文本t簇的相似度,这里指的是距离;

I(t,c) 是表示t簇是否属于类c,如果属于类c则为1,否则为0;

第二步:根据评分结果进行排序,选取前k个簇。

第三步:从这些簇中选取n个与测试集文本最近的文本,按照(1)式评分,判定该测试集文本类别,回归到传统的KNN算法。

改进算法中有领域半径r,指定邻域内最小文本数m,选取簇类个数k,从k簇中选取距离最小的n个文本这几个参数。根据试验表明,这几个参数需要经过多次试验,得出较优取值范围。

3 实验过程及结果分析

实验语料库选用谭松波博士收集整理的TanCorpV1.0中文文本分类语料库[5]。实验选取其中的三个类别,依次是财经、教育、科技,共540篇文档,每个类别中文档数均为180篇。

为了评估特征选择算法的有效性,实验在数据集上使用十折交叉验证法对六个类的文档进行测试。十折交叉验证法是将数据集分成十份,轮流将其中9份作为训练集,1份作为测试集进行实验,每次实验都会得出相应的准确率,十次结果的准确率的平均值即为对算法精度的估计。

现在通用的对聚类效果进行评估指标是查准率、查全率和F1值,查准率和查全率的公式分别为

$$P=Nrr/(Nrr+Nwr)\quad R=Nrr/(Nrr+Nrw)$$

Nrr表示正确聚类的个数,Nrw原本位于该类,被错误分出去的个数,Nwr原本不在该类却错误分到该类的个数。准确率和查全率反映了两个不同的方面,两者应该综合考虑,F1值是综合考虑这两者的一个评价指标,其公式如下:F1=2PR/(P+R)。F1值组合了查准率、查全率的思想,需要两者都很高时,才能得到一个较高的F1值。F1值越高说明聚类的效果也越好。在本次试验中k取3,n取15。实验结果如表1。

表1

	R=2 M=10		R=5 M=10		R=5 M=5	
	准确率%	召回率%	准确率%	召回率%	准确率%	召回率%
财经	0.684	0.622	0.759	0.767	0.713	0.693
教育	0.702	0.651	0.772	0,728	0.726	0.654
科技	0.647	0.682	0.746	0,711	0.734	0.731

很明显r越大,m越小,密度越小,聚类的个数就少,类中包含离散点就多,类中心与类中样本偏离较大,类和类中元素不能得到统一。而r越小,m越大,噪声点就多,类中包含样本点个数明显偏少。故本次实验取r=5和m=10较为合适,现在同样令k取3,分别实验n的不同取值对结果的影响。可以看到,k如果取聚类个数时,那就直接退化为原始KNN算法。

表2

	原始KNN	改进后	原始KNN	改进后	原始KNN	改进后
F1值	10	n=10	15	n=15	20	n=20
财经	0.830	0.712	0.859	0.890	0.853	0.872
教育	0.842	0.854	0.862	0.893	0.836	0.861
科技	0.871	0.856	0.869	0.887	0.845	0.822
所用时间	87. 4s	42.1s	96.2s	50.3s	136.5s	70.8s

由表2可以看到,改进算法F1值略优于原始算法,但在时间上改进算法明显好于原始算法,在n取20时,因为n取值较大,直接使用KNN算法得到的最近点和采用改进后算法得到的最近点相差不大,故准确率和召回率差别不明显,个别类别甚至优于改进后算法,若n值继续增大,则两者变化趋于平缓。而在n取10时,因为距离测试文本近的样本点有可能不在所选择聚类中,因此原始KNN算法F1值可能会略好于改进的算法,但在时间度量上却远远落后于改进后的算法,若n继续减小,聚类结果对之影响继续变小,两个算法F1值变化也趋于平缓。

4 结论

与KNN算法相比,改进算法大幅减少了距离的计算量,削弱了测试样本的分类受单个训练样本的影响,在分类性能和效率上较

(下转第9037页)

```
OracleDataAdapter oda = new OracleDataAdapter();
oda.SelectCommand = cmd;
DataSet ds = new DataSet();
oda.Fill(ds, tablename);
//list表遍历
foreach (DataRow theRow in ds.Tables[tablename].Rows)
{  zdz_struct tempstruct = new zdz_struct();
tempstruct.obtid  = "59287";
tempstruct.areacode = "广州";
if (theRow["ddatetime"].ToString() != null && theRow["ddatetime"].ToString() != "")
{ tempstruct.ddatetime = Convert.ToString(theRow["ddatetime"]);}
if (theRow["temp"].ToString() != null && theRow["temp"].ToString() != "")
{ tempstruct.temp = Convert.ToDouble(theRow["temp"]); }
if (theRow["hourrf"].ToString() != null && theRow["hourrf"].ToString() != "")
{tempstruct.hourrf = Convert.ToDouble(theRow["hourrf"]);}
zdzlist.Add(tempstruct);}
```

2 结论

社会经济的发展、导致用户的需求越来越多样化，如何真正做到“及时、准确、高效、快捷”、如何更好的落实科学发展观，值得我们思考和探索。

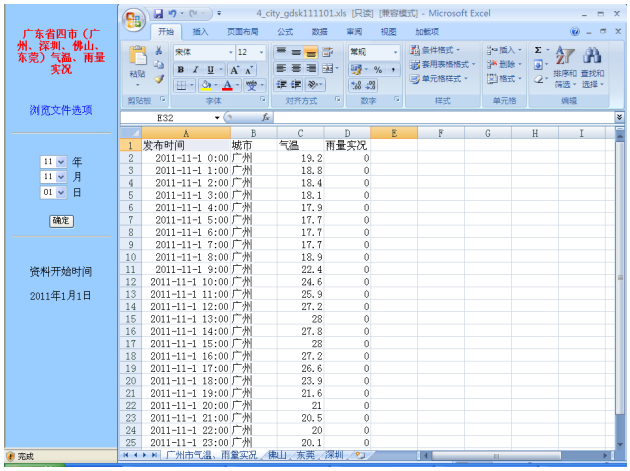

图3 系统效果图

参考文献：

[1] 杨长兴,刘卫国.C#程序设计[M].北京:中国铁道出版社,2008.
[2] 周存杰.C#网络编程实例教程[M].北京:北京希望电子出版社,2002.
[3] 罗斌,罗顺文.Visual C# 2005编程技巧大全[M].北京:中国水利水电出版社,2007.
[4] 章立民.Visual C# 2005程序开发与界面设计秘诀[M].北京:机械工业出版社,2006.

（上接第9034页）

KNN方法都要好。但改进算法受相关参数的影响,实验结果表明当这几个参数在一定的范围内选择时,能取得较好的分类性能,在效率提高的同时,性能受到了一定的影响,但与效率提高相比,这种影响还是可以接受的,下一步的工作是考虑如何使第一步的聚类结果能够更加有效,使之能够对第二步的消极影响降至较低。

参考文献：

[1] Joachims T.Text categorization with support vector machines:learning with many relevant features[C]//In the 10th European Conference on Machine Learning. NewYork:Springer:[s.n.],1998:137-142.
[2] Lewis D D.Naive(Bayes) at forty:the independence assumption in information retrieval[C]//In the 10th European Conference On Machine Learning.New York:Springer,1998:4-15.
[3] Yang Y,Liu X.A re-examination of text categorization methods[C]//In the 22nd Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval.New York:ACM Press,1999.
[4] 陈平,刘晓霞,李亚军.文本分类中改进型互信息特征选择的研究[J].微电子学与计算机,2008,25(6):194-197.
[5] 谭松波,王月粉.中文文本分类语料-TanCorpV1.0[EB/OL].http:// www.searchforum.org.cn/tansongbo/corpus.htm.