文章编号:1671-6833(2010)06-0116-03

基于多代表点的文本分类研究

陈可华

(宁德师范学院 计算机系,福建 宁德 352100)

摘 要:文本自动分类是一种有效的组织信息和管理信息的工具,传统分类方法一般在分类效果和运行效率上两者不可兼得,通过综合 Rocchio 和 KNN 2 种分类方法的优点,设计出一种基于多代表点的文本分类方法,该方法通过对各类挖掘出多个有效的代表点(真实或虚拟的),再使用基于这些代表点的 Rocchio 和 KNN 方法进行分类.实验表明,该方法以较少的训练时间达到令人满意的分类效果,并且能很好地解决不平衡类问题,实验结果显示,该方法能达到与 SVM 相当的分类效果.

关键词:文本分类;多代表点;Rocchio;KNN

中图分类号:TP391.1 **文献标识码**:A

0 引言

随着因特网等信息技术的发展,各类信息资源的增长都呈现海量特征,而其中文本数据始终占据重要地位.如何有效地组织、管理和使用这些文本信息成为当前的迫切需求,这促进了文本自动分类技术的迅速发展和广泛应用[1-3].

近年来,基于机器学习的文本自动分类技术研究取得了很大的进展,提出了很多有效的分类模型[3],如基于类中心的方法、Naïve Bayes 算法、支持向量机 SVM、KNN、神经网络、决策树、Boosting 等.

在这些分类方法中,基于类中心的 Rocchio 方法由于其简单快速而得到了广泛的应用,但是该方法的精度常常不能令人满意;KNN(k 近邻)也是一个常用的分类算法,并且在许多领域(简单情况和复杂情况)都显示出良好的性能,但是该方法的弱点就是它分类时的计算量较大.当它为一个未见实例分类时,它通常要遍历训练实例空间,以找到查询实例的 k 个最近的邻居.而其他大多数分类算法如决策树等都因为计算复杂度太高而不适用于大规模的场合,而且,当训练样本集增大时,都需要重新生成分类器,因而可扩展性差.

为了解决上述问题,提高分类效率但又不丢失分类的精度,笔者通过对文本分类的研究,结合基于类中心和 KNN 的优点,提出了一种新的简单有效的文本分类方法,通过给各类提取多个代表点,这些代表点可能是实际存在的文本也可能是虚拟的类中心,然后基于这些代表点对分类文档使用 Rocchio 和 KNN 等方法进行分类,从而能够利用较少的时间获得较高的分类精度,并且能很好的解决不平衡类分类问题.

1 相关研究

文本分类(也叫文档分类,本文中如不明确说明则是基于机器学习的文本自动分类)就是通过学习得到一个目标函数 f,能够将指定文档归入预先定义的几个类别中的一个或几个的过程[3].

当前对文本分类的研究主要是解决如下几个问题[2]:文本分类中的线性可行性问题、数据集偏斜(不平衡类)问题、标注瓶颈、多层分类、算法可扩展性等问题.文献[4]中对文本分类中常用分类器的可扩展性进行了分析.文献[5]使用外部辅助资源 Wikipedia 构建文本间的语义核来帮助文本分类.文献[6]通过对类中心描述的改进,设计了一个基于类中心的文本分类器.文献[7]旨在解决文本分类的时候无法多标签分类问题,通过构建一个元标签器自动检测文档应分配的标签数(类数).文献[8]通过对朴素贝叶斯方法为什么在大规模应用中无法有良好表现的分析,提出了两种改进的朴素贝叶斯分类方法.

收稿日期:2010-06-30;修订日期:2010-07-28

作者简介:陈可华(1974-)女,宁德师范学院计算机系讲师,主要从事数据库、数据挖掘方面的研究,E-mail:ckh2000985@sohu.com.

2　基于多代表点的文本分类模型

2.1　基于多代表点的文本分类模型

令 $C=\{C_i\}_{i=1}^{m}$ 表示预定义的类别集合,每个类中包含所属该类的文档集合. 每个类别包含若干个代表点,分别用 $T_{i,1},T_{i,2},\cdots,T_{i,j}$ 表示,$T_{i,j}$ 表示类别 i 中的第 j 个代表点,代表点挖掘方法在下一小节介绍. 则对于给定待分类文档 d,我们基于下述两种方法对其进行分类,记 $Sim(C_i,d)$ 为文档 d 与 C_i 的相似度,$Sim(T_{i,j},d)$ 为文档 d 与 C_i 中第 j 个代表点的相似度:

第 1 种方法通过计算文档与各类的相似度得到相似度最大的类标签,不过与 Rocchio 方法有所不同的是,此时类中心不只一个,而是由多个代表共同决定. 此时文档与类新相似度为通过如下公式计算:

$$NSim(C_i,d)=\sum_{j=1}^{\#T_i}(Sim(T_{i,j},d)\times W_{i,j}) \qquad (1)$$

式中:$W_{i,j}$ 表示第 j 个代表点在类 i 中的权重;$\#T_i$ 表示类 i 所包含的代表点数. 对于单标签分类问题,则返回最大的相似度的类标签;对于多标签分类问题,则返回相似度较大的若干个类标签.

第 2 种方法类似于 KNN 方法,不过此时待分类文档不是与训练实例表中的文档进行相似度计算,而是与所有代表点进行计算,其他过程与 KNN 类似,此处不赘述.

2.2　多代表点挖掘方法

我们使用 2 种方法进行代表点的挖掘. 基于聚类的方式首先对每个类别都进行聚类,聚成若干个簇,使用簇中心作为类的代表点,我们使用 K - Means[4] 算法进行文档聚类;基于文档密度的方式认为类中密度大的文档能代表该类的大部分性质,首先计算类中任意文档间的余弦相似度,然后通过如下公式计算每个文档的度.

$$D(D_i)=\sum_{j=1}^{\#C(d_i)}Sim(d_i,d_j) \qquad (2)$$

式中:$\#C(d_i)$ 表示文档 d_i 所属类别的文档数,最后返回度值较大的若干篇文档为该类的代表点.

3　实验与分析

3.1　文本分类数据集及评价指标

分别在以下 2 个数据集上验证我们模型的效果,中文数据集选用复旦文本分类语料库,英文数据集选用 20 - Newsgroups,保持训练集与测试集比例约为 7:3. 准确率、召回率和 $F1$ 值是评价分类性能的 2 种常用的指标[4]. 为了评估算法在整个数据集上的性能,使用以下 2 种平均的方法,分别称为宏平均和微平均. 宏平均是每一个类的性能指标的算术平均值,而微平均是每一个实例(文档)的性能指标的算术平均.

3.2　实验设计

使用 TR 表示使用 3.2 节中方法 1 的分类方法,TK 表示使用方法 2 的分类方法,并且在代表点挖掘方法上分别用 KM 和 DB 表示 K - Means 和基于密度(Density Based)2 种不同方法,所以通过组合有 4 种不同的分类方法,分别用 TR + KM,TR + DB,TK + KM 和 TK + DB 表示. 使用 Rocchio、KNN、朴素贝叶斯(NB)、SVM 作为基准模型进行比较.

3.3　参数选择

代表点个数的不同将会对分类结果产生很大的影响. 为了简化过程,设每个类别的代表点个数相同,分别验证代表点个数从 2 到 10 对分类结果的影响,由于方法 TK 还有参数 k 需要调整,则此处省略. 实验数据随机选择 20 - Newsgroups 中的 3 类和 4 类,限于篇幅只给出 3 类数据(分别是 Hardware,Forsale 和 Mideast)的准确率结果,其他未给出,但结果与此类似. 从图 1 可看出,每类 5 个代表点分类结果即达到很好效果,并且随着代表点个数的增加,分类准备率不增反而有稍许递减,笔者认为 5 个左右的代表点就能够很好地表示该类的性质,过多反而会带来噪音. 并且可以看到基于虚拟代表点的 KM 方法效果大大好于基于真实代表点的 DB 方法,这是因为 DB 方法在选择代表点后就丢失了该类其他文档的信息,而 KM 是通过综合簇中所有点形成质心,所以不会丢失过多关键信息.

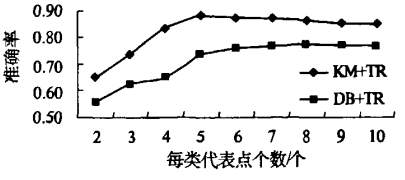

图 1　代表点个数对应的分类准确率

Fig. 1　Classification precision with the change of representative points number

3.4　结果及分析

使用 KM 方法挖掘代表点并设置每类代表点个数为 5,对于 TK 方法分别选择了 k 为 3 和 5,并设置基准模型 KNN 中 k 分别为 5,15 和 30. SVM 使用 libSVM[9] 包并设置所有参数为默认值. 评价指标中准确率、召回率和 $F1$ 值都为宏平均下的

准确率,召回率和 $F1$ 值. 表 1 和表 2 分别为这些方法在 20Newsgroups 和复旦数据集上的分类结果,表中分别用标识 Rocc,K,NB 表示 Rocchio,KNN,朴素贝叶斯方法. 从分类结果可以看出,该方法能表现出很好的性能. 对于 2 种分类方法 TR 和 TK,发现还是基于 KNN 的改进方法(TK)效果更好,这也从另一方面说明了 KNN 方法比 Rocchio 方法在分类效果上更好. 由于每类只有 5 个代表点(共 5×20 个代表点),对于 TK 方法的 k 值分别选择 3 和 5,认为让 k 过大对分类效果不会有太大提高,反而会因为代表点的增加给分类器带来更多干扰因素. 另外,从宏平均(大小类对宏平均贡献相同)结果可以看出,对于平衡类问题(20Newsgroups),笔者所提出方法最好表现与 SVM 相当;而对于不平衡类问题(复旦数据集),该方法表现更优,原因是传统方法在遇到不平衡类问题时会偏向大类别,使得分类效果会急剧下降,该方法通过对各个类提取相同数量的代表点,消除了大类会淹没小类的影响.

表 1 算法在 20 Newsgroups 上的分类结果

Tab.1 Classification results in 20 Newsgroups Dataset

评价指标	Rocc	K(5)	K(15)	K(30)	KM+TR	KM+TK(3)	KM+TK(5)	NB	SVM
微平均	0.479 2	0.802 7	0.837 5	0.838 1	0.738 1	0.908 1	0.909 3	0.826 7	0.907 2
准确率	0.429 6	0.806 2	0.837 3	0.838 8	0.841 7	0.898 8	0.901 2	0.857 3	0.904 2
召回率	0.487 9	0.827 3	0.839 2	0.839 0	0.782 9	0.919 0	0.919 1	0..842 7	0.914 5
$F1$	0.456 5	0.815 6	0.838 5	0.838 8	0.834 9	0.908 8	0.910 8	0.849 6	0.911 3

表 2 算法在复旦数据集上的分类结果

Tab.2 Classification results in FuDan Dataset

评价指标	Rocco	K(5)	K(15)	K(30)	KM+TR	KM+TK(3)	KM+TK(5)	NB	SVM
微平均	0.527 1	0.682 0	0.758 3	0.759 2	0.748 2	0.895 2	0.903 1	0.763 2	0.892 2
准确率	0.423 6	0.613 2	0.689 2	0.688 7	0.678 5	0.873 9	0.888 2	0.725 6	0.819 3
召回率	0.482 8	0.685 5	0.783 4	0.797 3	0.797 3	0.907 5	0.908 3	0.793 6	0.796 7
$F1$	0.463 9	0.662 7	0.745 9	0.762 8	0.736 3	0.889 2	0.896 7	0.751 9	0.805 1

4 结论

笔者通过对文本自动分类技术的研究,结合基于 Rocchio 方法和 KNN 方法的优点,提出了一种新的简单有效的文本分类方法,通过基于聚类和基于密度 2 种方法提取类的多个代表点,这些代表点可能是实际存在的文本,也可能是虚拟的类中心,并对这些代表点使用 Rocchio 和 KNN 的方法进行分类. 从而能够利用较少的时间获得较高的分类精度,并且能很好的解决不平衡类分类问题.

参考文献:

[1] SEBASTIANI F. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002,34(1):1-47.

[2] 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.

[3] 范明, 范宏建. 数据挖掘导论[M]. 北京:人民邮电出版社. 2006.

[4] YANG Y, ZHANG J, KISIEL B. A scalability analysis of classifiers in text categorization[C]// Proc. of the 26th ACM Int'l Conf. on Research and Development in Information Retrieval (SIGIR-03). Toronto: ACM Press, 2003:96-103.

[5] WANG P DOMENICONI C. Building Semantic Kernels for text classification using Wikipedia[C]// In Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, NY,2008:713-721.

[6] GUAN H, ZHOU J, Guo M. A class-feature-centroid classifier for text categorization[C]// Proceedings of the 18th international conference on World wide web. 2009:201-210.

[7] TANG L, RAJAN S, NARAYANAN V K. Large scale multi-label classification via MetaLabeler[C]// Proceedings of the 18th international conference on World Wide Web, 2009:211-220.

[8] ZHANG C, XUE GR, YU Y, et al. Web-scale classification with naive bayes[C]// In Proceedings of the

(下转第 125 页)

Structure Design of Multi – founctional Tents

MIAO Jian

(The Center of Henan Education Services, Zhengzhou 450002, China)

Abstract: Based on ergonomic principle and hiking rule, the multi – functional tent is designed with professional backpack, the mountaineering pack as an effective device, integrated with high – strength and light aluminium alloy frame, and the design of extensible, rotatable and foldable casing frames makes the supportive structure alternate and integratedly achieves its multiple functions, such as backpack, belt bag, fishing (raining) gear bag, baby frame carrier, fourgon, pram, folding seat, table, sunshade, mosquito net, tent, etc. Additionally it is equipped with anti – depression cushion and enlarged wheels for resting, pulling and travelling under slopes and various road conditions. The invention of multi – functional tent has extended the use of single – purposed traditional tent, serves the needs of three – people family to go out for shopping, outdoor recreation and it is also recommended to be used for emergency such as earthquake.

Key words: multi – functional tent; frame support; structure design

(上接第 118 页)

18th international conference on World Wide Web, 2009:1083 – 1084.

[9] CHANG C C and LIN C J, LIBSVM:a library for support vector machines [EB/OL]. Software available 2001. http://www.csie.ntu.edu.tw/ ~ cjlin/libs.

Research on Text Classification Based on Multiple Representative Points

CHEN Ke – hua

(Ningde Teacher College, Computer Department, Ningde,352100, China)

Abstract: Text classification is an effective tool of organization and management for information. Traditional classification methods are not good both in the effectiveness and in efficiency. This paper designed a method of classification based on multiple representative points, firstly mining a number of effective representative points to every category, and it can be true document or virtual point, then the methods of Rocchio and KNN can be working based on those points. Experiment results show that this classification method can achieve satisfactory results in less training time, and it can solve imbalance problem well, the results show that the method can achieve significant results similar to SVM.

Key words: text classification; multiple representative points; rocchio; KNN