一种文本特征选择方法的研究

陈素萍[1,2],谢丽聪[1]
(1.福州大学 数学与计算机科学学院,福建 福州 350002;
2.福建师范大学 协和学院,福建 福州 350007)

摘 要:在文本分类中,对高维的特征集进行降维是非常重要的,不但可以提高分类精度和效率,也可以找出富含信息的特征子集。而特征选择是有效降低特征向量维数的一种方法。目前常用的一些特征选择算法仅仅考虑了特征词与类别间的相关性,而忽略了特征词与特征词之间的相关性,从而存在特征冗余,影响了分类效果。为此,在分析了常用的一些特征选择算法之后,提出了一种基于mRMR模型的特征选择方法。实验表明,该特征选择方法有助于提高分类性能。

关键词:文本特征;文本分类;特征选择
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)02-0112-04

Research on Document Feature Selection

CHEN Su-ping[1,2], XIE Li-cong[1]
(1. College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350002, China;
2. Concord University College, Fujian Normal University, Fuzhou 350007, China)

Abstract: In text classification, it is very important to reduce the high-dimension of the feature vectors. It not only can improve the accuracy and efficiency of classification, but also can discover imformative feature subset. Feature selection is a valid method to reduce the dimention of feature vectors. Widely used feature selection algorithms have just considered the relationships between certain feature and a class, but neglected the relationships between features, then feature redundancy exists and it has an bad impact on text classification. In this paper, propose a feature selection method based on the minimum redundancy- maximum relevance framework. The experimental results show that this method is advantageous in improving the performance of text classification.

Key words: document feature; text classification; feature selection

0 引 言

随着计算机网络技术的发展,网络数据规模不断膨胀,如何从海量信息中搜索和管理这些资源成为一个重要的问题。近年来文本自动分类得到了广泛的关注和发展。在文本分类中,广泛使用Salton等人提出的向量空间模型(Vector Space Model)[1]来标引文本。然而,文本分类领域遇到的一个很大的挑战是高维的特征空间。对于大多数学习算法来说,不允许这样高的特征维数,并且有很多特征与分类无关,甚至有些误导分类的噪声数据。所以,如何进行特征集的降维操作是文本分类研究的一个关键问题。停词列表[2]和词干提取[3]成为该问题上的早期解决方法。

在文本分类中,目前常用的特征选择方法主要是基于统计理论和机器学习方法,比较著名的有文档频率方法、信息增益、期望交叉熵、互信息方法、X2统计量等,这些方法具有计算复杂度低,速度快的优点。

在特征选择方面美国卡内基梅隆大学的Yang Yiming教授的文章[4]较具代表性和总结性。在分析已有的几种特征选择方法的基础上,提出了一种新的特征选择评估函数。

1 相关研究

目前常用的文本特征选择算法的基本思想都是对特征集的每个特征进行评估。这样每个特征都获得了一个评估分,然后对所有的特征按照其评估分的大小进行排序,选取一定数目的最佳特征作为结果的特征子集。文中用 W 表示特征词,用 C_i 表示类别。

收稿日期:2008-06-25
基金项目:福建省B类科技发展基金(JB06023);中科院软件所开放课题基金(SYSKF0701);福州大学科技发展基金(2005-XQ-13, XRC-0511)
作者简介:陈素萍(1983-),女,福建福州人,硕士研究生,研究方向为特征选取、文本挖掘;谢丽聪,副教授,研究方向为数据挖掘、数据集成。

1.1 信息增益

在文本分类中,特征 W 的信息增益(Information Gain, IG) 如式(1) 所示。

$$IG(W) = P(W)\sum_t P(C_t \mid W)\log\frac{P(C_t \mid W)}{P(C_t)} + P(\overline{W})\sum_t P(C_t \mid \overline{W})\log\frac{P(C_t \mid \overline{W})}{P(C_t)} \quad (1)$$

式(1) 中,$P(W)$ 表示特征词 W 在文本集中出现的概率;$P(\overline{W})$ 表示特征词 W 在文本集中不出现的概率;$P(C_t \mid W)$ 表示在出现特征词 W 的情况下文本属于类 C_t 的概率;$P(C_t \mid \overline{W})$ 表示在不出现特征词 W 的情况下文本属于类 C_t 的概率。特征在文本中是否出现都将为文本分类提供信息,计算不同情况下的条件概率以确定提供的信息量的大小。信息增益是机器学习领域中使用较为广泛的特征选择方法。

1.2 期望交叉熵

期望交叉熵(Expected Cross Entropy, ECE)不考虑特征词未出现的情况。特征 W 的期望交叉熵如式(2) 所示。

$$ECE(W) = P(W)\sum_t P(C_t \mid W)\log\frac{P(C_t \mid W)}{P(C_t)} \quad (2)$$

期望交叉熵反映了文本类别的概率分布,以及在出现某种特征词的情况下文本类别概率分布之间的距离。

1.3 互信息

在文本分类中,特征 W 的互信息(Mutual Information, MI)如式(3)所示。

$$MI(W) = \sum_t P(C_t)\log\frac{P(W \mid C_t)}{P(W)} \quad (3)$$

互信息衡量的是某个特征词和类别之间的统计独立关系。用MI选择特征时,应该选择互信息大的特征。由于 MI 有利于低频特征,因此容易引起过学习(Over - fitting)。

1.4 X^2 统计量

在文本分类中,X^2 统计量(Chi - square, Chi)Chi用于衡量一个特征词和一个类别之间的统计独立关系。特征 W 的 Chi 值如式(4) 所示。

$$Chi(W, C_t) = \frac{N[P(W, C_t) * P(\overline{W}, \overline{C_t}) - P(W, \overline{C_t}) * p(\overline{W}, C_t)]^2}{P(W) * P(C_t) * P(\overline{W}) * P(\overline{C_t})} \quad (4)$$

式(2) 中,N 为文本集的总个数,$P(W, C_t)$ 为文本集中出现特征 W 并且属于类 C_t 的文本数除以 N,$P(W, \overline{C_t})$ 为文本集中出现特征 W 并且不属于类 C_t 的文本数除以 N,$P(\overline{W}, \overline{C_t})$ 为文本集中不出现特征 W 并且不属于类 C_t 的文本数除以 N,$P(C_t)$ 为类 C_t 的文本数除以 N,$P(\overline{C_t})$ 为非类 C_t 的文本数除以 N。特征 W 的全局 Chi 值如式(5) 所示。

$$Chi(W) = \sum_{i=1}^{m} Chi(W, C_i) \quad (5)$$

总体上,基于阈值的统计方法具有计算代价小,效率高的优点,尤其适合做文本分类中的特征选择,文献[4]中分析和比较了 DF, IG, MI 和 CHI 等 5 种方法,结合 LLSF 和 KNN 分类器,得出 IG 和 CHI 方法效果相对较好的结论。

然而,上面的这些评估函数考虑的是特征词和类别间的相关性,这样计算后选出 m 个最好的特征不一定是最好的 m 个特征。这些特征词之间会存在冗余,比如有个已被选择的特征词的评估值很大,那么和该特征词相关性比较大的某些特征词也很可能被选择,这样特征词间的冗余问题就出现了。所以,在选择和类别相关性比较大的特征词的同时也要考虑特征冗余的消除,以提高分类精度。下面,提出了一种新的特征选择方法。

2 新的特征选择方法

2.1 基本思想

在文本分类中,提出了一种基于 mRMR (Minimum Redundancy - Maximum Relevance)模型[5]的特征选择方法,其主要思想是将特征词与类别间最大相关性的选择标准和最小特征冗余的选择标准结合起来。与第一节中的常用特征选择方法相比,该方法多考虑了特征冗余的消除。

2.2 理论分析

首先,常用的最大相关性标准是按照公式(6)来选择特征的。其中 W_i 表示第 i 个特征词,C 表示类别集,S 表示要寻找的特征子集。

$$\max Q, Q = \frac{1}{|S|}\sum_{W_i \in S} I(W_i; C) \quad (6)$$

那么,像这样仅根据最大相关性标准选择到的特征词会存在冗余,以至对分类效果造成影响。所以也应该考虑到特征冗余的消除,于是应该添加下面的最小冗余标准[5] 来选择特征。

$$\min P, P = \frac{1}{|S|^2}\sum_{W_i, W_j \in S} I(W_i; W_j) \quad (7)$$

联合以上两个约束条件后得到的一个标准就是 mRMR(Minimum Redundancy - Maximum Relevance) 模型[5]。文献[6]中又以理论分析和基因选择的实验证明了该模型比仅考虑相关性标准的分类效果更好等。

使用差来联合这两个标准并简化后得到公式(8),

并按该式子逐一地选择特征词。在文本分类中，文中在计算特征词与类别间或特征词间的相关性时考虑特征词出现和不出现两种状态。其中 X 表示原始特征集，$X-S$ 表示未选择的特征集。

$$\max_{W_j\in X-S}[I(W_j;C)-\frac{1}{|S|}\sum_{W_i\in S}I(W_j;W_i)] \tag{8}$$

由于信息增益也是文本分类中广泛使用的效果较好的一种评估函数，所以在计算特征词与类别间的相关性 $I(W_j;C)$ 和特征词间的相关性 $I(W_j;W_i)$ 方面采用信息增益等相关理论来计算，如式(9) 和式(10) 所示。

$$I(W_j;C)=P(W_j)\sum_t P(C_t\mid W_j)\log\frac{P(C_t\mid W_j)}{P(C_t)}+P(\overline{W}_j)\sum_t P(C_t\mid \overline{W}_j)\log\frac{P(C_t\mid \overline{W}_j)}{P(C_t)} \tag{9}$$

$$I(W_j;W_i)=P(W_j,W_i)\log\frac{P(W_j\mid W_i)}{P(W_j)}+P(\overline{W}_j,W_i)\log\frac{P(\overline{W}_j\mid W_i)}{P(\overline{W}_j)}+P(W_j,\overline{W}_i)\log\frac{P(W_j\mid \overline{W}_i)}{P(W_j)}+P(\overline{W}_j,\overline{W}_i)\log\frac{P(\overline{W}_j\mid \overline{W}_i)}{P(\overline{W}_j)} \tag{10}$$

2.3 算法描述

那么，文中所提出的特征选择方法其具体算法流程可归结如下：

输入：原始特征集合 X，阈值 δ

输出：特征子集 S

步骤：

1) 初始化特征子集 $S=\{\}$；

2) 根据公式(9) 来计算 X 集合中每个特征词的 $I(W_j;C)$，并选择值最大的那个特征词并入到集合 S 中，作为第一个特征词。

3) 根据公式(8) 来选择下一个特征词，式(8) 中 $I(W_j;C)$ 和 $I(W_j;W_i)$ 的计算方法由式(9) 和式(10) 得到。

4) 重复步骤 3) 直到特征词个数达到阈值 δ。

在分类中，这种基于 mRMR 模型的特征选择方法一方面考虑到一般评估函数的特征词与类别的最大相关性问题，另一方面又减少了特征词间的冗余。

在接下来的实验中采用 KNN 分类器作为评价该算法的分类算法。在第 3 节中可以看到该特征选择方法得到的 KNN 分类器的分类精度比其它的分类精度高，也能够降低向量空间的维度，减少 KNN 分类器的计算量。

3 实验结果的比较

实验数据来源于复旦大学的中文语料库，共计 2816 篇，分为计算机、经济、环境、交通、教育、军事、体育、医药、艺术、政治十类，其中采用 934 篇文本作为训练集，1882 篇作为文本测试集，采用 KNN 作为分类器，其中 K 取 35。

3.1 评估指标

实验采用准确率和召回率评价分类性能的指标。

$$\text{分类的准确率}=\frac{\text{分类的正确文本数}}{\text{分类的实际文本数}}$$

$$\text{分类的召回率}=\frac{\text{分类的正确文本数}}{\text{分类应有的文本数}}$$

为了评估算法在整个数据集上的性能，有两种平均的方法可供使用，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，而微平均是每一个文档的性能指标的算术平均值。

3.2 实验结果的比较

对于同一个数据集它的准确率、召回率和微平均指标是相同的，但宏平均指标不同。当特征维数(即阈值 δ)取 1000 时，具体结果如图 1，图 2 所示。

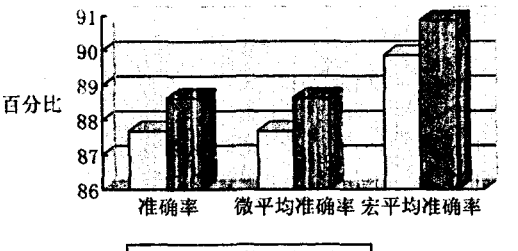

图 1 特征维数为 1000 时两种方法在准确率等方面的比较

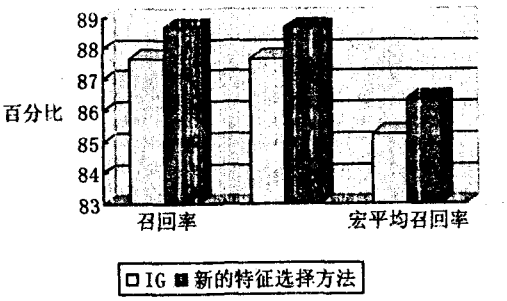

图 2 特征维数为 1000 时两种方法在召回率等方面的比较

从图中可见文中提出的特征选择方法会比信息增益的特征选择方法的分类性能更好些，而当特征维数较小时，两者的性能差异更加明显。当特征维数(即阈值 δ)取 800 时，其结果如图 3，图 4 所示。

当特征维数较小时有没有消除特征冗余对分类性能影响更大，而文中所提的特征选择方法比传统的特征选择方法多考虑了特征冗余的消除，所以在特征维数较小时新的特征选择方法的优势会更显著，而且消除了冗余之后所有特征的覆盖面也更大，更能代表其

原始特征空间。

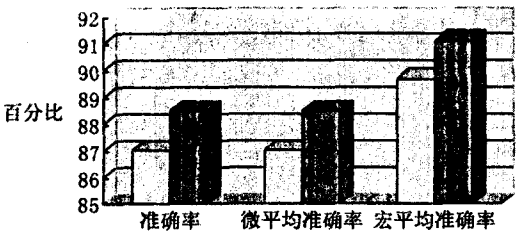

图 3　特征维数为 800 时两种方法在准确率等方面的比较

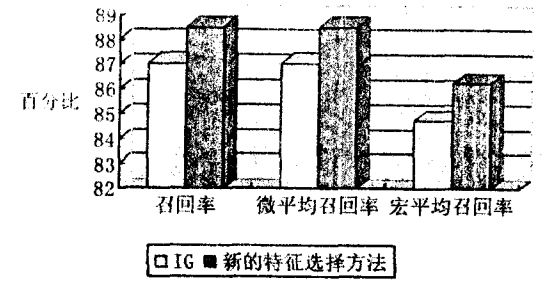

图 4　特征维数为 800 时两种方法在召回率等方面的比较

4　结束语

在文本分类中,如何进行特征选择是一个重要的研究问题。传统的特征选择方法主要集中在寻找与类别相关性大的特征上,而对特征冗余的问题没有给予足够的重视。如果没有消除特征冗余,那么选择出的特征子集并不能最大程度地代表其原始特征空间,显然会对分类性能产生影响。所以,提出了一种基于 mRMR 模型的特征选择方法,在考虑选择和类别相关性大的特征词的同时还考虑了特征冗余的消除。实验验证,其分类性能有所提高,虽然计算方法比传统的特征选择方法稍麻烦些。那么,研究该特征选择方法在不同语料集的分类情况以及改变公式(8)中 I 值的计算方法是否出现更好的分类效果,将是未来的研究工作。

参考文献:

[1] Salton G, Wong A, Yang C S. On the specification of term values in automatic Indexing[J]. Journal of Documentation, 1973,29(4):351 - 372.

[2] Fox C. Lexical Analysis and Stoplists[C]//Frakes W B, Baeze - Yates R. In Information Retrieval: Data Structure & Algorithms[s.l.]:P T R Prentice Hall,1992:102 - 130.

[3] Frakes W B. Stemming Algorithms[C]//Frakes W B, Baeze - Yates B. In Information Retrieval: Data Structure & Algorithms. [s.l.]:T P R Prentice Hall,1992:131 - 160.

[4] Yang Yiming, Pederson J O. A Comparative Study on Feature Selection in Text Categorization[C]//Proceedings of the 14th International Conference on Machine learning. Nashville: Morgan Kaufmann,1997:412 - 420.

[5] Ding C, Peng Hanchuan. Minimum redundancy feature selection from microarray gene expression data[C]// Proceeding of Second IEEE Computational Systems Bioinformatics Conference. LosA Lamitos, USA: IEEE Computer Society Press, 2003:523 - 528.

[6] Peng Hanchuan, Long Fuhui, Ding C. Feature Selection Based on Mutual Information Criteria of Max - Dependency Max - Relevance and Min - Redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2005,27(8):1226 - 1238.

(上接第 111 页)

方投影函数从边缘图像中定位出的人脸。

3　结束语

回顾了近年来在人脸检测与定位领域的一些常用方法,并在研究和分析各种人脸检测与定位算法的基础上,以计算简单、定位准确为原则,提出了一种将运动信息与边缘投影函数相结合的视频序列人脸检测与定位算法。边缘函数和投影函数检测人脸的缺点是易受背景中其它物体的干扰,但从视频序列中提取的特征对象图像具有背景简单、统一的特点,所以可利用边缘函数与投影函数进行后续的人脸检测。设计了双阈值 Sobel 算子进行边缘检测,该算子检测到的图像边缘清晰、细致、噪声少;提出了平方投影函数,该投影函数不但可区分均值相同的区域,而且可区分方差相同的区域。实验中使用的图像序列包括了室内环境、室外环境下的视频图像序列,运动目标包括单运动目标、具有可分离性的多运动目标。实验结果表明,该方法简单有效,能准确地进行人脸检测及定位。

参考文献:

[1] 李玉山.数字视觉视频技术[M].西安:西安电子科技大学出版社,2005.

[2] 刘党辉,沈兰荪.视频运动对象分割技术的研究[J].电路与系统学报,2002,7(3):77 - 85.

[3] 严云洋,郭志波,杨静宇.人脸识别特征抽取方法的研究进展[J].淮阴工学院学报,2007.16(3):24 - 30.

[4] 杨　莉,张　弘,李玉山.视频运动对象的自动分割[J].计算机辅助设计与图形学学报,2004,16(3):301 - 306.

[5] 严云洋,高尚兵,郭志波,等.基于视频图像的火灾自动检测[J].计算机应用研究,2008,25(4):1075 - 1078.