文章编号:1001 - 9081(2007)02 - 0433 - 03

一种改进的基于条件互信息的特征选择算法

王卫玲[1],刘培玉[1],初建崇[2]
(1. 山东师范大学 信息科学与工程学院,山东 济南 250014;
2. 海军航空工程学院 训练部,山东 烟台 264001)
(wangweiling0714@163.com)

摘 要:目前在文本分类领域较常用到的特征选择算法中,仅仅考虑了特征与类别之间的关联性,而对特征与特征之间的关联性没有予以足够的重视,这导致了特征之间预测能力的相互削弱,无法选出最有效的特征。提出了一种新的用于文本分类的特征选择算法(CMIM),它可以帮助选出区分能力强、弱相关的特征。经实验验证,CMIM 比传统的特征选择算法具有更好的性能。

关键词:特征选择;文本分类;条件互信息
中图分类号: TP311.13 **文献标识码**:A

Improved feature selection algorithm with conditional mutual information

WANG Wei-ling [1], LIU Pei-yu[1], CHU Jian-chong[2]
(1. *College of Computer Science and Engineering, School of Shandong Normal University, Jinan Jiangsu* 250014, *China*;
2. *Department of Training, Naval Aeronautical Engineering Institute, Yantai Shandong* 264001, *China*)

Abstract: Traditional feature selection algorithms have a common drawback, i. e. they do not consider the mutual relationships between features. It can result in that one feature's predictive power is weakened by others and the lost of efficiency. In this paper, we proposed a new feature selection method called Conditional Mutual Information Maximin (CMIM). It can select a set of individually discriminating and weakly dependent features. Simulation results demonstrate that the proposed method can improve the precision of text classification.

Key words: feature selection; text categorization; conditional mutual information

0 引言

在文本分类系统中,特征选择一直是一项关键技术和瓶颈技术。特征选择的目的在于减小文本的特征向量维数,去除冗余特征,保留有区分能力的特征。同时,具有区分能力的特征可以提高系统的效率和精度,所以文本分类系统中的特征选择部分是至关重要的。

目前,文本分类领域中开发的而且较常用的特征选择算法主要有:文档频率、信息增益、互信息、χ^2 统计等等。虽然采用了这些特征选择算法后可以大幅提高分类器的性能,但是由于它们只考虑了特征与类别之间的相关性,而忽略了特征之间的相关性,所以很容易出现以下问题:在有些情况下,某些特征之间的相关性很大,即它们之间很相似,但它们与类别的相关性也很大,于是这些相似的特征都被作为候选特征选入了最优特征子集,这就导致了特征子集中存在着大量的冗余特征,从而影响了分类器的性能。而这种情况在某些类别的训练集数目不足够多的情况下将会更加糟糕,因为在稀疏类别中的特征比那些在主要类别中特征的评估值要低,传统的特征选择算法往往会倾向于那些主要类别中的特征。

为了解决上述问题,本文提出了一种基于条件互信息理论的特征选择算法(Conditional Mutual Information Maximin, CMIM)。算法的基本思想是:在充分考虑到已选特征的条件下,使得新入选的特征与类别的条件互信息值最大。这样,当某个特征与已选的特征相关,即使这个特征与类别具有很强的关联性,CMIM 算法也不会将其选入特征子集,因为相比于已有的特征子集,它根本无法提供额外的与类别有关的信息。

1 相关理论

随着信息和信息科学对现代社会生活各方面影响的不断加大,人们对信息论的认识和价值估计也不断深化[5,6]。现在,信息论被广泛地应用在许多方面,诸如科学、工程、商业等等。在特征选择和分类学习中也采用了信息论的相关方法,如信息增益、互信息等。信息论是全局性的度量,它可以从全局上把握特征相关性和特征与类别的相关重要性程度。

1.1 熵

熵是通讯与信息理论中一个非常重要的概念,它是衡量一个随机变量取值的不确定性程度。而就数据集合而言,熵可以作为数据集合的不纯度或者说不规则程度的量度,所谓的不规则程度指的是集合中数据元素之间依赖关系的强弱。设 X 是一个离散随机变量,它可能的取值为 x 的概率 $P(X)$,那么定义:

$$H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

这里 $H(X)$ 就是随机变量 X 的熵,它是衡量随机变量取值的不确定性的量度。在随机试验之前,我们只了解各取值的

收稿日期:2006 - 08 - 22;**修订日期**:2006 - 11 - 10

作者简介:王卫玲(1979 -),女,山东烟台人,硕士研究生,主要研究方向:Web 挖掘、信息过滤; 刘培玉(1960 -),男,山东济南人,教授,博士生导师,主要研究方向:数据库、网络信息安全; 初建崇(1979 -),男,山东烟台人,助理工程师,主要研究方向:网络信息安全.

概率分布,而做完随机试验后,就确切地知道了取值,不确定性完全消失了。这样,通过随机试验我们获得了信息,且该信息的数量恰好等于随机变量的熵,在这个意义上,熵可以作为信息的度量。

1.2 互信息

在信息理论中为了更好地描述事物之间的普遍联系,引进了互信息(MI)的概念。对于两个随机变量X和Y,它们之间在某种程度上也是相互联系的,即它们之间存在着一定的统计依赖关系,互信息反映了两个随机变量之间相互依存关系的强弱。

互信息定义为:

$$I(X;Y) = -\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

互信息也是一种广泛应用的特征选择方法,$I(C;F_i)$可以从全局上衡量特征F_i与类别C之间关系,那些具有较高区分能力的单词都具有较高的互信息值。

但是由于互信息没有考虑词频,所以经常会倾向于选择低频词,因此单独使用MI的效果并不是很好,可以考虑与其他特征选择算法混合使用。

1.3 JMI及CMI

特征选择的主要目标是选择一定数量的特征,同时这些特征能够携带尽可能多的信息。这个目标可以被解释为最大化联合的互信息(JMI),$I(F_1,\cdots,F_K;C)$。有些文章中试图通过计算联合概率$P(f_1,\cdots,f_k,c)$来估计JMI的值,但是,这种方法在k值很大时往往会难以计算。因为如果特征子集中的每个变量可以取M个变量中的任何一个,那么对于任意向量$(F_1,\cdots,F_k,C)$会有M^K个不同的状态。

如果用以下的形式来表示联合互信息(JMI),就可以得到条件互信息(CMI)的定义:

$$I(F_1,\cdots,F_k;C) - I(F_1,\cdots,F_{k-1};C) = I(F_k;C\mid F_1,\cdots,F_{k-1}) \quad (1)$$

其中,$I(F_k;C\mid F_1,\cdots,F_{k-1})$表示CMI,它是用来衡量在$F_1,\cdots,F_{k-1}$已知的情况下,特征$F_k$与类别$C$之间的信息量。CMI提供了一个很好的选择新特征的方法,因为它很好地表达了新特征与已选特征之间的关系及它自身的类别区分能力。

定理 在特征选择中,JMI,$I(F_1,\cdots,F_K;C)$可以转化为求CMI,$I(F_k;C\mid F_1,\cdots,F_{k-1})$。

证明 将公式(1)表示为如下的形式:

$$I(F_1,\cdots,F_k;C) = I(F_1,\cdots,F_{k-1};C) + I(F_k;C\mid F_1,\cdots,F_{k-1})$$

假设当前的$k-1$个已选的特征最大化JMI,$I(F_1,\cdots,F_{k-1};C)$,那么下一个使CMI$(F_k;C\mid F_1,\cdots,F_{k-1})$最大化的特征,就应该被选入特征子集中以保证$K$个特征的JMI最大。这样,特征就一个个地被选入了特征子集。在每一步中,当$I(F^*;C\mid F_1,\cdots,F_{K-1})$最大时,特征$F^*$被选入特征子集。

因此,选一组最大化JMI的特征被转化为一个个地选入使CMI最大的特征。

证毕。

条件互信息虽然可以很好的表达特征与类别之间的关系以及特征与特征之间的关系,但是不幸的是,最大化CMI会遇到与JMI同样的问题,当特征的数目增加的时候,CMI同样难以计算。下面,我们提出一种算法来解决这个问题。

2 基于条件互信息的特征选择算法

直接计算CMI,$I(F^*;C\mid F_1,\cdots,F_{K-1})$,需要计算复杂的联合概率,这不但难以计算,同时也不具有健壮性。为了避免这个难题,我们倾向于通过把CMI分解成更简单的形式来避免复杂的联合概率的计算。首先,以少于K个特征的形式$I(F^*;C\mid \underbrace{F_i,\cdots,F_j}_{k-1})$来估算$I(F^*;C\mid F_1,\cdots,F_K)$。因为更多的信息将会降低不确定性,所以$I(F^*;C\mid F_1,\cdots,F_K) < I(F^*;C\mid \underbrace{F_i,\cdots,F_j}_{k-1})$,可以用下式来估算$I(F^*;C\mid F_1,\cdots,F_K)$的值:

$$I(F^*;C\mid F_1,\cdots,F_K) \approx \min I(F^*;C\mid \underbrace{F_i,\cdots,F_j}_{k-1}) \quad (2)$$

使得(2)最小化的那$K-1$个特征是已选特征子集中与F^*最相关的特征,这样的话,就使得F^*的类区分能力被削弱了。为了避免这样的情况,应该选择使得$\min I(F^*;C\mid \underbrace{F_i,\cdots,F_j}_{k-1})$尽可能大的特征。当一个特征能够较小地被已入选的特征影响并且它自身又是很重要的时候,这样的特征应该最大化$\min I(F^*;C\mid \underbrace{F_i,\cdots,F_j}_{k-1})$的值。这样就可以保证新特征既能够提供区分类别的信息又能够保证与其他已入选的特征的正交性。

在本文中,为了减少计算的时间,使用三元组的形式$I(F^*;C\mid F_i)$来估算CMI $I(F^*;C\mid F_1,\cdots,F_K)$。这个简单的形式衡量了特征$F^*$在选定特征$F_i$的情况下所包含的与类别有关的信息。将公式(2)中的右半部分换成更简单的三元组的形式$I(F^*;C\mid F_i)$,则公式变为:

$$I(F^*;C\mid F_i,\cdots,F_K) \approx I(F^*;C\mid F_i) \quad (3)$$

现在需要选择这样的特征F^*:它能够使得$\min I(F^*;C\mid \underbrace{F_i,\cdots,F_j}_{k-1})$的值最大。下面给出CMIM算法。因为考虑到互信息倾向于选择低频单词,我们先用词频对特征进行了初步的筛选,删除了部分低频词。

CMIM算法如下所示:

输入:n - 已选特征的数目
v - 所有特征的数目
输出:F - 选择的特征子集

1) 用DF删除部分低频单词
2) 设F为ϕ
3) 当$F_i = arg\ max_{i=1,\cdots,v} I(F_i;C)$时,将$F_i$加入F中
4) *repeat*
5) $m++$
6) 将F_i加入F中当 $F_i = \arg\max_{i=1,\cdots,v}\{\min_{F_j\in F} I(F_i;C\mid F_j)\}$
7) until $m = n$

由算法可以看出,*CMIM*的计算代价将远远小于计算*JMI*的代价。假设V是文本集中所有特征的数目,N是我们想要选择的特征的数目,D是训练文本的数目。为了计算一个F^*,需要对特征子集F中的每个特征F_i计算$I(F^*;C\mid F_i)$,这将会重复$|F|$次,每个特征的选择将会重复$(V-|F|)|F| = O(V^2)$次,则CMIM的时间复杂度为$O(NV^2)$。

3 实验与分析

实验中采用的语料共计9420篇,人工分为军事、体育、教

育、政治、经济、旅游、健康、娱乐 8 大类,其中采用 7 350 篇文本作为训练集,其余 2 070 篇文本作为测试集。

3.1　评估指标

实验采用的评估指标为 $F1$ 测试值,它综合考虑了文本分类的准确率与查全率,其具体计算公式如下:

$$F1\ 测试值 = \frac{准确率 \times 查全率 \times 2}{准确率 + 查全率}$$

$$准确率(precision) = \frac{分类的正确文本数}{实际分类的文本数}$$

$$查全率(recall) = \frac{分类正确文本数}{应有文本数}$$

3.2　实验结果及分析

我们将 CMIM 与 IG 的特征选择效果在 Naïve Bayes 和 SVM 两种分类器上进行了测试。图 1 给出了在这两种分类器上的比较结果。

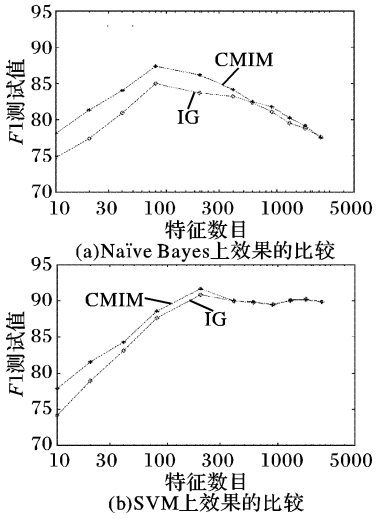

图 1　实验结果

通过图 1(a)、(b)的比较发现,在特征数目较小的时候,Naïve Bayes 与 SVM 的分类效果相差无几,但当特征数目逐渐变大的时候,Naïve Bayes 的性能逐渐变差。对于特征选择算法CMIM和IG来讲,当特征数目较小的时候(小于100),CMIM 的性能明显地要比 IG 好很多,随着特征数目的逐渐增多,这种优势就逐渐消失了,这时 CMIM 与 IG 的性能几乎等同。总体上来讲,CMIM 的性能比 IG 大约会提高 5% 左右。

4　结语

传统的特征选择研究主要集中在寻找类别相关的特征。虽然一些最近的研究指出了特征冗余的存在和影响,但是直接针对特征冗余的研究工作并不多。对此,本文提出了一种基于条件互信息的特征选择方法,由于直接利用条件互信息会因为计算量很大而无法实现,因此本文中对条件互信息进行了改进,将它分解为更简单的等价形式来减少计算量。经实验验证,CMIM 方法能够合理地选择出特征子集,提高了分类的准确度。

参考文献:

[1] GUYON I, ELISSEEFF A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2003, 27(3): 1157 - 1182.

[2] YU L, LIU H. Feature Selection for high - dimensional data: a fast correlation - based filter solution[A]. In Proceedings of the twenti-eth International Conference on Machine Learning[C]. Washington, 2003. 856 - 863.

[3] YU L, LIU H. Efficient Feature Selection via Analysis of Relevance and Redundancy[J]. Journal of Machine Research, 2004, 30(5): 1205 - 1224.

[4] 陈彬, 洪家荣, 王亚东. 最优特征子集选择问题[J]. 计算机学报, 1997, 2(20): 133 - 138.

[5] 常迥. 信息理论基础[M]. 北京: 清华大学出版杜, 1993.

[6] 朱雪龙. 应用信息论基础[M]. 北京: 清华大学出版社, 2001.

[7] WANG Y, WANG X-J. A New Approach to Feature Selection in Text Classification[A]. Proceeding of the Fourth International Conference on Machine Learning and Cybernetics[C]. Guangzhou, 2005. 355 - 360.

[8] SONG FX, LIU SH. A Comparative Study on Text Representation Schemes in Text Categorization[J]. Pattern Anal Applic, 2005, 30(8): 199 - 209.

(上接第 432 页)

2.3　系统运行情况

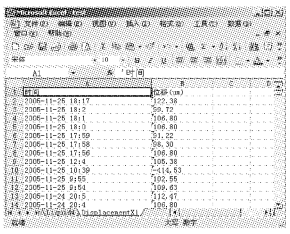

图 3　原始位移数据表

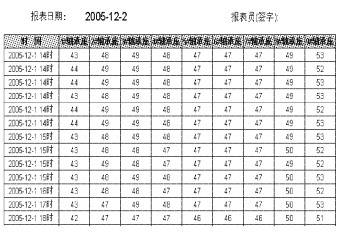

图 4　轴承座温度报表

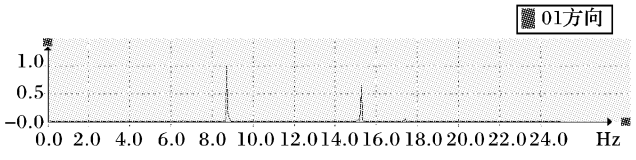

图 5　位振动信号频谱图

如图 3 ~ 图 5 所示,系统自从投入使用以来运行良好。状态监测与故障诊断系统在实际生产中发挥巨大作用,实时监测设备的运行状态,在线诊断设备故障,保障企业生产的正常进行,减少了企业的经济损失。

3　结语

对于多参数状态监测与故障诊断系统,结合系统的数据信息采用文件系统进行数据管理,同时采用内存对齐规则设计规划了系统数据的存放格式,这样既满足了设计的简单化又使系统在大数据量、高采样频率和大的数据分析处理过程中能够正常运转又符合了现场的要求,使整个系统很好的为工业现场服务。

参考文献:

[1] 萨师煊, 王珊. 数据库系统概论[M]. 第 3 版. 北京: 高等教育出版社, 2002.

[2] 祖淑芝, 王太勇, 邓学欣. 数据库技术在设备故障诊断系统中的应用[J]. 微处理机, 2005, (3): 24 - 26.

[3] 王实, 刘晓明. 深入浅出西门子 WinCC V6[M]. 北京: 北京航空航天大学出版社, 2004.

[4] 徐晓刚, 高兆法, 王秀娟. Visual C++ 6.0 入门与提高[M]. 北京: 清华大学出版社, 2000.

[5] 王华, 叶爱亮, 祁立学, 等. Visual C++6.0 编程实例与技巧[M]. 北京: 机械工业出版社, 1999.