文章编号: 1003-0077(2007)05-0056-05

面向文本分类的基于最小冗余原则的特征选取

张希娟,王会珍,朱靖波

(东北大学 信息学院 计算机软件与理论研究所,辽宁 沈阳 110004)

摘 要: 在文本分类中,为了降低计算复杂度,常用的特征选取方法(如 IG)都假设特征之间条件独立。该假设将引入严重的特征冗余现象。为了降低特征子集的冗余度,本文提出了一种基于最小冗余原则(minimal Redundancy Principle, MRP)的特征选取方法。通过考虑不同特征之间的相关性,选择较小冗余度的特征子集。实验结果显示基于最小冗余原则方法能够改善特征选取的效果,提高文本分类的性能。

关键词: 计算机应用;中文信息处理;条件独立性假设;最小冗余原则;特征选取;文本分类

中图分类号: TP391 **文献标识码**: A

Feature Selection Based on Minimal Redundancy Principle for Text Classification

ZHANG Xi-juan, WANG Hui-zhen, ZHU Jing-bo

(Institute of Computer Software and Theory, Northeastern University, Shenyang, Liaoning 110004, China)

Abstract: In text classification tasks, these well-known feature selection methods such as information gain adopt conditional independence assumption between various features. However, this assumption would result in serious redundancy problems among various selected features. To alleviate the redundancy problem within the selected feature subset, this paper proposed a method based on minimal redundancy principle (MRP) for feature selection, in which correlations between different features are considered in feature selection process, and a feature subset with less redundancy can be built. Experimental results showed that MRP method can improve the effectiveness of feature selection, and results in better text classification performance (in most cases).

Key words: computer application; Chinese information processing; conditional independence assumption; minimal redundancy principle; feature selection; text classification

1 引言

文本分类是根据事先提供的训练数据,构造高性能分类器,实现对新文本的自动分类过程[1]。目前已有很多机器学习技术被成功应用于文本自动分类过程,如朴素贝叶斯(naïve Bayes)[2]、最大熵(Maximum Entropy)[3]、K-Nearest-Neighbor[4]等模型。

在文本分类中,原始的特征空间可能具有非常高的维数,造成数据稀疏问题非常严重[4]。高维特征空间造成很多机器学习技术(如 Bayesian Belief Network, BBN)[5]的计算复杂度难以承受,除非引入不同特征之间条件独立性假设(Conditional Independence Assumption, CIA)[6]。目前文本分类中常用的特征选取方法都直接采用 CIA 假设,如信息增益(Information Gain, IG)[1]、Chi 统计(Chi-Statistics, CHI)[7]、文档频率(Document

收稿日期: 2007-04-15 **定稿日期**: 2007-06-28

基金项目: 国家自然科学基金资助项目(60473140);国家 863 高科技计划课题资助(2006AA01Z154);国家教育部新世纪优秀人才计划项目资助(NCET-05-0287);国家 985 工程计划项目资助(985-2-DB-C03)

作者简介: 张希娟(1984—),女,硕士生,主要研究方向为自然语言处理;王会珍(1980—),女,助教,主要研究方向为自然语言处理;朱靖波(1973—),男,教授,主要研究方向为自然语言处理。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Frequency, DF)[1,7]等。也有一些特征选取方法没有直接采用 CIA 假设,如 Sequential Forward Selection(SFS)[8]和 Wrapper[1,8]。SFS 和 Wrapper 方法虽然能够取得较好的特征选取效果,但由于没有直接采用 CIA 假设,造成计算复杂度非常高,只能适用于特征数较小的应用任务中(如 10～30 个特征)。在高维特征空间的文本分类应用任务中,SFS 和 Wrapper 方法的计算代价是不可忍受的。基于 CIA 假设的特征选取方法(如 IG)不但可以降低文本自动分类的计算代价,在某些情况下还可能提高分类的性能,主要原因在于特征选取方法可以通过去掉一些噪音特征、选择一个最好的特征子集实现降维目的[1,7,8]。

目前文本分类中常用的特征选取方法都没有直接考虑特征相关性,也没有考虑特征子集的冗余性。它们直接利用了特征之间的 CIA 假设,单独对每个特征进行评价,根据评价值从大到小进行排序,选择排名靠前的特征构造最佳特征子集。为此,本文主要研究通过考虑计算不同特征之间的相关性,减少特征子集的冗余度。实验结果显示基于最小冗余原则(Minimal Redundancy Principle, MRP)方法能够改善特征选取的效果,进而提高文本分类的性能。

2　最小冗余原则的特征选取方法

为了降低计算复杂度,目前文本分类中常用的特征选取方法[1],都引入了 CIA 假设。然而在实际应用中,特征之间的 CIA 特性是难以满足的,因为不同特征之间存在相关性(Correlation)。例如,在 Newsgroups 语料上,采用 IG 方法进行特征选取,排名前 10 个的特征基本上可以被分为六组: C1 = {"*God*","*Jesus*"}、C2 = {"*Windows*","*file*","*DOS*"}、C3 = {"*car*"}、C4 = {"*space*"}、C5 = {"*game*"}和 C6 = {"*scsi*","*drive*"}。其中,同一个组的特征对该组所体现的主题类别具有很强的分类能力,对其他主题类别的分类能力比较弱。从分类系统的设计角度看,以 C1 组为例,如果已经选择了"*God*",最好不要马上选择"*Jesus*",最好选择体现其他主题类别的特征,因为"*God*"和"*Jesus*"之间存在冗余(Redundancy)。同样对于 C2 组而言,虽然 C2 组中每个特征的 IG 评价值都很高,在前 10 个特征中占了 30 %的比例,但并不是最佳的选择,特征之间冗余度过高[9]。

基于上述考虑,为了降低特征集合的冗余度,又可以保证较低的计算复杂度,本文提出了一种基于最小冗余原则(Minimal Redundancy Principle, MRP)的特征选取方法。

最小冗余原则(MRP)方法的基本思想是选择新特征的同时,通过考虑新特征与已经确定的特征之间的冗余度,选择满足最小冗余度原则和最大评价值原则的候选特征作为新特征。在实际应用中,最小冗余度原则与最大评价值原则存在一定程度的冲突,MRP 方法本质上是寻求两者之间的一种均衡关系。为了实现最小冗余原则的特征选取方法,至少需要研究两个问题: 1)如何计算不同特征之间的冗余度;2)如何在特征评价函数中考虑冗余度因素。

平均互信息[9]是一种较好的冗余度测量函数,衡量两个随机变量之间的依赖程度。其值大小与依赖程度有关系。依赖程度越大,两者冗余度越高。由此,两个不同特征的冗余度计算如下:

$$I(X;Y) = \sum_{x\in\{0,1\},y\in\{0,1\}} p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)} \tag{1}$$

其中,1 和 0 分别表示特征出现与不出现,概率 $p(x)$, $p(y)$ 和 $p(x,y)$ 可以从训练数据中得到。实际应用中也可以采用条件互信息 $I(X;Y|C)$ 来计算两个不同特征之间的冗余度,计算复杂度比 $I(X;Y)$ 高一些。本文采用公式(1)来实现不同特征之间的冗余度测量。

最小冗余原则的特征选取,基本方法在于通过考虑冗余度因素来修正传统的评价函数。为了论述方便,假定目前已经确定包含 $m-1$ 个特征的子集 X_{m-1},采用基于最小冗余原则的评价函数 $J_{MRP}(\cdot)$,第 m 个特征的选取过程可以定义为:

$$\begin{aligned} x_m &= \arg\max_{x_j\in Y-X_{m-1}} J_{MRP}(x_j) \\ &= \arg\max_{x_j\in Y-X_{m-1}} \left[J(x_j) - \frac{\eta}{m-1}\sum_{x_i\in X_{m-1}} I(x_i;x_j)\right] \end{aligned} \tag{2}$$

在信息增益方法中,评价函数 $J(x)=IG(x)$,在 Chi 统计方法中,评价函数 $J(x)=CHI(x)$,在文档频率方法中,评价函数 $J(x)=DF(x)$。η 是一个预定义的固定系数,用于控制冗余度因素在评价函数中的影响程度。如果采用不同的具体评价函数,则对应的最小冗余原则的评价函数分别为:

$$J_{MRP-IG}(x_m) = IG(x_m) - \frac{\eta_{IG}}{m-1}\sum_{x_i\in X_{m-1}} I(x_i;x_m)$$

$$J_{MRP-CHI}(x_m) = CHI(x_m) - \frac{\eta_{CHI}}{m-1}\sum_{x_i\in X_{m-1}} I(x_i;x_m)$$

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$$J_{MRP\text{-}DF}(x_m) = DF(x_m) - \frac{\eta_{DF}}{m-1}\sum_{x_i \in x_{m-1}} I(x_i; x_m) \quad (3)$$

3 实验

3.1 分类器选择

在很多分类任务中,基于信息增益(Information Gain, IG)的特征选取方法性能表现最好[7],因此本文采用IG评价函数来构建比较分析实验。根据公式(3),对应的最小冗余原则方法采用评价函数 $J_{MRP\text{-}IG}$。信息增益的计算公式如下:

$$IG(t;C) = \sum_{t\in\{t,\bar{t}\}}\sum_{i=1}^{C} p(t,C_i)\log\frac{p(t,C_i)}{p(t)\,p(C_i)} \quad (4)$$

其中1和0分别表示特征出现与不出现。$P(t, C_i)$、$p(C_i)$和 $p(t)$均使用词频来计算概率值。

Peng等人[9]曾提出一种最小冗余最大相关(minimal Redundancy Maximal Relevance, mRMR)的方法用于候选特征的提取,然后再采用Wrapper方法进行最优特征子集的识别,取得了较好的效果。上述提到,Wrapper方法的计算复杂度非常高,并且完全依赖于具体的分类器本身。本文研究主要针对独立于具体分类器设计的特征选取方法。另外mRMR方法相当于本文提出的基于最小冗余原则方法的一个特例,即评价函数 $J_{MRP\text{-}IG}$,并且 $\eta=1$。

多项式朴素贝叶斯分类器[2]是一个比较常用的分类器,本文采用该分类器来比较分析IG方法和最小冗余原则方法MRP-IG的性能,采用Accuracy值作为评价方法。

$$C^* \equiv \arg\max_j p(C_j)\prod_{i=1}^{|V|} p(w_i \mid C_j)^{n(w_i)} \quad (5)$$

其中 $p(C_j)$表示类别的先验概率,$|V|$是特征总数,w_i是特征集合中的第 i 个特征。$n(w_i)$表示特征 w_i在给定文档中的词频数。$p(w_i|C_j)$表示特征 w_i 属于类别 C_j的概率。

3.2 语料

在本文的比较实验中,使用了三个公开的语料用于比较MRP-IG与IG方法的分类性能。分别是Newsgroups,WebKB和Indurstry Sector语料。对这三个语料的预处理过程均采用McCallum等人开发的Rainbow工具完成[10]。

1) Newsgroups语料。该语料由Ken Lang收集,分为20个类别,每个类别大约包含1 000个新闻文本。在文本预处理的过程中,去掉了UseNet头部,禁用词及在语料中只出现一次的词。预处理结束后剩余62 264个词。

2) WebKB语料。该语料是由来自Texas、Washington等大学计算机科学系收集的网页组成,这些网页被分为7类,在本文中,只使用了其中常用的4类: student, faculty, course和project类,共包含4 199篇网页。在预处理过程中,没有去禁用词和提取词干,去掉了在语料中只出现一次的词,剩余词汇23 803个。

3) Industry Sector语料。该语料包含71类6 400篇网页。预处理过程中去掉了禁用词和在语料中只出现一次的词。剩余词汇29 933个。

在实验过程中,WebKB语料使用随机选取的语料的35%作测试语料,其他语料均采用随机选取的语料的20%作测试语料。所有实验均采用五次交叉检验的结果。

3.3 实验结果与分析

图1～3比较了MRP-IG和IG方法在不同语料上的分类性能,为了观察冗余度因素在评价函数中的影响程度,本文比较了η分别为0.5,1,2,3的时候,MRP-IG和IG方法的分类性能。

图1是MRP-IG方法和IG方法在Newsgroups语料上分类性能的比较结果,两种方法都是取所有特征时达到最佳性能84.596%,总体趋势非常接近,只有当特征数目较少的时候(<100),MRP-IG方法稍微好于IG方法。随着特征数目的增加,两者的性能接近一样。为了表明MRP特征选取方法在Newsgroups语料上的有效性,本

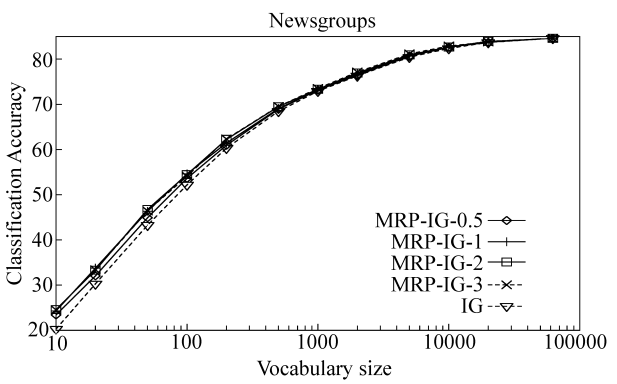


图1 MRP-IG与IG方法在Newsgroups语料上分类性能的比较

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

文针对使用 IG 方法选取出的前 10 个特征,在 η 取不同的值的时候,对 MRP-IG 方法和 IG 方法的特征选取结果进行了对比。

表 1 是 MRP-IG 方法在 η 分别取 0.5,1,2,3 的时候,与 IG 方法的前 10 个特征的选取结果的对比。表 1 显示,MRP-IG 的方法有效的降低了冗余特征出现的几率,而且随着 η 的增大,效果更加明显。例如在{"*God*","*Jesus*"}组中,当"*God*"被选中之后,"*Jesus*"被选中的几率下降,在{"*Windows*","*file*","*DOS*"}组中,当"*Windows*"被选中以后,"*file*"和"*DOS*"被选中的几率下降的非常明显。而且 MRP-IG 的方法在降低冗余特征被选中的几率的同时并没有将非冗余特征排除,"*space*"的排名并没有下降多少,而"*game*"的排名不仅没有下降反而有所上升。表 1 显示,最小冗余原则方法在选择新特征的时候,考虑了新特征与已经确定的特征之间的冗余度,选择的新特征与已有特征集合之间具有最小的冗余度。

表 1　MRP-IG 方法在 Newsgroups 语料上对 IG 前 10 个特征的选取结果

Features	IG	MRP-IG-0.5	MRP-IG-1	MRP-IG-2	MRP-IG-3
God	1	1	1	1	1
Jesus	4	10	13	15	18
Windows	2	2	2	2	2
File	8	22	44	291	1 854
DOS	9	15	33	89	328
Car	3	3	4	5	6
SCSI	6	5	5	4	5
Drive	10	11	12	13	16
Space	5	6	6	9	9
Game	7	4	3	3	3

图 2 是 MRP-IG 和 IG 方法在 WebKB 语料上分类性能的比较结果。可以看出,MRP 方法的优势很明显,而且 η 值越小优势越显著。在特征数为 5 000 时,IG 达到最佳性能 86.862%,在特征数为 2 000 时,MRP-IG-0.5,MRP-IG-1 和 MRP-IG-2 分别达到最佳性能 87.626%,88.004%和 87.976%。但在初始阶段特征数小于 500 的时候,MRP-IG-0.5 的分类性能最佳,本文认为原因是在该语料的预处理过程中,没有去掉禁用词。本文发现去掉禁用词会降低该语料的分类性能,因为在该语料中许多禁用词都是一些类别的指示词,例如"*my*"是 *student* 类的指示词,对 *student* 类有很好的区分能力。这个问题在其他研究中也被提到过[2]。但是禁用词与很多特征相关度很高,例如"*my*"和"*class*",而"*class*"是 *course* 类的指示词,但一旦特征"*my*"被选中以后,"*class*"被选中的几率就会变小,而且 η 越大选中的几率就越小。因此,冗余度因子的值越小,冗余度因素所起的作用越小,禁用词对指示词("*my*"对"*class*")的影响就会越低,分类性能越好。

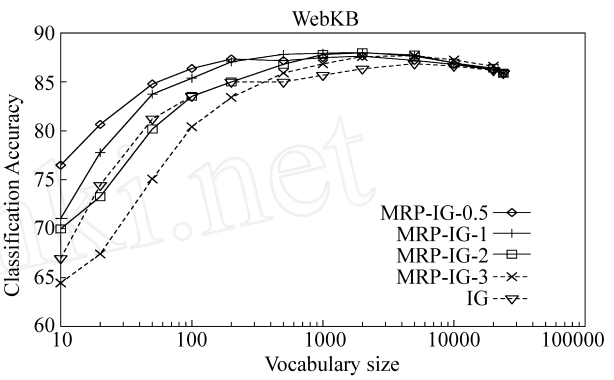


图 2　MRP-IG 与 IG 方法在 WebKB 语料上分类性能的比较

图 3 是 MRP-IG 和 IG 方法在 Industry Sector 语料上分类性能的比较结果。图 3 显示,当特征数超过 100 以后,最小冗余原则方法分类性能明显高于 IG 的方法,而且 η 值越大,分类效果越好。在特征数从 100 到 10 000 的过程中,MRP-IG-3 的方法比 IG 的方法平均高 4 个百分点。所有方法都是在特征数 10 000 个的时候达到最佳性能,IG 为 74.598%,MRP-IG-3 为 75.776%。而当特征数目少于 100 的时候,MRP-IG 的方法与 IG 的方法性能

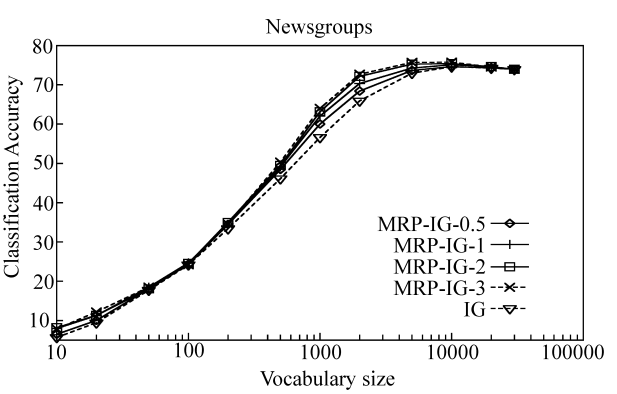


图 3　MRP-IG 与 IG 方法在 Industry Sector 语料上分类性能的比较

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

很接近,本文认为可能的原因是 Industry Sector 语料的类别非常多(71 类),因此当特征数目比较少的时候,特征之间的冗余程度非常小,冗余度因素所起的作用不大。例如使用 IG 方法选择出来的前 10 个特征,可以分成八组:C1 = {"*src*","*gif*","*image*"}、C2 = {"*images*"}、C3 = {"*thu*"}、C4 = {"*gas*"}、C5 = {"*fund*"}、C6 = {"*menu*"}、C7 = {"*water*"}、C8 = {"*coal*"},其中只有三个特征之间存在冗余。为了表明 MRP-IG 的方法在特征数较少的时候仍然有效,本文针对 C1 组的特征,对 MRP-IG 和 IG 的选取结果进行了对比。表 2 是 MRP-IG 和 IG 的方法对 C1 组三个特征的选取结果。从表中可以看出,一旦"*src*"选定,最小冗余原则方法能有效降低"*gif*"和"*image*"被选中的几率。

表 2 在 Industry Sector 语料上 MRP-IG 和 IG 方法对"src"、"gif"和"image"特征的选取结果

Features	IG	MRP-IG-0.5	MRP-IG-1	MRP-IG-2	MRP-IG-3
src	1	1	1	1	1
gif	3	8	14	27	37
image	6	13	25	43	83

4 结论

目前常用的特征选取方法(IG,CHI 和 DF)利用了特征之间的 CIA 假设,没有考虑特征集合的冗余度问题,造成特征集合冗余度过高,影响分类性能。为了解决上述问题,本文重点研究了特征之间的冗余问题,提出了基于最小冗余原则的特征选取方法。该方法通过考虑计算不同特征之间的相关性,减少特征子集的冗余度,来改善特征选取效果,进而提高分类性能。在三个文本分类语料上的实验结果显示,最小冗余原则方法能很好的降低冗余特征被选中的几率,从而降低特征集合的冗余度,改善特征选取的效果,相对于 IG 的方法,能更好的提高分类性能。

参考文献:

[1] F. Sebastiani.. Machine learning in automated text categorization[J]. ACM computing surveys., 2002, 34(1): 1-47.

[2] A. McCallum, K. Nigam. A comparison of event models for naïve bayes text classification [A]. In: AAAI-98 Workshop on Learning for Text Categorization [C]. 1998.

[3] 陈文亮,朱慕华,朱靖波,姚天顺. 基于 Bootstrapping 的文本分类模型[J]. 中文信息学报,2005,19(2):86-92.

[4] D. Lewis, R. Schapire, J. Callan, and R. Papka. Training Algorithms for Linear Text Classifiers[A]. In: Proceedings of ACM SIGIR[C]. 1996, 298-306.

[5] G. F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks [J]. Artificial Intelligence, 1990, 42:393-405.

[6] S. Cooper. Some inconsistencies and Misnomers in probabilistic information retrieval[A]. In: Proceedings of the 14th ACM SIGIR International Conference on Research and Development in Information Retrieval [C]. 1991.

[7] Y. Yang, J. O. Pedersen. A comparative study on feature selection in text categorization [A]. In: Proceedings of ICML-97, 14th International Conference on Machine Learning[C]. 1997. 412-420.

[8] A. Appice, M. Ceci, S. Rawles, P Flach. Redundant feature elimination for multi-class problems[A]. In: ACM International Conference Proceeding Series[C]. 2004.

[9] Hanchuan Peng, Fuhui Long, and Chris Ding. 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).

[10] McCallum, Andrew Kachites. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering [DB/OL]. Published in http://www.cs.cmu.edu/~mccallum/bow. 1996.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net