基于交叉覆盖算法的文本分类

王倩倩,段 震,张燕平
(安徽大学 计算智能与信号处理重点实验室,安徽 合肥 230039)

摘 要:分类是文本信息搜索和挖掘的核心内容,被广泛应用于搜索引擎的设计以及数据挖掘的研究中。首先对文本进行分词,对分词的结果采用 x^2 统计量的方法提取特征,再使用前向神经网络的交叉覆盖算法作为分类器进行文本分类。实验表明,x^2 统计量可大规模降低特征维数,在此基础上结合交叉覆盖算法的优秀分类能力,可在特征维数较低的情况下获得一个性能较好的文本分类器。

关键词:文本分类;x2 统计量;交叉覆盖算法
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2007)06-0113-03

Text Classification Based on Cross Cover Algorithm

WANG Qian-qian, DUAN Zhen, ZHANG Yan-ping
(Key Lab. of Intelligent Computing & Signal Processing, Anhui University, Hefei 230039, China)

Abstract: Text classification is the key point in text information searching and mining and is widely used in the design of search engine and data mining. Use the method of x^2 statistic to extract text's characteristics after processing the text and then use cross cover algorithm to design a classifier. The result of experiment shows that the method of x^2 statistic can decrease the dimensions effectively and the cross cover algorithm has good classification ability that can achieve a satisfactory classifier.

Key words: text classification; x^2 statistic; cross cover algorithm

0 引 言

近年来,随着科技和社会的发展,尤其是 Internet 的出现并普及使用,文本信息资料呈现爆炸式增长,为了在海量的信息中找到有价值的信息,人们迫切需要实现文本的自动分类。文中将 CHI 方法和交叉覆盖算法相结合,给出一种有效的文本分类方法。

文本分类从本质上来说是对文本模式特征的识别过程。根据一般模式识别的工作流程,可以把整个文本分类的过程分为三个阶段:预处理、特征选择和分类,即首先读入文本,在字典的支持下把文本切割为有意义的词组序列,然后根据一定的特征选择算法,提取出文本的特征值,整理成学习样本集后进行学习得到分类器,从而判断未知文本的类别。

收稿日期:2006-08-08
基金项目:"九七三"计划国家重点基础研究(2004CB318108);国家自然科学基金(60475017, 60135010);安徽省自然科学基金(050420208)
作者简介:王倩倩(1982-),女,安徽六安人,硕士研究生,研究方向为计算智能;张燕平,教授,硕士生导师,研究方向为人工神经网络、智能算法及其应用。

1 预处理

首先读入文本信息并将其表示为可供计算机处理的形式。一类主要的表示方法是 Salton 提出的向量空间法,即将文本信息用向量的形式表示为:$\boldsymbol{d}_i=(w_{i1}, w_{i2}, \cdots, w_{in})$,其中 $\boldsymbol{d}_i$ 表示文本,n 表示特征项向量空间的维数,w_{ij} 表示有意义的词组即特征项的权重,具体应用中可使用词组是否出现和词组出现频率两种计算方法。一般文本分类会使用单个字或词来作为特征项,已有的文本分类的研究结果表明,使用词组来作为特征项的效果要优于使用字作为特征项。因此,要把文本转化为向量空间中的向量,首先就必须分词。文中采用 2002 现代汉语词典作为词典,使用双向扫描法对文本进行分词。通过分词可以得到大量有意义的词组,但词组数目过多会导致分类器的运算强度过大,且不同词组对分类的影响度是不同的,因此需要采用合适的特征选择算法来找出需要的词组集。

2 特征选择

特征选择的目的是在分词所得到的大量词组中找出某一真子集,选择标准是此真子集可以显著增加文本分类的准确性,此真子集不应该改变原始空间的性

质,而应该是原始空间中的最重要的特征,此真子集组成了文本分类的特征集[1]。

文本分类中,用于特征选择的方法主要有[2]:特征频度(Term Frequency),文档频度(Document Frequency),特征熵(Term Entropy),互信息(Multi - Information),信息增益(Information Gain),x^2 统计量(Chi - square),几率比(Odds Ratio) 等。在文献[3] 中指出当特征值维数较低时,x^2 统计量的方法具有较好的效果。x^2 统计量方法中,特征值 t_k 的 CHI 权重如式(1) 所示:

$$x^2(t_k,c_i) = \frac{n[P(t_k,c_i)\times P(\bar{t}_k,\bar{c}_i) - P(t_k,\bar{c}_i)\times P(\bar{t}_k,c_i)]^2}{P(t_k)\times P(c_i)\times P(\bar{t}_k)\times P(\bar{c}_i)} \quad (1)$$

其中 n 表示总的文本数,$P(t_k,c_i)$ 表示学习样本集中出现特征 t_k 并属于类型 c_i 的文本的概率,$P(\bar{t}_k,\bar{c}_i)$,表示样本中不属于类型 c_i 的文本中不出现特征 t_k 的文本的概率,$P(t_k,\bar{c}_i)$ 表示样本中不属于类型 c_i 的文本中出现特征 t_k 的文本的概率,$P(\bar{t}_k,c_i)$ 表示样本中属于类型 c_i 的文本中不出现特征 t_k 的文本的概率,$P(t_k)$ 指出现特征 t_k 的文本概率,$P(c_i)$ 指 c_i 类文本出现的概率。$x^2(t_k,c_i)$ 度量了 t_k、c_i 的相关程度。CHI 值越大,t_k 和 c_i 就越相关,特征 t_k 对文本分类的影响就越大。特征在计算中可以简化该式为式(2)[3]:

$$x^2(t_k,c_i) = \frac{n\times(AD - CB)^2}{(A + C)\times(B + D)\times(A + B)\times(C + D)} \quad (2)$$

A 表示 t_k 和 c_i 同时出现的文本数,B 表示 t_k 出现但 c_i 不出现的文本数,C 表示 c_i 出现但 t_k 不出现的文本数,D 表示 t_k、c_i 均不出现的文本数。综合考虑特征对各类文本的影响,在实用中采用另一改进的全局 CHI[1] 值,如式(3) 所示:

$$x^2(t_k) = \max_{i=1}^{m}\{x^2(t_k,c_i)\} \quad (3)$$

使用特征选择之后,可以得到各个词组的 CHI 值,并将其由大到小排列,根据时空复杂度的需要选取一定数量的词组作为特征词组,放入特征库中。

3 文本训练与分类

多层前向神经网络的交叉覆盖算法是根据神经元的几何意义[4]提出的,它根据样本的特征来构造性地建立神经网络模型,在一定意义上解决了多年来一直未解决的作为分类器的多层前向网络的设计问题。文中采用交叉覆盖算法作为分类器来进行文本分类。

根据特征选择中所获得的特征库来读入学习样本,取得特征向量。由这些特征向量及其分类属性值(属于哪一类文件)得到样本集并按如下的步骤进行学习[5,6],让机器自己去总结这些样本各自在特征上的相似和差异之处,构造出对应的交叉覆盖网络,从而对新的样本给出较为准确的判断。构造样本 X 的覆盖过程如下:

给定样本集 $k = \varepsilon(x^i,y^i), i = 1,\cdots,p$

(1) 若 $|x^i|$ 不相等,$x^i \in R^n$,则做变换 $f(x):R^n \to S^{n+1}$

$$f(x) = (x,\sqrt{R^2 - |x|^2}), R \geqslant \max_i\{|x^i|\}$$(一般取 R 为 1.1 ~ 1.2 倍)。

(2) 从 1 开始取类别号($i = 1$) 构造覆盖 $C(i)$。

(3) 若样本集 k 中无尚未被覆盖的点,则算法结束,否则任取样本集中一未被覆盖的点 $a^1(|a^1| = R)$,求得阈值 θ^1,可得以 a 为中心,以 θ^1 为阈值的覆盖 $C(a^1)$,θ^1 求法如下:

$d^1(a^1) = \max\{d(x,a^1)\}, x \notin C(a^1), x \in k$,$d(x,y)$ 表示 x 和 y 的距离。

$d^2(a^1) = \max\{d(x,a^1) \mid d(x,a^1) > d^1(a^1)\}$, $x \in C(a^1)$

$$\theta^1 = \frac{d^1(a^1) + d^2(a^1)}{2}$$

(4) 求覆盖 $C(a^1)$ 的重心,同(1) 使其映射到球面上,设投影点为 a^2,再如(3) 求得其阈值 θ^2,得到新的覆盖 $C(a^2)$,若 $C(a^2)$ 比 $C(a^1)$ 覆盖更多的点,则 $a^2 \to a^1, \theta^2 \to \theta^1$。循环操作直到 $C(a^2)$ 不能覆盖更多的点。

(5) 求 a^2 的平移点[5] a^3,如上方法求得覆盖 $C(a^3)$,若 $C(a^3)$ 比 $C(a^2)$ 覆盖更多的点,则 $a^3 \to a^1$, $\theta^3 \to \theta^1$,转(4),否则就求得了 $C(i)$ 类的一个覆盖,如果 i 小于已给定的类别数,则 $i = i + 1$,转(3),否则学习结束。

设已求得覆盖组 $C(i)(i = 1,\cdots,n)$,取三层神经网络,隐层取 n 个神经元,每个神经元为一个覆盖,其激励函数取正线性函数。输出层取 k 个神经元,第 i 个神经元的输入为覆盖第 i 类点的覆盖的输出,其激励函数为或门。这样的三层网络,就可以完成分类。

使用测试样本对训练好的分类器进行测试,步骤如下:

① 同学习方法步骤(1),将每个测试样本映射到球面上,对每一个样本 x,计算:

$d(x,C(i)) = \max_{C\in C(i)}\{\sigma(K(\omega_c,x) - \theta\}$,其中:$\omega_c$ 是 C 的中心,θ 是阈值。其中:

$$\sigma(x) = \begin{cases} x & x \geqslant 0 \\ 0 & x < 0 \end{cases}$$

② 求 $\max_{1\leqslant i\leqslant k} d(x,C(i))$ 所对应的 i,确定样本 x 的

类别;对于拒识的样本,采用就近原则确定样本的类别。

4　实验和结论

文中所使用的实验数据是中文自然语言处理开放平台中提供的中文文本分类语料库,选取了计算机、交通、教育、经济、体育、艺术、政治七类共 2162 篇文本,并在各类别中按 1:1 的近似比例随机抽取学习样本集和测试样本集。文本首先通过分词,共得到 79360 个词组,然后通过 x^2 统计量方法进行特征降维,从中取出 CHI 值较高的 1500 个词组作为特征库,特征值只占词组总数的 1.8901%。实验结果如下所示,其中表 1 以词组在文本中是否出现作为特征向量中各维的值(0 或 1),表 2 以词组在文本中出现的频率作为特征向量中各维的值,以精确率和召回率作为评价标准[7]。

表 1　以词组是否出现作为向量值

类别	学习数	测试	精确率	召回率
计算机	99	101	87.3585%	92.0792%
交通	99	115	91.3462%	82.6087%
教育	109	111	75.9398%	90.9910%
经济	162	163	90.4348%	63.8037%
体育	219	231	92.2080%	92.0780%
艺术	124	125	75.4718%	96.0000%
政治	270	235	91.4163%	90.6383%
总正确率	1082	1081	86.8640%	

从实验结果可以看出,在文中所使用的方法中,以词组在文本中出现的频率作为特征向量中各维的值,其效果较以词组在文本中是否出现作为特征向量中各维的值有明显提高。从文本各类别的精确率和分类总正确率上可以看出,使用 CHI 方法降维,可以在特征向量维数较低时仍能取得较好分类效果,从而可以大大降低分类的时空开销。可见,将 CHI 方法与交叉覆盖算法结合可得到一个效果较好的文本分类器。

表 2　以词组出现频率作为向量值

类别	学习	测试	精确率	召回率
计算机	99	101	92.2330%	94.0594%
交通	99	115	87.4016%	96.5217%
教育	109	111	91.1504%	92.7928%
经济	162	163	93.2836%	76.6871%
体育	219	231	95.5947%	93.9394%
艺术	124	125	93.7985%	96.8000%
政治	270	235	89.5161%	94.4681%
总正确率	1082	1081	91.9519%	

参考文献:

[1] 刘丽珍,宋瀚涛.文本分类中的特征抽取[J].计算机工程,2004,30(4):14-15.

[2] 陈　涛,谢阳群.文本分类中的特征降维方法综述[J].情报学报,2005,24(6):600-604.

[3] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]//In:Proceeding of the 14th International Conference on Machine Learning(ICML'97). San Francisco: Morgan Kaufmann Publishers,1997:412-420.

[4] 张　铃,张　钹.M-P 神经元模型的几何意义及其应用[J].软件学报,1998,9(5):334-338.

[5] 张　铃,张　钹,殷海风.多层前向网络的交叉覆盖算法[J].软件学报,1999,10(7):737-742.

[6] 吴　涛,张燕平,张　铃.前向神经网络交叉覆盖算法的一种改进[J].微机发展,2003,13(3):50-52.

[7] 程泽凯,林士敏.文本分类器准确性评估方法[J].情报学报,2004,23(5):631-636.

(上接第 112 页)

5　结　语

文中将 Hilbert R-树构造方法与 k-min 聚类技术相结合,提出了一种先对数据采用 k-min 聚类方法进行聚类,然后分别对 k 个聚类采用 Hilbert R-树构造方法构造 R-树的新的压缩算法——HilCluster。经分析得出该压缩算法时间消耗低、存储利用率高,而且查询效率高;实验数据进一步表明,无论是点查询性能还是区域查询性能,由 HilCluster 算法构造的 R-树比前面提到的两种压缩算法都有较大优势,尤其是在数据分布不均匀的情况下,优势更加明显。

事实上,Hilbert R-树构造方法本身也属于一种聚类方法,文中就是将两种聚类方法有机结合起来应用于 R-树的构造过程。然而 k-min 方法需要人为地给出一个参数(k 值),将数据集合分为 k 个聚类,这也不一定完全符合实际数据分布的特点,因此引进一种能自动聚类的聚类算法很有必要。

参考文献:

[1] Kamel I, Faloutsos C. On packing R-trees[C]//In: Proceedings of CIKM. Washington, DC, USA: [s. n.], 1993: 490-499.

[2] 王　颖,汪晓岩.基于递归聚类索引树的剪枝相似检索算法[J].合肥工业大学学报:自然科学版,2000,23(4):597-600.

[3] 黄继先,鲍光淑.一种面向 GIS 的静态 R-树数据组织方法[J].中南大学学报:自然科学版,2005,36(3):491-495.

[4] 谢昆青.空间数据库[M].北京:机械工业出版社,2004:118-123.

[5] 张明波,陆　锋.R 树家族的演变和发展[J].计算机学报,2005,28(3):289-300.