

一种基于贪婪覆盖的文本分类方法

张燕平,徐庆鹏,苏守宝,邢 猛

(安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要:文本分类是信息检索和数据挖掘中的重要主题之一。文中提出了一种基于贪婪覆盖算法的文本分类方法,首先对文本进行分词,分词的结果用 CHI 统计量的方法提取特征,使用 TF-IDF-ICSD 进行特征权重计算。对贪婪覆盖算法采用另一种选取初始点的方法来构建分类器,用复旦大学语料库作为测试数据集,并与 BP 算法相比较。实验结果表明文本提出的方法是有效的。

关键词:文本分类;CHI 统计量;TF-IDF-ICSD;贪婪覆盖算法

中图分类号:TP311.5

文献标识码:A

文章编号:1673-629X(2009)01-0074-03

A Text Categorization Method Based on Greedy Cover

ZHANG Yan-ping, XU Qing-peng, SU Shou-bao, XING Meng

(Ministry of Education Key Lab. of Intelligent Computing & Signal Processing,
Anhui University, Hefei 230039, China)

Abstract: Text classification is one of the key topics in information retrieval and data mining. A new text categorization technique based on greedy cover algorithm (GCA) was presented in this paper. The method can be conducted as following, text segmentation, feature extraction using CHI statistic, calculating feature weighting with TF-IDF-ICSD, constructing classifier for GCA by employing another initial point. The proposed method was experimented on some test dataset taken from the Corpus of Fudan University. The test results show that the proposed method is feasible and effective compared to BP neural network algorithm.

Key words: text classification; CHI statistic; TF-IDF-ICSD; greedy cover algorithm

0 引 言

文本分类是信息检索和数据挖掘中的重要主题之一,被广泛应用于多个领域:信息检索、搜索引擎、文本数据库、数字化图书馆等。因此,对文本分类技术的研究具有现实的意义。文本自动分类,目前已有许多成熟的方法。文献[1]对一些常见的分类算法作了讨论,支持向量机作为文本自动分类方法被广泛应用,它的主要优点是将降维和分类两个问题集中处理,且训练速度与 Rocchio 算法相当;神经网络自 1995 年应用于文本自动分类之后,发展迅速,典型的有 BP 算法,但在处理海量数据时,时间开销过大;K 近邻算法(KNN)是一种基于实例的文本分类方法,对于一个待分类文本,计算它与训练样本集中每个文本的文本相

似度,根据文本相似度找出 K 个最相似的训练文本,它作为一种常用的算法,在许多领域都显示出良好的性能,然而,在文本分类中,KNN 的一个弱点是它分类时的计算量较大,当它为一个未知实例分类时,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。此外,还有文本聚类的方法,如:群体智能的 Web 文档聚类算法[2]。

覆盖算法[3]提出以来,被应用于股票预测[4]、文本分类[5]、图像识别[6]等方面,也有许多改进研究[7,8]。覆盖算法初始点的选取在该算法中具有非常重要的地位,受贪婪式覆盖算法[9]的启发,给出了另一种初始点的选择方法,并将贪婪覆盖算法应用于文本分类,取得了较好的实验效果。

1 预处理

分类过程中,计算机无法直接处理文本信息,预处理时将文本表示成可供计算机处理的形式。文中使用由 Salton 提出的向量空间法,即将文本信息以向量的形式表示为: $D_i = (t_1, w_1; t_2, w_2; \dots; t_n, w_n)$,其中 D_i

收稿日期:2008-04-28

基金项目:国家重点基础研究 973 计划资助项目(2004CB318108; 2007CB311003);国家自然科学基金资助项目(60675031)

作者简介:张燕平(1962-),女,教授,硕士生导师,研究方向为人工智能、神经网络、机器学习及应用;苏守宝,博士,副教授,研究方向为群智能与模式识别。

为某一文本, t_i 为有意义的特征词或词组, w_i 为特征词或词组对应的权重, n 表示特征项向量空间的维数。把文本转化为向量形式, 首先要对其进行分词, 文中分词程序使用中国科学院计算技术研究所的汉语词法分析系统 ICTCLAS3.0^[10]。分词后得到大量的词组, 但这些词组会导致分类器的运算时间过长。另外, 不同词组对分类的影响程度也是不同的, 因此需要采用合适的特征选择算法选出对分类最有用的特征词集。

2 特征选择

特征选择的目的, 就是选出最能代表某篇文章或某类的特征词或词组, 以达到用较少的特征词来表示某类文本。在文本分类中, 特征选择的方法主要有: 信息增益 (Information Gain), 互信息 (Multi-Information), 特征频度 (Term Frequency), 特征熵 (Term Entropy), 文档频度 (Document Frequency), χ^2 统计量 (CHI), 几率比 (Odds Ratio) 等。文中使用 χ^2 统计量, 文献[11]中指出当 χ^2 统计量对于特征维数较低时, 有很好的效果。 χ^2 统计公式为:

$$\chi^2(t_i, c_j) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (1)$$

其中: N 表示文本总数; A 表示 t_i 和 c_j 同时出现的文本个数; B 表示 t_i 出现但 c_j 不出现的文本个数; C 表示 t_i 不出现但 c_j 出现的文本个数; D 表示 t_i 和 c_j 都不出现的文本个数。这样计算某个特征词可能同时出现在几个类中, 为使其应用于多类中, 一种方法是取其均值, 另一种方法是取其最大值。文中使用后者, 如式

$$\chi^2(t_i) = \max_{j=1}^m \{ \chi^2(t_i, c_j) \} \quad (2)$$

计算完成后, 可根据实验需要, 选取一定数量的特征项放入特征词库中, 以便于进一步对其处理。

3 特征权重计算

特征权重计算算法有多种, 各有优劣。文中使用一种改进型的 TF-IDF, 来计算特征词的权重, 即 TF-IDF-ICSD^[12]。文献[12]指出, 特征项的类间分布信息用下面的公式来表示:

$$\text{ICSD}(t_i) = \frac{\sqrt{\sum_j [tf(t_i, \bar{c}_j) - \frac{\sum_j tf(t_i, \bar{c}_j)}{NC}]^2 / NC}}{\sum_j tf(t_i, \bar{c}_j)} \quad (3)$$

其中, NC 是类别个数, j 的取值范围是 $(1, 2, \dots, NC)$ 。 $tf(t_i, \bar{c}_j)$ 表示为第 i 个特征项 t_i 在第 j 类 \bar{c}_j 上的平均词

频。

$$tf(t_i, \bar{c}_j) = \frac{\sum_k \omega_{ijk}}{|\bar{c}_j|} \quad (4)$$

其中, ω_{ijk} 是特征项 t_i 在 \bar{c}_j 类中第 k 篇文档中的词频, k 的取值范围是: $(1, 2, \dots, |\bar{c}_j|)$ 。

结合 TF-IDF 算法, 可得 TF-IDF-ICSD 的计算公式:

$$w(t_i, \bar{c}_j)_{\text{icsd}} = \frac{tf(t_i, \bar{c}_j) \times \log(N/n_i + L)}{\sqrt{\sum_{t \in \bar{c}} [tf(t, \bar{c}) \times \log(N/n_i + L)]^2}} \times \text{ICSD}(t_i) \quad (5)$$

由上式可知, 当 $w(t_i, \bar{c}_j)_{\text{icsd}}$ 值越小时, 该特征项的分类能力越弱。

4 交叉覆盖算法

根据 M-P 神经元的几何意义^[13], 提出的多层前向网络的交叉覆盖设计算法^[3]针对学习样本的特征构造神经网络。它的主要思想是先求一个领域覆盖 c_1 , 它只覆盖 k_1 中的点, 而不覆盖其它不属于 k_1 的点, 然后将被 c_1 覆盖的点删去。对余下的点求另一领域覆盖 c_2 , 它只覆盖 k_2 的点, 然后将被 c_2 覆盖的点删去, ……如此交叉进行覆盖, 直到所有的点全部被删除为止。

设学习样本共有 N 类, 记为: $X = \{X_1, X_2, \dots, X_N\}$ 。则构造第 k 类学习样本的球形领域的方法是: 在第 k 类点中任取一点 a_i , 设 a_i 到最近的异类点之间的距离为 d_1 , a_i 到最远且距离小于 d_1 的同类点之间的距离为 d_2 , 可得覆盖领域的半径为 $r = (d_1 + d_2)/2$, 覆盖中心为 a_i 。此外, 可通过求覆盖领域的重心或平移来调整覆盖中心使之可以覆盖更多的样本点, 按照这样的方法可求出样本的全部覆盖。

识别方法: 给定一个样本, 若它被某类覆盖领域所覆盖, 即可确定其类别, 否则若它不属于任何类别覆盖的覆盖领域时, 则按就近原则确定其类别。

5 贪婪式覆盖算法

由交叉覆盖算法可知, 在求每一个覆盖时, 初始点的选择不同, 覆盖的结果也不同。当然测试的准确性也不相同。要想使覆盖领域覆盖更多的点, 即覆盖个数较少, 关键之一就是寻找覆盖的中心点和下一覆盖的初始点。文献[9]中, 作者针对初始点的选择弱点加以改进, 取得了很好的效果。文中将用另一种方法对贪婪覆盖算法 (Greedy Cover Algorithm, GCA) 求第一个覆盖时的初始点, 并将其应用于文本分类。具体算法参考文献[3, 13], 此处省去了求平移的过程, 文中所

用贪婪覆盖算法如下:

(1) 求第一个覆盖。

1) 求每类的重心点 $a_i (i = 1, 2, \dots, n, n$ 为某类样本个数)。若 a_i 非某样本点, 则以 a_i 最近的同类点为中心求覆盖, 记为 $C_j (j = 1, 2, \dots, t, t$ 为类别个数)。

2) 求得 C_{\max} 作为第一个覆盖, C_{\max} 为 C_j 中覆盖样本个数最多的一个覆盖。

(2) 先以遇到的最后一个异类样本点为中心求新的覆盖, 记为 C_1 。

(3) 再依次从已有的覆盖中心构建新的覆盖 C_2 , 如 C_2 比 C_1 覆盖点数量多, 则用 C_2 代替 C_1 。

(4) 由(2)、(3)步中选出覆盖个数最多的一个覆盖, 并保存到相关数据(如覆盖中心点、特征词及权重、半径和类型等)。

(5) 在数据集中删除选中覆盖包含的点。

(6) 返回第(2)步, 反复执行, 直到所有样本点都覆盖完毕。

注: ① 由于学习样本过多, 所求覆盖也越来越多, 在实验第(3)步中随机选取 15 个覆盖构建新的覆盖。

② 对拒识样本, 采用就近原则确定样本的类别。

6 实验和结论

从复旦大学提供的中文文本分类语料库中选取 Art, Space, Computer, Environment, Agriculture, Economy, Sports 等七类共 2370 篇文本, 随机抽取近似 1:1 的样本进行学习和测试。由于名词最能表现文本内容和类别, 初次选取名词、名动词、名形容词三种词性为特征词组, 然后去除单个词和 χ^2 统计结果为零的词, 共得到 55232 个特征项, 再分别从中取出 χ^2 统计值较高的 1500、2000 和 2500 个词组作为特征词库。分别对各个特征词库做 10 次实验, 取其均值。实验中发现当取 2000 维时效果最好, 正确率可达 89%, 而 1500 维是 84%, 2500 维是 87%。实验结果与 BP 算法^[14]作了比较, 见表 1。

表 1 算法 GCA 与 BP 的实验比较
(BP 算法迭代次数为 2000)

测试数据集	学习	测试	查准率%		查全率%	
			BP	GCA	BP	GCA
Art	113	113	89.47	93.86	90.02	94.69
Space	128	128	70.67	88.03	83.41	80.47
Computer	133	132	79.14	85.27	70.00	82.71
Environment	262	263	78.74	86.59	77.99	86.59
Agriculture	195	195	88.77	90.10	81.08	88.71
Economy	119	119	87.43	86.99	76.84	89.92
Sports	236	236	80.72	89.96	88.62	94.92
总正确率	1185	1185	81.80	88.61		

文本分类的评价标准是一个重要指标, 它体现分类结果的优劣。文中使用两种^[15]标准来评价实验结果:

(1) 精度/查准率 (precision), 即分类器在某类别中做出的正确分类个数与分类器在该类别上做出的所有分类个数的百分比, 精度越高表明分类器在该类上出错的概率越小。

(2) 查全率/召回率 (recall), 即分类器在某类别中做出的正确分类个数与该类实际应有文本个数的百分比, 查全率越高表明分类器在该类上可能漏掉的分类越少。

实验结果如表 1 所示, 使用 BP 算法也能达到较好的效果, 由于要处理的数据量过大, 使其学习时间过长。而用 GCA 作为分类器, 学习时间短, 且具有良好的性能。贪婪覆盖算法追求用较少的覆盖, 来覆盖较多的样本点, 在某种程度上, 影响了分类精度。但由于覆盖个数少, 缩短了测试时间, 这正是贪婪覆盖算法的优点。实验结果表明, 将贪婪覆盖算法应用于文本分类是可行的。

参考文献:

- [1] 张雪英. 基于机器学习的文本自动分类研究进展[J]. 情报学报, 2006, 25(6): 730-739.
- [2] 吴 斌, 傅伟鹏, 郑 毅, 等. 一种基于群体智能的 Web 文档聚类算法[J]. 计算机研究与发展, 2002, 39(11): 1429-1435.
- [3] 张 铃, 张 钺, 殷海风. 多层前向网络的交叉覆盖设计算法[J]. 软件学报, 1999, 10(7): 737-742.
- [4] 张燕平, 张 铃, 吴 涛, 等. 基于覆盖的构造性学习算法 (SLA) 及在股票预测中的应用[J]. 计算机研究与发展, 2004, 41(6): 979-984.
- [5] 王倩倩, 段 震, 张燕平. 基于交叉覆盖算法的文本分类[J]. 计算机技术与发展, 2007, 17(6): 113-115.
- [6] 张燕平, 张 铃, 段 震. 构造性核覆盖算法在图像识别中的应用[J]. 中国图像图形学报, 2004, 9(11): 1304-1308.
- [7] 吴 涛, 张 铃, 张燕平. 机器学习中的核覆盖算法[J]. 计算机学报, 2005, 28(8): 1295-1301.
- [8] 赵 姝, 张燕平, 张 媛, 等. 基于交叉覆盖算法的改进算法[J]. 微机发展(现更名为: 计算机技术与发展), 2004, 14(11): 1-3.
- [9] 宋 杰, 程家兴, 许中卫, 等. 一种改进的贪婪式覆盖算法[J]. 计算机技术与发展, 2006, 16(8): 113-115.
- [10] 中科院汉语词法分析系统[EB/OL]. 2008-03-18. <http://www.i3s.ac.cn>.
- [11] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]//In: Proceeding of the 14th In-

(下转第 80 页)

则的纹理区域,能得到理想的结果,而对于不规则的背景(例如树叶),恢复的结果一般,但图像的清晰度明显优于双线性内插结果和 MAP 重建结果。

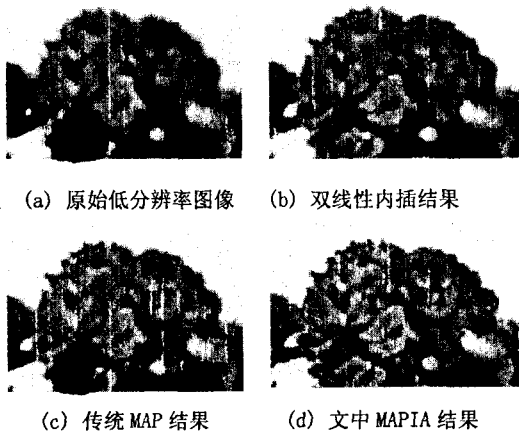


图 1 无高斯噪声的图像重建结果

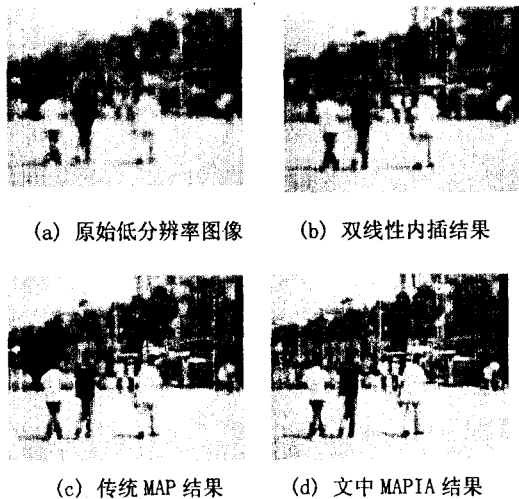


图 2 有高斯噪声的图像重建结果

5 结束语

在最大后验概率算法的基础上,结合图像类推的特点,提出了能适应序列图像的超分辨率重建的 MAPIA 算法,通过实验证明文中方法重建图像相对于双线性内插和传统的 MAP 的重建图像峰值信噪比值有明显的提高,图像边缘信息更加清晰丰富,细节信息也比较突出,从而改变了图像类推技术目前只用于单

幅图像处理的局限。同时也存在着不足之处。一方面,文中 MAPIA 方法是在传统 MAP 的框架内进行的超分辨率图像处理,MAP 是基于模型的方式,模型方式要求图像特征满足特定的假设条件,这些条件与真实图像会有差异,这种差异又会被带入训练图像对之间关系的学习中,以致生成的目标图像个别地方会产生一定的人工痕迹;此外,由于 MAPIA 算法是两种处理方法的结合,相对于一般方法的运算量相对还是比较大的,下一步的研究工作是如何减少 MAPIA 算法的运算量。

参考文献:

- [1] Tsai R Y, Huang T S. Multiframe image restoration and registration[J]. Advances in Computer Vision and Image Processing, 1984, 1: 317 - 339.
- [2] Stark G, Oskoui P. High-resolution image recovery from image plane arrays, using convex projection[J]. Journal of Optical Society of America (a Series A), 1989, 6(11): 1715 - 1726.
- [3] Sezan M I, Stark H. Image restoration by the method of convex projections: part 2 - applications, and numerical results [J]. IEEE Transactions on Medical Imaging, 1982(12): 95 - 101.
- [4] Schultz R R, Stevenson R L. A Bayesian approach to image expansion for improved definition[J]. IEEE Transactions on Image Processing, 1994(33): 233 - 242.
- [5] Borman S, Stevenson R L. Simultaneous multiframe MAP super-resolution video enhancement using spatiotemporal priors [C] // IEEE Int. Conf. Image Processing, Kobe, Japan: [s. n.], 1999: 469 - 473.
- [6] Hertzmann A, Jacobs C E, Oliver N, et al. Image analogies [C] // In: Proc. of the 28th Annual Conf. on Computer Graphics and Interactive Techniques SIGGRAPH 2001. LA California: ACM Press, 2001: 327 - 340.
- [7] 沈海. 一种基于类推思想的图像分割方法[J]. 计算机工程与应用, 2006, 42(9): 45 - 47.
- [8] 古元亨, 吴恩华. 基于图像类推的超分辨率技术[J]. 软件学报, 2008, 19(4): 994 - 1003.
- [9] 刘晓天. 基于 MAP 技术的图像超分辨率复原研究与实现 [D]. 长沙: 国防科学技术大学, 2004.

(上接第 76 页)

- ternational Conference on Machine Learning (ICML'97). San Francisco: Morgan Kaufmann Publishers, 1997: 412 - 420.
- [12] 苏力华. 基于向量空间模型的文本分类技术研究[D]. 西安: 西安电子科技大学, 2006.
- [13] 张铃, 张钺. 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.

- [14] 韩力群. 神经网络教程[M]. 北京: 北京邮电大学出版社, 2007: 59 - 78.
- [15] Yang Yiming, Liu Xin. A re-examination of text categorization methods[C] // Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99). New York, USA: ACM Press, 1999: 42 - 49.