一种基于RBF神经网络的XML文本分类方法

刘 锋,唐 佳,仲 红
(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:为了快速有效地组织和分析海量的XML文本信息,XML文本的自动分类必不可少。文中提出了一种基于RBF神经网络的分类方法,并运用改进型的CHI统计量方法进行特征提取,对传统的加权公式进行了一些改进,再运用资源优化神经网络(RON)进行训练,做了必要的实验分析。实验结果表明该分离器有较高的分类质量,提高了分类的效率,有较高的分类准确性,满足了XML文本自动分类的要求。

关键词:XML文本分类;CHI统计量;RON;RBF神经网络
中图分类号:TP311.5 **文献标识码**:A **文章编号**:1673-629X(2009)08-0034-03

A Text Categorization Method Based on RBF Neural Network

LIU Feng, TANG Jia, ZHONG Hong
(School of Computer Science & Technology, Anhui University, Hefei 230039, China)

Abstract: In order to organize and analyse the mass information of the XML text messages quickly and efficiently, the automatic classification of XML text is essential. Proposes an classification method based on RBF neural network, and uses an improved CHI statistic methods to extract feature, improves the traditional weighted formula. The results show that the separator has a higher quality of the classification, to improve the categorizing effectiveness and to meet the XML text categorization.

Key words: XML classification; CHI statistic; RON; RBF; neural network

0 引 言

XML(eXtensible Markup Language,可扩展标记语言)是一套定义语义标记的规范,其目标是能够定义计算机和人都方便识别的数据类型,它继承了SGML自定义标记的优点,具有更好的数据描述性。XML文档所特有的良好数据存储格式、强大的可扩展性以及更好的信息搜索效果等优点都使得XML文档得到越来越多的重视,被作为一种新的格式文件广泛地使用[1]。目前,符合XML规范的数据已经大量存在于当前的信息社会里,尤其是电子商务、Web服务、数字图书馆等。对于如此海量信息的有效管理也随之成为研究的热点。

文中就基于RBF神经网络算法[2],并根据理论描述使用C++语言构建了一种神经网络分类器。实验表明,神经网络分离器可以取得优良的分类效果。

收稿日期:2008-12-03;**修回日期**:2009-03-02
基金项目:国家自然科学基金(60773114);安徽省重点自然科学研究项目(KJ2007A43)
作者简介:刘 锋(1962-),男,博士,教授,研究方向为软件工程、并行分布计算、计算机网络。

1 XML文本提取

文献[3]使用了Xpath查询语言实现对XML文档的查询,Xpath使用位置路径表达式,用轴指出节点搜索方向,支持通配符,提供运算符和"与"和"或"等表达式及一些函数,在XML文档中可以快速准确地查找到所需信息。

2 文本预处理

2.1 中文分词

由于中文信息没有特定的词组边界[4],因此从特定文本里提取有效的关键词就较英文资料困难许多。文中采用中国科学院计算技术研究所的汉语词法分析系统ICTCLAS3.0[5]进行分词。

2.2 特征提取

因为文本分类数据量很大[6],为了减少特征项即文本维数和减少影响文本分类的噪音,需要进行特征提取,提取出对文本分类贡献大的特征项,同时删除对文本分类贡献小的特征项。特征提取就是提取出最能代表某篇文章或某类的特征项,以达到降维的效果而减少文本分类的计算量[7]。文献[8,9]中介绍了一些典型特征提取方法。文中采取文档频度(Document

Frequency) 的方法[10]:根据出现词条的文档数量的多少来选取特征词;使用了 $\chi 2$ 统计量(CHI)。

$$x^2(t,c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$

其中:N 表示文本总数;A 表示 t 和 c 同时出现的文本个数;B 表示 t 出现但 c 不出现的文本个数;C 表示 t 不出现但 c 出现的文本个数;D 表示 t 和 c 都不出现的文本个数。

文中使用文献[9] 中改进型的 CHI 方法:

$$x^2(t,c)_{new} = \log(N/N_{ft}) \times \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (1)$$

其中:N、A、B、C、D 的含义同 $\chi 2$,N_{ft} 为出现 t 的文档数。它基于如下的假设:如果词条出现的文档数接近训练集中所有的文档数时,即 $N_{ft} \to N$ 时 $\log(N/N_{ft}) \to 0$,此类词条应该过滤掉,并且适当地提高了低频词的权重。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。文中使用前者,即:

$$x^2(t) = \frac{\sum_{i=1}^{m} x^2(t,c)}{m} \quad (2)$$

计算出所有特征词的统计值后,按从大到小进行排序,然后根据需要从上到下选取一定数量的特征词构建文本分类的特征词库。

3　文本表示

计算机不能直接处理文本文件[11],特征项的权值计算之后,要把文本文档进行向量化。即将文本信息以向量的形式表示为:

$D_i = (t_1, w_1; t_2, w_2; \cdots; t_n, w_n)$,其中 D_i 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。特征权重计算算法有多种,各有优劣。文中使用一个比较常见的 TF - IDF 公式:

$$w(t_i, \bar{c}_j) = \frac{tf(t_i, \bar{c}_j) \times \log(N/n_i + 0.1)}{\sqrt{\sum_{i \in \bar{c}} [tf(t_i, \bar{c}_j]) \times \log(N/n_i + 0.1)]^2}}$$

其中,NC 是类别个数,j 的取值范围是,$(1,2,\cdots, NC)$,N 为所有文档数目,n_i 为含有词 t_i 条的文档数目。$tf(t_i, \bar{c}_j)$ 表示为第 i 个特征项 t_i 在第 j 类 $\bar{c}_j$ 上的平均词频。

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

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:$(1,2,\cdots, |\bar{c}_j|)$。根据 TF - IDF 公式,文档集中包含某一词条的文档越多,说明它区分文档类别属性的能力越低,其权值越小;另一方面,某一文档中某一词条出现的频率越高,说明它区分文档内容属性的能力越强,其权值越大。

4　分类算法

常见的归纳学习算法有[12]朴素贝叶斯、贝叶斯网络、k 近邻、神经网络、决策树、决策规则、支持向量机等方法。选用 k 近邻和 RBF 神经网络两种方法作为比较,实验表明两种方法都是行之有效的。

4.1　k 近邻算法

该算法的基本思路是[13]:在给定新文本后,考虑在训练文本集中与该新文本距离最近(最相似)的 K 篇文本,根据这 K 篇文本所属的类别判定新文本所属的类别,具体的算法步骤如下:

1) 根据特征项集合重新描述训练文本向量。

2) 在新文本到达后,根据特征词分词新文本,确定新文本的向量表示。

3) 在训练文本集中选出与新文本最相似的 K 个文本,计算公式为:

$$\mathrm{Sim}(d_i, d_j) = \frac{\sum_{K=1}^{M} W_{iK} \times W_{jK}}{\sqrt{(\sum_{K=1}^{M} W_{iK}^2)(\sum_{K=2}^{M} W_{jK}^2)}}$$

其中 M 为特征词总数。K 值的确定目前没有很好的方法,一般采用先定一个初始值,然后根据实验测试的结果调整 K 值。

4) 在新文本的 K 个邻居中,依次计算每类的权重,计算公式如下:

$$P(\boldsymbol{x}, C_j) = \sum_{d \in \mathrm{KNN}} \mathrm{Sim}(\boldsymbol{x}, \boldsymbol{d}_i) y(\boldsymbol{d}_i, C_j)$$

其中,$\boldsymbol{x}$ 为新文本的特征向量,$\mathrm{Sim}(\boldsymbol{x}, \boldsymbol{d}_i)$ 为相似度计算公式,与上一步骤的计算公式相同,而 $y(\boldsymbol{x}, C_i)$ 为类别属性函数,即,如果 $\boldsymbol{d}_i$ 属于类 C_i,那么函数值为 1,否则为 0。

5) 比较类的权重,将文本分到权重最大的那个类别中。

4.2　径向基函数神经网络

径向基函数[14](Radial Basis Function, RBF)神经网络是由 J. Moody 和 C. Darken 于 20 世纪 80 年代末提出的。最基本的径向基函数(RBF)神经网络的构成包括 3 层,其中每一层都有着完全不同的作用[15]。输入层由一些感知单元组成,它们将网络与外界环境连接起来;第二层是网络中仅有的一个隐层,它的作用是

从输入空间到隐层空间之间进行非线性变换，在大多数情况下，隐层空间有较高的维数；输出层是线性的，它为作用于输入层的激活模式提供响应。由于径向基函数网络模拟了人脑中局部调整、相互覆盖接受域(或称感受域，Receptive Field)的神经网络结构，因此，RBF网络是一种局部逼近网络，现已证明它能以任意精度逼近任一连续函数[16]。建立RBF神经网络模型的关键在于2个方面[8]：

(1)RBF神经网络隐层数据中心确定；

(2)输出权值 w_{jk} 的学习调整。

5 实验和结论

在一个具有400篇中文文本的语料库上测试系统实现的分类算法，并对其效率和结果进行比较分析。语料库中的文本都是新闻电讯稿，采自新浪网，这些文本都有一些共性，文章不长而且具有一定的相似性。实验结果如表1所示。

表1 文中算法GCA与k近邻的实验比较

算法	学习	测试	查准率	查全率
k近邻	70%	30%	0.924	0.922
GCA	70%	30%	0.939	0.947

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。文中使用两种标准来评价实验结果[9]：

(1)精度/查准率(precision)：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)：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Knn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RON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] Meng X F，Zhou L X，Wang S. State of trends in database research[J]. Journal of Software，2004，15(12)：1822－1836.

[3] 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.

[4] 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.

[5] 中科院汉语词法分析系统[EB/OL].2008－03－18. http://www.i3s.ac.cn.

[6] 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.

[7] 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.

[8] 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].中国图象图形学报，2004，9(11)：1304－1308.

[9] 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.

[10] Yang Yiming，Liu Xin. A Re－examination of Text Categorization Methods[C]//Proceedings of SIGIR－99，22nd ACM International Conference on Research and Development in Information Retrieval. [s.l.]：[s.n.]，1999：42－49.

[11] 张 铃，张 钹，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.

[12] 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.

[13] 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.

[14] 韩力群.人工神经网络教程[M].北京：北京邮电大学出版社，2007：59－78.

[15] 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.

[16] 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：计算机技术与发展)，2004，14(11)：1－3.