有限混合模型在文本分类中的应用研究

周 瑛[1,2], 张 铃[2]
(1. 南京大学,江苏 南京 210093;
2. 安徽大学,安徽 合肥 230039)

摘 要:通过对覆盖算法(CA)结果的分析,将覆盖某一类样本的每个覆盖看成一个 Gauss 分布,利用有限混合模型的极大似然拟合,用期望最大化算法(EM 算法)来对覆盖算法进行优化处理。算法的迭代过程,就是不断调整各覆盖的中心、"半径"以及其线性组合系数,逐渐趋向最优解的过程。目的是为了提高覆盖算法的精度。应用于文本分类的实验证明,通过 EM 方法对均值、方差和线性组合系数进行迭代计算,将所求得的参数用于测试时所得到的平均精度都高于原覆盖算法的最高分类精度以及 SVM 处理同类数据的分类精度。

关键词:有限混合模型;EM 算法;覆盖算法;文本分类
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2010)06-0018-03

Study of Application of Finite Mixture Model in Text Classification

ZHOU Ying[1,2], ZHANG Ling[2]
(1. Nanjing University, Nanjing 210093, China;
2. Anhui University, Hefei 230039, China)

Abstract: Analyze the results of Cover Algorithm. It considers every coverage which is included in one class of samples as a Gauss distribution. Then, with the help of maximum likelihood estimation of finite mixture of models, one could optimize the Cover Algorithm with the expectation maximization algorithm (EM algorithm). The process of the iterative algorithm is the optimized process that adjusts continuously the center, radius of every coverage and their linear combined coefficient. The aim is to improve the examination precision of Cover Algorithm. Such a model has been used on text classification database and their results have shown that the new parameters, which have been got through the iterative calculation of the mean value, square deviation and the linear combined coefficients by EM algorithm, have got the higher examination precision than the precisions of the original Cover Algorithm and SVM in processing the same database.

Key words: finite mixture model; EM algorithm; cover algorithm; text classification

0 引 言

张铃教授利用 M-P 神经元的新的几何意义,提出了一种前向神经网络的新的学习算法——覆盖算法[1,2]。该算法的主要思想是构造一个网络,使得对于给定的样本集进行分类等价于求出一组领域,对所给定的样本集中的点,能按分类的要求用所覆盖的领域将它们分隔开来。根据这个思想,算法首先将原空间的样本点向高维空间投影。在投影后,每个样本点都落在一个超球面上,再根据投影后的位置来构造神经网络。这种方法可迅速地、构造性地得到对于训练数据几乎完全正确分类的神经网络,而不必像传统的 BP 算法那样反复地进行迭代训练而未必会有好的结果。该算法被成功地应用在金融预测、模式识别、手写汉字、文本分类、网络上图像检索等问题中[3,4]。文中通过对覆盖算法(CA)结果的分析,将覆盖某一类样本的每个覆盖看成一个 Gauss 分布,利用混合模型的极大似然拟合,用期望最大化算法来对覆盖算法进行优化处理,目的是为了提高覆盖算法的精度。文中具体到 Gauss 混合模型的极大似然拟合形式及 EM 算法实现,最后将该迭代算法应用在覆盖算法中。应用于文本分类的实验证明,用所求得的参数用于测试时所得到的平均精度都高于原覆盖算法的最高分类精度以及 SVM 处理同类数据的分类精度。

收稿日期:2009-09-30;**修回日期**:2009-12-31
基金项目:安徽省哲学社会科学规划基金(AHSKF07-08D13);安徽省人文社会科学研究基金(2009sk038)
作者简介:周 瑛(1968-),女,安徽无为人,教授,博士后,研究方向为模糊理论及应用、神经网络、信息检索;张 铃,教授,博士生导师,研究方向为人工智能理论、机器学习理论和方法、智能计算技术、神经网络技术等。

1　有限混合模型的极大似然拟合及其求解方法

1.1　有限混合模型的定义

设 $X_1,\cdots,X_n$ 是样本量为 n 的独立同分布(independent and identically distributed, i.i.d.)随机样本,其中 X_i 是 p 维随机变量,其概率密度函数为 $p(x_i)$。假设样本是这样产生的:先以概率 π_j 决定其所属类别 ω_j,接着根据概率密度 $p_j(x_i)$ 生成一个具体的样本。于是,对于一个给定的样本 X_i,其产生的概率可写成形式:

$$p(x_i;\Theta)=\sum_{j=1}^{c}p_j(x_j;\theta_j)\pi_j \tag{1}$$

其中 Θ 包含模型中的所有参数。

利用混合模型来分类,有两个基本的问题需要解决:一是混合模型的参数估计;另一个是混合模型的分量数的确定。前者可采用期望最大(expectation maximization, EM)算法;一般情况下,后者的处理较为复杂,但在实验中,分量数通过基本的覆盖算法已经确定。

1.2　混合模型的极大似然拟合

虽然估计混合密度的参数的方法有多种,如矩方法、ML 方法和 Bayes 方法,但一般很难得到参数估计的解析公式。即使对 Gauss 混合密度,ML 方法也得不到混合比例、分量均值和方差的封闭形式解[5]。而在 Dempster et al. 的 EM 算法提出之后,ML 方法发展迅速,现已成为混合模型拟合的主流方法。

假设样本集 D 中有 n 个样本:$x_1,x_2,\cdots,x_n$。由于样本是独立抽取的,因此公式(2)成立:

$$p(D;\Theta)=\prod_{k=1}^{n}p(x_k;\Theta) \tag{2}$$

在公式(2)中,把 $p(D;\Theta)$ 看成是参数 Θ 的函数,称为样本 $x_1,x_2,\cdots,x_n$ 的似然函数(likelihood function),为方便分析,实际应用中总是使用似然函数的对数函数。参数向量 Θ 的极大似然估计,就是使 $\ln L(\Theta)$ 达到最大值的那个参数 $\hat{\Theta}$,称 $\hat{\Theta}$ 是 Θ 的极大似然估计(maximum likelihood estimator, MLE)。为了得到最大值,$\hat{\Theta}$ 必须满足的条件是似然函数对的梯度必须为零,即 $\hat{\Theta}$ 是方程(组)$\partial\ln L(\Theta)/\partial\Theta=0$ 的解。

现在考虑混合密度的对数似然函数方程组的解。设 $X_1,\cdots,X_n$ 是独立地抽自式(1),对数似然函数为:

$$\ln L(\Theta)=\sum_{i=1}^{n}\ln p(x_i;\Theta)=\sum_{i=1}^{n}\ln\Big(\sum_{j=1}^{c}p_j(x_i;\theta_j)\pi_j\Big) \tag{3}$$

直接对 $l(\Theta)$ 关于 Θ 求微商并令其为 0,通过运算可以发现 Θ 的极大似然估计值 $\hat{\Theta}$ 必须满足:

$$\sum_{i=1}^{n}\beta_j(x_i;\hat{\Theta})(\partial\ln p_j(x_i;\theta_j)/\partial\theta_j)\mid_{\theta_j=\hat{\theta}_j}=0 \tag{4}$$

及

$$\hat{\pi}_j=\frac{1}{n}\sum_{i=1}^{n}\beta_j(x_i;\hat{\Theta}) \tag{5}$$

对 $j=1,\cdots,c$ 成立,其中令

$$\beta_j(x_i;\hat{\Theta})=\frac{\hat{\pi}_j p_j(x_j;\hat{\theta}_j)}{\sum_{l=1}^{c}\hat{\pi}_l p_l(x_i;\hat{\theta}_l)} \tag{6}$$

为 x_i 抽自第 j 个分量的后验概率(posterior probability)。当分量密度具体指定时,方程组(4)、(5)和(6)可用迭代方法计算,即给定 Θ 的一个初始值 Θ^0,代入方程组(6)的右边,再通过(4)、(5)式计算出 Θ 的一个新估计值 Θ^1,它又可代入(6)的右边,如此直到收敛。

2　应用 EM 算法对覆盖算法进行优化

将原覆盖算法中覆盖第 i 类样本点的 n_i 个覆盖中的每一个覆盖看成一个 Gauss 分布,每个覆盖的中心和覆盖半径分别看成是 Gauss 分布的均值和方差,而第 i 类的样本分布则可用这 n_i 个 Gauss 分布的混合模型进行模拟。为简单起见,只考虑这 n_i 个 Gauss 分布具有对角协方差矩阵 $\Sigma_j=\sigma_j^2 I$ 的情况。通过一定数量的训练样本来估计每个正态分布的参数及混合比例,训练的目的是找出对训练样本来说使误差最小的参数,即对分类器进行优化。训练结束后,再将测试样本代入该混合模型,从而计算出每个测试样本属于某一类别的概率,概率最大者,即为该类。因此,第 i 类的判别函数可变为:

$$F_i(x)=\sum_{j=1}^{g}\alpha_j\cdot\frac{1}{(2\pi\sigma_j^2)^{p/2}}\exp\Big(-\frac{(x-\mu_j)^2}{\sigma_j^2}\Big) \tag{7}$$

设已求出对第 i 类点的覆盖组 $C=\{C_1,C_2,\cdots,C_g\}$,且 C_i 的中心为 a_i,半径为 r_i。将每个覆盖视为一个具有 Gauss 分布且参数未知的模型,则覆盖组所对应的有限混合概率模型为:

$$F_i(x)=\sum_j\alpha_j\cdot\varphi(x,\mu_j,\Sigma_j)\quad j=1,\cdots,g \tag{8}$$

其中 $\varphi(x,\mu_j,\Sigma_j)$ 是 p 维正态分布密度函数。

$$\varphi(x,\mu_j,\Sigma_j)=\frac{1}{(2\pi\sigma_j^2)^{p/2}}e^{\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right)} \tag{9}$$

下面利用极大似然拟合法进行拟合,具体算法如下:

算法 1 改进的覆盖算法 FMMCA(Finite Mixture Model Cover Algorithm)

(1) for $s=1$ to c (设样本共有 c 个类别,每个样本是 p 维向量)

(2) 取第 s 个类别的覆盖个数 g、第 j 个覆盖中的

样本个数 d_j 及覆盖中心 a_j、半径 r_j

(3) initialize $\alpha_j^{(0)} = \frac{d_j}{\sum_{j=1}^{g} d_j}$、$\mu_j^{(0)} = a_j$、$\sigma_j^{(0)} = r_j$、$F_s = 0$

(4) for $k = 1$ to m // m 为迭代次数

(5) $\varphi_{ij}^{(k-1)} = \frac{1}{(2\pi\sigma_j^{(k-1)^2})^{p/2}} e^{(-\frac{(x_i-\mu_j^{(k-1)})^2}{2\sigma_j^{(k-1)^2}})}$

(6) $\beta_{ij}^{(k-1)} = \frac{\alpha_j^{(k-1)} \cdot \varphi_{ij}^{(k-1)}}{\sum_{j=1}^{g} \alpha_j^{(k-1)} \cdot \varphi_{ij}^{(k-1)}}$

(7) $\alpha_j^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \beta_{ij}^{(k-1)}$ //n 为第 s 类的学习样本个数

(8) $\mu_j^{(k)} = \frac{\sum_{i=1}^{n} \beta_{ij}^{(k-1)} \cdot x_i}{\sum_{i=1}^{n} \beta_{ij}^{(k-1)}}$

(9) $(\sigma_j^2)^{(k)} = \frac{\sum_{i=1}^{n} \beta_{ij}^{(k-1)} \cdot |(x_i - \mu_j^{(k)})|^2}{\sum_{i=1}^{n} \beta_{ij}^{(k-1)}}$

(10) end

(11) 计算第 s 类的判别函数

$$F_s(x) = \sum_{j=1}^{g} \alpha_j \cdot \frac{1}{\sqrt{2\pi} \cdot \sigma_j} \exp(-\frac{(x-\mu_j)^2}{\sigma_j^2})$$

(12) end

3 实验数据及结果

文中的实验数据是从中文自然语言处理开放平台(http://www.nlp.org.cn)中下载的文本分类语料库，该语料库由中科院计算技术研究所提供。含有环境、计算机、交通、教育、经济、军事、体育、医药、艺术、政治 10 个类别共 2799 篇文章，训练语料和测试语料基本按照 1:1 的比例来划分，其中学习语料有 1370 篇文档，测试语料含 1429 篇文档。每篇文章先用分词程序(由中科院计算技术研究所提供的汉语词法分析系统 ICTCLAS)进行分词，再用统计程序(用 Java 语言编程)进行处理[6]。经过去除停用词、稀有词、统计名词和动词出现的频率的处理后，每篇文章可表示成向量形式。矩阵中的元素 d_{ij} 采用公式 (10) 来进行计算，d_{ij} 表示第 j 个文档中第 i 个词的权重，由于词条和文本的数量都很大且词的分布很广，而单个文本中出现的词比较有限，因此一般为稀疏矩阵[7]。d_{ij} 表示第 j 个文档中第 i 个词出现的频度，经过归一化处理后，d_{ij} 的值介于 0 和 1 之间。文档间的距离采用余弦距离来计算。

$$W(t,d) = \frac{tf(t,d) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t \in d} [tf(t,d) \times \log(N/n_t + 0.01)]^2}} \quad (10)$$

实验 1：对训练集中经过处理的 1370 篇文本，用覆盖算法 CA 来训练，得到分类器 1。用测试集中的经过处理的 1429 篇文本进行测试。共进行 5 次实验，取平均值作为最后的结果。

实验 2：对实验 1 中的训练集文本及训练结果用算法 FMMCA 进行优化(EM 的迭代次数取 20)，得到分类器 2。用实验 1 中相同的测试样本进行测试，共进行 5 次实验，取平均值作为最后的结果。实验结果及用 SVM 对同类数据的分类结果如表 1 所示。

表 1 不同分类器的分类正确率比较

类别	正确率%		
	分类器 1	分类器 2	SVM
环境	90.5	93.9	93.3
计算机	91.4	94.2	93.2
交通	95.3	97.5	96.4
教育	95.2	98.1	96.1
经济	93.1	95.9	91.9
军事	90.2	92.9	89.7
体育	92.3	94.6	92.2
医疗	93.6	96.3	95.8
艺术	96.8	98.4	96.7
政治	90.6	93.2	88.8

从表 1 中知道经过迭代后测试样本的平均正确率明显高于未处理的样本，而它们的学习、测试时间均无显著变化。与实验 1 分类器的最高精确度相比，经过迭代后，三个数据库的精度都高于它的最高分类精度以及 SVM 处理同类数据的精度[8]。

4 结束语

有限混合模型为广泛的随机现象进行统计建模提供了一种数学方法。将覆盖某一类样本的每个覆盖看成一个 Gauss 分布，利用混合模型的极大似然拟合，用期望最大化算法对覆盖算法进行优化处理。算法的迭代过程，就是不断调整各覆盖的中心、"半径"以及其线性组合系数，逐渐趋向最优解的过程。实验证明通过 EM 方法所求得的参数用于样本测试时所得到的平均精度都高于原覆盖算法的最高分类精度以及 SVM 处理同类数据的分类精度，进而验证了这种方法的优化作用。

参考文献：

[1] Zhang Ling. A Geometrical Representation of McCulloch-

(下转第 24 页)

对于初始特征向量用基于 SVM 和遗传算法的特征选择方法选出近似最优特征子集。计算并记录所有特征分量的分类权值。遗传算法中的个体由长度为 77 的二进制字符串表示，代表一个用户样本。初始种群规模 $N_0 = 50$，交叉概率 $P_c = 0.6$，变异概率 $P_m = 0.02$。对于适应度函数的参数 F 指标，在 992 个用户样本中随机抽取 800 个作为训练样本训练 SVM，剩下的 192 个样本作为测试样本，测试模型正确率与召回率，计算出 F 值。SVM 选用 C - SVM，核函数取 RBF 核函数，通过实验确定最优参数 $C = 25$，$\gamma = 0.3$。对于适应度函数的两个参数 C_1 以及 C_2，首先设置罚函数参数 $C_2 = 0$ 时的实验结果如表 1 所示。

表 1 C_1 取不同值时特征子集以及 F 值

C_1	特征子集维度	F - measure
0.05	37	0.9012
0.10	35	0.9276
0.15	28	0.9221
0.20	34	0.9102
0.25	31	0.9089

由此可见，当分类权值参数 C_1 取值为 0.1 时，分类准确率最高。F - measure 指标为 0.9276。取 $C_1 = 0.10$，SVM 核函数取 RBF 核函数，罚函数 C_2 在不同取值的情况下的实验结果如表 2 所示。

表 2 C_2 取不同值时特征子集以及 F 值

C_1	C_2	特征子集维度	F - measure
0.10	0.05	30	0.9356
0.10	0.10	24	0.9132
0.10	0.15	17	0.8821
0.10	0.20	12	0.8610

罚函数 C_2 取值为 0 时的最优 F 值为 0.9276。当 C_2 取值大于 0 时，遗传算法优先选择特征维度小的个体。当 $C_2 = 0.05$ 时，特征子集维度由 35 减少到 30，不但增加了支持向量机的运行效率，减少了系统运行时间，而且还具有最好的分类准确率，获得了更优的 F 值。但是随着 C_2 值的增大，特征数大大减少，虽然运行时间减少，但是严重影响了分类性能。

4 结束语

结合 SVM 和分类权值的评价准则提出一种基于遗传算法的特征选择，并将其应用于消费欺诈预警系统。理论和实验分析表明，基于该特征选择方法的预警系统有效地预测了恶意欠费用户，获得了令人满意的结果。

参考文献：

[1] 舒 宁，马洪超，孙和利. 模式识别的理论与方法[M]. 武汉：武汉大学出版社，2004.

[2] 董 梅，胡学钢. 基于多特征选择的中文文本分类[J]. 计算机技术与发展，2007，17(7)：117 - 119.

[3] Vapnik V N. The Nature of Statistical Learing Theory[M]. New York：Springer - Verlag，1995.

[4] 邓乃扬，田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京：科学出版社，2004.

[5] 柏海滨，李 俊. 基于支持向量机的入侵检测系统的研究[J]. 计算机技术与发展，2008，18(4)：137 - 139.

[6] 王 辉. 主成分分析及支持向量机在人脸识别中的应用[J]. 计算机技术与发展，2006，16(8)：24 - 26.

[7] Liu H，Setiono R. A probabilistic approach to feature selection filter solution[C]//In：Proceedings of International Conference on Machine Learning. [s. l.]：[s. n.]，1996：319 - 327.

[8] Liu H，Motoda H. Feature Selection for Knowledge Discovery and Data Mining[M]. [s. l.]：Kluwer Academic Publishers，1998.

[9] Goldberg D E. Genetic Algorithm in search，optimizing & machine learning[M]. USA：Addison - Wesley，1989.

[10] 王小平，曹立明. 遗传算法——理论、应用与软件实现[M]. 西安：西安交通大学出版社，2002.

(上接第 20 页)

Pitts Neural Model and Its Applications[J]. IEEE Trans, on Neural Networks，1999，10(4)：925 - 929.

[2] 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.

[3] 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.

[4] 周 瑛，张 铃. 基于概率的覆盖算法的研究[J]. 计算机技术与发展，2006，16(2)：29 - 31.

[5] Dempster A P，Laird N M，Rubin D B. Maximum likelihood from incomplete data using the EM algorithm(with discussion)[J]. J. R. Stat . Soc. Ser. B，1977，39：1 - 38.

[6] 周 瑛. 神经网络作为分类器的算法研究及在信息检索中的应用[D]. 合肥：安徽大学，2006.

[7] 苏新宁，杨建林. 数据挖掘理论与技术[M]. 北京：科学技术文献出版社，2003.

[8] 邹 涛，王继成，黄 源. 中文文档自动分类系统的设计与实现[J]. 中文信息学报，1999，13(3)：26 - 32.