

# 基于覆盖算法的煤炭供应商评测模型

胡光杰,张燕平,陈洁

(安徽大学 智能计算与信号处理重点实验室,安徽 合肥 230039)

**摘要:**阐述了当前进行煤炭供应商评测的方法以及这些方法的弊端。针对这种弊端,根据煤炭供应商评测的特点,利用前向神经网络的交叉覆盖算法及其改进算法对煤炭供应商供货质量进行了评测,在实验中将其与统计理论中加权平均的方法进行比较,证明取得了不错的效果,同时证明了核覆盖算法对交叉覆盖算法的改进。

**关键词:**供应商评测;交叉覆盖算法;核函数;核覆盖算法

**中图分类号:**TP182

**文献标识码:**A

**文章编号:**1673-629X(2007)01-0006-03

## Coal Provider Evaluation Based on Covering Algorithm

HU Guang-jie, ZHANG Yan-ping, CHEN Jie

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

**Abstract:** States disadvantages of traditional coal provider evaluation methods. In allusion to these disadvantages, and according to the characteristics of coal provider evaluation, uses the Alternative Covering Design Algorithm and its improved algorithm to evaluate the coal providers. In the experiment, compare the result with statistical theory methods. The experiment shows good effect of the evaluation, and also shows the improvement of the Alternative Covering Design Algorithm by Kernel Based Covering Algorithm.

**Key words:** provider evaluation; alternative covering design algorithm; kernel function; kernel based covering algorithm

### 0 引言

在火电厂能源供给上,煤炭供应商的选择占有重要的地位。因此,一种有效的煤炭供应商评测模型很有必要。目前,普遍只能以单个属性,如煤价或人为经验来对供应商进行评测,这样做虽有一定道理,但是并不能完全真实有效地反映供货商的供货能力。在对煤炭供应商评测时,更合理的方式应该是:综合考虑煤质、煤价和煤量等各方面因素进行评估,如煤质中的水分、灰分、挥发分、发热量、硫分,煤量中的途损、计划兑现率,以及运费、标煤单价等。

文献[1,2]根据神经元的几何意义提出了多层前向神经网络的交叉覆盖算法。该算法根据样本数据的结构,构造性地建立神经网络模型,在一定意义上解决了多层前向神经网络的设计问题。该算法已经广泛运用于金融预测、入侵检测、车牌识别、信号样式识别等方面。

### 1 相关知识

#### 1.1 交叉覆盖算法

根据 M-P 神经元的几何意义<sup>[1]</sup>,给定一输入集  $K$

$= \{x^1, x^2, \dots, x^k\}$ , 设  $K$  分为  $s$  个子集  $K^1 = \{x^1, x^2, \dots, x^{m(1)}\}, \dots, K^s = \{x^{m(s-1)+1}, x^{m(s-1)+2}, \dots, x^k\}$ , 用一个三层网络构造分类器,使对给定的样本集能进行符合要求的分类,即等价于求出一组领域,这组领域将不同类的点分隔开来,使属于  $K^i$  的点的输出均为  $y_i = \{0, \dots, 0, 1, 0, \dots, 0\}$  (即其第  $i$  个分量为 1, 其余分量为 0 的向量),  $i = 1, 2, \dots, s$ 。

交叉覆盖算法的基本思想是:先求一个领域  $C^1$ , 它只覆盖  $K^1$  中的点,而不覆盖  $K^2$  中的点,然后将被  $C^1$  覆盖的点删去,对余下的点求另一领域  $C^2$ , 它只覆盖  $K^2$  的点,而不覆盖  $K^1$  的点,然后将  $C^2$  被覆盖的点删去……如此交叉进行覆盖,直到只剩下最后一类点为止,将最后一类点作为一个覆盖。

#### 1.2 核函数

1995 年 Vapnik 提出了 SVM 支持向量机学习方法<sup>[3,4]</sup>, 对线性可分情况给出一个用规划方法解得的最大间隔解,当样本是线性不可分时,通过某种非线性映射函数将  $x$  从所在的输入空间  $X$  映射到一高维特征空间  $H$  上 (Hilbert 空间),再在高维空间  $H$  上建立优化超平面。这样虽然使得向量集更容易划分,但同时增加了计算的复杂度,而核函数正好巧妙地解决了这个问题。一般支持向量机的最终决策函数值仅仅依赖于变换后的 Hilbert 空间中的内积  $\Phi(x_i) \cdot \Phi(x_j)$ , 因此只要选取核函数为高维空间的一个点积即可:

$$K(x_i, x_j) = (\Phi(x_i) \cdot \Phi(x_j))$$

收稿日期:2006-04-14

基金项目:973 计划资助项目(2004CB318108);国家自然科学基金(60475017);安徽省自然科学基金(050420208)

作者简介:胡光杰(1983-),男,安徽六安人,硕士研究生,主要研究方向为智能算法及应用;张燕平,教授,硕士生导师,研究领域为人工神经网络、机器学习、数据挖掘。



这样就大大降低了计算的复杂度,实际上,甚至不需要知道具体的映射是什么,只要选定函数  $K(\cdot, \cdot)$  就够了。核函数  $K$  的选取需要满足 Mercer 条件<sup>[3,4]</sup>。目前并没有固定的核函数,对于不同的问题往往选取不同的核函数。目前主要研究的核函数有以下三种:

① 多项式核函数:

$$K(x, x_i) = [(x, x_i) + 1]^d$$

② 径向基函数:

最通常采用的核函数为高斯函数

$$K_r(|x - x_i|) = \exp\left\{-\frac{(|x - x_i|)^2}{\delta^2}\right\}$$

③ Sigmoid 函数:

$$K(x_i, x_j) = \tanh(v(x_i^T * x_j) - c)$$

文献[5]中指出了基于核函数的 SVM 机与三层前向神经网络是等价的,指出对任给样本集,均存在一映射  $F$ ,在此映射下,  $F(K)$  是线性可分的,并给出了求解核函数的构造性方法。文献[6~8]中将核函数法与覆盖算法结合,提出了核覆盖算法。核覆盖算法结合了核函数法与覆盖算法的优点,有效地解决了核函数法中对多分类计算量过大的问题,以及普通覆盖算法中只有局部求优的问题。

### 1.3 核覆盖算法

假设输入样本空间已被映射到特征空间上,特征空间中的距离为:

$$\begin{aligned} d(x_i, x_j) &= \sqrt{\|\Phi(x_i) - \Phi(x_j)\|^2} = \\ &= \sqrt{\Phi(x_i) \cdot \Phi(x_i) - 2\Phi(x_i) \cdot \Phi(x_j) + \Phi(x_j) \cdot \Phi(x_j)} \\ &= \sqrt{K(x_i, x_i) - 2K(x_i, x_j) + K(x_j, x_j)} \end{aligned}$$

算法步骤:

第一步,找初始点  $a$ ,可取为所有样本的重心。

第二步,从点  $a$  开始覆盖。

第三步,求以  $a$  为圆心,域值为  $r$  的覆盖领域  $C(a)$ 。其中  $C(a)$  的半径  $r$  可按如下方法求得:找离  $a$  最近的异类点,其距离记  $d_1$ ;找离  $a$  最远的距离小于  $d$  的同类点,其距离记为  $d_2$ 。覆盖领域的半径为  $r = (d_1 + d_2)/2$ 。

第四步,求领域  $C(a)$  的重心,并记为  $a$ 。

第五步,重复第三到第五步,直到覆盖的点不再增加为止。这样就得到一个覆盖  $K^1$  中点的局部最大覆盖  $C^1$ 。其覆盖  $K^1$  中的部分记为  $K^{1i}$ 。

第六步,找一个不同类点  $a$  再开始覆盖。令  $T \leftarrow K^1/K^{1i}$ ,  $K^1 \leftarrow K^2$ ,  $K^2 \leftarrow T$ 。

若只剩最后一类点,则将最后一类点作为一个覆盖。否则回第三步。

## 2 覆盖算法在煤炭供应商评测中的应用

### 2.1 应用思路

火力发电厂在进行锅炉设计时都有一个锅炉设计标准,其中包括煤质方面的一些标准值,在进行煤炭供应商的选择时,当然是其提供的煤质越靠近这些标准值越好。如果把这个标准看作一个向量,即看成欧式空间中的一个

点,那么在空间中越靠近这个标准点的向量无疑是越好的。交叉覆盖算法的基本思路正是利用向量间的距离进行分类,而核覆盖是将样本投影到线性空间中,再利用线性空间中向量的距离进行分类,所以,通过覆盖算法对煤炭供应商进行评测,符合煤炭供应商评测的实际情况。

### 2.2 数据源简介

本实验数据来源于淮北市电厂和淮南市洛河电厂这两个火力发电厂的燃料管理系统中生成的实时数据,这些数据的真实性、实时性将对实验提供强有力的支持。实验选取了在煤炭供应商的评测时应该综合考虑的煤质、煤价及煤量方面的数据,包括:水分、灰分、挥发分、发热量、硫分、厂矿热值差、运费以及标煤单价。其中运费以及标煤单价当然是供应商评测应该重点考虑的指标,水分、灰分、挥发分、发热量和硫分则是评测煤炭质量时的几个专业术语,下面分别解释这几个数据的意义:

水分是指煤炭中包含的水分,煤中水分过大将不利于加工和运输,燃烧时会影响热稳定性和热传导,炼焦时会降低焦产率和延长焦化周期。

灰分指煤在燃烧后留下的残渣,灰分高,说明煤中可燃成分较低,发热量就低。

挥发分指煤中有机物和部分矿物质加热分解后的产物,其大小与煤的变质程度有关,煤炭变质程度越高,挥发分产率就越低。

煤的发热量,又称为煤的热值,即单位质量的煤完全燃烧所发出的热量。煤的发热量是煤按热值计价的基础指标。煤作为动力燃料,主要是利用煤的发热量,发热量愈高,其经济价值愈大。

硫分指煤炭中硫的含量,硫是煤炭中的有害元素,硫分低于 1% 方可用于燃料。

厂矿热值差是指厂方化验所得热值与矿方提供的热值的差值,是衡量煤炭供应商信誉度的重要指标。

本实验选取这八维数据作为样本的属性,从数据库中提取了两个电厂从 2004 年 7 月到 2005 年 6 月共 423 条数据。

### 2.3 数据预处理

首先对样本的各维属性进行归一化,设  $X_i^j$  表示第  $i$  个样本第  $j$  维的值,采用如下方法进行归一化:

$$X_i^j = X_i^j / \max_j(X_i^j)$$

学习样本的分类是按如下方法进行的:对于选取的八维属性,由专家给出样本各个属性的权值

$$W_i (i = 1, \dots, 8, \sum_{i=1}^8 W_i = 1)$$

设标准样本为  $X_0$ ,可按加权平均的方法计算样本与标准样本的差值:

$$d = \sum_{j=1}^8 (W_j * X_i - W_j * X_0)$$

根据样本与标准样本的差值,可将学习样本分类。

为了更好地测试算法的分类效果,首先将学习样本分



为三类进行测试,即对煤炭供应商分为好、中、差三类,差值较小的三分之一作为第一类,差值较大的三分之一作为第三类,其余的作为第二类。然后将样本分为四类进行测试。

#### 2.4 实验结果

从实验结果(见表 1)可以看到,样本分为三类时,交叉覆盖算法的分类准确率为 88.33%,而核覆盖算法分类的准确率达到 91.67%;样本分为四类时,分类准确率有所下降,分别为 71.67% 和 73.33%。可以看出,当样本分为三类时,利用覆盖算法对煤炭供应商进行评测,取得了不错的效果;分为四类时,准确率有所下降。笔者认为有以下原因:首先,算法本身还有待进一步的改进;其次,对学习样本的分类是由专家根据经验进行的,当分类增多时,没有一个确定的分类标准,也一定程度上造成了测试准确率的下降。

表 1 实验结果

分类数	算法	学习样本数	测试样本数	正确分类样本数	准确率	覆盖数
三类	交叉覆盖	361	60	53	88.33%	38
	核覆盖	361	60	55	91.67%	35
四类	交叉覆盖	361	60	43	71.67%	50
	核覆盖	361	60	44	73.33%	53

从实验中还可以看出,无论样本是分为三类还是四类时,核覆盖算法的分类准确率都较交叉覆盖算法有所提高,其对算法分类准确率的提升效果还是很明显的。覆盖算法处理海量数据更为有效,当知识库中累计了更多的知

识(即学习样本更多)时,相信预测的准确率将更高。

### 3 结 论

利用前向神经网络的覆盖算法及其改进算法核覆盖算法对煤炭供应商的供煤情况进行了评测,在选取数据的时候选取了煤炭供应商评测时最应综合考虑的煤质、煤价等方面数据,实验结果与统计理论中加权平均的方法进行了比较,证明取得了不错的效果。

#### 参考文献:

- [1] 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.
- [2] 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.
- [3] Vapnik V N. Statistical Learning Theory[M]. New York: John Wiley & Sons,1998.
- [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] 张 铃. 基于核函数的 SVM 机与三层前向神经网络的关系[J]. 计算机学报,2002,25(7):696-700.
- [6] 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.
- [7] 赵 姝,张燕平,张 媛,等. 基于交叉覆盖算法的改进算法——核平移覆盖算法[J]. 微机发展,2004,14(11):1-3.
- [8] 张燕平,张 铃,段 正. 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.

(上接第 3 页)

### 3 结束语

如今许多应用都需要处理连续的数据流,而不仅仅是传统的固定存储的数据集合。在当今的网络监控、电信数据管理、传感器数据监控等应用中,数据采取的是多维的、连续的、快速的、随时间变化的流式数据的形式,对数据的访问也是多次和连续的,并要求即时的响应。笔者根据这种流式数据的特征设计了一种新的基于数据流的数据模型,并就今后如何进行数据流管理系统的研究提出了一些新的看法。

#### 参考文献:

- [1] Arasu A, Babcock B, Babu S, et al. STREAM: The Stanford Data Stream Management System[EB/OL]. 2004. <http://dbpubs.stanford.edu/pub/2004-20>.
- [2] Abadi D J, Carney D. Aurora: a new model and architecture for data stream management[J]. VLDB Journal, 2003, 12(2):120-139.
- [3] Chandrasekharan S, Cooper O. TelegraphCQ: Continuous data-flow processing for an uncertain world[C]// In Proc of the 1st Conf: on Innovative Data Systems Research. Asilomar, USA:

[s. n.], 2003:269-280.

- [4] Babcock B, Babu S, Datar M, et al. Models and Issues in Data Streams Systems[C]//In: Popa L. Proc of the 21st ACM SIGACT - SIGMOD - SIGART Symp, on Principles of Database Systems. Madison: ACM Press, 2002:1-16.
- [5] Babu S, Widom J. Continuous Queries over Data Streams[J]. SIGMOD Record, 2001,30(3):109-120.
- [6] 萨师煊,王 珊. 数据库系统概论[M]. 第 3 版. 北京:高等教育出版社,2000:13-121.
- [7] Datar M, Gionis A, Indyk P, et al. Maintaining Stream Statistics Over Sliding Windows[J]. SIAM Journal on Computing, 2001,31(6):1794-1813.
- [8] Arasu A, Babu S, Widom J. The CQL Continuous Query Language: Semantic Foundations and Query Execution[R/OL]// Technical report, Stanford University, 2003. <http://dbpubs.stanford.edu/pub/2003-67>.
- [9] Motwani R, Widom J, Arasu A. Query processing, approximation, and resource management in a data stream management system[C/OL]// In: Proc. of the 1st Biennial Conf. on Innovative Data Syst. Res (CIDR), 2003. <http://newdbpubs.stanford.edu/pub/2002-41>.