第 25 卷第 5 期
2008 年 10 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 25 No. 5
Oct. 2008

文章编号: 1000−8152(2008)05−0929−04

基于标准化高斯pLSA协同过滤的用电量预测模型

刘粤钳[1,2], 姚红玉[3]

(1. 湖州师范学院 人文学院, 浙江 湖州 313000; 2. 中国传媒大学 文学院, 北京 100024;

3. 湖州师范学院 教育科学与技术学院, 浙江 湖州 313000)

摘要: 现有的电力负荷预测算法在中长期预测时存在不同程度的局限性. 究其原因, 是因为影响复杂非线性系统输出的变元过多, 难以用解析的方法对其进行描述. 本文提出利用概率潜在语义分析使历史随机数据呈现出各种有规律的示象(aspect), 结合对内容的协同过滤技术去建立用电量预测模型, 从而利用统计学习的方法避开了对影响系统输出的隐含变元的寻找与刻画. 采用MATLAB进行数值仿真实验的结果表明该算法相比于神经网络和灰色预测在准确度方面具有优势.

关键词: 概率潜在语义分析; 协同过滤; 示象模型; 用电量预测模型

中图分类号: TP273　　**文献标识码**: A

Load-forecasting model based on normalized Gaussian pLSA collaborative filtering

LIU Yue-qian[1,2], YAO Hong-yu[3]

(1. Humanity School, Huzhou Teachers College, Huzhou Zhejiang 313000, China;
2. Literature School, Communication University of China, Beijing 100024, China;
3. Huzhou Teachers College School, Educational science and technology, Huzhou Zhejiang 313000, China)

Abstract: To some extent the existing long-term load-forecasting algorithms have their limitations because the variables influencing the output of the complex non-linear system are too many to be described. By combining the probabilistic Latent Semantic Analysis (pLSA) that can cluster random data into respective aspects and content-based collaborative filtering, a novel load forecasting model based on normalized Gaussian probabilistic latent semantic analysis collaborative filtering is proposed in order to avoid seeking and describing of the hidden variables mentioned above. Simulating experiments via MATLAB show that this method gains the advantage in accuracy over neural network and grey prediction.

Key words: probabilistic latent semantic analysis; collaborative filtering; aspect model; load forecasting model

1 引言(Introduction)

社会用电量的预测结果的准确与否直接影响到电力产业乃至整个国民经济的发展, 目前采用基于用户隐性偏好分析模型的预测方法还不多见.

用电量的历史数据中隐含着长期稳定的用电户的隐性偏好信息. 概率潜在语义分析pLSA(probabilistic latent semantic analysis)协同过滤的基本思想是: 假定系统中存在对输出起作用的无法准确描述的隐含变量, 计算使若干个有规律的示象(aspect)呈现; 从而避开对诸多潜在因素的分析.

本研究尝试采用标准化高斯pLSA协同过滤技术, 利用电力负荷及用电量的历史数据集训练获得中长期用电量的预测模型, 实验证明该算法相对于神经网络和灰色预测法具有较高的准确度.

2 示象模型(Aspect model)

Hofmann于1999年[1]提出的pLSA是一种无监督学习, 已成功应用于文件分类、语音识别及网页搜索等方面.

pLSA的核心思想是示象模型(aspect model). 该模型使隐含的类变量集合$z_k \in \{z_1, z_2, \cdots, z_K\}$与每一次观测值(某特定文本中某词的出现概率)相关. 由此, 词w_j文d_i同现的联合概率产生式为

$$\begin{cases} P(d_i, w_j) = P(d_i)P(w_j|d_i), \\ P(w_j|d_i) = \sum\limits_{k=1}^{K} P(w_j|z_k)P(z_k|d_i). \end{cases} \tag{1}$$

假设d_i和w_j独立, 则$P(w_j|d_i)$可视作K个概率面$P(w_j|z_k)$的凸组合.

考虑通过等价地反转图1(a)中D和Z之间的箭头得到图1(b)来确定模型的参数, 于是式(1)中联合概率的参数可由式(2)得到

$$P(d_i, w_j) = \sum_{k=1}^{K} P(z_k)P(d_i|z_k)P(w_j|z_k). \tag{2}$$

收稿日期: 2006−02−19; 收修改稿日期: 2007−05−22.

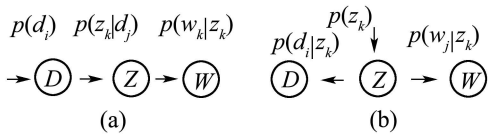

图 1 pLSA算法示意图

Fig. 1 pLSA algorithm sketch

式(1)的假设模型可理解为由概率群分布函数$P(\cdot|z_k)$的凸组合近似表达的所有$P(\cdot|d_i)$(其中$1 \leqslant i \leqslant N$)的条件概率集如图2.

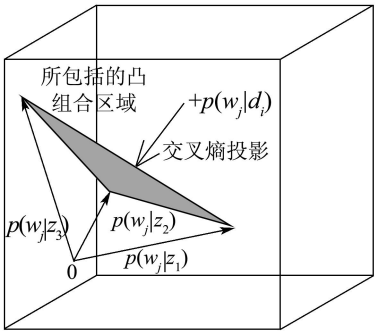

图 2 示象模型中由类条件概率函数所包括的概率四面体和凸组合区域示意图

Fig. 2 Sketch of the probability simplex and a convex region spanned by class-conditional probabilities in the aspect model

即便引入潜在变量是离散的, 在w上的所有概率群分布函数空间内仍可获得连续的潜在空间. 模型参数的估测可通过交替运行EM算法的E步骤和M步骤直到满足某一收敛的终止条件(或利用早停止技术)来实现.

3 pLSA协同过滤(pLSA-based collaborative filtering)

目前, 基于记忆的方法在协同过滤中占优势, 但基于模型的技术(如聚类、贝叶斯网和依存网)也颇受关注. pLSA协同过滤基于潜在因素模型以引出用户社区或项目群体观念, 并通过提供概率语义, 为用户偏好建立统计模型.

本研究将项目定义为用户对用电量的选择, 关注预测用户将会选择多少用电量及将会如何对用电量项目进行评比, 因此只考虑自由预测和隐性评比, 即用户处于项目选择的控制之下的情形. 自由预测中模型中元素的依存结构如图3所示.

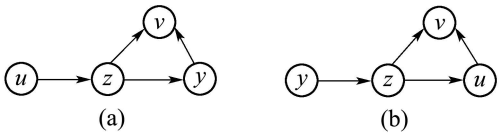

图 3 引入评比变量v后的pLSA扩展依存模型

Fig. 3 Graphical representation of dependency structure of the pLSA model to include a rating variable v

令$\mathcal{H}$为参数化的模型空间, 参数θ为$\mathcal{H}$中一特定模型, 引入损失函数$L : \mathcal{X} \times \mathcal{H} \to \mathcal{R}$. 其中$\mathcal{X} = \mathcal{U} \times \mathcal{V} \times \mathcal{Y}$, $\mathcal{V}$在隐性评比中被当作空集. 对给定的观察(u, v, y), $L((u, v, y), \theta)$愈小, 则θ 与观察越一致.

再引入风险函数

$$R(\theta) = \sum_{u,v,y} P(u, v, y) L((u, v, y); \theta). \tag{3}$$

将经验损失减到最小, 得

$$R^{emp}(\theta) = \frac{1}{N} \sum_{\langle u,v,y \rangle} L((u, v, y); \theta). \tag{4}$$

其中N为被观察三元组的总数.

4 基于标准化高斯pLSA协同过滤的用电量预测模型(Load forecasting model based on normalized Gaussian pLSA collaborative filtering)

pLSA模型可视为以隐性偏好为数据的协同过滤的特例[2]. 定义用户与项目对为用电户和用电量对, 基于标准化高斯pLSA协同过滤的用电量预测模型如下.

4.1 pLSA预测模型概述(Summarization of pLSA-based predictive model)

假定用户、用电量对(u, y)独立. 引进隐含变量集Z, u和y为有条件独立, 假定所有可能z的集合有限且其大小为k. 则模型为

$$P(u, y; \theta) = \sum_z P(y|z) P(z|u) P(u). \tag{5}$$

由Bayes定理

$$P(y|u; \theta) = \sum_z P(y|z) P(z|u). \tag{6}$$

$P(z|u)$和$P(y|z)$分别需要$(k-1) \times n$和$(m-1) \times k$个独立参数来描述. 以调整模型复杂度的方式(如交叉验证cross-validation)挑选k. 因为k通常比项目和用户数小许多, 该模型将促使用户结群进入用户社区, 使用电量项目结成相关项目群. 为避免可能因数据稀疏而产生的过度拟和, 采用使规则化风险函数而不是经验风险函数最小化的退火EM算法[3](tempered expectation maximization)修改E步骤来优化参数.

至此, 问题的关键变为如何计数变量v的可能度量标准, 及如何使类条件分布参数化, 为此引入标准化高斯pLSA预测模型如下.

4.2 标准化高斯pLSA预测模型(Normalized Gaussian pLSA forecasting model)

4.2.1 用户评比的标准化处理(Normalization of user's ratings)

第1步 减去评比μ_u; 在所有用户的评比中记录单个用户的差异, 并对每个用户校准中立部分.

第2步　将每个用户评比的变量标准化为1; 动态地调整其排列使评比在用户间更具有可比性.

平滑变量的估测

$$\sigma_u^2 = \frac{\sum\limits_{\langle u,v,y\rangle}(v-\mu_u)^2 + q\sigma^2}{n_u+q}. \tag{7}$$

其中σ^2为等级的所有变化, n_u是用户u可用到的等级数, q是控制平滑强度的自由参数(实验中令$q=8$).

4.2.2 标准化高斯pLSA预测模型(Normalized Gaussian pLSA forecasting model)

因为$v\in\{-1,1\}$是二值类变量, 引入成功概率参数$\pi_{y,z}\in[0,1]$并定义$P(v|y,z)\equiv\pi_{y,z}$. 用下式使类变量的条件概率参数化[3]:

$$P(v|y,z)\equiv\pi_{y,z}^v. \tag{8}$$

其中: $\sum\limits_{v\in\mathcal{V}}\pi_{y,z}^v=1$, 引入位置参数$\mu_{y,z}\in\mathbb{R}$和等级参数 $\sigma_{y,z}\in\mathbb{R}^+$.

$$\begin{cases} P(v|u,y)=\sum\limits_z P(z|u)P(v;\mu_{y,z},\sigma_{y,z}), \\ P(v;\mu,\sigma)=\dfrac{\exp[-\dfrac{(v-u)^2}{2\sigma^2}]}{\sqrt{2\pi}\sigma}. \end{cases} \tag{9}$$

考虑到具有变量$\sigma_{y,z}^2$的正常分布的噪声对评比的破坏, 期望得到的反应可如下计算:

$$\mathrm{E}[v|u,y]=\sum_z P(z|u)\mu_{y,z}. \tag{10}$$

4.3 模型的更新算法(Updating algorithm)

采用Bayes法更新资料集X来更新已存在的pLSA模型, 使其符合变化的数据集, 则参数集θ可由后验概率$P(\theta|X)$最大化原则来估测:

$$P(\theta|X)=\frac{\prod\limits_{u_i\in\mathcal{U}}\prod\limits_{y_j\in\mathcal{V}}P(u_i,y_j|\theta)P(\theta)}{P(X)}. \tag{11}$$

式中先验概率$P(\theta)$是为了求出θ的变异程度. 有

$$\theta=\arg\max_\theta\log(P(X|\theta))+\log(g(\theta)). \tag{12}$$

其中$g(\theta)$为先验概率在$P(w_j|z_k)$与$P(z_k|d_i)$相互独立假设下的简化式, 令$\{\alpha_{j,k},\beta_{k,i}\}$为Dirichlet分布参数, 有

$$g(\theta)\propto\prod_{k=1}^K[\prod_{j=1}^M P(w_j|z_k)^{\alpha_{j,k}-1}\prod_{i=1}^N P(z_k|d_i)^{\beta_{k,i}-1}]. \tag{13}$$

在EM算法的E步骤中计算出后验概率的期望函数值$R(\hat{\theta}|\theta)$, 在M步骤中使用θ来计算出新的最大后验概率, 通过迭代便可计算出区域最佳值. 本研究将实验验证此法.

5 实验(Experiments)

5.1 实验设计(Design)

将某省年度总用电量分为: 居民生活用电量、工业用电量、商业农林及其它用电量, 依据这3大类别, 以某省10年各类中每个独立用户u用电量y的历史数据集作为训练数据集, 得到对应的基于潜在类变量z空间的评比v和相应的用电量; 再用标准化高斯pLSA模型进行训练, 得到每个类别的预测模型, 最后汇总为总用电量.

以某省1992~2001年的用电量数据作为实验用数据集, 预处理后, 得到共2300万个居民用电户档案, 1273个项目, 173296个工业用电户档案, 2125个项目, 商业农林及其它用电户档案为241593个, 3346个项目, 在CPU为P4 3.2 GHz(800 MHz FSB, 2MB Cache)Dell台式机利用MATLAB7.3进行仿真实验.

项目等级排序算法[4]为: τ为用电量项目的一个排列, 且$\tau(y)$为对应于τ的用电量项目y的等级. 则最高等级化的用电量项目y有$\tau(y)=1$, 第2个项目$\tau(y)=2$等等. 用于训练的用电量项目的评比不包含在排列队列之中. 用$\bar{v}$表示全部投票的均值, 对τ使用下式:

$$R(u,\tau)=\sum_{\langle u',v,y\rangle:\ u=u'}\frac{\max(v-\bar{v},0)}{2^{\dfrac{\tau(y)-1}{\alpha-1}}}. \tag{14}$$

显然用户将会从顶部开始筛选, 直至找到某用电量项目或者放弃. 上式可直观地理解为用户会在某等级注意到某用电量项目的概率被模型化为半衰常数α(本实验$\alpha=5$)的指数分布.

用户群的总得分由下式测量:

$$R=100\frac{\sum\limits_u R(u,\tau_u)}{\sum\limits_u\max_{\tau'}R(u,\tau')}. \tag{15}$$

用“去一法”[4]对观察进行评比: 当给定用户选择的项目数$M\geqslant 2$时, 则从评比中随机排除一个评比, 以此对用户集上的损失函数平均化, 便可获得对风险的估测.

为研究预测的准确性, 不断改变M值以逼近用户所能容忍的最小值. 对不同的随机种子重复执行“去一法”30次. 最后得到的是30次的平均结果.

5.2 实验结果(Results of experiments)

5.2.1 模型的准确性实验(Testing the accuracy of the model)

用BP神经网络和灰色预测法[5~7]预测某省2002 ~ 2004年的年度用电量[8,9], 并对比本算法预测的结果, 实验结果如表1所示.

表 1 采用3种方法预测某省年用电量及相对误差(10^8 kWh/%)

Table 1 Annual load forecasting results for some province using 3 methods and their relative errors

年份	2002	2003	2004
实际值	1245.14	1505.11	1820.09
BP 神经网络	1208.25/–2.962	1467.10/–2.525	1777.91/–2.317
灰色预测	1221.59/–1.891	1477.05/–1.864	1788.76/–1.721
标准化高斯 pLSA	1259.55/ 1.157	1517.54/ 0.826	1837.98/ 0.983

5.2.2 模型参数对模型性能的影响(Influences of parameters of the model on performance of the model)

本研究中用户选择的最小项目数M取值范围为1∼10, 实验结果如图4所示, 随着M增大模型性能不断改善, 当$M = 10$时, 平均相对误差最小.

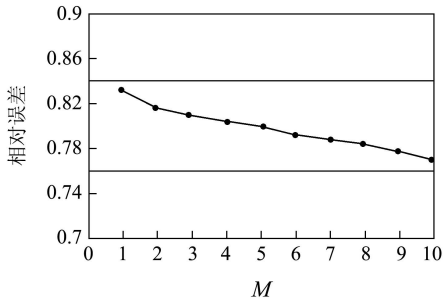

图 4 M取不同值时的相对误差

Fig. 4 Comparative error as M has different values

图5的实验表明用户社区数$k = 60$左右预测准确性最佳, 对所有用户的计算时间大约用了93 s. 退火EM迭代100次左右, 能达到较高的准确度.

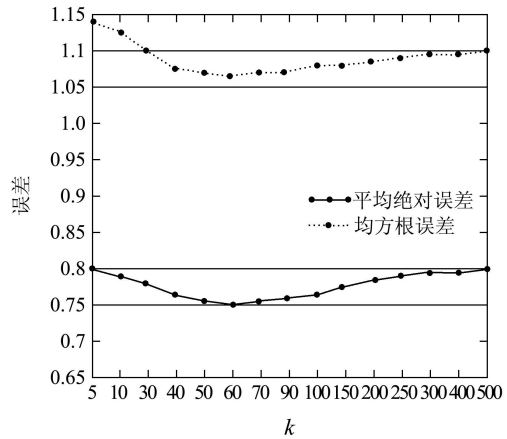

图 5 k取不同值时的平均绝对误差与均方根误差

Fig. 5 MAE and RMS as k has different values

5.2.3 模型的鲁棒性实验(Testing the robustness of the model)

某省自2004年4月1日起逐步推行分时电价, 价格的导向作用会使日电力负荷分布发生变化. 采用上述的模型更新算法. 以2004年5月到9月采用分时电价后的用户24小时电力负荷分布数据作为更新数据集, 用更新后的模型来预测2004年12月某日的全省24小时电力负荷分布情况, 并与实际情况加以对比. 实验结果如图6所示.

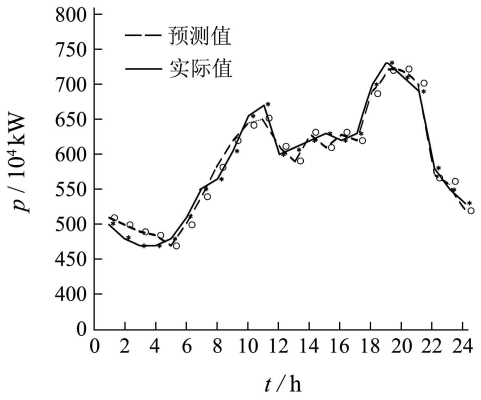

图 6 某省分时电价后24 h/日电力负荷变化趋势图

Fig. 6 24 h/day load varying trend in some province after carrying out time-sharing electrovalence policy

6 结语(Conclusion)

本算法E步骤与M步骤的计算复杂度均为$O(k \cdot N)$; 当动态更新训练数据时, 其计算复杂度为$O(n_u \cdot k)$, 此时计算复杂度与用户与项目数无关, 与用户已经评比的项目数有关.

本文运用标准化高斯pLSA协同过滤模型来预测用电量, 仿真试验表明该算法相比于目前常用的中长期预测算法(如神经网络和灰色预测)具有较高的准确度和较好的鲁棒性. 标准化高斯pLSA预测模型作为一种基于统计的新型预测算法, 其应用有待进一步深入研究.

参考文献(References):

[1] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis[J]. *Machine Learning*, 2001, 42(1): 177 – 196.

[2] HOFMANN T, PUZICHA. Latent class models for collaborative filtering[C] //*Proceedings of the International Joint Conference on Artificial Intelligence*. San Fransisco: Morgan Kaufmann Publishers Inc., 1999: 688 – 693.

(下转第937页)

5　结论(Conclusion)

针对热轧冷却过程中的带钢温度模型换热系数具有非线性、时变、难以用数学模型描述的综合复杂特性, 本文提出了一种混合智能参数辨识方法. 基于实际工业运行过程数据的仿真实验研究结果表明本文提出的混合智能参数辨识方法大幅度提高了冷却过程带钢温度预报模型精度, 对冷却控制具有重要意义.

参考文献(References):

[1] GUO R M, HWANG H T. Investigation of strip cooling behavior in the run-out section of hot strip mill[J]. *Journal of Mater Processing Manufacturing Science*, 1996, 4(4): 339 – 351.

[2] CHAI T Y, TAN M H, CHEN X Y. Intelligent optimization control for laminar cooling[C] *// Proceeding of the 15th IFAC World Congress*. Barcelona, Spain: Elsevier Science Ltd, 2002: 181 – 186

[3] FLETCHER R, XU C. Hybirid methods for nonlinear least squares[J]. *IMAJ Numer Anal*, 1979, 7: 371 – 389.

[4] TJOA I B, BIEGLER L T. Simultaneous solution and optimization strategies for parameter estimation of differential-algebraic equation systems[J]. *Ind Engng Chem Res*, 1991, 30: 376 – 385.

[5] 谭明皓, 柴天佑. 基于案例推理的层流冷却过程建模[J]. 控制理论与应用, 2005, 22(2): 248 – 253.
(TAN Minghao, CHAI Tianyou. Modeling of the laminar cooling process with case-based reasoning[J]. *Control Theory & Applications*, 2005, 22(2): 248 – 253.)

[6] KOLODNER J L. An introduction to case-based reasoning[J]. *Artificial Intelligence Review*, 1992, 6(1): 3 – 34.

[7] PAL S K, DE P K, BASAK J. Unsupervised feature evaluation: a neuro-fuzzy approach[J]. *IEEE Transactions on Neural Networks*, 2000, 11(2): 366 – 376.

作者简介:

片锦香　(1974—), 女, 博士, 研究方向为复杂工业建模与优化控制, E-mail: jxpian@hotmail.com;

柴天佑　(1947—), 男, 教授, 工程院院士, 研究领域为自适应控制、智能控制与综合自动化系统, E-mail: tychai@mail.neu.edu.cn.

(上接第932页)

[3] HOFMANN T. Latent semantic models for collaborative filtering[J]. *ACM Transactions on Information Systems*, 2004, 22(1): 89 – 115.

[4] BREESE J S, HECKERMAN D, KARDIE C. Empiricial analysis of predictive algorithms for collaborative filtering[C] *// Proceedings of the 14th Conference on Uncertainity on Aritificial Intelligence*. San Fransisco: Madison, Wisconsin, Morgan Kaufmann Publishers Inc., 1998: 43 – 52.

[5] 罗滇生, 姚建刚, 何洪英, 等. 基于自适应滚动优化的电力负荷多模型组合预测系统的研究与开发[J]. 中国电机工程学报, 2003, 23(5): 58 – 61.
(LUO Diansheng, YAO Jiangang, HE Hongying, et al. Research and development of multi-model combining load forecasting system based on self-adaptive rolling optimization[J]. *Proceedings of the Chinese Sociaty for Electrical Engineering*, 2003, 23(5): 58 – 61.)

[6] 张大海, 江世芳, 史开泉. 灰色预测公式的理论缺陷及改进[J]. 系统工程理论与实践, 2002, 22(8): 140 – 142.
(ZHANG Dahai, JIANG Shifang, SHI Kaiquan. Theoretical defect of grey prediction formula and its improvement[J]. *System Engineering Theory and Practice*, 2002, 22(8): 140 – 142.)

[7] 韩敏, 韩冰. 一种通用学习网络自适应算法及其在预测控制中的应用[J]. 控制理论与应用, 2006, 23(6): 900 – 907.
(HAN min, HAN bing. Adaptive algorithm of universal learning network and its application to predictive control[J]. *Control Theory & Applications*, 2006, 23(6): 900 – 907.)

[8] 中国电力年鉴1992～2001[M]. 北京: 中国电力出版社, 1992～2001.

[9] 中国统计年鉴1992～2001[M]. 北京: 中国统计出版社, 1992～2001.

作者简介:

刘粤钳　(1974—), 男, 湖州师范学院人文学院讲师, 中国传媒大学博士研究生, 研究方向为自然语言理解, E-mail: liuyueqian@126.com;

姚红玉　(1973—), 女, 理学博士, 湖州师范学院教育科学与技术学院副教授, 研究方向为机器学习, E-mail: yaohongyu@126.com.