Associated Clustering and Classification method for Electric Power Load Forecasting

Quansheng Dou[1,2], Kailei Fu[2], Haiyan Zhu[2], Ping Jiang[2], Zhongzhi Shi [1]

[1]Key Laboratory of Intelligent Information Processing; Institute of Computing Technology; Chinese Academy of Sciences; Beijing; 100080
[2]School of Computer Science and Technology; Shandong Institute of Business and Technology, Yantai 26400)

douqs@ics.ict.ac.cn

Abstract — In the process of power load forecasting, electricity experts always divide the forecasting situation into several categories, and the same category uses the same forecasting model. There exists such a situation that some load curve which domain experts consider belonging to the same category has shown the different characteristics, but some load curve which belongs to different category seems very similar, and usually able to gather into a category by clustering. For this problem, the definition of associated matrix was proposed in this paper, and based on this conception the associated clustering-classification algorithm was proposed, We applied this algorithm to data sample classification for power load prediction, the experiment showed that the classification results obtained by our method were more reliable.

Key words — Electricity Load Prediction; Classification; Wavelet Anlysis

1. Introduction

Load forecasting is an important component for power system energy management system. Precise load forecasting helps the electric utility to make unit commitment decisions, reduce spinning reserve capacity and schedule device maintenance plan properly. Besides playing a key role in reducing the generation cost, it is also essential to the reliability of power systems. The system operators use the load forecasting result as a basis of off-line network analysis to determine if the system is vulnerable. If so, corrective actions should be prepared, such as load shedding, power purchases and bringing peaking units on line.

Classification and clustering are two important research areas of data mining. To map data into some given classes, classification depends on prior knowledge, and clustering is to make samples in the same cluster similar enough, while samples belonging to different clusters should have enough difference. Recently, [1]~[2] use granularity computation to solve classification problems, and with the improvement of granularity computation theory these methods will develop further. [3]~[4] use ant colony optimization etc. to search the classification rules, these algorithms are the combination of data mining and intelligence computation. [5] proposes a new classification algorithm based on the combination of supported vector machine and non-supervisor clustering, and gets better results when it is applied in web page classification. [6] systematically summarizes the clustering method. All these researches represent the newest development in this area.

In the process of power load forecasting, electricity experts always divide the forecasting situation into several categories, the same category uses the same forecasting model. There exists such a situation that some load curve which domain experts consider belonging to the same category has shown the different characteristics, but some load curve which belongs to different category seems very similar, and usually able to gather into a category by clustering. In other words, the prior knowledge is very likely uncoordinated with similarity measure. [7] analyzed this issue by granularity theory and put forward classification algorithm based on information granularity principle. This has strong theoretical and practical significance. Aimed at the above problem, this paper proposes the Associate Clustering-Classification Algorithm to ensure the consistency of classification and clustering. The algorithm in sample classification of power system load forecasting is applied, and better results are obtained. The detail of the Associate clustering-classification method will be described in the following.

2. Associated Clustering and Classification method

Let $U = \{x_1, x_2 ..., x_k\}$ be a sample set, δ be a cluster operation, and it forms into a Cluster Genealogy G under action of δ. We cut Cluster Genealogy G, divide U into independent subset , and get $\delta(U) = \{G_1, ...G_m\}$. We classify U with Priori knowledge and obtain classification $C = \{C_1, ...C_n\}$.For $\forall C_i \in C \quad i = 1, 2, ..., n$ is divided into m sub-sets by $\delta(U)$, we have $C_i = \bigcup_{j=1}^{m} C_{ij}$ where m is the number of sub-sets in $\delta(U)$, and

$C_{ij} = C_i \cap G_j$.

Definition1. Suppose U is a sample space, δ is a cluster operation, $G = \delta(U) = \{G_1, ...G_m\}$ and $C = \{C_1, ...C_n\}$ is classification of U. We call matrix $\Lambda = \left\{\lambda_{ij}\right\}_{n\times m}$ an Associated Matrix of classification C based on G. Here, $\lambda_{ij} = \frac{|C_{ij}|}{|C_i|}$ and $C_{ij} = C_i \cap G_j$, $|C_{ij}|$ and $|C_i|$ represent the number of elements in the collection C_{ij} and C_i. We call each row $R_i \quad i = 1, ..., n$ of matrix Λ an Associated Vector.

Definition2. On the basis of definition 1, let $R_i = (r_{i1}, r_{i2}, ..., r_{im}) \quad i = 1, ..., n$,where R_i is the Associated Vector. If there are s components which satisfy $r_{ij} \geq \frac{1}{m}, \quad j = 1, 2, ...s$, then vector R_i is called s items dominant or multi-term dominant, and if vector R_i has only one r_{ij} which satisfies $r_{ij} > \frac{1}{m}$,then R_i is called one term dominant. If $r_{ij} > 0.8$ and makes R_i one term dominant, then R_i is called one term sufficiently dominant.

Algorithm.1 Associated Clustering and Classification Algorithm

Step1. Classify C according to prior-knowledge, and get the initial classification $C = \{C_1, C_2, ..., C_n\}$

Step2. Implement cluster operation according to Euclidean distance on U and get cluster genealogy G. Cut from top of G, and get two branches, each of which forms one class. Get the Associated Matrix:

$$\Lambda = \begin{pmatrix} \lambda_{1G1} & \lambda_{1G2} \\ \cdots & \cdots \\ \lambda_{nG1} & \lambda_{nG2} \end{pmatrix} \qquad (2)$$

Step3. At some time , suppose Associated Matrix

$$\Lambda = \begin{pmatrix} \lambda_{1G1} & \lambda_{1G2} \cdots & \cdots \lambda_{1Gl} \\ \cdots & & \cdots \\ \lambda_{nG1} & \lambda_{nG2} \cdots & \cdots \lambda_{nGl} \end{pmatrix} \qquad (3)$$

Inspect each column $(\lambda_{1Gj}, \lambda_{2Gj}, ..., \lambda_{nGj})^T$ of Λ, If there are two or more components λ

larger than μ, and $|G_j| > \tau |U|$, where $0 < \tau < \mu < 1$ is the threshold parameter, $|G_j|$ and $|U|$ represent the number of elements of G_j and U respectively. Cutting at the top of the cluster genealogy G_j, form the new branch, and revise the Associated Matrix. Implement step3 repeatedly until there is only one $\lambda_{iGk} > \mu$ in each row of Λ or $|G_j| < \tau |U|$.

Step4. For each column $(\lambda_{1Gj}, \lambda_{2Gj}, ..., \lambda_{nGj})^T$ of the matrix Λ, if there are two or more components λ sufficiently dominant in their rows, set these components be $\lambda_{hGj}, \lambda_{h+1Gj}, ...\lambda_{pGj}$, then combine $\lambda_{hGj}, \lambda_{h+1Gj}, ...\lambda_{pGj}$ into one category. Revise matrix Λ.

Step5. Analyze each row $Row_i = (\lambda_{iG1}, \lambda_{iG2}, ...\lambda_{iGm})$ of matrix Λ. If Row_i is one term sufficiently dominant, then take C_i as one class individually. Otherwise, set the threshold κ, Suppose there are l components larger than κ in Row_i, and they are $\lambda_{is}, \lambda_{is+1}, ...\lambda_{is+l-1}$ respectively. C_i Will be divided into l categories based on $G_s, ..., G_{s+l-1}$. For $\forall x \in C_i - \bigcup_{r=0}^{l-1} G_{s+r}$, according to the principle of minimum distance to the collection center, add them into some class of C_i

Step 1 and 2 of the algorithm implement classification and clustering operation on the sample collection. The emphasis is that the prior-knowledge used by classification and the measure function used by clustering are essentially the same, otherwise, it is not worth harmonizing.

Step3 of the algorithm can ensure that there is only one classification C_i, whose most samples appear in a clustering G_j. If there are two or more classifications, whose most elements are in the same clustering called G_j, cut up G_j on the top of the clustering genealogy. Finally, if there are still two or more classifications whose most elements are in the same clustering G_j, the number of samples in G_j must be below a certain size. These classifications were combined into one class in step4.

In step5, if a row Row_i of the Associated Matrix is one term sufficiently dominant, C_i should be set as one class individually. Steps 3 and 4 have ensured that there can be no more than two categories C_i whose majority of samples appear in a clustering G_j.If Row_i is not a one

term sufficiently dominant vector, and most samples of C_i distribute in $G_s,...,G_{s+l-1},C_i$ should be divided into l classes according to $G_s,...,G_{s+l-1}$. Samples out of $G_s,...,G_{s+l-1}$ in C_i should be added to a certain classification in C_i with the principle of minimum distance.

By analyzing the performance process of the above algorithm, it's easy to see that when the algorithm is finished, Associated Matrix Λ has k rows which are one term sufficiently dominant and $s-k$ rows which are multi-term dominant. Here, $k \geq 0$ and $s \leq n$, n is the number of classifications obtained by priori knowledge. And no columns in Λ can make all rows sufficiently dominant. The above can be shown in the following formula（4）:

$$\Lambda = \begin{pmatrix} 1 & 0 & \cdots & & & \\ 0 & 1 & \cdots & & & \\ \cdots & \cdots & & & & \\ \hline \lambda_{i1} & 0 & \lambda_{ij} & \cdots & 0 & k+1 \\ 0 & \cdots & \lambda_{i+1,j} & \lambda_{i+1,m} & \cdots & \\ \cdots & \cdots & & & & s \end{pmatrix} \begin{matrix} \\ \\ \\ k \\ \\ \\ \end{matrix} \tag{4}$$

Two kinds of standards are involved here. One is the priori knowledge of domain experts. Because many complex factors affect the change of power load, and some reasons which cause power load changing is not clear, the priori knowledge used by experts on power, often just reflect the variation of load roughly. The other is that characteristics of load change can be objectively identified by clustering, but the reasons why the samples cluster into a class are not yet determined. This makes the clustering method can not be directly used on predicting. For this reason, the next best thing is to take a relatively compromise. Associated clustering-classification algorithm is an exactly compromise method between classification and clustering.

3. Description of the problem of power system load forecasting

Load forecasting is a traditional research field of power system [9]~[11]. In the process of power load forecasting, electricity experts always divide the forecasting situation into several categories, the same category uses the same forecasting model. So a reasonable classification is the basis for effective forecast. Generally, domain experts classify the samples relying on their experience. In this paper, 96-points data samples of a Chinese power company in recent years were classified by the expert's experience and associated clustering-classification algorithm described in the previous section. Here, the forecasting models used by the different classification methods are the same.

First of all, the samples are classified. For different categories, Daubechies wavelets are used to extract the feature of load data.

Let $\{p(t)\}$ $t=1,...,96$ be the load value of 96 points one day. Let $C_0(t)=p(t)$, wavelet decomposition is shown as follows:

$$\left\{\begin{array}{l} C_0 = p(t) \\ C_j[k] = C_{j-1}\bar{h}[2k] \\ D_j[k] = C_{j-1}\bar{g}[2k] \end{array}\right\} \quad j = 1,2,...,L \qquad (5)$$

In the formula $\bar{h}[-k] = h[k]$, $\bar{g}[k] = g[-k]$, $g[k] = (-1)^{k-1}h[k-1]$. $h[k]$ is the low-pass filter, $g[k]$ is the high-pass filter, and L is the decomposition level. $C_j[k]$, $D_j[k]$ $j = 1,2,...,L$ are low-pass signal (features) and high-pass signal (noise) of the *j*-layer wavelet transform respectively. By the wavelet transform, the 96-points time series are broken down into two parts, feature and noise. The dimension of the low-pass signal C_{j+1} and high-pass signal D_{j+1} obtained in each of the decomposition is half of the dimension of C_j. Let C_0 be the 96 points load data initially, the dimensions of C_3, D_3, D_2, D_1 are 12,12,24,48 after three wavelet decomposition. So the dimension of $\{C_3, D_3, D_2, D_1\}$ remains 96. Here, the previous 12 components C_3 contain the overall volatility information $\{p(t)\}$,i.e. the characteristic component while $D_3, D_2,$ and D_1 are high-frequency information, i.e. the noise component of $\{p(t)\}$ at different spatial scales. $\{p(t)\}$ can be obtained by reconstruction of vector $\{C_3, D_3, D_2, D_1\}$.

We can obtain the temperature information from the meteorological station and analyze the relationship between the temperature and the feature components. As the temperature changes, the feature component values show a certain discipline. We can regress this law and get the polynomial relations between the temperature and features components. So the feature components can be forecast according to the change of real sense temperature.

It is impossible to determine the relationship between temperature and noise with regression approach, because noise components show splattering distribution to the temperature. As described above, noise component is constituted by the high frequency information on different scales of space obtained by three-layer wavelet decomposition to 96 points data. Its vector length is 84. We use the following method to determine the noise components:

Let $D_i = \{d_{i1}, d_{i2}, ..., d_{i84}\}$, i=1,2,…,q be the noise component of 96 point data for one day.

Let $f(d) = \sum_{i=1}^{q}\sum_{j=1}^{84} | d - d_{ij} |$, where q is the total number of samples for the classification. The

noise component $\bar{d}_j$, j=1,2,…84 can be determined by solving the optimization problem of $\min f(d)$.

As mentioned above, we can predict the feature $\{C_3\}$ through the temperature. Reconstruction of C_3 and $\{D_3, D_2, D_1\}$ will get the electricity load forecasting value on that day.

4. Forecasting results of different classification methods

Classify the samples by date type according to prior knowledge, denote as $C = \{C_1, C_2, ..., C_n\}$. For each of the 96 points historical load data denoted as $(p_0, p_1, ..., p_{95})$, let $\Delta p = (\frac{p_1 - p_0}{p_0}, \frac{p_2 - p_1}{p_1}, ..., \frac{p_{95} - p_{94}}{p_{94}})$. Cluster Δp by Euclidean distance to form cluster genealogy, and reclassify the above classification with associated clustering-classification algorithm. The original classification $\{C_1, ...C_n\}$ is reclassified to get the classification $C' = \{C'_1, ...C'_m\}$. Analyze the final associated matrix

$$\begin{pmatrix} 1 & 0 & ... & & \\ 0 & 1 & ... & & \\ ... & ... & & & \\ \hline \lambda_{i1} & 0 & \lambda_{ij} & ... & 0 \\ 0 & ... & \lambda_{i+1,j} & \lambda_{i+1,m} & ... \\ ... & ... & & & \end{pmatrix}$$

For all the rows which are one term sufficiently dominant in associated matrix Λ. Classification results based on priori knowledge are about the same as the results of clustering, and the method used for forecasting is consistent with that mentioned above.

For all the rows as $(0, ..., \lambda_{ij}, ...0..., \lambda_{im}, ..., 0)$, which are multi-terms dominant in Associated Matrix Λ, suppose $\lambda'_{is}, \lambda'_{is+1}, ...\lambda'_{is+l-1}$ are dominant items. There exists classification $C'_{is}, C'_{is+1}, ...C'_{is+l-1}$ in the new classification set C' corresponding with them. Extract feature and noise in these classification respectively, and regress relationship between the feature and temperature, and reconstruct with the corresponding high-frequency signal to get predictor $P'_{is}, P'_{is+1}, ...P'_{is+l-1}$. Because the reason why C_i is classified into

$C'_{is}, C'_{is+1}, ...C'_{is+l-1}$ is unknown, a specific forecast can only start from a priori knowledge. Forecast should be taken as follows on the corresponding situation of the row:

$$P_i(t) = \sum_{j=s}^{s+l-1} \lambda_{ij} P_{ij} \qquad (8)$$

To verify the effectiveness of the method described in this paper, we use the historical data of the previous two years as the learning sample, and respectively predict the load of next year by different classification methods. In the process of load forecasting, for some special holidays such as the Chinese Spring Festival and New Year's Day, it has no sense to regress for lacking of historical data, and the forecast can only be made by historical trends.

Table 1 lists several groups of statistical results of the experiment. In table 1, statistics type A is the percentage of the data points whose errors are less than 1%. B is the percentage of points whose errors are between 1% and 3%. C is the percentage of points whose errors are larger than 3%. D is the average of root-mean-square error between predictions and the actual data.

Table1. Predictions based on different classification. Error described as: A, B, C, D are equal to the percentage of points whose errors are less than 1%, the percentage of points whose errors are between 1% and 3% ,the percentage of points whose errors are larger than 3% and the average of error respectively, in the forecast data points.

Year	Error type	Forecasting result of classification according to the prior knowledge	Forecasting result of Associated Clustering and Classification
2007	A	68%（23827 points）	76%（26630points）
	B	17%（5957points）	14%（4906points）
	C	15%（5256points）	10%（3504points）
	D	3.02%	2.41%
2008	A	77%（27055points）	81%（28460points）
	B	15%（5270points）	13%（4568points）
	C	8%（2811points）	6%（2108points）
	D	2.16%	1.88%
The first half of 2009	A	75%（13032points）	80%（13901points）
	B	16%（2780points）	15%（2606points）
	C	9%（1564points）	5%（869points）
	D	2.25%	1.82%

From Table 1, we can see that forecasting results based on the new classification method are significantly better than the original classification based on experience. It is not difficult to see that through the above analysis, the load data often have different characteristics objectively in the classification based on priori knowledge. Considering classification whose features is different from samples as one classification to regress is the main reason for the large error of regression curve. The associated clustering-classification algorithm in this paper avoids this problem to a

certain extent, and forecast accuracy has been improved significantly. It also validated that the method proposed in this paper better solved the inconsistent problem between the priori knowledge and the similarity measure function.

5. Conclusion

Classification and clustering are two important research areas of data mining; however in the process of power load forecasting, the classification results based on priori knowledge and the clustering results are not consistent. For this problem and the practical application background of power system, the definition of associated matrix has been proposed in this paper, and based on this concept, the associated clustering-classification algorithm has been proposed. We applied this algorithm to data sample classification for power load prediction, the experiment showed that the classification results obtained by our method were more reliable.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (No.60970088,60775035), the National High-Tech Research and Development Plan of China (No. 2007AA01Z132), National Basic Research Priorities Programme(No. 2007CB311004) and National Science and Technology Support Plan (No.2006BAC08B06), Dean Foundation of Graduate University of Chinese Academy of Sciences(O85101JM03).

References

[1] T Yao, Y Y Yao. granular computing approach to machine learning. Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery: Computational Intelligence for the E-Age, 2 Volumes, November 18-22, 2002, Orchid Country Club, Singapore. 2002

[2] Y Y Yao, J T Yao. Induction of classification rules by granular computing. In:Proceedings of The Third International Conference on Rough Sets and Current Trends in Computing, pp.331~338 2002

[3] Parpinelli R S, Lopes H S, Freitas A A. Data mining with an Ant Colony Optimization Algorithm. IEEE Trans. on Evolutionary Computation. 2002

[4] Sousa T, Silva A, Neves A. A particle swarm data miner. Fernando Moura Pires, Salvador Abreu (Eds.): EPIA 2003, LNAI 2902, pp. 43–53, 2003

[5] Xiao li Li, Ji min Liu, Zhong zhi Shi A Chinese Web Page Classifier Based on Support Vector Machine and Unsupervised Clustering. Chinese Journal of Computers. V24.No1.2001

[6] Rui Xu. Survey of Clustering Algorithms IEEE Transactions on Neural Networks, Vol.16, No. 3, May 2005

[7] Dong bo Bu, Shuo Bai, Guo jie Li. Principle of Granularity in Clustering and Classification . Chinese Journal of Computers V25.No8.2002

[8] Jing Zhang, Rui Song Wen xian Yu, Sheng ping Xia , Wei dong Hu. Construction of Hierarchical Classifiers Based on the Confusion Matrix and Fisher's Principle.
Journal of Software. Vol.16, No.9 2005 1000-9825/2005/16(09)1560

[9] Alfares H. K.; Nazeeruddin M. Electric load forecasting: literature survey and classification of methods International Journal of Systems Science, Volume 33,Number 1, 1 January 2002 , pp. 23-34(12)

[10] K Metaxiotis, A Kagiannas, D Askounis, J Psarras Artificial intelligence in short term electric load forecasting: a state-of-the-art survey for the researcher Energy Conversion and management Volume 44, Issue 9, June 2003, Pages 1525-1534

[11] Chong qing Kang, Qing Xia, Bo ming Zhang. Review of power system load forecasting and its development . Automation of Electric Power Systems V28.No17.2004