# Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (13): 2443-2449, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.2443.2449


# Heat-Supply Network State Prediction Based on Optimum Combination Model of Data Mining

[1]Xiufang Wang, [1]Yan Wang, [1]Hongbo Bi and [2]Running Gao
[1]School of Electrical Information Engineering, Northeast Petroleum University, Daqing, China
[2]180 w. Big Springs Rood. Apt.23, Riverside, PC: 92507, California, USA

**Abstract:** At present, a massive portion of data stored in the heat-supply network management system has formed the data grave which does not embody the intrinsic properties of data. To solve this problem, it is particularly important to take effective mining methods which reuse the existing historical data to improve the current system. In this paper, we first dispersed the continuous attribute information based on both entropy and importance of attribute to preprocess the data of heat-supply network. Then we exploited three kinds of algorithm for data mining, namely, classification and prediction based on the decision tree, cluster analysis based on the K-mean partition and association rules mining based on the frequent itemset model. Finally, we established forecasting model combining the results of three aforementioned mining schemes. The model was then embedded into the prediction module of present system and the results demonstrated the proposed scheme can improve the prediction performance efficiently.

**Key words:** Data preprocessing, decision tree classification, clustering analysis, frequent itemset, combination forecasting model

## INTRODUCTION

A large amount of historical data of heat-supply network management system stored in large databases has evolved into the "data grave" (Wang, 2002; Zhang *et al*., 2008) which people hardly have the opportunity to access, causing the occupation of considerably amount of memory space and redundancy of massive data. Therefore, the Data Mining (DM) techniques on the massive inventory data which describes the inner relationship of data, are desirable to improve the network detection method. As a result, the network pipeline would be safer and more effective and the error detection level would also be improved. Knowledge discovery based on massive inventory data can be used to predict the operation condition of heat-supply network.

DM algorithms have been widely used in various fields recently. However, the work about their applications on prediction of the leakage alarm and detection of the operating status for the heat-supply pipeline has been seldom reported. The reasons are two folds. First the online leakage monitoring method is still under development and the prediction algorithm cannot achieve a satisfying accuracy. On the other hand, the risk of using DM is considerable, due to the fact that the resource investment may not be rewarded appropriately. Hence, DM algorithm based on the heat-supply network historical data is a new direction worth exploring with risks yet potential value. DM study of heating inventory not only reuses the historical data which enhances the value of inventory information, but also lays the foundation for the new prediction method based on extracted hidden information in data. It also shows potential in improving heat-supply network monitoring and prediction (Chen *et al*., 2004).

**Data preprocessing:** Data preprocessing is the preparation of DM, on the one hand we should ensure the effectiveness and correctness of DM, on the other hand, we must adjust the format and content of data to meet the standard of DM. Data processing includes, data cleaning, data integration, data transformation, data reduction etc. These preprocessing methods not only improve the quality of model, but also reduce time on the actual mining.

**Rough set theory:** Rough set theory is a mathematical tool which was proposed by Pawlak (1982) (Chen and Li, 2010; Bean and Kambhampati, 2008), which is used to handle the uncertain and inaccurate problems. Based on remain of the original knowledge, it improves the

**Corresponding Author:** Xiufang Wang, School of Electrical Information Engineering, Northeast Petroleum University, Daqing, China Tel: 086-0459-6503160

classification efficiency by reducing the knowledge theory and deleting the redundant information. This method can also analyze and infer information to mine the implicit knowledge, revealing the potential rules.

Suppose the tuple S = {U, A, V, f} as an information system, where U = {$x_1$, $x_2$, …, $x_n$} is the nonempty finite sets of object, $x_i$ (i = 1, 2, …, n) as the research object, A is the attribute collection and A = C∪D, C = {a|a∈C} as condition attributes, where a is a simple attribute of the set C, D = {d|d∈D} is the decision attribute set and C≠Φ, D≠Φ, C∩D = Φ, V = ∪ is the attribute range, f as the mapping of U×A→V which is called "information function" (Li *et al*., 2010). When the decision attribute is only one for the single decision table, in general, the decision table can be transformed into a single decision table.

**Discretization based on information entropy:** The discretization of continuous attributes based on information entropy is a top-down splitting technique which searches the optimal partition by remaining the information entropy of decision attributes of original decision table unchanged to that of the conditional attributes. According to the decision table S = {U, A, V, f} defined in the rough set theory and A = C∪D, the range $V_a$ of the condition attribute $V_a$ has the maximum $V_{amax}$ and the minimum $V_{amin}$, the method of information entropy discretization based on attribute a is showed as following: First, we determine the split point (denoted as split_point) which can be divided into two subsets in the range of attribute a. Then based on the expected information needs of a for tuple D, it is given as following:

$$\text{Info}_a(D) = \frac{|D_1|}{|D|}\text{Entropy}(D_1) + \frac{|D_2|}{|D|}\text{Entropy}(D_2) \quad (1)$$

In the Eq. 1, $D_1$ meets the condition a≤split_point and $D_2$ meets the condition a≤split_point, |D| is the number of tuple D. In the given set, entropy function Entropy is calculated by the distribution of tuples in the set. If $D_1$ contains classes $C_1$, $C_2$, …, $C_m$, the entropy of $D_1$ is:

$$\text{Entropy}(D_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

The expected information need can be gotten:

$$\text{Info}_a(D) = \frac{|D_1|}{|D|}\left(-\sum_{i=1}^{m} p_i \log_2(p_i)\right) + \frac{|D_2|}{|D|}\left(-\sum_{i=1}^{n} p_i \log_2(p_i)\right) \quad (2)$$

When choosing the split point of attribute A, attribute values of the minimum expected information is required which is equivalent to producing the property value of maximum information gain (Miao and Li, 2008).

**Discretization based on attribute importance:** This method is based on the classification ability of attributes to measure the importance of condition attributes which can be used to judge how the classification system changes by removing itself: The larger classification changed caused by removing attribute, the higher the attribute importance will be. Otherwise, the importance of the attribute will be lower. This algorithm is described as follows. First, according to the attribute importance we should sort the condition attributes $v_i$ (i = 0, 1, …, n); when the property statuses are same, we sort the condition attribute by the breakpoint number from the most to the least. Then we consider the presence of each attribute which meets $v_i$∈A and each breakpoint $C_j$ (j = 0, 1, …, $i_j$) in $v_i$ exists. Change the smaller attribute value adjacent with $C_j$ to a larger value, if the system doesn't introduce conflict, $C_{v_i} = C_{v_i}/\{c_j\}$; otherwise, the changes of attribute value will be recovered. The proposed algorithm removes the redundancy by judging the breakpoint value to get the simplified set.

**Algorithm simulation:** In this paper, attribute information in heat exchanger station of No. 1 is used as example, heating data of a heating period is used as iteration. (Data includes the heating information from October 14th, 2011 to April 15th, 2012, the real-time data is updated every 5 seconds, excluding four days we test running and stop the system. The effective information includes 3127680 records, reaching the standard of data mining.). Extracted continuous attribute value of daily heating information and distribution characteristics are shown in Fig. 1.

Eight attributes of the heating network include temperature, pressure of supply water, pressure of return water, temperature of supply water, temperature of return water, instantaneous heat, water supply instantaneous flow and water return instantaneous flow. These discrete attribute values are obtained by adopting the discretization algorithm, the structure is shown as the Fig. 2.

## DATA MINING

DM is the process which extracts implicit content which is unknown in advance, but containing the potential significance of the information and knowledge in the data which is large, incomplete, noisy, fuzzy and random (Wang and Wang, 2009; Kantardzic, 2003). The

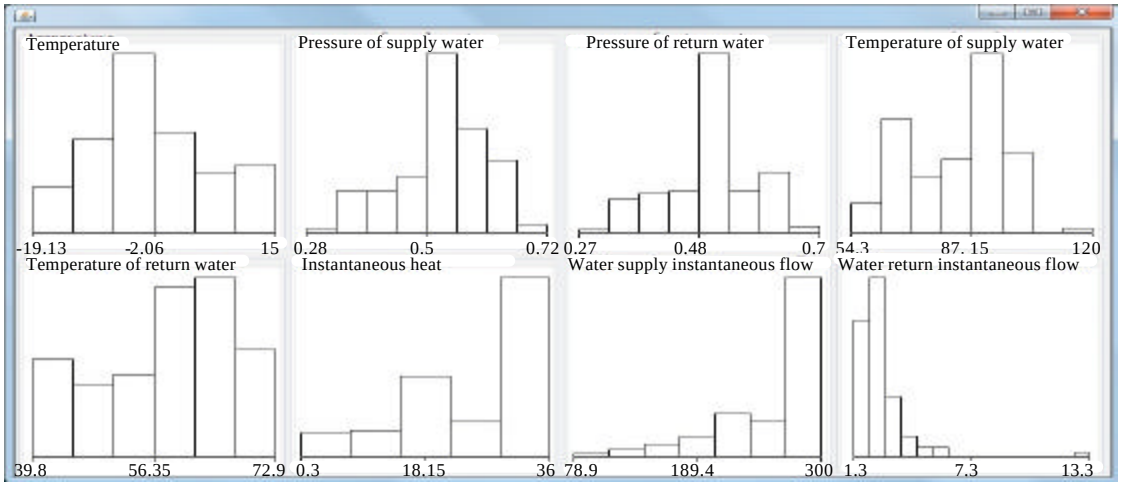


Fig. 1: Distribution map of network's continuous parameter

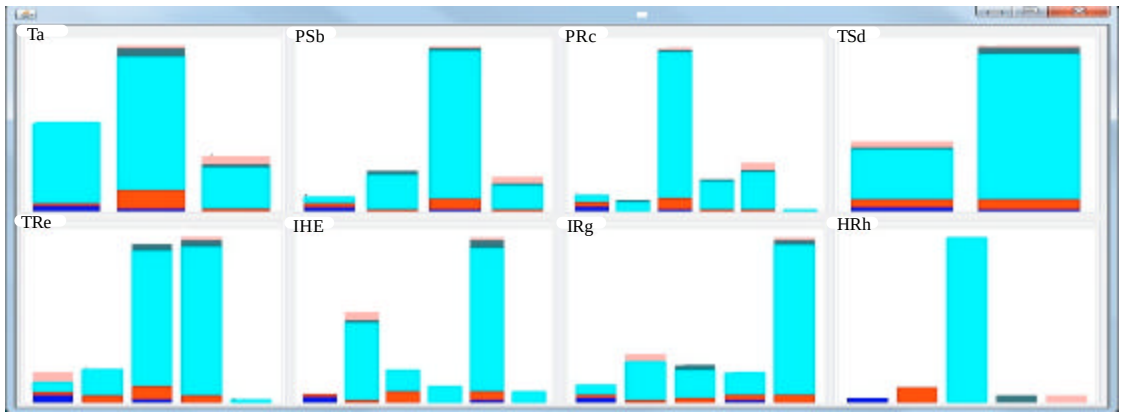


Fig. 2: Distribution map of network' dispersion parameter

task of DM is to find the implicit application patterns in the database. According to the function, the method of DM model is divided into two broad categories generally: descriptive model and forecasting model. The classification model type and the association rules belong to forecasting model and the clustering analysis belongs to descriptive model.

**The classification prediction based on decision tree:** ID3 which created a precedent of decision tree algorithm is one of the earliest and most influential methods in international machine learning (Li and Zhang, 2011). D is assumed to be a training instance set and |D| is the number of training instances in D. $C_i$ (i$\epsilon$ {1, 3, ..., m}) is the number of samples which corresponds to categorical attributes item I, $p_i$ is the probability $p_i = C_i/|D|$ of $C_i$ in D. The expected information needed by classifying instances in D is:

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (3)$$

We divide the training set according to attribute A and assume that A has values of V which is $\{a_1, a_2, \ldots, a_v\}$, then classify instances in D including the values of V which is $\{D_1, D_2, \ldots, D_v\}$ after dividing D which is according to attribute A, we need to know the information:

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(D_j) \quad (4)$$

We can get the complete information of division according to attribute A by arranging the above formula:

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \left( -\sum_{i=1}^{m} p_i \log_2(p_i) \right) \quad (5)$$

Therefore, the information gain of attribute A is InfoGain(A) = Info(D)-Info$_A$(D$_j$). According to information theory knowledge, the smaller the expected information is, the greater the information gain is, thus, the higher the purity is; on the contrary, the purity is much lower.

The classification pattern of ID3 algorithm is simple and is robust against noise data, but the inductive bias of the algorithm can cause the fact that the tree branch with shoal depth appears earlier than the deeper one. In order to improve the original defects of ID3 algorithm, we add an additional correction parameter f(n) = 1/n. Where n represents the number of value of each decision attribute and the larger value of n is, the smaller the correction function f(n) is. Then we define a new standard of gain:

$$\text{InfoGain}'(A) = f(n)\text{InfoGain}(A) \quad (6)$$

We can avoid bias and the corrected gain will decrease while the attribute increases which will result in the corrected gain won't be selected as the upper attribute easily.

**Clustering analysis based on K-means:** Clustering analysis (Han *et al*., 2011) is mainly used for similarity class cluster research which is essentially to search a sign which can reflect the affinity-disaffinity relationship between samples or between variables objectively, then according to the sign samples or variables, samples and variables are divided into several classes. The principle of work is: Choose K points randomly from the data set as the initial clustering center, then calculate the sample's distance from the center of the cluster and returns the samples to the class which is the nearest the cluster center.

Calculate the mean value of each cluster data object new formed to get the center of a new cluster, if two adjacent cluster centers have no difference, the sample adjustment is completed and the clustering criterion function is converged.

The algorithm uses squared error sum function (Jolla L., 1992) as the standard to evaluate clustering performance. A given data set X which only contains the descriptive attribute, excludes the class attribute. Determine the initial clustering midpoint $\{c_1, c_2, \ldots, c_k\}$, after the completion of cluster for the first time, all kinds of subset $\{X_1, X_2, \ldots, X_k\}$, where $c_i \in X_i (i = 1, 2, \ldots, k)$. The number of samples in each cluster subset is denoted as $n_1, n_2, \ldots, n_k$ and the mean value of each cluster subset is $\{m_1, m_2, \ldots, m_k\}$. Compare it with the initial clustering midpoint $\{c_1, c_2, \ldots, c_k\}$: If two comparisons remain unchanged, the clustering center remains constant. Otherwise, assign the mean value to the center point and then continue the next iteration. The squared error sum criterion function can be expressed as following:

$$E = \sum_{i=1}^{k} \sum_{p \in X_i} \|p - c_i\|^2 \quad (7)$$

If the criterion function converges, we should terminate the cluster analysis; otherwise, continue iterations of the clustering process called K-means.

**Association rules mining based on frequent itemsets model:** Association rules mining is to search the information hidden behind the data, that is, to dig out the relationship that relies on each other to meet the dependency from a large number of data. Apriori algorithm is the most famous in numbers of frequent itemsets mining algorithms.

The basic idea of the algorithm is: first generate a one-dimensional frequent itemset called collection $L_1$, then use $L_1$ to generate a two-dimensional frequent itemset called collection $L_2$ and use $L_2$ to generate a two-dimensional frequent itemset called collection $L_3$. It will not stop, until it can't generate higher dimensional frequent itemset. The most important step is to use the frequent itemset $L_{k-1}$ to generate a frequent itemset called collection $L_k$: First, connect $L_{k-1}$ to itself to generate candidate itemset K-dimensional data itemset $C_k$ and then make frequency statistics for the data itemset in $C_k$ and discard the data itemset whose frequeny is lower than the minimum support to form the set of frequent itemset.

The algorithm with an iterative way layer by layer has the advantage of straightforward theoretical derivation and is easy to implement. However, this algorithm scans the database too many times; and the I/O load of system is very large under the limited memory capacity, thus, the efficiency is low. Therefore, the database is partitioned into multiple segments in Apriori algorithm which can reduce the times of scanning database by using Partitions (Gordon and Vichi, 2001) algorithm to make it scan the database twice at most which reduces the number of I/O in the system and improves the efficiency of the algorithm's implementation to a certain extent.

Mining portfolio model

**Combined model:** The key of the combination forecasting model (Batesand and Granger, 1969) is the determination

Table 1: Mining results of combining simulation

| Mining model | Coverage ratio (%) | Temperature Ta (Celsius degree) | Pressure of supply water PSb(MPa) | Pressure of return water PRc(MPa) | Temperature of supply water Tsd Celsius degree) | Temperature of return water Tre (Celsius degree) | Instantaneous heat Ihf (GJ $h^{-1}$) | Water supply instantaneous flow Isg ($m^3$ $h^{-1}$) | Water return instantaneous flow Irh ($m^3$ $h^{-1}$) |
|---|---|---|---|---|---|---|---|---|---|
| Decision tree | 44.58 | -8.8860 | 0.5550 | 0.5170 | 92.1880 | 66.6480 | 35.7480 | 298.9970 | 2.2810 |
| Classification | 13.56 | -3.3500 | 0.5550 | 0.5170 | 92.1880 | 57.6350 | 16.6340 | 253.5120 | 2.2810 |
| | 7.60 | 10.5920 | 0. 454 | 0. 447 | 69.3270 | 44.4070 | 16.6340 | 220.1760 | 2.2810 |
| | - | - | - | - | - | - | - | - | - |
| K-means | 25.00 | -8.4247 | 0.6253 | 0.5812 | 97.5250 | 67.9083 | 32.3167 | 300.0000 | 2.7917 |
| Clustering | 23.00 | -4.5109 | 0.5644 | 0.5252 | 94.6636 | 67.2515 | 34.9485 | 298.8848 | 2.5121 |
| analysis | 16.00 | -0.1867 | 0.4860 | 0.4594 | 77.2333 | 54.9000 | 22.5917 | 262.5458 | 2.6542 |
| | 14.00 | 8.9775 | 0.6277 | 0.6087 | 67.9200 | 50.7000 | 16.5900 | 228.9100 | 2.5350 |
| | 9.00 | 2.3708 | 0.3518 | 0.3387 | 69.8246 | 44.8869 | 8.0300 | 161.0062 | 3.9408 |
| | - | - | - | - | - | - | - | - | - |
| Frequent | 44.58 | -8.886 | 0.5550 | 0.5170 | 92.1880 | 66.6480 | 35.7480 | 298.9970 | 2.2810 |
| Itemsets rule | 13.56 | -3.350 | 0.5550 | 0.5170 | 92.1880 | 57.6350 | 16.6340 | 253.5120 | 2.2810 |
| | 7.60 | 10.592 | 0. 454 | 0. 447 | 69.3270 | 44.4070 | 16.6340 | 220.1760 | 2.2810 |
| | - | - | - | - | - | - | - | - | - |

Table 2: Real value of the normal average

| Mining model | Coverage ratio (%) | Temperature Ta (Celsius degree) | Pressure of supply water PSb(MPa) | Pressure of return water PRc(MPa) | Temperature of supply water Tsd Celsius degree) | Temperature of return water Tre (Celsius degree) | Instantaneous heat Ihf (GJ $h^{-1}$) | Water supply instantaneous flow Isg ($m^3$ $h^{-1}$) | Water return instantaneous flow Irh ($m^3$ $h^{-1}$) |
|---|---|---|---|---|---|---|---|---|---|
| Decision tree Classification | 44.92 | -8.8860 | 0.5550 | 0.5170 | 92.1880 | 44.4070 | 35.7480 | 298.9970 | 2.2810 |
| K-means Clustering analysis | 100.00 | -2.1691 | 0.5295 | 0.5001 | 83.1371 | 58.1872 | 26.3225 | 262.1526 | 2.6523 |
| Frequent itemsets rule | 44.92 | -8.8860 | 0.5550 | 0.5170 | 92.1880 | 44.4070 | 35.7480 | 298.9970 | 2.2810 |

of weight coefficients, so we introduce the design of combined model to determine the weight coefficients. In the composite model of heat supply network state prediction, we define X as the value of network's running status. The actual vector-valued is $(X_1, X_2,\ldots,X_8)^T$, $X_1'$, $X_2'$, $X_3'$ represent the results of three kinds of prediction models which includes decision tree classification prediction, K-means clustering analysis and frequent itemsets rules, respectively. The elements in the vector $W = (\lambda_1, \lambda_2, \lambda_3)^T$ are the weight of corresponding model, respectively. Where, the decision tree classification prediction and frequent itemsets mining association rules of data are the same, as a result, the proportion is the same, that is $\lambda_1 = \lambda_3$ Thus, we get the combination forecasting model:

$$X = \sum_{i=1}^{3} \lambda_i X'_i = \lambda_1 X'_1 + \lambda_2 X'_2 + \lambda_3 X'_3 \quad (8)$$

Fitting offset:

$$e_{ji} = X_j - X'_{ji}, \quad i = 1,2,3 \quad j = 1,2,\cdots,8 \quad (9)$$

$$e_j = \sum_{i=1}^{3} \lambda_i e_{ji} = \sum_{i=1}^{3} \lambda_i (X_j - X'_{ji}) \quad (10)$$

According to the optimal weight of combined forecasting method, we establish the mathematical model which is shown as following:

$$\begin{cases} \min Q = \sum_{j=1}^{8} e_j^2 \\ st. \sum_{i=1}^{3} \lambda_i = 1, \lambda_i \geq 0 \end{cases} \quad (11)$$

The combination type is a constrained optimization problem and the method of solving constrained optimization problems needs gradient information, to obtain the continuously differentiable objective function or differentiable constraint condition. According to this constraint combination, we can get the weight of heat supply network's combination forecasting model.

**Simulation experiment:** Decision tree classification, clustering analysis and frequent itemsets are three kinds of mining algorithms which predict the network status value of heat exchange station. Their mining results are listed in Table 1.

Bring the three status value of the mining algorithms from the above list back to the above formula, we can calculate the weight proportion of each algorithm:

$\lambda_1 = 0.317$, $\lambda_2 = 0.366$ and $\lambda_1 = 0.317$. The state prediction model of heat supply network is $X = 0.317X'_1 + 0.366X'_2 + 0.317X'_3$, where $X'_1$, $X'_2$, $X'_3$ represent the prediction models of decision tree classification prediction, k-means clustering analysis and frequent itemsets rule, respectively. The normal average situation of mining algorithms is listed in Table 2.

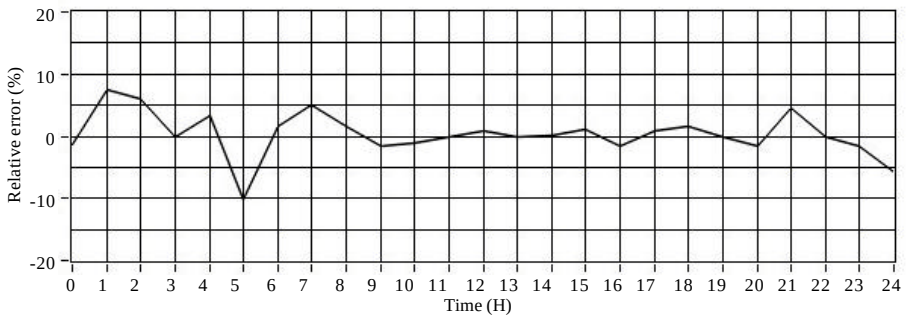


Fig. 3: Actual and the predicted curve based on pressure model of supply water

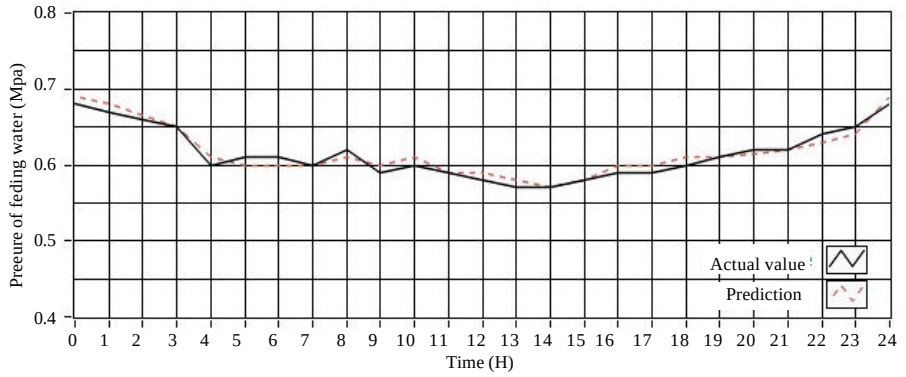


Fig.4: Relative error curve of pressure model

Table 3: Predictive value of the normal average

| Temperature Ta (Celsius degree) | Pressure of supply water Psb (MPa) | Pressure of return water Rc (MPa) | Temperature of supply water Tsd (Celsius degree) | Temperature of return water Tre (Celsius degree) | Instantaneous heat Ihf (GJ $h^{-1}$) | Water supply instantaneous flow Isg ($m^3$ $h^{-1}$) | Water return instantaneous flow Irh ($m^3$ $h^{-1}$) |
|---|---|---|---|---|---|---|---|
| -6.4276146 | 0.545667 | 0.5108146 | 88.8753706 | 49.4505532 | 32.298267 | 285.5119496 | 2.4168958 |

Based on the combination forecasting model, we can achieve the normal average prediction listed in Table 3.

## PERFORMANCE ASSESSMENT

According to the combination model, we forecast the heating status of heat exchange station1# on November 29th, 2012. Fig. 3 and 4 show the actual, predicted curve of water pressure and the error curve in combination model.

According to the daily observation, the comparison between the actual and predicted curve of supply pressure shows that, actual supply pressure of heat exchange station1# is maintained at [0.576, 0.689 Mpa] and the predicted value is maintained at [0.578, 0.69 Mpa]. Range of fluctuation is not big and the relative error of the predicted and actual values is maintained at [-10.03, 7.21%] which is within the acceptable tolerance. According to the experiment of comparison of various parameters, it can be learned that the relative error is less than 10%, the error maintains between 1 and 5%. Prediction results are more accurate than before and the situation can lead to deviation bigger easily only when a few trends change quickly. Therefore, the combination forecasting model can conform to the prediction effect of the overall parameters.

## CONCLUSION

In the stage of data preprocessing which depends on the characteristics of network state, we use the equal distance classification, discretization of information

entropy based on rough set and attribute importance. These three methods realize the discretization of different state values. Optimized algorithms for mining data based on these classical data mining methods, include decision tree classification, clustering, frequent itemset pattern. They not only improve the efficiency and accuracy of the original algorithm effectively, but also create the prediction model $X = 0.317X'_1+0.366X'_2+0.317X'_3$ from the mining results. Compared with the results of the original experiment, the proposed scheme demonstrates that the error rate remains below 10% which is within the tolerance range. And the results show that the algorithm based on combination forecasting model has a higher accuracy than that of the single forecasting model.

## ACKNOWLEDGMENT

This study is supported by the National Science and Technology Support Plan Major Project under Grant No. 2012BAH12B03.

## REFERENCES

Bates, J.M. and C.W.J. Granger, 1969. The combination of forecasts. Oper. Res. Q., 20: 451-468.

Bean, C. and C. Kambhampati, 2008. Autonomous clustering using rough set theory. Int. J. Autom. Comput., 5: 90-102.

Chen, B.X., J.S. Yu and H.W. Guan, 2004. The feasibility study of the weather forecast based on the application of data mining. Applied Sci. Technol., 31: 48-50.

Chen, J.H. and X.M. Li, 2010. A method for discretization of continuous attributes based on the conditional entropy of rough set. Sci. Technol. Eng., 10: 3730-3748.

Gordon, A.D. and M. Vichi, 2001. Fuzzy partition models for fitting a set of partitions. Psychometrika, 66: 229-247.

Han, J., M. Kamber and J. Pei, 2011. Data Mining: Concepts and Techniques. 3rd Edn., Morgan Kaufmann Publishers, USA., ISBN-13: 9780123814791, Pages: 744.

Kantardzic, M., 2003. Data Mining: Concepts, Models, Methods and Algorithms. 1st Edn., Tsinghua University Press, China, ISBN-13: 9787302307143.

Li, H., D. Yan, Q. Li and L. Han, 2010. A new discretization algorithm for continuous attributes based on rough set theory. Appl. Res. Comput., 27: 77-78.

Li, S.H. and Z.W. Zhang, 2011. ID3 decision tree algorithm in weak. J. Changchun Univ., 21: 67-69.

Miao, D.Q. and D.G. Li, 2008. Rough Set Theory, Algorithm and Application. 1st Edn., Tsinghua University Press, China, ISBN: 9787302165521.

Wang, D.H., 2002. The analysis of on-line monitoring system for city heating network. Master's Thesis, Dalian University of Technology, China.

Wang, H.Y. and H.Y. Wang, 2009. The research progress and development tendency of data mining. Technol. Plaza, 9: 237-238.

Zhang, J., G.B. Cui and Y.Q. Chen, 2008. Discussion on the heating status and heating plan of Anshan. Sci. Technol. Innov. Herald, 25: 150-150.