International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.4 (2015), pp.53-62
http://dx.doi.org/10.14257/ijmue.2015.10.4.06

Wind Data Preprocessing Algorithm Based on Extracting Isolated Points

Guo Xiaoli, Yu Ying, Wang Ling, Qu Zhaoyang, Wang Yongwen

College of Information Engineering, Northeast Dianli University, Jilin Province
243589657@qq.com,173647508@qq.com,simle2867ling@gmail.com,qzywww@mail.nedu.edu.cn,995802183@qq.com

Abstract

Now existed methods of wind data preprocessing methods have bad influence on data authenticity because of "valuable isolated points" filtered out. Actually, these valuable isolated points contain very helpful information for the wind prediction. Therefore, we proposed a novel wind preprocessing algorithm called S-time, which focus on valuable isolated points extraction. Extracting several special points, so as to restore the data authenticity and reduce the memory space; intuitive trend can be converted by raw data via extracting "special points", thus improve the efficiency of forecasting; During the process, we proposed a novel method for isolated points detection， which is proved that more effective to identify isolated points. We extract several power values as experimental data, it can quickly calculate generated electricity according to the extracted points. The experiment proves that S-time is effective by comparing the relative error, which is only 0.58%. At the same time, proved that can improve the accuracy of power prediction by comparing with stratified sampling.

Keywords: *data preprocessing; isolated points; wind prediction*

1. Introduction

Nowadays, more and more attention is paid on wind energy—a kind of clean and renewable energy. While developing the wind power, we should also keep a watchful eye on the ability of real-time data processing, in order to make the wind power develop healthily and rapidly, taking the road of sustainable development is the ultimate goal. Now wind power data is usually applied in wind power prediction, which is benefit for reducing the shock of wind power on the grid, and improving the economy of the grid operations. As we know, effective wind data preprocessing is the key to wind power forecasts. Provide clean, accurate data for, data mining, thus reduce the amount of data processing, and then deduce the valuable information.

Although there are a lot of methods of data preprocessing [1-6], few of which apply in the wind power data. Wind data preprocessing existed mainly study on attribute reduction, missing values, isolated points, but offer few reference value for prediction. To solve these problems, we proposed a wind data preprocessing algorithm based on extracting special points: S-time. As wind data is time series data, and most of prediction methods focus on real-time prediction, we add time slice to data analyze. There are some features of wind power data, such as having a big amount of data, the overall fluctuation is large, having no obvious periodicity, but some regulations are still there. The algorithm is mainly visually display the overall trend of the raw data by extracting the data center of the clusters and isolated points, which make the overall distribution of the raw data clearly showed, provide reliable and efficient reference value for prediction, while reducing the amount of data , reduced storage space. When observe the data carefully, we can find most of point are dense except certain "noise", regard these dense point as a

ISSN: 1975-0080 IJMUE

cluster, then we will get many clusters, it is called an isolated point which are not included in the cluster—"noise", then the value of center point is the average value of the entire cluster, which can instead of the cluster points' value. For isolated points, don't just reject, because data preprocessing is based on the authenticity of the data provided to restore the value of data mining, and isolated points have great influence on the analysis results, so it is necessary to find out the accurate outlier. This algorithm is to make the trend visualized clearly, which via extracting the center of clusters, isolated points, not only the overall distribution of the original data clearly shown, provides a reference to the prediction value, but also reduce the required memory space for storing data. If the processed data is power value, you can quickly calculate the amount of electricity through the points after extraction. Extracting data as experimental data from power data, the relative error was proved with the original data generation capacity is only 0.58%.

2. Related Work

Data preprocessing work is extremely important for the entire data mining tasks, many failures of data mining tasks are caused by the low quality of the data. As the importance of data preprocessing makes it as a research focus. In general, data preprocessing can be divided into four parts: data cleaning, data integration, data transformation and data reduction. Existed methods of wind data preprocessing don't receive sufficient attention. The mainly methods are as follows: Zhou Songlin applied PCA to the wind power input variables in order to extract the main component, which can reduce the dimension of input variables [1]. Chen Zehuai proposed a data processing based on RBF neural network, (1) complement on the vacant data (2) find and fix on the anamorphic data [2]. R.R.B applied the principle of wavelet singularity to isolated points detection on wind data [3]. Zhang Femg applied stratified sampling to reduction, which is nine times as accurate as a simple random sampling. Through compare with stratified sampling to verify S-time can improve the accuracy of power prediction.

Clustering algorithm based on density is a focus of study, the basic idea of the algorithm is: by measuring the number of objects included in the cluster, to form a cluster by clustering algorithm. Representatives algorithms: DBSCAN algorithm, OPTICS algorithm. DBSCAN algorithm has a high density zone for clusters and the clusters which are random shape can be found in the spatial database with the "noise". Representative OPTICS sorting algorithm generates a parametric clustering based on the density of the structure of a database, you can get the same density-based clustering results through this sort of information and parameter settings contained Make the idea of density clustering algorithm integrate into the wind data preprocessing, the characteristics of wind power data are large fluctuations, have no obvious periodicity, but in a certain period of time the data changes remain stable, forming a data cluster by clustering, the data fluctuate small within a cluster, the value located in the center position of the cluster within the cluster can be regarded as the average data, screening of the center point to represent the value of entire cluster, thus reducing the amount of data.

Outlier detection has always been a hot research of data mining field. Liao Guoqiong presented a local stream outlier detection algorithm based on distance LSOD and global stream outlier detection algorithms based on approximate estimates GSOD [5]. LSOD greatly reduces the storage space requirements of the stream data and the time of node update. GSOD reduces the communication and computation load center node; In the reference [6], the author came up with a outlier detection algorithm based on the average density, detect isolated points by comparing isolated points with average density, the detection of this algorithm is more automate, and normally it doesn't depend on the input parameters by user, but in large-scale data, it need consider the sample to determine the average density and the average distance.

3. S-time Algorithm Description

Process several times on the original data, filter the "special points" from the original data, thereby complete the pretreatment.

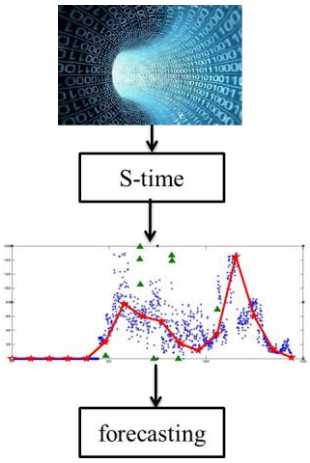

Figure 1. The Whole Process

3.1 Screen the Center Points of the Cluster and Suspicious Isolated Points

This article makes the idea of density clustering algorithm integrate into the wind data preprocessing. First apply cluster method which based on density to forming clusters on the original data, and then filter several important points: the center point of the cluster, the maximum point and the minimum point, suspicious isolated points.

The region which has a sufficiently high density is divided into clusters, it is defined as the maximum congregation of density linked. These definitions are shown first.

For an object in space, if in a given neighborhood of radius *Eps* contain at least the minimum number of *MinPts* object, then the object is called a core object.

Given a set of objects S, if p is in the neighborhood of q and q is a core object, we say that the object p q starting from the object is directly density-reachable.

If a neighborhood of the point p contains more than *MinPts* points to create a new object p cluster core. Then, iteratively find objects directly from these core density-reachable objects, this process may involve some density up clusters merge. When a new point can't be added to any cluster, the process ends.

Screen cluster the center point of the generated cluster, the maximum point and the minimum point and suspicious isolated points.

Pseudo-code：

Input：Database S, the minimum number *MinPts* and radius *Eps*;
Output: a list;

Begin
C ← defined a bunch；
For all point **do**
m ← a point of S；
If m > *MinPts*
C ← m;
Else
list←m;
End
Maxi,Mini←Maxmum and Minmum of C
m′ ←center point of C;
list←m′ , Max and Min;
Return list;

From the description, we know that in the original data clustering, selecting *Eps* and *MinPts* parameter values will directly affect the clustering results and determine isolated points. Inappropriate parameter values will regard some points which great degree of dispersion as mistaken isolated points, so after a cluster screening, these isolated points are suspicious isolated points. In fact, which itself belongs to the class of data objects is determined by the properties of its own, as isolated points are the real points, and have a great influence on the data analysis, so need find the identified isolated points. In this article, if the value of *MinPts* and *Eps* too large, it will incorrectly judged the isolated points as clusters within points, which will directly affect the outcome of isolated points identified, so it is crucial to select *Eps* and *MinPts*. That 1/25 value of set data is taken as *MinPts* data is an effective way [8].

3.2 Determine Isolated Points

Here is the basic concept of an isolated point:

In probability theory, the appearance of isolated points can be seen as a small probability event, therefore its probability is less than 5%, if the existing data follow a normal distribution, according to the normal distribution "3δ" principle $P\{\mu-3\sigma<X\leq\mu+3\sigma\}=99.7\%$, only 0.3% of the points fall outside of the distribution, so if a point falls outside the range of 99.7%, it can be regarded as an isolated point.

From Chebyshev inequality, let the random variable X with mean μ and variance δ^2, then for any small $\varepsilon>0$, $P\{|X-\mu|\geq\varepsilon\}\leq\frac{\delta^2}{\varepsilon^2}$,no matter what the distribution of X, for any normal number k has $P\{|X-\mu|\geq k\delta\}\leq\frac{1}{k}$,when k=3, $P\{|X-\mu|\geq 3\delta\}\leq\frac{1}{9}$, if the data follow a normal distribution, from a normal distribution "3δ principle".

$P\{|X-\mu|\geq 3\delta\}=0.3\%$, for random variables X, if $|X-\mu|\geq 3\delta$, then X is the isolated point. Seen this formula to determine whether a point is isolated point need for the mean and variance, in the reference [7], the author considers that isolated point will significantly affect the mean and variance, so use median to estimate the variance. $\gamma=0.6745\delta$(Assuming the whole data follow random variable $\zeta\in N(0,\delta^2)$, γ is median，δ is variance), as the outcome is by estimation, there must be deviation. But, in this article, we don't need to have this concern, as the data from this article is with time, we can make the suspicious points correspond to the timeline's specific location, pack data between the two adjacent points, after excluding isolated points, calculate the mean of the packed, this moment, the outcome is excellent accurate, calculate the variance according to $D(x)=\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N-1}$, In the aspect of accuracy, this method is better than the method that median estimate variance.

3.3 Characterization of S-time Preprocessing Algorithm

After S-time, we screen several special points, such as the center point, regarding which as average, think that if the data to be processed is power value, not only can visually show a clear trend of power values, but also can quickly calculate the amount of electricity based on the extraction of isolated points and the center points of the cluster. The special points after extraction represents by box diagram, power values inside the box marked as the center point of the clusters, the center value of cluster can be expressed as the average of the cluster, which can represent average power value of a period of time, t_1

, $t_2,\ldots,t_n$ are cluster center value corresponding to the x-axis time, t_u is the final value of the x-axis. According to, we can rapidly conclude the amount of wind power.

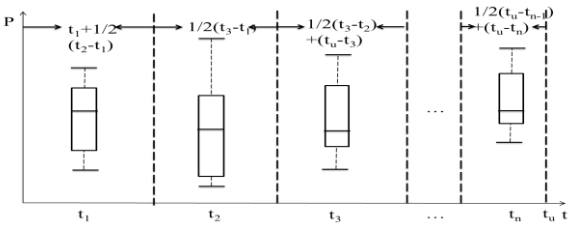

Figure 2. Showing by Box Figure

$$E = \sum \bar{P}t = \bar{P}_1[t_1 + 0.5(t_2 - t_1)] + \bar{P}_2[0.5(t_2 - t_1) + 0.5(t_3 - t_2)] + \ldots + \bar{P}_n[0.5(t_n - t_{n-1}) + (t_u - t_n)]$$

Wind power data preprocessing algorithm based on S-time restore the authenticity of the data while reducing the amount of data, reduces the memory space that required for data storage. Combining the thought of clustering with data preprocessing, screening valuable information, offer help for the next job. In the process of determining the isolated-point, use the mean instead of the median to find the variance, improving the accuracy of the calculation, so isolated points determined is more accurate, by screening valuable information to calculate the amount of wind power generation.

4. Experimental Analysis

Experimental data from a wind farm in Jilin Province, the purpose of this article is to facilitate the forecasting of wind power data, forecast of wind power is mainly about wind velocity and wind power, we select power values to analyze and verify. Data collected once per second, extract a total of 86,400 data, as characteristic of wind power data is fluctuations in the steady period, too much data will increase the difficulty, so transform the average value of data within 1minite, improve the collection frequency at the same time can reduce the amount of data to 1440.First need to determine the validity of the sample data extracted, that is, whether the sample data is missing, invalid non-numeric data. If there is missing or non-numeric and invalid, in this paper, adopt the method of Mean value interpolation to fill and replace.

4.1 Preliminary Screen of Raw Data

In the paper, apply MATLAB, make use of clustering algorithm based on density, preliminary screening suspicious isolated points and the center point of clusters, at the same time, exporting the maximum and minimum values within clusters. Figure 3 is the power distribution of the data.

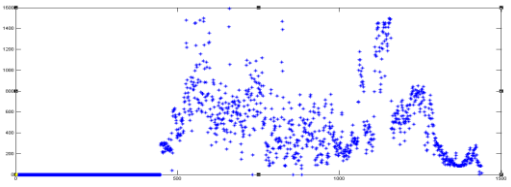

Figure 3. Wind Power Distribution

Wind data is collected in chronological order, is timing data, the x-axis represents time in Figure 3, y-axis represents the power value corresponding to the time, as can be seen

from Figure 3, although wind power data is large fluctuations, cycle is not obvious, but it is not no rules to follow, within a certain period of time, the power values collected are small fluctuations, apply the algorithm referred in 3.1 to screen special points: center points、isolated points、max and min, here we take *MinPts*=57，*Eps*=0.0875[8], the results shown in Figure 4.

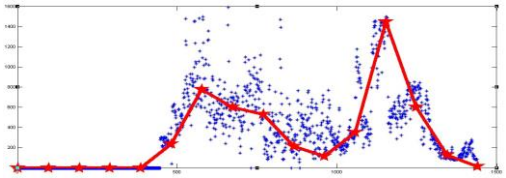

Figure 4. Preliminary Screening Point

The red stars are the center of the clusters, a total of 16 center points, maximum、minimum of 16 clusters in order.

Table 1. Max and Min of Cluster

Number	Max	Min
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	1467.26	0
7	1486.62	271.07
8	1085.08	283.45
9	1219.24	84.43
10	763.73	55.66
11	717.54	92.54
12	1087.42	120.01
13	1496.93	329.43
14	858.08	181.09
15	357.03	85.98
16	282.23	3.03

There are 22 suspicious isolated points: (482,39.40)(651,241.05)(658,1413.59)(659,1589.47)(660,1054.02)(696,144.28)(697,211.81)(730,0)(731,188.12)(821,1075.74)(822,1466.45)(823,1390.53)(824,991.81)(856,0)(883,0)(973,767.66)(1056,349.95)(1057,697.07)(1061,1175.22)(1106,555.74)(1159,702.41)(1186,286.34).

4.2 Secondarily Screen Special Points

Put suspicious isolated points corresponding to the timeline, pack data between two adjacent isolated points, through data packed, the overall total data is divided into 15 parts, due to suspicious isolated points have been removed, its mean-variance will not be affected by isolated points, calculate the mean and variance of each data within package, table1、table2 are respectively represent the variance calculated through mean and the standard deviation through median, Figure 4 shows two ways to calculate the standard deviation:

Table 2. Mean, Variance, Suspicious Outlier

No	ave	Var	SD	Suspicious
C_1	267.24	1585.19	39.81	(482,39.40)
C_2	700.47	71052.52	266.56	(651,241.05）
C_3	589.84	7941.94	89.12	(658,1413.59)
				(659,1588.47)
				(660,1054.02)
C_4	574.98	17373	131.80	(696,144.28)
				(697,211.81)

C_5	674.45	27665	166.33	(730,0) (731,188.12)
C_6	502.66	54366.67	233.17	(821,1075.74) (822,1466.45) (823,1390.53) (824,991.81)
C_7	354.23	6915.76	83.16	(856,0)
C_8	291.03	34576.51	185.95	(883,0)
C_9	379.72	28881.44	169.95	(973,767.66)
C_{10}	296.32	5335.33	73.04	(1056,349.95)
C_{11}	929.40	29576.94	171.98	(1061,1175.22)
C_{12}	537.25	70946.95	266.36	(1106,555.74)
C_{13}	1267.2	38429.81	196.04	(1159,702.41)
C_{14}	514.34	4769.45	69.06	(1186,286.34)
C_{15}	345.28	64700.84	254.36	(1186,286.34)

Table 3. Median, Variance, Suspicious Isolated Points

No	M	SD	Suspicious
C_1	0	0	(482,39.40)
C_2	632.90	426.89	（651,241.05）
C_3	608.23	410.25	（658,1413.59）（659,1588.47）（660,1054.02）
C_4	588.57	396.99	（696,144.28）（697,211.81）
C_5	689.92	465.35	（730,0） （731,188.12）
C_6	480.15	323.86	（821,1075.74）（822,1466.45）（823,1390.53）(824,991.81)
C_7	323.34	218.09	（856,0）
C_8	243.54	164.27	（883,0）
C_9	388.69	262.17	（973,767.66）
C_{10}	278.53	187.87	（1056,349.95）
C_{11}	994.04	670.48	（1061,1175.22）
C_{12}	437.41	295.03	（1106,555.74）
C_{13}	1325.6	894.09	（1159,702.41）
C_{14}	504.66	340.39	（1186,286.34）
C_{15}	233.43	157.45	（1186,286.34）

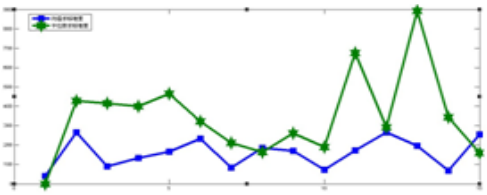

Figure 5. The Results Comparison Chart

From Figure 5 and Table 2, 3, we the standard deviation calculated by two methods have a big difference, now through isolated points determined to judge the accuracy of the two methods.

After finding the mean and variance of $C_i\{C_1, C_2,..., C_{15}\}$, detect the suspicious isolated points $P_i\{p_{i1}, p_{i2}, \ldots, p_{ij}\}$, for p_{ij}, outlier detection using the proposed algorithm to detect the isolated points, calculate $\min\{|p_{ij}-\mu_{i-1}|, |p_{ij}-\mu_i|, |p_{ij}-\mu_{i+1}|\}$, and compare it with the size $3\delta_{ig}$, when g=1, represent that find variance through mean, when g=2, represent that find variance through median. First, take out a suspicious outlier (482,39.40), calculate $\min\{|39.40-267.24|, |39.40-700.47|\}=227.84$, $3\delta_{11}=119.44$, $3\delta_{12}=0$, as 227.84>119.44, judge (482,39.40) is an isolated point by mean find variance.

According to the above process, respectively compare $|p_{ij}-\mu_i|$ with $3\delta_{ig}$ by mean find variance and median find variance. If judged isolated points by mean find variance:(482,39.40)(658,1413.59)(659,1588.47)

(660,1054.02)(730,0)(822,1466.45)(823,1390.53)(856,0)(1057,697.07),and by median find variance:(482, 39.40).

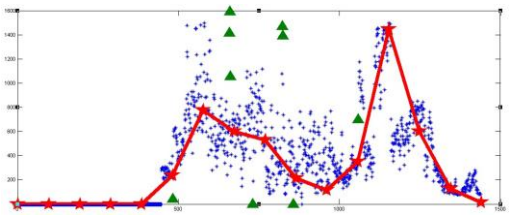

Figure 6. Isolated Points Identified by the Method of Median Seeking Variance

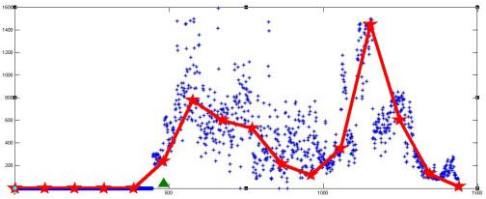

Figure 7. Isolated Points Identified by the Method of Mean Seeking Variance

In the picture, the red stars are the center of the clusters, labeled by green triangle are isolated points. From the comparison between Figure 6 and Figure 7, we can distinctly see that mean-variance is more accurate than median-variance.

4.3 Verify the Effectiveness of S-time Algorithm

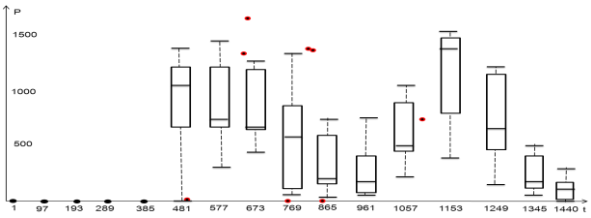

Figure 8. Showing by Box Figure

In order to verify the validity of this algorithm, this paper via calculating generated electrical energy to analysis and verify. According to the distribution in figure 8, on the basis of 3.3, red points are power values of isolated points. Assuming generation of one day is E, the generation of the original data is $E_1 = \sum_{m=1}^{1440} P_m t$, Where P is the power value, computing power generation according to the special extraction $E_2 = \bar{P} t$, put the data into it, E_1=576842Wh，E_2=573496Wh，△E=3346Wh，relative error=△E/E_1=0.58%, then prove that the algorithm has authenticity and effectiveness.

4.4 Verify S-time can Improve the Accuracy of Power Prediction

S-time algorithm proposed to reduce the amount of data, while preserving the value information of the data, which belongs to the scope of data reduction. For data reduction of single dimension, the common method is data sampling, in the current sampling methods, hierarchical sampling has a higher precision [4], combined with distribution of wind power data, we can know that this method is suitable for wind power data.

Therefore, in this article, through compare hierarchical sampling [4] with S-time to verify S-time can improve the accuracy of power prediction, respectively put "special points" filtered after 3.2 and data after hierarchical sampling, after BP neural network prediction model to predict the power of data, to verify whether it has improved accuracy. Let hierarchical sampling ratio of 1:15, the common method of evaluate prediction accuracy is mean absolute error (MAPE), according to the formula, $MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{P_t^{'} - P_t}{P_t}\right|$, calculate MAPE, where P_t ' is predicted value, P_t is actual value.

Figure 9 is comparison of predictive value among S-time、hierarchical sampling and the actual value, Figure 10 is MAE of different data processing algorithms.

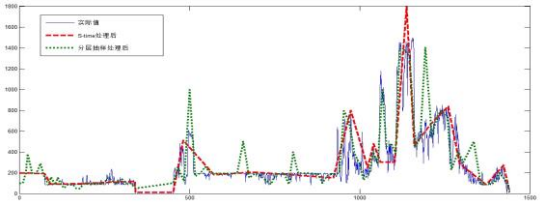

Figure 9. Comparison of Power Values

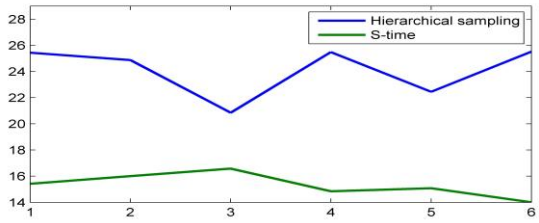

Figure 10. Comparison of MAPE (per 4 Hours)

Calculate MAPE, $MAPE_1$=0.15,$MAPE_2$=0.27(the subscript 1、2 respectively treated with S - time algorithm, hierarchical sampling). As can be seen from figure 9, use data after S-time process as test data has a higher the goodness of fit with the actual value. From Figure 10, we can know that in different periods the average absolute error after S-time process is more accurate than stratified sampling. Through the above comparison shows that S-time data preprocessing algorithm can improve the accuracy.

5. Conclusion

In this paper, against the deficiency of existing wind data processing, proposed a data processing method based on special points extraction S-time algorithm, combining the thought of clustering based on density with preprocessing, preliminary filtering the special points, then determine isolated points via the method of outlier determination that this paper proposed for secondary screening, finally, use the special points to represent the whole trend of wind power. Through the experiment, the method can accurately filter out the special points of data, the extraction of special points make data distribution more intuitive, can effectively improve the prediction efficiency, at the same time, as reduce the amount, greatly reduce the required memory space. By comparing with stratified sampling to verify the S-time wind data preprocessing algorithm can improve the prediction accuracy of power.

References

[1] S.-l. Zhou and J.-h. Su, "Forecasting of wind power with principal component analysis and ANN [J]", Power System Technology (**2011**).
[2] Z.-h. Chen and Z.-g. Wu, "Application of RBF neural network in medium and long—term load Forecasting [J]", Proceedings of the CSU—EPSA (**2006**).
[3] R. R. B Lira, "Application of wavelet and neural network models for wind speed and power generation forecasting in a Brazilian experimental wind park [J]", Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA (**2009**).
[4] F. Zhang, "Measurement technique for high-speed network based on stratified sampling [J]", Journal of Jilin University (**2004**).
[5] G.-q. Liao and J. Li, "Distance-based isolated point detection for distributed RFID data streams [J]", Journal of Computer Research and Development (**2010**).
[6] H.-j. Shi and X.-y. Li, "Average density-based outliers detection [J]", Journal of University of Electronic Science and Technology of China (**2007**).
[7] L.-d. Cai and Y. Fu, "A robust isolated points detection method--Variance from Median [J]", Computer Science (**2006**).
[8] M. Daszykowski, B. Walczak and D L. Massart, "Looking for natural patterns in data [J]", Chemometrics and Intelligent Laboratory Systems (**2010**).
[9] J. Wen, Z. Yuan and Z. Li, "Study on modeling of wind generator based on data reduction preprocessing [C]", IEEE Innovative Smart Grid Technologies – Asia (**2012**).
[10] T. Mathaba and X. Xia, "Short-term Wind Power Prediction using Least-Square Support Vector Machines [J]", Power Engineering Society Conference and Exposition in Africa(Power Africa) (**2012**).
[11] G. Wang and J. Li, "Boundary detection methos in support of k-outlier degree [J]", Computer Engineering and Applications (**2011**).
[12] F. Shao and Z. Yu, "Principle and algorithm of data mining [M]", 2nd ed, Beijing: Science Press (**2009**).
[13] S. Fan, J R. Liao and R. Yokoyama, "Forecasting the Wind Generation Using a Two-Stage Network Based on Meteorological Information [J]", IEEE Trans on Energy Conversion (**2010**).
[14] C. W. Potter and M. Negnevitsky, "Very short-term wind forecasting for Tasmanian power generation", IEEE Trans. On Power Systems [J], (**2011**).
[15] C. Sanjay and P. Sun, "SLOM: A New Measure for Local Spatial Outliers [J]", Knowledge and Information Systems (**2010**).

