Journal of Applied Science and Engineering, Vol. 21, No. 1, pp. 33–44 (2018) *DOI: 10.6180/jase.201803_21(1).0005*

Anomaly Detection Analysis for District Heating Apartments

Wenhui Sun, Dongyan Chen* and Wei Peng

School of Control Science and Engineering, Shandong University, Jinan, 250061, P.R. China

Abstract

With the increasing popularity of the heat meter readings for district heating apartments in China, the analysis of the generated big data is becoming a critical problem. With the nonlinear of the district heating household dataset, this archive describes a kernel Gaussian mixture cluster (KGMC) based data mining algorithm within which the original data in low-dimensional space are projected into the high-dimensional space to do clustering and identify anomalies. At the meantime, this article adopts Gaussian kernel function to prevent the curse of dimensionality. According to the implementation of the experiment with Spark, the data from 18 zones of 17,000 apartments belonging to 6 substations have been studied and four kinds of anomalies have been identified. With the detection and correction of abnormal actions, 5.4% of the demand of heat will be proactively reduced in heating areas in China. Meanwhile, with the comparison, the proposed KGMC can outperform K-means and Gaussian Mixture Model (GMM) methods in terms of detection rate (DR) and false positive rate (FPR).

Key Words: District Heating Apartments, Anomaly Identification, Kernel Gaussian Mixture Cluster, Gaussian Kernel Function

1. Introduction

In China, the area adopting district heating systems has been up to 6.11 billion square meters before 2015 [1], and residential buildings account for about 70%. District heating systems always consist of three parts: heat generation, substations and district heating network which also has drawn much attention for the potential of optimization. Since 2015, devices called automatic meter readings have been installed aiming to refer the government to the monthly billing of actual heat usage. In one hour, for one apartment there're about ten values need to be measured which include heat, flow, and supply and return temperatures, etc. For the heating company which supplies heat to 500,000 square meters and takes charge of 10,000 apartments, there are over 100,000 values to be collected and stored [2]. But the generated big dataset in the field of district heating apartments have not been deeply analyzed, therefore, it's difficult to manage and optimize the heating system for energy saving. Compared with the United States, Japan and other developed countries in the same latitude, the building heating energy consumption per unit area in China is 2–3 times more [1]. Furthermore, the damages caused by the anomalies in district heating buildings will be transferred via the substation to the heat generation and lead to the further decline in energy efficiency. To optimize the heat-supply regulation and improve the efficiency of the heating system, the anomalies in the district heating system according to the apartments' heating data should be mined, analyzed, and solved. However, to author's knowledge, few studies have applied data mining and other techniques to detect the anomalies of instruments and actions according to the data of district heating apartments.

*Corresponding author. E-mail: chendysd@163.com

There are a lot of articles existing for the improvement of district heating systems. Baldvinsson and Nakata [3] presented a method based on high spatial resolution for designing and operating a district heating system with low temperature and studying its feasibility and energy saving performance. Considering the influence to economic and energy, Wang and Sipilä [4] comprehensively compared the consumptions of energy, costs and flexibility of group and building substation. To design the district heating systems and the parameters, Zaunbrecher et al. [5] constructed a local district heating network, concerning multiple possibilities about network design, security of supply, and choice of energy source. To control and predict heat loads of individual consumers, Shamshirband et al. [6] constructed an adaptive neuron-fuzzy inferences system (ANFIS) for district heating systems. [7,8] investigated heat transfer rate based on the effectiveness analysis of heat exchangers. The concept of anomaly detection is to determine the performance of a district heating system which is unacceptable or incorrect from the expected action [9]. Although for nearly two decades anomaly detection has been a famous research area, only few simple tools and algorithms are commonly employed in district energy aspects. Some literatures [10–14] have described the outliers of substations, and the methods are designed to identify the differences from the preceding heat demands which are thought to be normal. But if the data of meter readings are already more or less faulty, this does not work. Within district heating application, many anomaly detection methods relying on statistical methods [15] which require accurate heat demands to be compared with are employed. However, a majority of the heating networks do not operate optimally—i.e., heat meter readings are not completely correct to be compared with. Numerous comprehensive methods and reviews are described in Chapter 10 of paper [16] for fault detection of heat exchangers. But all of them are linear and parameter variable models that are not appropriate for the nonlinear data of district heating apartments. Isermann [17] used limit checking method which requires experienced operators to set thresholds manually to detect anomaly. The thresholds which are loose or tight cannot be adopted. At the meantime, the anomalies of the district heating apartments' data are not properly paid attention, district heating network are not intensively supervised and the big data analysis is difficult to accomplish.

Because of the lack of prior knowledge of the district heating household data and the difficulty or impossibility obtaining either labeled or purely normal data, this paper adopts unsupervised learning or clustering algorithm, assumes the anomalies are well separated from the data points that are normal. The excessive amount of district heating data make machine learning algorithms needed rather than sample data analysis methods. In recent times, the mainly discussed clustering anomalies detection methods are statistical-based [18], distance-based [19], model-based [20] and density-based [21]. As a model based clustering algorithm, GMM [22] which employs several Gaussian sub-distributiosns to do the clustering, optimize the adaptation of the model and the actual data, get the probability that the point belongs to each cluster, has been used in many fields. But with the significant nonlinearity of district heating household data, this paper presented KGMC approach which combines GMM and kernel function [23] to project the original data into the high-dimensional space to do the clustering. In this case, a new machine learning method which called KGMC is employed to do the anomaly detection for the data of district heating apartments.

This paper will be accomplished by answering the following three research questions:

- According to the significant nonlinearity and volume big dataset of district heating apartments' data, what kind of method can be adopted to identify the anomalies?
- What types of anomalies can be identified in district heating apartments and what are the proportions of the energy that can be saved owning to the solution of the anomalies?
- In which aspect that the proposed KGMC approach is superior to other approaches?

This paper is structured into four sections. In section 2, the existing methods and the proposed idea are described. Section 3 reports the abnormal features in district heating data in China that have never been analyzed

before and the comparison of three clustering algorithms, while conclusions are covered in section 4.

2. Methodologies

In this section, three clustering techniques employed in this study for anomalies detection are briefly explained. K-means, Gaussian Mixture Model and KGMC are respectively carried out.

2.1 K-means Clustering

K-means [24] which belongs to the category of partitioning methods is selected for this study. One of the required input parameters is *k* which means the number of sections that the dataset should be divided into. Every cluster is represented by the centroid which stands for the mean value of the vectors belong to this cluster. Before main iterative procedure of this method, a set-up phase is accomplished where *k* vectors of the dataset being the initial centroids are random options. Every iteration consists of two steps. In step one, we assign each vector to the cluster so that its centroid is the nearest one. In step two, centroids are recalculated by calculating the mean of the vectors within each cluster. The iterations will not be stopped until the *k* centroids do not change.

2.2 Gaussian Mixture Model based Clustering

A Gaussian Mixture Model based clustering belongs to model-based methods which use specific models to do the clustering and try to optimize the adaptation of the model and the actual data. It is recalled soft clustering and represents a composite distribution where points are drawn from one of *k* Gaussian sub-distributions, each with its own probability.

Gaussian probability density function is defined in Eq. (1).

$$P(x;\mu,\Sigma)=\frac{1}{\sqrt{2\pi|\Sigma|}}\exp\left[-\frac{1}{2}(x-\mu)^T\sum{}^{-1}(x-\mu)\right] \quad (1)$$

$$\Pr(x)=\sum_{k=1}^{K}\pi_k P(x;\mu_k,\Sigma_k) \quad (2)$$

where x is a vector of some variables. Every component which means one cluster is characterized by mean vector μ and variance Σ. Generally, μ and Σ can be replaced by the mean and the variance of the samples respectively. The vector x will belong to the *i*th cluster when probability P is greater than the threshold of that. In Eq. (2), probability density function can also be described as the sum of that of the components. K which represents the total number of components, needs to be set in advance. π_k is the prior probability of the *k*th component and $\sum_{k=1}^{K}\pi_k=1$. As long as k is large enough, the Mixture Model will become complicated enough and it can be used to approximate arbitrary continuous probability density distribution. Gaussian function has good computational performance for big dataset, so that it will be adopted in this paper.

With sample data, we can calculate the model parameters (π_k, μ_k and Σ_k) with unknown classification. Using the parameters above, the probability distribution will generate given data with the largest probability which means the largest Likelihood Function $\sum_{i=1}^{N}\log p(x_i)$, where N is the number of samples.

With the following EM algorithm, $p(x_i)$ can be solved. Subsequently, the clusters will be classified.

Expectation step:

The probability that the *i*th sample x_i belongs to the *k*th Gaussian component C_k is estimated as the expectation step like

$$p(x_i\in C_k)=\frac{\pi_k P(x_i\mid\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j P(x_i\mid\mu_j,\Sigma_j)} \quad (3)$$

Maximization step:

Then the parameters of the *k*th Gaussian component can be computed as Eqs. (4), (5) and (6).

$$\mu_k=\frac{\sum_{i=1}^{N}p(x_i\in C_k)x_i}{\sum_{i=1}^{N}p(x_i\in C_k)} \quad (4)$$

$$\Sigma_k=\frac{\sum_{i=1}^{N}p(x_i\in C_k)(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{N}p(x_i\in C_k)} \quad (5)$$

$$\pi_k = \frac{\max\left\{0, \sum_{i=1}^{N} p(x_i \in C_k) - V/2\right\}}{\sum_{j=1}^{K} \max\left\{0, \sum_{i=1}^{N} p(x_i \in C_j) - V/2\right\}} \tag{6}$$

where $V = \frac{1}{2}D^2 + \frac{3}{2}D$ and D means the dimensions of random variable x.

Iteratively, EM algorithm will stop until Likelihood Function converges to the stable value.

2.3 Kernel Gaussian Mixture Clustering

2.3.1 Kernel Gaussian Mixture Model

With the significant nonlinearity of district heating data, GMM based clustering may not efficiently classify the different features within different operations. In the following steps, the data is mapped to the higher dimensional feature space via the function $\phi(x)$.

Then the probability density function will be defined as

$$P(\phi(x); \tilde{\mu}, \tilde{\Sigma}) = \frac{1}{\sqrt{2\pi\left|\tilde{\Sigma}\right|}} \exp\left[-\frac{1}{2}(\phi(x) - \tilde{\mu})^T \sum{}^{-1} (\phi(x) - \tilde{\mu})\right] \tag{7}$$

$$\Pr(\phi(x)) = \sum_{k=1}^{K} \tilde{\pi}_k P(\phi(x); \tilde{\mu}_k, \tilde{\Sigma}_k) \tag{8}$$

where $\tilde{\mu}$ and $\tilde{\Sigma}$ is the mean vector and variance of $\phi(x)$ within each component.

Thereby, in the high-dimensional feature space, EM algorithm will be modified as follows.

Expectation step:

The probability that the ith sample $\phi(x_i)$ belongs to the kth Gaussian component $\tilde{C}_k$ is estimated as the expectation step like

$$\tilde{p}(\phi(x_i) \in \tilde{C}_k) = \frac{\tilde{\pi}_k P(\phi(x_i) \mid \tilde{\mu}_k, \tilde{\Sigma}_k)}{\sum_{j=1}^{K} \tilde{\pi}_j P(\phi(x_i) \mid \tilde{\mu}_j, \tilde{\Sigma}_j)} \tag{9}$$

Maximization step:

The parameters of the kth Gaussian component can be computed as Eqs. (10), (11) and (12).

$$\tilde{\mu}_k = \frac{\sum_{i=1}^{N} \tilde{p}(\phi(x_i) \in \tilde{C}_k)\phi(x_i)}{\sum_{i=1}^{N} \tilde{p}(\phi(x_i) \in \tilde{C}_k)} \tag{10}$$

$$\tilde{\Sigma}_k = \frac{\sum_{i=1}^{N} \tilde{p}(\phi(x_i) \in \tilde{C}_k)(\phi(x_i) - \tilde{\mu}_k)(\phi(x_i) - \tilde{\mu}_k)^T}{\sum_{i=1}^{N} \tilde{p}(\phi(x_i) \in \tilde{C}_k)} \tag{11}$$

$$\tilde{\pi}_k = \frac{\max\left\{0, \sum_{i=1}^{N} \tilde{p}(\phi(x_i) \in \tilde{C}_k) - \tilde{V}/2\right\}}{\sum_{j=1}^{K} \max\left\{0, \sum_{i=1}^{N} \tilde{p}(\phi(x_i) \in \tilde{C}_j) - \tilde{V}/2\right\}} \tag{12}$$

where $\tilde{V} = \frac{1}{2}\tilde{D}^2 + \frac{3}{2}\tilde{D}$ and $\tilde{D}$ means the dimensions of high-dimensional feature space.

With the transformation, there will be the curse of dimensionality. To avoid that, kernel function will transform the m-dimensional inner product into n-dimensional input dimension like

$$K_{ij} = K(x_i, x_j) = \left\langle \phi(x_i), \phi(x_j) \right\rangle \tag{13}$$

where x belongs to low-dimensional (n) feature space, $\phi(x)$ belongs to high-dimensional (m) feature space. Then we can calculate the inner product matrix instead of computing all the function mappings of the original data. We don't need to know the form and parameters of $\phi(x)$.

Here we can use one commonly used kernel function which called Gaussian kernel function to do the transformation. The function can be described as

$$K_{ij} = K(x_i, x_j) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\lambda^2}\right) \tag{14}$$

where λ is used to adjust the width of function.

With the kernel function, the Gaussian density function for the nonlinearly mapped data [25] can be esti-

mated as

$$\Pr(\phi(x)) = \sum_{k=1}^{K} \tilde{\pi}_k P(\phi(x); \tilde{\mu}_k, \tilde{\Sigma}_k) = \frac{1}{(2\pi)^{\tilde{D}/2} \prod_{s=1}^{\tilde{D}} (\alpha_{ks})^{1/2}} \exp\left[-\frac{1}{2} \sum_{s=1}^{\tilde{D}} \frac{x_s^2}{\alpha_{ks}} \right] \quad (15)$$

where α_{ks} is the sth eigenvalue of kernel matrix of the kth Gaussian component and X_s is computed as

$$x_s = \xi_{ks} T \eta_i \quad (16)$$

where ξ_{ks} means the eigenvalue of kernel matrix of the kth Gaussian component; η_i denotes the ith column vector of kernel matrix of the kth Gaussian component. Then substituting sample data to Eqs. (8) and (15), the $\tilde{\pi}_k$, $\tilde{\mu}_k$, $\tilde{\Sigma}_k$ will be calculated.

2.3.2 Kernel Gaussian Mixture Clustering Based Anomaly Detection

The flow diagram of the anomaly detection can be viewed in Figure 1.

The steps of the kernel Gaussian anomaly detection can be described as follows.

(1) Collect N district heating data set $\{x_1, x_2, ..., x_N\}$, where one value means a vector contains heat power, instant flow, and differential temperature of the incoming and return water.
(2) Calculate Gaussian kernel function K with Eq. (14).
(3) Initialize iteration number to be 1, and with the result of K-means, set the number of components to be 8.
(4) Calculate the posterior probability $\Pr(\phi(x))$ with Eqs. (15) and (16).
(5) Substitute sample data to Eqs. (8) and (15), calculate the parameters as $\tilde{\pi}_k$, $\tilde{\mu}_k$, $\tilde{\Sigma}_k$.
(6) Estimate whether likelihood function is converged, if not, iterate (4) and (5).
(7) Label the data with its component and its statistical parameters according to Eq. (9).
(8) Analysis situation of every component and find the anomalous actions.

3. Experiments and Analysis

In order to find out some kinds of anomalous instruments and actions when using district heating to help the heating company with heat-supply regulating, this evaluation is performed. In the meantime, with the nonlinearity of district heating data, the anomaly detection's performance of the proposed KGMC will be compared with conventional GMM and K-Means methods.

3.1 Measured Data Sets and Software Tools

In this paper, the data are extracted from meter readings of real apartments which belong to 18 zones and 6 substations in Jinan, 60 degrees north latitude and 117 degrees east longitude in China. That covers an area of 1.5 million square meters and consists of about 17,000 apartments. Heating seasons begins at November 15th, ends at March 15th, therefore, 90 million vectors recorded at one-hour intervals in three heating seasons from 2013 to 2016 can be available. The dataset contain heat power, instant flow, and differential temperature of the incoming and return water and so on. In China, a substation supply hot water to one or several zones and zones belong to the same substation have similar features through the data analysis. Subsequently, the clusters and the anomalous actions within the same substation have been discussed below.

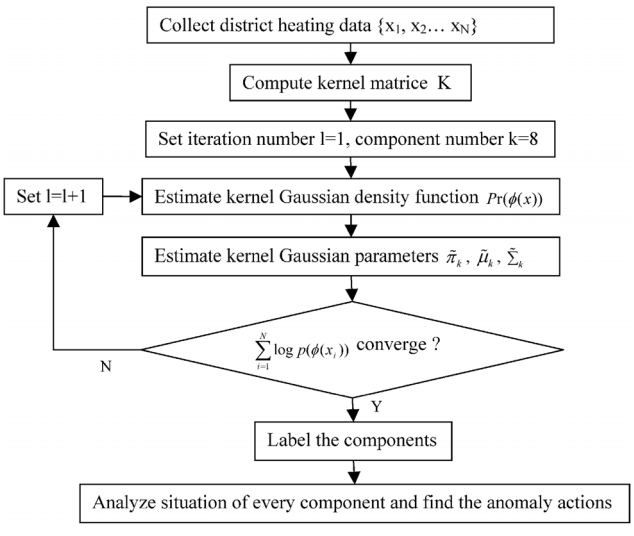


Figure 1. The flow diagram of the KGMC anomaly detection algorithm.

According to Figure 2, the big dataset of district heating data is analyzed on the platform which is called industrial big data ingestion and analysis platform (IBDP) [26]. Since the dataset in this article belong to relational data, only related sections of IBDP are interpreted and the architecture and procedure are presented below. Data source: some of the data originate from the database, and some from the files. Data ingestion layer: Sqoop is a tool that transfers dataset between Hadoop systems and enterprise data storages. Data analytics layer: as the core of the system, big datasets are processed with clustering algorithm by Spark module which is a general and fast computing engine designed for large-scale data processing. Data storage layer: high-throughput access to application data and analysis results are supplied by a distributed file system HDFS. Process flow: the district heating datasets are imported into HDFS directly or via Sqoop and then analyzed by Spark, finally, the clustering results stored in HDFS can be analyzed for specific application.

3.2 Number of Clusters

K-means clustering is the fast and sample method to decide the number of clusters. So, this method which has a good performance is used to calculate the number, and the basic parameters of that are set below. The number of iterations is set to be 10, the number of application running is set to be 5.

To evaluate the performance of the K-means model for heat load clustering and select the most suitable K (the number of clusters), the way is to find the minimum SSE (Sum of Squared Error)

$$\text{Argcost}_{\min} = \arg_s \min \sum_{i=1}^{K} \sum_{x \in S_i} \left\| x - \mu_i \right\| \tag{17}$$

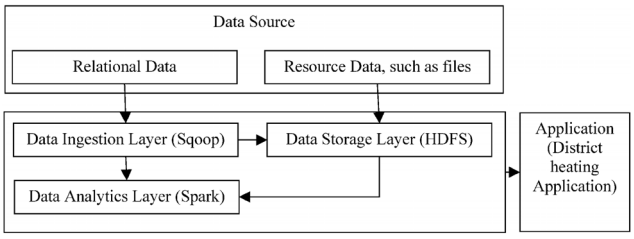


Figure 2. The flow diagram of the software tools.

where $S = \{S_1, S_2 \ldots S_k\}$, S_i means the set of the ith cluster. μ_i is the mean of points in S_i. As shown in Figure 3, the argcost for 3 zones within different substation changes along with the number of clusters. It easily can be seen that the *argcost* will decrease as k increases, but it is hard to explain the phenomenon with too many clusters. If the k is set 8, the *argcost* declines the fastest. So with 8 clusters of the dataset, some results can be analyzed below.

3.3 Anomalies of Actions and Instruments

The district heating data of one zone have been separated into several certain clusters and each cluster may have some significant features (Table 1). The indicators involved in every vector are heat power, instant flow, and differential temperature of the incoming and return water. In Table 1, room_id represents the id of the unique apartment in the database, cluster means the cluster id of all the clusters and dist means the distance between the vector and the centroid. Every vector in the table is the one which owns the smallest distance from the centroid of its own cluster. So it can represent its cluster best.

From the results in Table 1, we can see the results are grouped into eight general clusters. With the small amount of vectors in category 0, 1, 2 and 3 which account for about 3.98%, 3.9%, 2.5% and 1.9% respectively, it can be concluded that the vectors may represent the anomalous actions. With the experience and the examination,

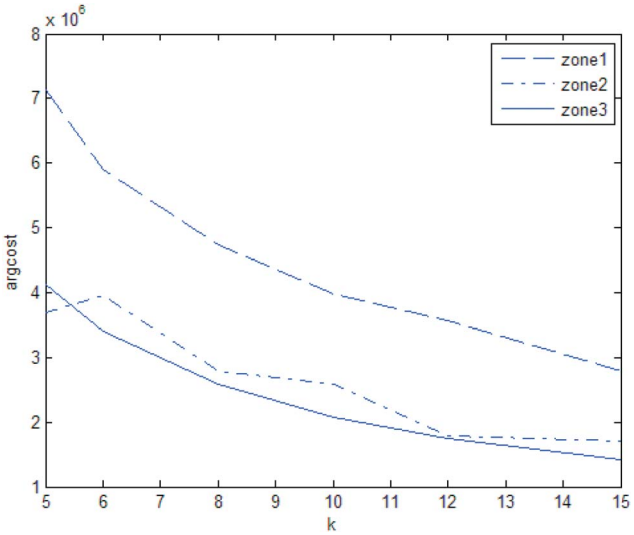


Figure 3. The argcost for different num of clusters.

Table 1. K-means clustering results of one zone

Room_id	Heat_power (kW)	Instant_flow (m^3/h)	Diff_temp (°C)	Cluster	Dist	Num
3060	6.552042	0.810641	17.01	0	4.59E-06	32308
3076	0	0	-3.76	1	0.002932	31632
12413	5.020213	0.58084	-9.86	2	0.045354	20261
3470	2.553635	0.079954	26.68	3	5.59E-04	15477
3115	4.595001	0.568395	7.15	4	3.61E-05	236141
3379	2.966605	0.484653	5.52	5	3.92E-05	139781
3073	3.808116	0.25913	12.21	6	2.19E-04	90068
3707	4.891801	0.458012	9.36	7	4.80E-05	245188

we can further recognize and explain the anomalies in the district heating system within Figures 4–7. To specify the characteristics of abnormal situations, the x axis represents fitting points taken at different times.

From Figure 4, we can observe that when x equals to 9, the point represents the feature of category 0. It in-

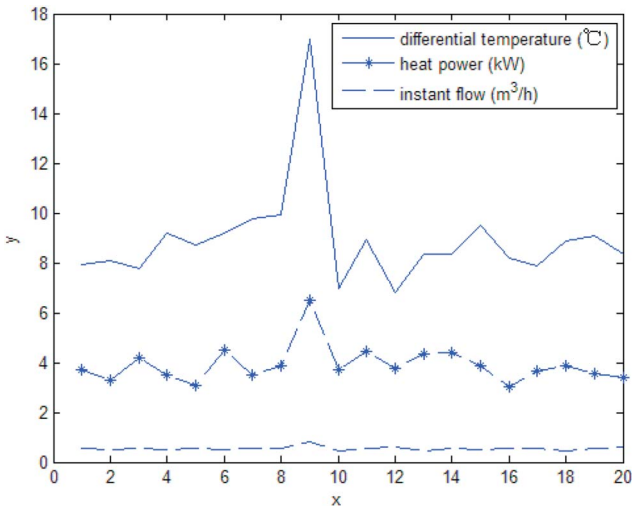


Figure 4. Feature of cluster 0.

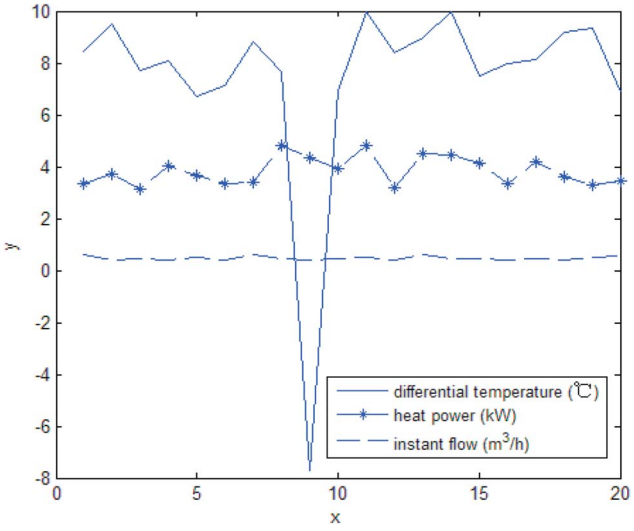


Figure 6. Feature of cluster 2.

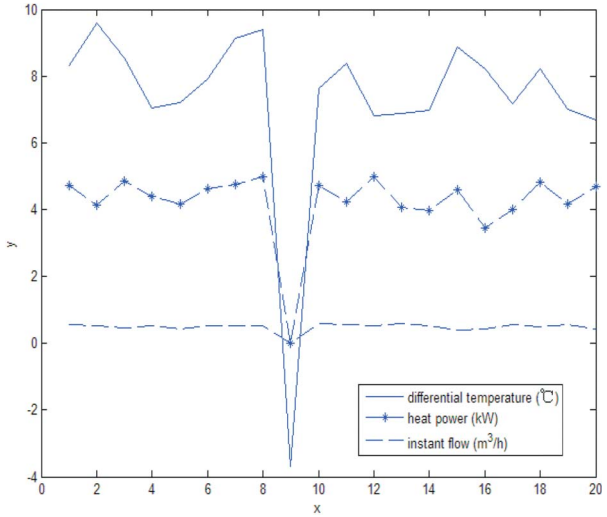


Figure 5. Feature of cluster 1.

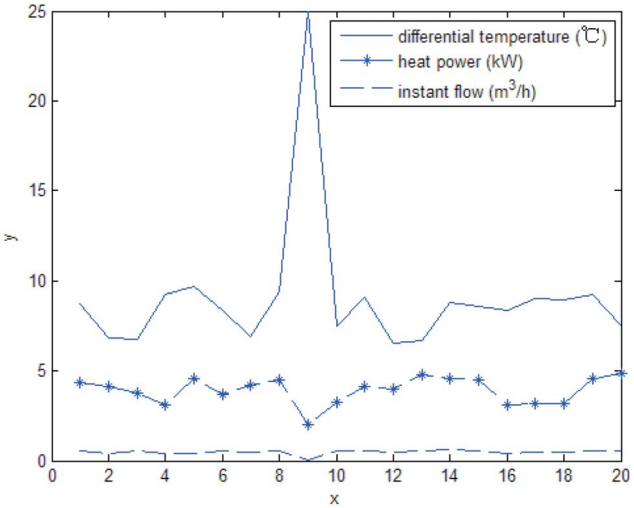


Figure 7. Feature of cluster 3.

cludes the anomalous situation that differential temperature is larger than 17 °C, and instant flow is larger than 0.8 m^3/h. Because of its higher differential temperature and lower temperature of return water, it can be inferred that someone have taken the hot water from the pipeline unauthorized. That behavior will lead to still water in the cooling water pipeline which will decrease the temperature of return water. If the water taken from pipes is up to a certain amount, the flow of hot water in all the apartments under the same exchanger station will be reduced. For the efficient utilization of heat energy and safety of pipes, this behavior is forbidden in China. There's another situation in this category that differential temperature is larger than 17 °C, and instant flow is in the normal range. The window maybe kept opened or the heat exchanger is installed without permission. For avoiding the influence, the heat company has to heat much more water and lead to more energy consumption.

From Figure 5, we can observe that when x equals to 9, the point represents the feature of category 1. It includes the anomalous situation that differential temperature is lower than 0 °C, instant flow is between 0 m^3/h and 1 m^3/h, and heat power is equal to 0 kW. In the meantime, it includes the dataset that differential temperature is greater than 5 °C and heat power is lower than 0 kW. Through the comparison with the dataset from heat meters in Turkey, this cluster is raised by the inaccurate heat meters in the zone.

From Figure 6, we can observe that when x equals to 9, the point represents the feature of category 2. It includes the anomalous situation that differential temperature is lower than 0 °C, but instant flow and heat power both keep in the normal range. This means the temperature measurement probe was installed incorrectly as temperature probe of intake pipe is installed on the return pipe and that of return pipe is opposite. For this category, the absolute value of the differential temperature is applied to the calculation of heat power that lead to the plus power in the heat meter.

From Figure 7, we can observe that when x equals to 9, the point represents the feature of category 3. It includes the anomalous situation that differential temperature is too high to be more than 20 °C, instant flow contains some that lower than 0.1 m^3/h, and some in the normal range. Different to category 0, this high differential temperature is due to the high temperature of incoming water and regular temperature of return water. This means something wrong occurred to the temperature probe of the intake pipe.

From Figure 8, we can observe that within category 4 and category 7, the dataset mean the normal circumstances of the customer secondary heat supply systems. The heat power is almost 5 kW, instant flow is between 0.4 m^3/h and 0.6 m^3/h, and differential temperature is between 6.5 °C and 10 °C. As the temperature increases, the substations will adjust instant flow and temperature of incoming water to maintain the stability of heat power and temperature in buildings. Under this circumstances, the temperature in the rooms can be up to 23 °C and meet the demand of normal lives.

Within category 5 and category 6, the dataset mean the low quantity and smaller changes which occur in the normal range.

To sum up, features of eight clusters have been listed in Table 2 and among that four anomalous actions have been presented in category 0~3. By correcting the abnormal behaviors, such as opening windows and increased convections in category 0~3 within 18 zones, all the values of heat power will become normal. Moreover, heat consumption will be reduced by 5.4% derived from the change of related heat power to produce the same ther-

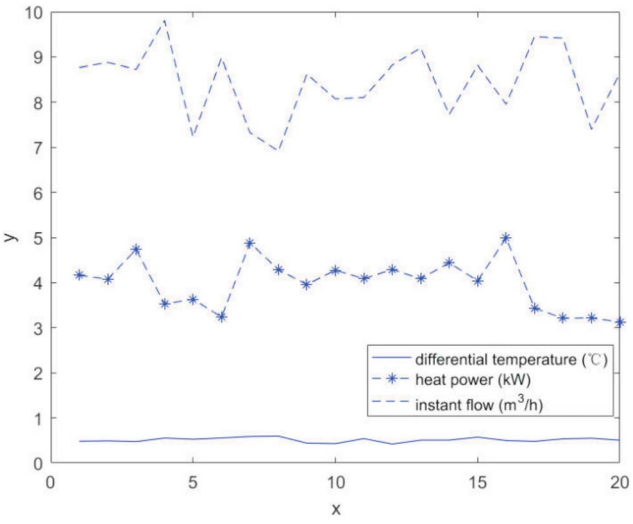


Figure 8. Feature of cluster 4 and cluster 7.

Table 2. Features of the clusters

Cluster	Anomaly interpretation	Instant flow (m^3/h)	Differential temperature (°C)	Heat power (kW)
0	Leakage, opening windows, extra heat exchanger	> 0.8 or nomal	> 17	> 6.2
1	Inaccurate heat meters	0~1	< 0	0
		nomal	> 5	< 0
2	Temperature probe installation error	nomal	< 0	nomal
3	Temperature probe damage	< 0.1 or nomal	> 20	< 3.38
4~7	Normal dataset	nomal	nomal	nomal

mal comfort expressed in Eq. 18. As Ji'nan is a typical winter heating area located in the middle eastern of China, we can assume that the results of this study can be widely applied in heating areas in China. Denote reduced proportion by η, actual heat demand by Q_{real}, heat demand without abnormal actions by Q_{ideal}.

$$\eta = \frac{Q_{real} - Q_{ideal}}{Q_{real}} \times 100\% \tag{18}$$

3.4 Comparison of Clustering Algorithms

According to K-means, K is also set 8 in GMM and KGMC. Before convergence, the maximum change in log-likelihood is considered to set to be 1e-4 and the maximum number of performed iterations is set to be 100. Initial model is an optional starting point from which to start the EM algorithm, and a random starting point is constructed from the data. Table 3 shows the clusters of one zone in Jinan which belong to a different substation from the ones in the previous section. The parameters involved are the same with the data for K-means clustering. The weight in Table 3 stands for the ratio of the number of the vectors in this cluster to the number of all the vectors. Mean vector indicates the average of dataset in this cluster.

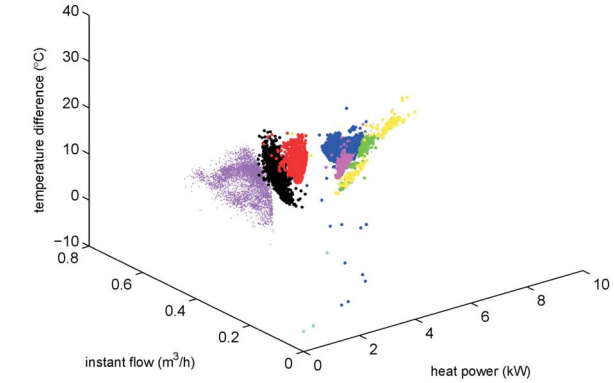
From weight in Table 3, it can be inferred that cluster 0, cluster 1, cluster 2 and cluster 3 may be abnormal according to the extremely small value of the quantity. From the characteristics of the dataset in following columns, the anomalies as the ones mentioned in the previous sector are proved. Category 0 means someone have taken hot water from the pipelines unauthorized or kept windows opened all the time. Perhaps several blockages occurred in the pipelines of apartments in category 1. Category 2 means temperature measurement probes have been installed inversely. In category 3, temperature of incoming water is much higher than that in other categories, hence, maybe the temperature probes of intake pipes have been broken.

The clustering results of GMM and KGMC methods are plotted in Figures 9(a), (b) respectively. The comparison can be recognized much more intuitively. Axis in the three-dimensional coordinates represent three indicators of the vector. Four clusters with small amount of data show the anomalies.

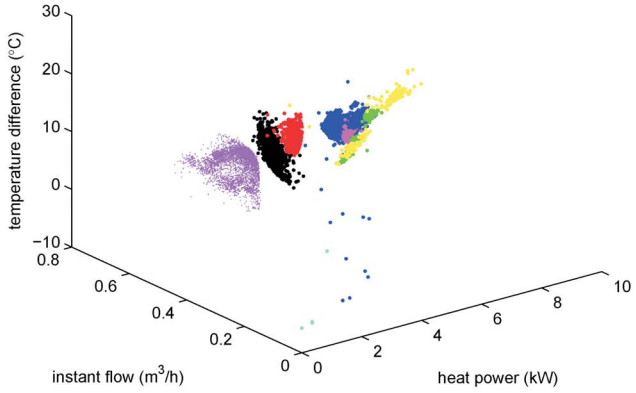
To clearly identify the effectiveness of the proposed

Table 3. GMM and KGMC results of one zone

Cluster	Weight for GMM	Mean vector for GMM: heat_power (kW)/instant_flow(m^3/h)/diff_temp (°C)			Weight for KGMC	Mean vector for KGMC: heat_power (kW)/instant_flow (m^3/h)/diff_temp (°C)		
0	0.0672	4.73	0.39	18.33	0.072	7.5	0.82	18.38
1	0.0304	0.004	-2.77E-5	-7.44	0.0292	0.02	-1.87E-4	-8.04
2	0.0193	4.25	0.38	-9.37	0.0165	4.25	0.36	-9.77
3	0.0072	1.96	0.09	21.74	0.0123	1.57	0.07	20.45
4	0.2326	3.43	0.3	9.79	0.253	3.62	0.36	10.99
5	0.2252	4.63	0.39	11.93	0.206	4.8	0.41	12.11
6	0.1201	3.60	0.46	6.89	0.104	3.52	0.38	6.93
7	0.2976	4.20	0.52	7.076	0.306	4.20	0.52	7.076



(a) GMM cluster result



(b) KGMC cluster result

Figure 9. Cluster result of the two methods.

method KGMC, every record which consists of various features plus one class label that represents normal one or one of the abnormal types have been described in Table 1 and Table 3. Detection rate (DR) and false positive rate (FPR) of three methods: K-means, GMM and KGMC have been analyzed. By the comparisons between the examination of residential thermal behaviors and the data shown in Table 1, K-means performs extremely poorly on triggering alarms on anomalous actions and avoiding incorrect alarms on normal events. Significant amount of normal data in cluster 4, 6 are wrongly detected as anomalous ones. Some anomalous data belong to cluster 1 have been detected as cluster 2. Accordingly, K-means is insensitive to the value of instant flow. For GMM method, the boundary of every partition cannot be qualified to be an ellipsoid in the original description space, in terms of the nonlinearity of district heating dataset. Hereby, comparing the experience with the data shown in Table 3 and Figure 9, a significant partition of normal points have been misclassified as anomalous ones and some anomalous data haven't been detected, especially some ones belong to cluster 0 haven't been identified. In contrast, KGMC has the best performance with 90.5% anomaly detection rate but only 4.1% false positive rate. Multiple classes have been isolated within the high-dimensional kernel feature space in place of the original measurement space, so that both the nonlinearity of processes and multimodality of operations can be extremely taken into accounted. In summary, detection rate and false positive rate of every method for two zones which belong to different substations have been given in Table 4.

4. Conclusions

Few researchers have applied data mining and other techniques to detect the instrumentation and action problems on the basis of the dataset of district heating apartments. Under this circumstance, KGMC approach is proposed and utilized in this paper. Original data in low-dimensional space are projected into high-dimensional space for clustering. As shown in the experimental results, the detection rate and false positive rate have been improved to the different extent compared to K-means and GMM. Four kinds of anomalies have been identified: abnormal heat behaviors, inaccurate heat meters, exceptional temperature probes of intake pipes, and inverse temperature probes of intake and return pipes. As the accuracy, the proposed model KGMC will assist in helping heating company do the heat-supply regulation, proactively detect anomalies, improve energy efficiency and thermal comfort with the decrease of 5.4% heat demand in heating areas in China.

Table 4. Comparison of anomaly detection results among K-means, GMM and KGMC

	K-means		GMM		KGMC	
	DR	FPR	DR	FPR	DR	FPR
Zone 1	70.6%	9.8%	86.57%	7.44%	90.5%	4.14%
Zone 2	73.2%	8.5%	87.3%	6.93%	91.22%	4.68%

References

[1] http://news.ces.cn/gongre/gongreshichang/2015/10/13/77971_1.shtml.

[2] http://www.china-heating.com/news/2016/30125.html.

[3] Baldvinsson, I. and Nakata, T., "A Feasibility and Performance Assessment of a Low Temperature District Heating System–a North Japanese Case Study," *Energy*, Vol. 95, pp. 155–174 (2016). doi: 10.1016/j.energy.2015.11.057

[4] Wang, P. and Sipilä, K., "Energy-consumption and Economic Analysis of Group and Building Substation Systems—a Case Study of the Reformation of the District Heating System in China," *Renewable Energy*, Vol. 87, pp. 1139–1147 (2016). doi: 10.1016/j.renene.2015.08.070

[5] Zaunbrecher, B. S., Arning, K., Falke, T., et al., "No Pipes in My Backyard: Preferences for Local District Heating Network Design in Germany," *Energy Research & Social Science*, Vol. 14, pp. 90–101 (2016). doi: 10.1016/j.erss.2016.01.008

[6] Shamshirband, S., Petkovic, D., Enayatifar, R., et al., "Heat Load Prediction in District Heating Systems with Adaptive Neuron-fuzzy Method," *Renewable and Sustainable Energy Reviews*, Vol. 48, pp. 760–767 (2015). doi: 10.1016/j.rser.2015.04.020

[7] Yeh, H. M., "Heat Transfer Performance in Double-pass Flat-plate Heat Exchangers with External Recycle," *Journal of Applied Science and Engineering*, Vol. 17, No. 3, pp. 293–304 (2014). doi: 10.6180/jase.2014.17.3.10

[8] Lee, C. Y., Huang, H. H., Lee, S. M., et al., "Numerical Simulation of the Heat Transfer Characteristics of Low-Watt Thermosyphon Influence Factors," *Journal of Applied Science and Engineering*, Vol. 17, No. 4, pp. 423–428 (2014). doi: 10.6180/jase.2014.17.4.09

[9] Haves, P., "Overview of Diagnostic Methods," Proceedings of the Workshop on Diagnostics for Commercial Buildings: From Research to Practice, pp. 16–17 (1999).

[10] Ahonen, M., Hyvärinen, J., Kuismin, J. and Pakanen, J., *Fault Detection Methods for District Heating Substations*, Research notes 1780. Oulu: VTT Building Technology (1996).

[11] Sandin, F., Gustafsson, J. and Delsing, J., *Fault Detection with Hourly District Energy Data*, Stockholm: Swedish District Heating Association (2013).

[12] Kiluk, S., "Algorithmic Acquisition of Diagnostic Patterns in District Heating Billing System," *Appl Energy*, Vol. 91, pp. 146–155 (2012). doi: 10.1016/j.apenergy.2011.09.023

[13] Delsing, J. and Svensson, B., "En ny method för funktionsdiagnos och felsökning av fjärrvärmecentraler utifrån värmemängdsdata (a new method for operation diagnosis and fault detection in substations by using heat meter readings)," Luleå: EISLAB, Luleå University of Technology (2001).

[14] Sandin, F., Gustafsson, J., Eklund, R. and Delsing, J., "Basic Methods for Automated Fault Detection and Energy Data Validation in Existing District Heating Systems," 13th International Symposium on District Heating and Cooling, Copenhagen (2013).

[15] Gadda, H. and Werner, S., "Fault Detection in District Heating Substations," *Applied Energy*, Vol. 157, pp. 51–59 (2015). doi: 10.1016/j.apenergy.2015.07.061

[16] Isermann, R., *Fault-diagnosis Applications*, Springer, Berlin Heidelberg, p. 354 (2011).

[17] Isermann, R., *Fault-Diagnosis Systems–an Introduction from Fault Detection to Fault Tolerance*, Springer, Berlin Heidelberg, p. 475 (2006).

[18] Yamanishi, K. and Takeuchi, J., "Discovering Outlier Filtering Rules from Unlabeled Data: Combining a Supervised Learner with an Unsupervised Learner," Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 389–394 (2001).

[19] Ramaswamy, S., Rastogi, R. and Shim, K., "Efficient Algorithms for Mining Outliers from Large Data Sets," *ACM SIGMOD Record, ACM*, Vol. 29, No. 2, pp. 427–438 (2000).

[20] Arning, A., Agrawal, R. and Raghavan, P., "A Linear Method for Deviation Detection in Large Databases," Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 164–169 (1996).

[21] Ester, M., Kriegel, H. P., Sander, J., et al., "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), Vol. 96, No. 34, pp. 226–231 (1996).

[22] Li, R., Wang, Z., Gu, C., et al., "A Novel Time-of-use Tariff Design Based on Gaussian Mixture Model," *Applied Energy*, Vol. 162, pp. 1530–1536 (2016).

[23] Jeong, Y. S. and Jayaraman, R., "Support Vector-based Algorithms with Weighted Dynamic Time Warping Kernel Function for Time Series Classification," *Knowledge-Based Systems*, Vol. 75, pp. 184–191 (2015).

[24] Zahra, S., Ghazanfar, M. A., Khalid, A., et al., "Novel Centroid Selection Approaches for KMeans-clustering Based Recommender Systems," *Information Sciences*, Vol. 320, pp. 156–189 (2015).

[25] Wang, J., Lee, J. and Zhang, C., "Kernel Trick Embedded Gaussian Mixture Model," *Lecture Notes in Computer Science*, Vol. 2842, pp. 159–174 (2003).

[26] Ji, C., Liu, S., Yang, C., et al., "IBDP: an Industrial Big Data Ingestion and Analysis Platform and Case Studies," 2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI), IEEE, pp. 223–228 (2015).

Manuscript Received: Mar. 2, 2017
Accepted: Jan. 5, 2018