TELKOMNIKA, Vol.14, No.3, September 2016, pp. 1150~1156
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v14i3.3805

A New Semi-supervised Clustering Algorithm Based on Variational Bayesian and Its Application

Shoulin Yin[1], Jie Liu*[2], Lin Teng[3]
Software College, Shenyang Normal University,
No.253, HuangHe Bei Street, HuangGu District, Shenyang, P.C 110034 - China
Corresponding author, e-mail: nan127@sohu.com[1], 352720214@qq.com[2], 1532554069@qq.com[3]

Abstract

Biclustering algorithm is proposed for discovering matrix with biological significance in gene expression data matrix and it is used widely in machine learning which can cluster the row and column of matrix. In order to further improve the performance of biclustering algorithm, this paper proposes a semi-supervised clustering algorithm based on variational Bayesian. Firstly, it introduces supplementary information of row and column for biclustering process and represents corresponding joint distribution probability model. In addition, it estimates the parameter of joint distribution probability model based on variational Bayesian learning method. Finally, it estimates the performance of proposed algorithm through synthesized data and real gene expression data set. Experiments show that normalized mutual information of this paper's new method is better than relevant biclustering algorithms for biclustering analysis.

Keywords: *biclustering algorithm, variational bayesian, joint distribution probability, semi-supervised clustering*

1. Introduction

In the given data matrix, clustering technology [1] divides data into several groups according to the similarity of data. In the division group, the similarity of the each group in data set is very high, at the same time, similarity of the data in different groups is as small as possible. Clustering technique is the most basic research content in data mining field.

There are some classical clustering algorithm including k-means [2, 3], spectral clustering [4] and probability modeling methods [5] based on mixed model. These algorithms that divide data into different groups according to row and column of data matrix are single clustering algorithms. However, when the dimension of the matrix is very high, classification performance of the single clustering algorithm is greatly restricted. When adopting biclustering algorithm to classify data in the matrix, the similarity between row and column in matrix should be taken into consideration at the same time. Therefore, it can improve the performance of clustering algorithm commendably. In recommended systems, each row of the matrix represents a user, and each column represents a commodity. Traditional recommendation methods tend to find similar candidates according to the similarity of users. It, nevertheless, predicts those goods based on similar candidates. With biclustering method, it can finds similar goods and users. In gene expression analysis of bioinformatics, each row in a matrix data represents a gene and each column represents a condition, such as normal state, abnormal state, cancer disease state, etc,. Meanwhile, biclustering algorithm can divide similar gene into different groups on the similar conditions, which can better analyze the patient's pathology.

The main idea of traditional biclustering method is that it clusters the row and column of matrix through traditional clustering respectively based on the single clustering method, and then merges clustering results. Typical clustering method contains coupled two-way clustering [6], fuzzy c-means clustering [7] and Bi-Correlation Clustering algorithm (BCCA) [8] etc. In order to avoid limitations of the traditional clustering and better improve the efficiency of the clustering algorithm, for example, SB Huang [9] proposed a fuzzy co-clustering algorithm which minimized distances between objects and centers of clusters in each feature space. ZF Yang [10] represented fuzzy C-means clustering algorithm based on the improved quantum particle swarm optimization. The local search ability and quantum gates update strategy were improved by making full use of the advantages of fast convergence of quantum particle swarm

Received March 14, 2016; Revised May 31, 2016; Accepted June 15, 2016

optimization (QPSO). And Zhou Y [11] showed a semi-supervised method to improve the clustering results, which adjusted the similarity matrix based on Bayesian information, and fixed the class labels on the reference of the pairwise constraints at last. Although, these methods are proposed, there are still some convergence problems.

So in this paper, it presents a semi-supervised clustering algorithm based on variable decibels Bayesian. This new algorithm is on the basis of the matrix data, auxiliary information are introduced to its rows and columns respectively. The structure of the new scheme is as Figure 1. The input data in the algorithm includes matrix data and auxiliary information of columns and rows. Auxiliary information of columns and rows can be represented by adjacency matrix of network. The construction of this paper is as follows. In section 2, we introduce the method of new scheme, it contains probabilistic model and the process of new semi-supervised clustering algorithm based on variable decibels Bayesian. In section 3, we make experiments to show the efficiency of our new scheme. Section 4 gives a conclusion.

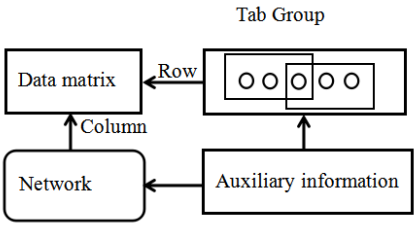

Figure 1. Structure of new scheme

2. The Method of New Semi-supervised Clustering Algorithm

Assuming $X \in R^{M \times N}$ is a *M*×*N*-dimension matrix. Auxiliary information of column and row is W^h and W^c respectively. *K* and *L* are the number of groups in the rows and columns respectively. *K* and *L* are known when making double clustering analysis. In addition, h_m represents the latent variable of *m-th* row. z_n represents the latent variable of *n-th* columns. And $h = (h_1 \ldots h_M)$, $z = (z_1 \ldots z_N)$.

We introduce Gaussian distribution $N(x \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} e^{\frac{(x-\mu)^2}{2\sigma^2}}$, *t* distribution $\Gamma(x, \mu, v) = \frac{\Gamma(v/2+1/2)}{\Gamma(v/2)} (\frac{\lambda}{\pi v})^{\frac{1}{2}} \{1 + \frac{\lambda(x-\mu)^2}{v}\}^{-v/2-1/2}$ and Γ distribution $\Gamma(x, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ to our new algorithm. Where Γ function is $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$ in the *t* distribution.

Let $\varphi(x) = \frac{d \log \Gamma(x)}{dx}$, $\delta(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$. In this part, the reason why we introduce Gaussian distribution and *t* distribution is that they make a contribution to solve the following prior probability, prior distribution and posterior distribution.

2.1. Probabilistic Model Based on Auxiliary Information

Firstly, we establish the new probabilistic model for our scheme. If matrix *X*, lurking variable *h* and *z* are regarded as data information, the joint probability can be expressed by the following formula:

$$p(X, z, h, \theta \mid \theta_0) = p(X \mid \theta, z, h)\, p(z \mid w^h)\, p(\theta \mid \theta_0) \quad (1)$$

Where θ is parameter of the joint distribution, θ_0 is super parameter vector. Under the condition of known θ, *h* and *z*, conditional distribution of *X* can be expressed by Gaussian distribution:

$$p(X \mid \theta, z, h) = \prod_{m=1}^{M} \prod_{n=1}^{N} p(X_{mn} \mid \theta_{h_m z_n}); p(X \mid \theta_{lk}) = N(X \mid \mu_{lk}, (s_{lk})^{-1}) \quad (2)$$

Where $\theta_{lk} = (\mu_{lk}, s_{lk})$. Formula(2) aims to ensure that double cluster elements contain the same value in each group. Under the condition of known w^z and w^h, w^z and w^h is the super parameters of the prior knowledge. On account of the fact that auxiliary information is graph structure, we can use Markov random field to calculate the probability of $p(h \mid w^h)$ and $p(z \mid w^z)$ respectively. Therefore, prior probability of lurking variable *h* and *z* can be represented by:

$$p(h \mid w^h) = 1/C_h \cdot e^{\sum_{i=1}^{M} \sum_{j=1}^{M} w_A^h \cdot W_{ji}^h \delta(h_i, h_j) + \sum_{l=1}^{L} w_l^h \sum_{m=1}^{M} \delta(h_m, l)} \quad (3)$$

$$p(z \mid w^z) = 1/C_z \cdot e^{\sum_{i=1}^{N} \sum_{j=1}^{N} w_A^z \cdot W_{ji}^z \delta(z_i, z_j) + \sum_{k=1}^{K} w_l^z \sum_{n=1}^{N} \delta(z_n, k)} \quad (4)$$

Where C_h and C_z are standardized coefficient, and they can normalize $p(h \mid w^h)$ and $p(z \mid w^z)$ as decimals between 0 and 1. If lurking variable *h* and *z* can make more connection sharing through auxiliary graph information w^z and w^h, then it can obtain larger conditional probability, which shows that it can introduce auxiliary information into biclustering algorithm by probability distribution.

This paper assumes that prior distribution of parameter θ is conjugate prior distribution, namely:

$$p(\theta \mid \theta_0) = \prod_{l=1}^{L} \prod_{k=1}^{K} p(\theta_{lk} \mid \theta_0); p(\theta_{lk} \mid \theta_0) = N(\mu_{lk} \mu_0, (\xi_0 s_{lk})^{-1}) G(s_{lk} \mid \alpha_0, \beta_0) \quad (5)$$

2.2. The New Semi-supervised Clustering Algorithm Based on Variable Decibels Bayesian

In this subsection, we mainly illustrate the new method through the improved formulas. This paper adopts matrix *X* and θ_0 to calculate *z*, *h* and θ in formula (1), then it uses joint distribution to conduct double clustering analysis. However, when using Bayesian method to study the above parameters, the computational complexity is very high. So it needs a kind of approximate Bayesian learning method. This paper adopts Variational Bayes method and assumes that posterior distribution of parameters are independent of each other. So we can get the follows:

$$q(\theta, z, h) = (\prod_{l=1}^{L} \prod_{k=1}^{K} q(\theta_{lk})) \cdot (\prod_{n=1}^{N} q(z_n)) \cdot (\prod_{m=1}^{M} q(h_m)) \quad (6)$$

Where $q(\cdot)$ is posterior distribution of Variational Bayesian. Then, we optimize Variational Bayesian distribution through minimum true posterior distribution and Kullback-Leibler difference of Variational Bayesian distribution according to the known information. The minimization method is equivalent to the previous of minimization Bayesian free energy [12] (F[q]), namely:

$$\begin{aligned} F_t(X \mid \theta_0) &= -\log \sum_z \sum_h \int_\theta p(X, z, h, \theta \mid \theta_0) d\theta \\ &\le -\sum_z \sum_h \int_\theta q(z, h, \theta \mid \theta_0) \log \frac{p(X, z, h, \theta \mid \theta_0)}{q(z, h, \theta)} d\theta = F[q] \end{aligned} \quad (7)$$

The Bayesian free energy, which is also called the variational stochastic complexity and corresponds to a lower bound for the Bayesian evidence, is a key quantity for model selection.

We assume that $\overline{z}_n^k = q(z_n = k)$, $\overline{Z}_k = \sum_n \overline{z}_n^k$, $\overline{h}_m^l = q(h_m = l)$, $\overline{H}_l = \sum_m \overline{h}_m^l$. $\overline{X}_{lk} = \sum_m \sum_n \overline{h}_m^l \overline{z}_n^k X_{mn}$, $\overline{X}_{lk}^2 = \sum_m \sum_n \overline{h}_m^l \overline{z}_n^k X_{lk}^2$. So we could get the posterior distribution of Variational Bayesian of μ_{lk} and s_{lk}.

$$q(\mu_{lk}) = T(\mu_{lk} \mid \overline{\mu}_{lk}, \frac{\alpha_{lk}}{\beta_{lk}} \xi_{lk}, \alpha_{lk}); q(s_{lk}) = \Gamma(s_{lk} \mid \alpha_{lk}, \beta_{lk}) \tag{8}$$

And α_{lk}, β_{lk}, $\overline{\mu}_{lk}$ and ξ_{lk} can be defined as follows:

$$\alpha_{lk} = \alpha_0 + \frac{1}{2} \overline{H}_l \overline{Z}_k \tag{9}$$

$$\beta_{lk} = \beta_0 + \frac{1}{2} \{\xi_0 \mu_0^2 + \overline{X}_{lk}^2 - \xi_{lk} \overline{\mu}_{lk}^2\} \tag{10}$$

$$\overline{\mu}_{lk} = \frac{\xi_0 \mu_0 + \overline{X}_{lk}}{\xi_{lk}} \tag{11}$$

$$\xi_{lk} = \xi_0 + \overline{H}_l \overline{Z}_k \tag{12}$$

In addition, the posterior distribution of Variational Bayesian $q(z_n)$ of z_n is as:

$$q(z_n = k) = e^{\gamma_{nk}} / \sum_{k=1}^{K} e^{\gamma_{nk}} \tag{13}$$

γ_{nk} can be calculated by the follow formula:

$$\begin{aligned} \gamma_{nk} = & \sum_{j=1}^{N} w_Z^z W_{nj}^z \overline{z}_j^k + w_k^z - \frac{1}{2} \sum_{l=1}^{L} \overline{H}_l \frac{\alpha_{lk}}{\xi_{lk}(\alpha_{lk} - 1)} - \frac{1}{2} \sum_{l=1}^{L} \frac{\alpha_{lk}}{\beta_{lk}} \sum_{m=1}^{M} \overline{h}_m^l (\overline{\mu}_{lk} - X_{mn})^2 \\ & + \frac{1}{2} \sum_{l=1}^{L} \overline{H}_l \{\varphi(\alpha_{lk}) - \log \beta_{lk}\} \end{aligned} \tag{14}$$

Just as the above, posterior distribution of h_m is:

$$q(h_m = l) = e^{\eta_{ml}} / \sum_{l=1}^{L} e^{\eta_{ml}} \tag{15}$$

η_{ml} is as follows:

$$\begin{aligned} \eta_{ml} = & \sum_{j=1}^{M} w_A^h W_{mj}^h \overline{h}_j^l + w_l^h - \frac{1}{2} \sum_{k=1}^{K} \overline{Z}_k \frac{\alpha_{lk}}{\xi_{lk}(\alpha_{lk} - 1)} - \frac{1}{2} \sum_{k=1}^{K} \frac{\alpha_{lk}}{\beta_{lk}} \sum_{n=1}^{N} \overline{z}_n^k (\overline{\mu}_{lk} - X_{mn})^2 \\ & + \frac{1}{2} \sum_{k=1}^{K} \overline{Z}_k \{\varphi(\alpha_{lk}) - \log \beta_{lk}\} \end{aligned} \tag{16}$$

Finally, we use Variational Bayesian method to get $q(\theta)$, $q(z)$ and $q(h)$, and put *z*, *h* and θ into formula (1) to analyze Biclustering.

3. **Experimental results and analysis**.

Experiment 1.

In this paper, Variational Bayesian semi-supervised Biclustering method is abbreviated as VBSB. We make comparison to Bayesian method (BM) and k-means method. Performance evaluation criterion uses normalized mutual information(NMI), also we use several experiments to illustrate the new algorithm.

Parameter setting of experiments is $w_k^z = w_l^h = 1$, $\alpha_0 = 2$, $\beta_0 = 1$, $\mu_0 = 1$, $w_A^z = w_A^h$. They are closely to the true values ranging from an acceptable value. We compare the three algorithm's performance of synthetic data sets.

Let *K=3, L=2*, each group of biclustering contains 20 rows and 50 columns. For column *k*, $z_n^* = k(50(k-1)+1 \le n \le 50k)$. For row *l*, $h_m^* = k(20(l-1)+1 \le m \le 20k)$. After generating the groups, it can produce *m*×*n* matrix X_{mn} by Gaussian distribution $N(\mu_{lk}, \sigma^2)$. Where $l = h_m^*$, $k = z_n^*$, $\mu_{11} = 1$, $\mu_{12} = 0$, $\mu_{13} = 1$, $\mu_{21} = 0$, $\mu_{22} = 0$, $\mu_{23} = 0.5$. We use the three parameters R_{in}, R_{out} and R_s to generate the corresponding auxiliary information network. Where R_{in} represents the proportion of edge in one internal group. R_{out} denotes the proportion of edge within mixed groups. R_s is the proportion of nodes with labels. For column auxiliary information W^z, it produces *3×50×49×* R_{in} internal sides and *2×50×50×* R_{out} mixed groups sides. For row auxiliary information W^h, it produces *2×20×19×* R_{in} internal sides and *20×20×* R_{out} mixed groups sides. To realize a semi-supervised learning algorithm, it randomly removes 50 and 20 sides for column and row respectively with no labels in node after generating auxiliary network W^z and W^h.

During the experiments, we select (0.1,1,4), (0.15,1,3) and (0.1,0.75,4) as the value of (R_{out}, R_s, σ^2). For each value, it randomly produces 10 data and selects the average value. Figure 2 and Figure 3 is NMI performance comparison of the row and column clustering with the value (0.1,1,4). In addition, with the increasing of auxiliary information weight, NMI of VBSB algorithm increases first and then decreases.

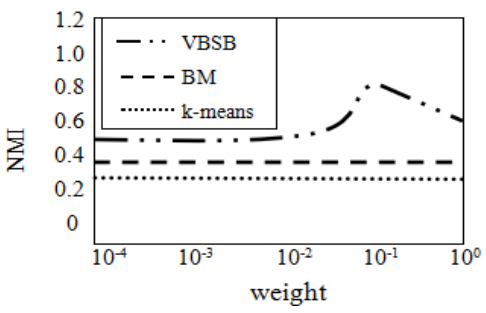

Figure 2. Row clustering comparison with the value (0.1,1,4)

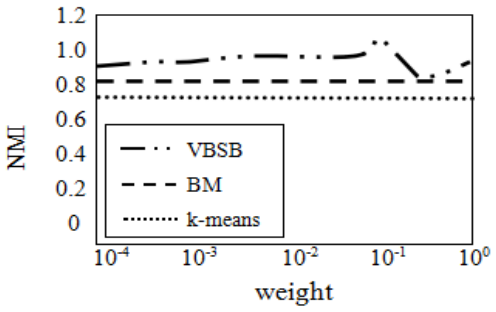

Figure 3. Column clustering comparison with the value (0.1,1,4)

From Figure 2, we can know that BM and k-means method reaches a plateau at 0.38 and 0.3 respectively. When weight is 0.1, the NMI of VBSB is approximately 0.8, which is the highest. Similarly, the NMI of VBSB is superior to BM and k-means in Figure 3.

Then we select (0.15,1,3) and (0.1,0.75,4) as the value of (R_{out} , R_s , σ^2). Repeat the above experiments. And we get the Figure 4, 5, 6, 7.

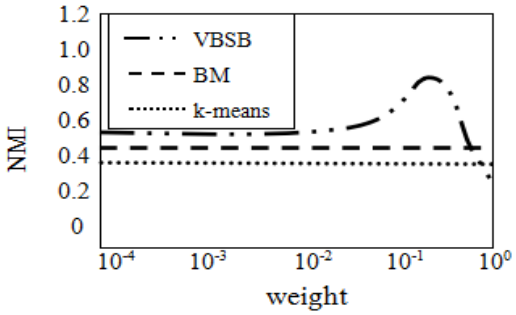

Figure 4. Row clustering comparison with the value (0.15,1,3)

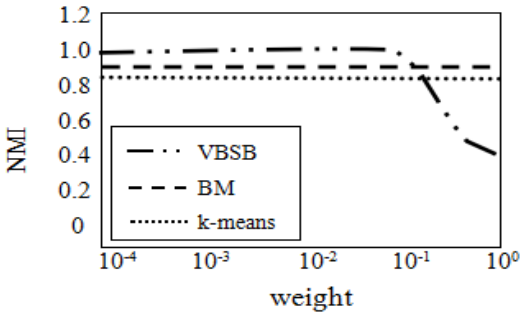

Figure 5. Column clustering comparison with the value (0.15,1,3)

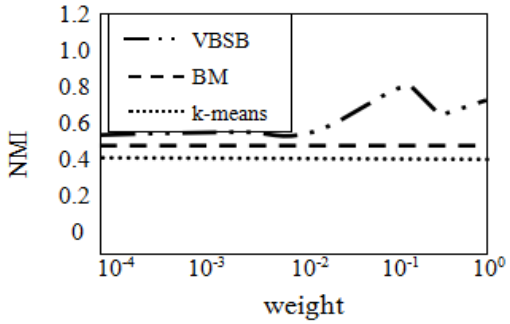

Figure 6. Row clustering comparison with the value (0.1,0.75,4)

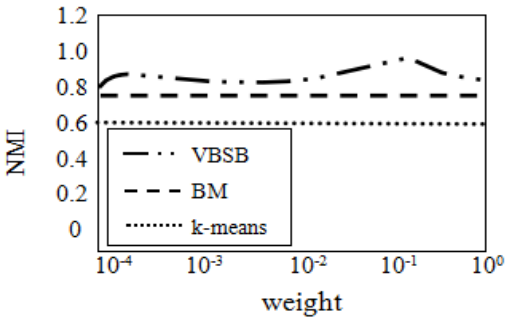

Figure 7. Column clustering comparison with the value (0.1,0.75,4)

From Figure 4-7 we can know that they have the similar results. The NMI performance of VBSB algorithm obviously exceeds that in other two algorithms with reaching plateau at constant level. In Figure 4, 5, if the weight of auxiliary information is higher, so the NMI performance of VBSB algorithm is predicted to experience a decreasing trend. Therefore, the value of σ^2 plays a significant influence on the algorithm. Figure 6 also shows that NMI performance comparison of the row clustering with the value (0.15,1,3) in VBSB is superior to BM which only accounts for approximately 0.5 and k-means with nearly 0.4. The analysis is similar to Figure 7. In addition, VBSB algorithm has a fast convergence rate. So the new algorithm has been proved.

Experiment 2.

In order to verify the efficiency of VBSB algorithm, we make an intrusion detection experiment with our new method and make a comparison with reference [13]. The method in [13] is that it uses part of marked data from the sample data set and generates the Seed set for initializing the cluster. By calculating the Euclidean distance between marked point in sample data set and the average value of labeled data in each cluster and getting the initial center point, it effectively avoids the Blindness and randomness when choosing initial clustering center by traditional clustering algorithm. We introduce performance indicators to compare the performance of algorithms including detection rate (DR) and false positive rate (FPR). In this experiment, we select representative 5000 data. And other data are selected as in [13]. Through testing the DR and FPR of 5000 aggressive data with different algorithms, we can measure the detection effect of each algorithm. We adopt VBSB and the method in [13] and get the detection results as Figure 8.

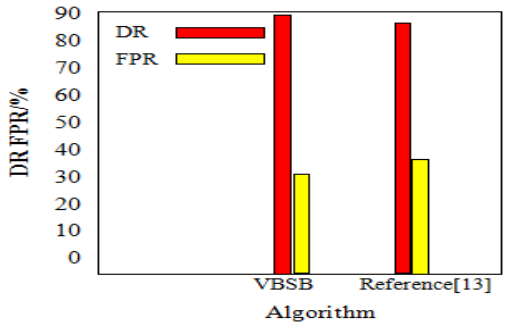

Figure 8. Comparison with VBSB and reference [13]

It is clearly from Figure 8 that DR of VBSB nearly is 89% over that of reference [13] about 87%. Nevertheless, for FPR, VBSB only accounts for roughly 30% and reference [13] with 35%. Therefore, the results show that the new semi-supervised clustering algorithm based on Variational Bayesian is very effective for the detection with a lower false positive rate.

4. Conclusion

This paper puts forward a semi-supervised clustering algorithm based on Variational Bayesian. The algorithm not only contains the target matrix, but also introduces the auxiliary information of row and column into its process. The proposed algorithm combines target matrix with the auxiliary information as a joint distribution probability model, then it adopts Variational Bayesian learning method to estimate parameters in this model. At the end of this paper, we make experiments through synthetic data sets and compare the VBSB algorithm to Bayesian method and k-means method to verify the good performance of proposed algorithm. In the future, we are expected to study more advanced semi-supervised clustering algorithms to improved our method and apply them into practical engineering applications.

References

[1] Petrini F, Feng WC, Hoisie A, et al. *The Quadrics network (QsNet): high-performance clustering technology*. Hot Interconnects 9, 2001, IEEE. 2001: 0125.
[2] Jia RY, Guan YY, Ya-Long LI. Parallel k-means clustering algorithm based on MapReduce model. *Computer Engineering & Design.* 2014.
[3] Tianhua Liu, Shoulin Yin. An Improved K-Means Clustering Algorithm for Kalman Filter. *ICIC Express Letters, Part B: Applications.* 2015; 6(10).
[4] Long B, Zhang Z. *Spectral clustering for multi-type relational data: US8185481 B2*. 2014.
[5] Lucchini T, D'Errico G, Contino F, et al. Towards the Use of Eulerian Field PDF Methods for Combustion Modeling in IC Engines. *Computer Simulation.* 2014; 7(1).
[6] Getz G, Levine E, Domany E. *Coupled two-way clustering of gene microarray data.* Proceedings of the National Academy of Sciences. 2000; 97.
[7] Lu C, Xiao S, Gu X. Improving fuzzy C-means clustering algorithm based on a density-induced distance measure. *Journal of Engineering.* 2014; 1.
[8] Bhattacharya A, De RK. Bi-Correlation Clustering Algorithm (BCCA) for determining a set of co-regulated genes. *Bioinformatics.* 2009; 25.
[9] Huang SB, Yang XX, Shen LS, et al. Fuzzy co-clustering algorithm for high-order heterogeneous data. *Journal on Communications*. 2014.
[10] Yang ZF, Shi HS, School SE, et al. Fuzzy C-means clustering algorithm based on improved QPSO. *Modern Electronics Technique*. 2014.
[11] Zhou Y, Wang Y, Chen D, et al. Semi-supervised spectral clustering algorithm based on bayesian decision. *Journal of Computational Information Systems.* 2015; 11(4): 1333-1342.
[12] Watanabe K, Shiga M, Watanabe S. Upper bound for variational free energy of Bayesian networks. *Machine Learning.* 2009; 75(2): 199-215.
[13] Xia ZG, Wan L, Cai SY, et al. A semi-supervised clustering algorithm oriented to intrusion detection. *Journal of Shandong University.* 2012.