Information Sciences 533 (2020) 138–149

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Group anomaly detection based on Bayesian framework with genetic algorithm☆

Wanjuan Song[a,b,c], Wenyong Dong[a,d,*], Lanlan Kang[e]

[a] School of Computer Science, Wuhan University, Wuhan 430070, China
[b] College of Computer, Hubei University of Education, Wuhan 430205, China
[c] Hubei Education Cloud Service Engineering Technology Research Center, Wuhan 430205, China
[d] School of Software, Nanyang Institute of Technology, Nanyang 473004, China
[e] College of applied science, Jiangxi University of Science and Technology, Ganzhou, 341000, China

ARTICLE INFO

Article history:
Received 17 December 2019
Revised 28 February 2020
Accepted 30 March 2020
Available online 13 May 2020

Keywords:
Group correlation
Genetic algorithm
Anomaly group detection
Logistic normal distribution
Variational inference

ABSTRACT

Anomaly detection is an important application field of evolutionary algorithm. Unlike traditionly anomaly detection, group anomaly detection aims to discover the anomalous aggregate behaviors in data points. Over past decades, a large number of promising methods have been successfully applied for group anomaly detection. However, they inherently neglect the correlations among groups in data points, limiting their abilities. This paper presents a correlated hierarchical generative model, which can model the intricate correlations hidden in groups by introducing a logistic normal distribution to capture the correlations among groups. With the proposed model, we construct a full variational Bayesian framework, which can data-adaptively optimize the model parameters of the proposed model. The model is designed and trained using Genetic Algorithm (GA), which helps automating the use of generative model. Further, a new score function is proposed as an anomaly criterion to estimate final anomaly groups in data points. Several experiments on synthetic data and real astronomical star data from Sloan Digital Sky Survey demonstrate the effectiveness of proposed method compared with the-state-of-art methods, in terms of average accurac (AP) and area under the Receiver Operating Characteristic(ROC) curve(AUC).

1. Introduction

Evolutionary algorithm is used to solve many optimization problems, and anomaly detection is an important application field. Many evolutionary algorithms, such as Partical Swarm Optimization(PSO) and GA, can help us optimize parameters of the models in many applications [6,7,11,20,25]. In the process of anomaly detection, there are many optimization problems, such as the establishment of optimal generation model and the optimization of super parameters in the model. Traditional anomaly detection attempts to find small part data in all given data, which is significantly different from most data. These part of data (also called point anomalies) does not conform to the general model of data set, but often contains some impor-

☆ This research work is supported by the National Natural Science Foundation of China under grant nos. 61672024, 61170305 and 60873114, and National Key R&D Program of China (Nos. 2018YFB0904200 and 2018YFB2100500).

* Corresponding author at: School of Computer Science, Wuhan University, Wuhan 430072, China.
E-mail addresses: key_swj@whu.edu.cn (W. Song), dwy@whu.edu.cn (W. Dong).

https://doi.org/10.1016/j.ins.2020.03.110
0020-0255/

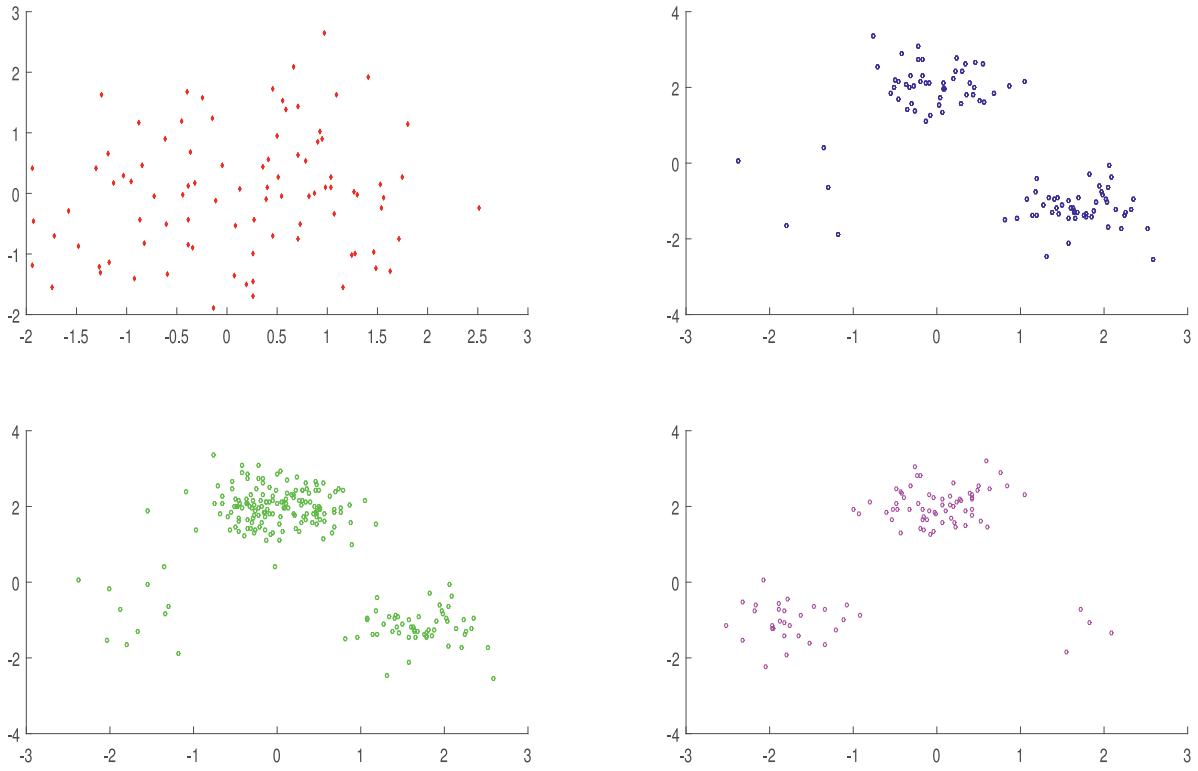

Fig. 1. Illustration for the two forms of group anomaly. The red group (upper left) is a point-based anomaly group, while the magenta group (bottom right) is a distribution-based group. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tant information, e.g., the abnormal information of the body indicating the disease. Traditional anomaly detection has been widely used in various fields [1,6,7,11,25,29,31,32,36,41], i.e., card fraud detection, network intrusion detection, disease diagnosis, fault detection, terrorism prevention. However, traditional anomaly detection can not model the anomalous aggregate behaviors of data points, e.g., abnormal behavior in social media.

Unlike traditional anomaly detection, group anomaly detection is to detect anomalous collections of individual data points. Due to strong practicability, group anomaly detection [5,15] has attracted much attention in various real applications [3–5,13,16,17,19,34,37,38], e.g., anomalous behavior in social media, the special galaxies in astronomical data, the anomalous currents in turbulence data, the abnormal topic in text classification and scene classification. Because Bayesian framework can handle hidden variables better,many researchers use Bayesian model in related area. Group anomaly detection is often divided into two categories according to reference [13] of Guevara et al.: point-based anomalies and distribution-based anomalies. The former is to find the anomalous groups, of which the all data points are anomalous, while the latter searches such anomalous groups that the data points in the groups are seemingly regular but the groups is obviously different from other groups according to the proportion of topic distribution. To vividly depict the two categories, Fig. 1 shows the two forms of anomaly group above mentioned. There are four groups of data points distinguished by different colors. As shown in Fig. 1, the red group(upper left) represents point-based anomaly, because every point in this group is anomaly relative to the other three groups. Further, the group in bottom right corner(color in mageta) is distribution-based anomalies, since its distribution is distinguished from the blue group(upper right) and the green group(bottom left), in spite of its regular data points. Therefore, there are two anomaly groups, e.g., the red group and the magenta group. Regardless of which category of the above two is mentioned. The first key element to discover anomalous groups is understanding what is normality. Similar to most of existing methods, this paper employs the probabilistic simplex describing the normal topic distributions.

This paper focuses on the problem of distribution-based anomaly group detection. Although most of existing methods [4,13,26,37,38,40] have been successfully applied for group anomaly detection. However, they inherently neglect the correlations among groups, e.g., the correlation between two different distributions in top right corner in Fig. 1. Failure to capture these correlation may loss useful information to improve the performance of anomaly group detection. To capture the correlation among groups for anomaly group detection, the contributions of this paper are as follows:

- We presents a correlated hierarchical generative model (CHGM) to model the correlation among groups for anomaly group detection. The proposed model employs a logistic normal distribution to model topic and utilizes a Gaussian distribution to depict data characteristics.
- We construct a full variational Bayesian framework, which can data-adaptively optimize the model parameters of the proposed model with Genetic Algorithm.
- We design a novel scoring strategy to score each group, and then determine anomaly groups from given data set.
- Experiments on synthetic data set and real astronomical galaxy data set from Sloan Digital Sky Survey demonstrate the superiority of the proposed method compared with the-state-of-art methods.

The chapters of this paper are organized as follows. The related work of anomaly group detection is reviewed in Section 2. The related background knowledge is given in Section 3. The proposed model is presented in Section 4. Variational inference for the model parameters of CHGM is given in Section 5. The experimental results are provided in Section 6.

2. Related work

While detecting anomalies,high-dimensionality data causes a lot of trouble. It not only increases computational overhead and memory requirements, but also affects their performance in real applications. Finding the low dimensional space or dimension reduction method for high-dimensional observed data has been a challenging problem in the field of anomaly detection. Many researchers [6–9,12,28] are devoted to this problem. Chen et al. [6] propose a novel image outlier detection method by combining autoencoder with Adaboost (ADAE),which combine Adagrad with Proximal Gradient Descent to optimize learning objective. This method has been successfully applied in the highdimensional image datasets. Chen et al. [7] use the projection pursuit method to find the best projection direction based on the dynamic evolution algorithm, and project the high-dimensional data to the low-dimensional data space. Owing to the inefficiency of the Euclidean distance difference in high-dimensional data, Radovanovi et al. [28] propose a new distance-based method to detect anomalies in high-dimensional data using reverse nearest neighbor number as anomaly indicator. Motivated by the inaccuracy of anomaly detection for usually high dimensionality of the observational space, Wang et al. [9] describe anomaly point in a new framework,which interprets and discovers anomaly with non-negative matrix factorization (NMF) from the perspective of latent subspace. Ju et al. [12] use latent variable as anomaly sign in L1-norm-based probabilistic PCA model to detect anomalies, which is robust to anomalies and also makes good use of spatial information in 2D data. Banerjee et al. [8] redefine anomaly point using sparse coding from a more microscopic perspective. According to the definition, atoms in anomalies should appear in normal points with a lower probability. Using this property as anomaly index, anomalies in high-dimensional data are detected.

The data descriptions of group anomalies and point anomalies are different, so, the detection methods [8,9,12,28] for point anomalies cannot be directly applied to group anomaly detection. Up to now,some solutions are proposed to detect anomaly groups. Xiong et al. improve a Mixed Gauss Mixture Model (MGMM) [38] to detect anomaly groups based the traditional LDA model. MGMM can deal with the condition of multiple non-anomalous topic distributions and real vector-valued observations. Considering the local and global topics, Xiong et al. propose a more flexible generating model [37] for GAD,which can using topic generator to generate topics adaptively. Yu et al. propose a hierarchical Bayesian model [40] for GAD in social media, which can dynamically find groups and detect anomaly groups. This method is different from the two methods [37,38] mentioned above, which can grouping according to certain criteria before detection. Muandet et al. propose a class of support measures (OCSMMs) [26],which extends a class of support vector machines (SVMs) to measure space. Guevara et al. propose three new discriminant and nonparametric DD models (SMDD) to simulate the vectors in the support kernel method [13]. The DD models describe the data set of probability metrics by optimizing the volume set of probability metrics,which can be applied in GAD. Chalapathy et al. use a deep generative model [5] by Adversarial autoencoder (AAE) and variational autoencoder (VAE) to detect abnormal groups. The methods for GAD have been used in the field text classification and scene recognition [6,17,22,27,30,34]. However, the data in actual application tends to contain some relevancies, the above approaches cannot perfectly model the data because they assume that topics are independent with each other.

3. Preliminary

To more easily understand the proposed method, in this section, we introduces related background knowledge, including notation and traditional latent Dirichlet allocation model.

3.1. Notation

Assuming that there are M groups for given data set and $G_1, \ldots, G_M$ means different groups. Each group $G_i (i = 1, \ldots, M)$ consist of N data points, which is denoted by $x_{i,j}$ $(j = 1, \ldots, N, x_{i,j} \in R^f)$, where f is the dimensionality of points features. Each group shares K topics and each data point $x_{i,j}$ belongs one of K topics, which is expressed as $z_{i,j}$, $z_{i,j} \in \{1, \ldots, K\}$. In order to solve muti-mode problem about non-anomalous topic distributions, we introduce K-dimensional probabilistic simplex $\mathbb{S}^K$ to express the set of non-anomalous topic distribution proportion ($\mathbb{S}^K = \{s \in R^K | s_k \geq 0, \sum_{k=1}^{K} s_k = 1\}$). s_k represents the

distributing proportion of topic k and the sum of distributing proportion for K topics is 1. Let $y_t \in \mathbb{S}^k$, for all $t = 1, \ldots, T$, and $y = \{y_1, \ldots, y_T\}$ denotes the set of T possible non-anomalous distribution proportions of the K different topics in the M groups.

3.2. Traditional latent dirichlet allocation model

Latent Dirichlet allocation (LDA) model is a topic model which discovers underlying topics in a collection of documents and infers word probabilities in topics. Originally, LDA are widely used in text classification [3,30,34] and image classification [5,17,35]. The generative process of LDA is described in Algorithm 1. In Algorithm 1, $Dir(\alpha)$ represents the Dirichlet

Algorithm 1 The generative process of LDA.

for $i = 1$ to M **do**
  choose topic distribution θ_i of the ith document G_i.
    $\theta_i \sim Dir(\alpha), \theta_i \in \mathbb{S}^K$
  for $j = 1$ to N **do**
    for the jth word $x_{i,j}$
    Choose a topic $z_{i,j}$ of the jth word in the ith document. $z_{i,j} \sim Mult(\theta)$.
    Choose word distribution $\varphi_k (k \in \{1, 2, \ldots K\}, k = z_{i,j})$ from Dirichlet.
    Choose a word $x_{i,j}$ of the ith document from $Mult(\varphi_k)$.
      $P = Mult(\varphi_k), x_{i,j} \sim P(.|z_{i,j}, \varphi), \varphi = \varphi_1, \ldots, \varphi_K$
  end for
end for

distribution with non-negative hyper-parameters $\alpha (\alpha \in \mathrm{R}_+^K)$. $Mult(\theta)$ denotes the multinomial distribution with parametric θ.

In the past several decades, due to its simplicity in parameter estimation and ability to avoid over-fitting, LDA have been extended for GAD, such as, MGMM [38], FGM (Flexible Genre Model) [37] and DGM (Deep Generative Model) [5]. The proposed method in this paper is also based on LDA framework.

4. The proposed correlated hierarchical generative model

In this section, we give the proposed correlated hierarchical generative model. To do that, we first analysis the necessary of capturing the correlations in real applications, e.g., text and image, and then introduce logistic normal distribution, which is a key element of the proposed model.

4.1. Correlations in real applications

To model real applications, there are intrinsic correlation in data set. For example, in most text corpora, it is natural that the latent topics are highly correlated. We can even judge the type of documents by the degree of correlation among topics. For instance, an article about sports culture may involve basketball and ball-game stars, but not astrophysics. If track and field and star Kobe Bryant appear together, this document is more likely to be a sports news. And if Kobe Bryant appears with entertainment stars in an article, it's more likely to be an entertainment story. Further, in image scene classification, as shown in Fig. 2, it can be easily observed that sky, tree, sea and sandbeach co-exist in the same coast scene. Obviously, a scene exhibits a strong semantic correlation. Therefore, we argue that correlation in real applications is common, and should be effectively modeled in considering generative model.

Although LDA, as a Bag-of-Words (BOW) model, has been widely unitized in various applications, it assumes that all of the words are independent of each other. This is because it imposes a Dirichlet distribution prior on the topic proportion [3], which essentially assumes that topics are mutually independent. The limitation of Dirichlet distribution cause the ineffectiveness of modeling correlation of topics. It should be note that the other similar hierarchical probabilistic generative models (MGMM, FGM) are also based on a Dirichlet distribution, hence hindering their performance.

4.2. Logistic normal distribution

Before giving a correlated hierarchical generative model, we first introduce logistic normal distribution, which is the key point to model correlation of topics.

The definition of logistic normal distribution [2] is similar to that of lognormal distribution. Assuming R^d represents d-dimensional real value space, P^d means positive quadrant of R^d, $\mathbb{S}^d$ represents d-dimensional positive simplex.

Lognormal distribution is defined as follows. Supposed $v \sim \mathcal{N}^d(\mu, \Sigma), v \in \mathrm{R}^d, w = e^v, w \in \mathrm{p}^d$, the inverse process is $v = \log w$, so w obeys lognormal distribution and can be expressed as $w \sim \Lambda_d(\mu, \Sigma)$.

Logistic normal distribution is defined as follows:

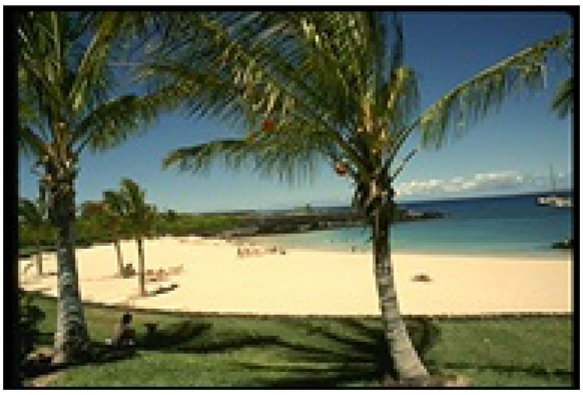

Fig. 2. Coast Scene. This is a cosat scence which includes trees, beaches, oceans and sky.

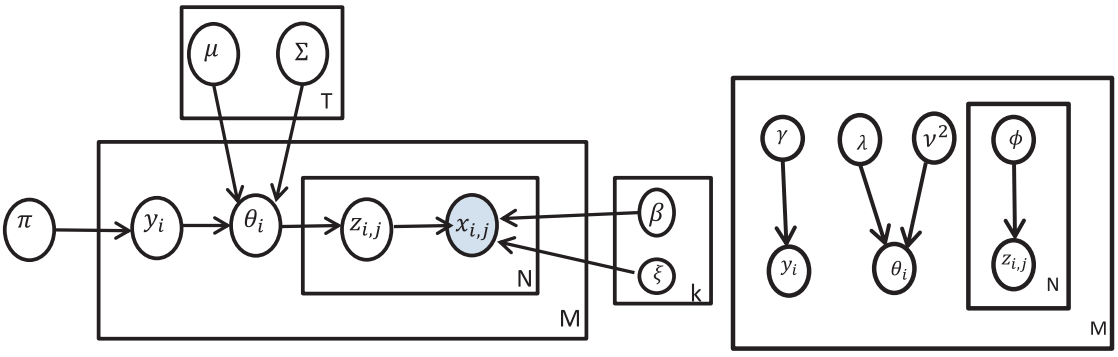

Fig. 3. (Left)Graphical model for CHGM.(Right)Graphical model representation of the variational distribution used to approximate the posterior.

Supposed $v \sim \mathcal{N}^d(\mu, \Sigma), v \in \mathrm{R}^d, v = \log \frac{u}{1-\sum_{j=1}^{d} u_j}$, and then, we obtain $u = \frac{e^v}{1+\sum_{j=1}^{d} e^{v_j}}$. Thus, u obeys the logistic normal distribution ($u \in \mathbb{S}^d$) and can be expressed $u \sim \mathcal{L}_d(\mu, \Sigma)$. Logistic normal distribution is often used to describe the proportion of components, e.g., the geological composition of minerals. In our model, we utilize logistic normal distribution to generate topic distribution, whose covariance reflects the correlations between topics.

4.3. The proposed correlated hierarchical generative model

A graphical representation [21,42] of CHGM is given in Fig. 3. The generative process of the CHGM is stated in Algorithm 2.

Algorithm 2 Generative process of CHGM.

for $i = 1$ to M **do**
  for each group
  draw a genre $y_i \sim Mult(\pi), \pi \in S^T, y_i \in \{1, \ldots, T\}$
  draw a topic distribution θ_i according to the genre y_i:
    $\theta_i \sim \mathcal{L}(\mu_{y_i}, \Sigma_{y_i}), \theta_i \in \mathbb{S}^K$
  for $j = 1$ to N **do**
    for each point
    draw a topic type $z_{i,j} \in Mult(\theta_i), z_{i,j} \in \{1, 2, \ldots, K\}$
    generate a data point feature $x_{i,j}$ according to topic type $z_{i,j}$:
      $x_{i,j} \sim P(x_{i,j}|\beta, \xi, z_{i,j}) = \mathcal{N}(\beta_{z_{i,j}}, \xi_{z_{i,j}}), x_{i,j} \in \mathrm{R}^f$
  end for
end for

The variables in the generative model are interpreted as follows.$Mult(\pi)$ denotes multinomial normal distribution with a T-dimensional vector π, $\mathcal{L}(\mu_{y_i}, \Sigma_{y_i})$ denotes logistic normal distribution with a K-dimensional vector μ and a $K \times K$ covariance matrix Σ. $\mathcal{N}(\beta_{z_{i,j}}, \xi_{z_{i,j}})$ denotes normal distribution with a mean vector β and covariance matrix ξ. y_i denotes

the ith group genre of non-anomalous;θ_i denotes the topic distribution of the ith group;$z_{i,j}$ denotes the topic of the jth data point of the ith group; $x_{i,j}$ is the jth data point of the ith group.

4.4. Group anomaly detection strategy

With the CHGM model, it is natural to define an anomaly scoring function to evaluate each group in order to detect final anomaly groups. Some studies [37,38] usually use marginal probability $p(G_i|\Theta)(\Theta = \{\pi, \mu, \Sigma, \beta, \xi\}, G_i = \{x_{i,j}\})$ as an anomaly indicator for each group. However, existing scoring functions are just appropriate for point anomalies or point-based anomaly groups, not for distribution-based anomaly groups. Further detailed definition of scoring function can be founded in Section 5.2, which can be used as anomaly criterion for GAD.

It is difficult to calculate the marginal probability because the prior distribution and the posterior distribution are not conjugate and interdependent distribution in the CHGM model. Many researchers put forward some methods of parameter solution and optimization [10,18,33,39]. This paper uses variational EM to compute the marginal distribution $p(G_i|\Theta)$.

5. Optimization

Conventional approaches that resort to sum-product belief propagation or sampling algorithms often face with high computational cost. To reduce the computational burden [23,24], we use a variational inference and GA to optimize the paramaters of the proposed model. As for GA, we refer to a large number of references, in which GA has a good effect. In the experiment, we also use Differential Evolution Algorithm and GA to carry out a comparative experiment, so we finally choose GA as our optimization algorithm.

We encounter several computational problems when using this model to detect anomaly groups. As shown in Fig. 3, given the observations and latent variables, we can write the complete likelihood of the ith group as follows:

$$\begin{aligned} p(y_i, \theta_i, z_i, G_i|\pi, \mu, \Sigma, \beta, \xi) &= p(y_i|\pi)p(\theta_i|\mu_{y_i}, \Sigma_{y_i})\prod_{j=1}^{N} p(z_{i,j}|\theta_i)P(x_{i,j}|\beta, \xi, z_{i,j}) \\ &= Mult(y_i|\pi)\mathcal{L}(\mu_{y_i}\prod_{j=1}^{N} MUlt(z_{i,j}|\theta_i)\mathcal{N}(x_{i,j}|\beta, \xi, z_{i,j}) \end{aligned} \tag{1}$$

where, $z_i = \{z_{i,j}\}, G_i = \{x_{i,j}\}(i = 1, \ldots, M, j = 1 \ldots, N)$, the meaning of other symbols in formula (1) are explained in Section 4.3.

By integrating out θ_i, and summing out y_i and z_i, we get the marginal likelihood of the ith group:

$$p(G_i|\Theta) = \sum_{t=1}^{T}\pi_t \int_{\theta_i} \mathcal{L}(\theta_i|\mu_t, \Sigma_t)\prod_{j=1}^{N}\sum_{k=1}^{K}\theta_{ik}\mathcal{N}(x_{i,j}|\beta, \xi, z_{i,j})\, d_{\theta_i} \tag{2}$$

5.1. Variational inference

In order to solve formula (2), all the parameters in formula (2) must be estimated. To learn the hyper parameters $\Theta = \{\pi, \mu, \Sigma, \beta, \xi\}$, using maximum likelihood estimation, we want to calculate formula (3).

$$\text{argmax}_{\Theta}\prod_{i=1}^{M} P(G_i|\Theta) \tag{3}$$

It is intractable to compute because of the integral in formula (2), whose simplify form is necessary. Considering the logarithm form of formula (3), we can get new objective function as formula (4).

$$\text{argmax}_{\Theta}\sum_{i=1}^{M} \ln P(G_i|\Theta) \tag{4}$$

Then, introducing distribution of latent variables $q_i(y, \theta, z)$ and according to the Jensen inequality, we can get the lower bound of it and be shown in formula (5). If and only if the $q_i(y, \theta, z) = p(y, \theta, z|G_i, \Theta)$ holds, the equality sign in the inequality holds.

$$\ln p(G_i|\Theta) \geq \mathbb{E}_{q_i}[\ln p(y, \theta, z, G_i|\Theta)] - \mathbb{E}_{q_i}[\ln q_i(y, \theta, z)] \tag{5}$$

Therefore, the following formula (6) is the problem we are asking for.

$$\text{argmax}\sum_{i=1}^{M} \mathbb{E}_{q_i}[\ln p(y, \theta, z, G_i|\Theta)] - \mathbb{E}_{q_i}[\ln q_i(y, \theta, z)] \tag{6}$$

Since it is difficult to find the conditional probability distribution and corresponding expectations, variational inference is introduced to solve it. Considering the CHGM model shown in Fig. 5(right) and according the mean field assumption, we assume all the hidden variables are independent with each other.

$$q(y_i, \theta_i, z_{i,j}|\gamma, \lambda, \nu^2, \phi) = \prod_{i=1}^{M} q(y_i|\gamma) q(\theta_i|\lambda, \nu^2) \prod_{j=1}^{N} q(z_{i,j}|\phi) \tag{7}$$

where $\gamma \in \mathbb{S}^T, \phi \in \mathbb{S}^K, \lambda$ is a k-dimensional vector and ν^2 is a $K \times K$ covariance matrix. $\{\gamma, \lambda, \nu^2, \phi\}$ are the variational parameters. The specific form of the variational parameters is as follows.

$$q(y_i|\gamma) = Mult(\gamma), q(\theta_i|\lambda, \nu^2) = \mathcal{L}(\lambda, \nu^2), q(z_{i,j}|\phi) = mult(\phi) \tag{8}$$

We can rewrite the part of formula (6) with the variational parameter.

$$L(\gamma, \lambda, \nu^2, \phi; \pi, \mu, \Sigma, \beta, \xi) = \mathbb{E}_q[\ln p(y_i, \theta_i, z_i, G_i|\Theta)] - \mathbb{E}_q[\ln q(y_i, \theta_i, z_i|\gamma, \lambda, \nu^2, \phi)] \tag{9}$$

So, the objective function can be expressed as follows formula (10).

$$\text{argmax}_{\gamma, \lambda, \nu^2, \phi, \Theta} \sum_{i=1}^{M} L(\gamma, \lambda, \nu^2, \phi; \Theta) \tag{10}$$

Continuing expand $L(\gamma, \lambda, \nu^2, \phi; \Theta)$, we have follow form:

$$\begin{aligned} L(\gamma, \lambda, \nu^2, \phi; \Theta) &= \mathbb{E}_q[\ln p(y_i, \theta_i, z_i, G_i|\Theta)] - \mathbb{E}_q[\ln q(y_i, \theta_i, z_i|\gamma, \lambda, \nu^2, \phi)] \\ &= \mathbb{E}_q[\ln p(y_i|\pi)] + \mathbb{E}_q[\ln p(\theta_i|\mu_{y_i}, \Sigma_{y_i})] + \sum_{j=1}^{N} \mathbb{E}_q[\ln p(z_{i,j}|\theta)] \\ &\quad + \sum_{j=1}^{N} \mathbb{E}_q[\ln p(x_{i,j}|z_{i,j}, \beta, \xi)] - \sum_{t=1}^{T} \mathbb{E}_q[\ln q(y_t|\gamma_t)] \\ &\quad - \sum_{k=1}^{K} \mathbb{E}_q[\ln q(\theta_k|\lambda_k, \nu_k^2)] - \sum_{k=1}^{K} \mathbb{E}_q[\ln q(z_k|\phi_k)] \end{aligned} \tag{11}$$

EM algorithm [4] is used to estimate all the parameters through maximize the function L. EM algorithm has two stages during its iterative process: Expectation step (E-step) and Maximum step (M-step). The optimal variational parameters are firstly calculated in the E-step, and then the variational parameters are kept fixing, the optimal model parameters are estimated in the M-step. When the iterative process is terminal, the final parameters set $\{\gamma, \lambda, \nu^2, \phi; \Theta\}$ results are used in the proposed model. For the seven expected formulas of L expansion, if the parameters are constrained, we use Lagrange multiplier method to find partial derivatives. The calculation results of the parameters are as follows.

$$\begin{aligned} \pi^* &= \frac{\sum_{i=1}^{M} \gamma_{i,t}}{\sum_{t=1}^{T} \sum_{i=1}^{M} \gamma_{i,t}} \\ \mu^* &= \frac{1}{M} \sum_{i=1}^{M} \lambda_i, \\ \Sigma^* &= \frac{1}{M} \sum_{i=1}^{M} \mathbf{I}(\nu_i^2) + (\lambda_i - \mu^*)(\lambda_i - \mu^*)^T, \end{aligned} \tag{12}$$

For parameters and,we need to calculate formula (13).

$$\text{argmax}_{\beta, \xi} \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{K} \phi_{i,j,k} \ln \mathcal{N}(x_{i,j}|\beta_k, \xi_k) \tag{13}$$

Let $\phi_(i, j, k)$ be the proportion of topic distribution, we fit the Gaussians in GMM model to solve β and ξ.

5.2. Optimizing hyper-parameters using GA

Genetic algorithm [14] is proposed by imitating the process of biological evolution. Genetic algorithm imitates the individuals in the biological community by compiling chromosomes, and simulates the reproduction and evolution of all things in the biological community by selecting and cross mutating chromosomes. Therefore, the final individuals obtained by genetic algorithm are basically excellent individuals in the population. On this basis, genetic algorithm is widely used in the optimization of various algorithms. The algorithm framework of genetic algorithm is shown in Algorithm 3. We employ GA to optimize hyper-paramaters to avoid local optimization solution,which helps us solve parameters Θ.

Algorithm 3 Genetic algorithm.

1: Randomly generate an initial population.
2: Select an evaluation function in advance and calculate the fitness value of each individual in the population.
3: Determine whether the iteration stop condition is met. If the algorithm stop condition is met, otherwise continue to execute the program.
4: According to a certain selection method, a certain number of individuals are selected from the population and put into the matching pool.
5: The next generation population was generated by crossing and mutating the individuals in the matched pool.
6: Retrun to step 2.

Table 1
Description of data sets. Four correlated data sets and one independent data set.

Data sets (color in Fig. 4)	Topic distribution
cor-data1(blue)	(0.03,0.53,0.43),(0.07,0.06,0.87)
cor-data2(green)	(0.01,0.22,0.77),(0.06,0.66,0.28)
cor-data3(yellow)	(0.10,0.60,0.30),(0.15,0.05,0.79)
cor-data4(magenta)	(0.55,0.19,0.26),(0.30,0.01,0.69)
ind-data(red)	(0.33,0.64,0.03),(0.33,0.03,0.64)

5.3. Anomaly scoring function

Following the previous process, all the parameters of the model can be estimated and then, the complete model can be established. Some information of the model can provide us anomaly scoring evidence. Generally, the likelihood value $-\ln p(G_i|\theta)$ of ith group in CHGM can be used as the scoring function for point anomalies or point-based anomaly groups. However, owing to its focusing on point anomalies rather than group-level anomalies, it cannot help us acquiring expected detecting effect for distribution-based anomaly group detection. In order to finding proper scoring function for group-level anomalies, we thought that only the topic distribution of each group can be used as scoring function based on the definition of anomaly group. First, the posterior distribution of the topic is inferred from the given data, and then the expected likelihood of the topic distribution is calculated, which can be defined as follows.

$$\mathbb{E}_{\theta_i}[-\ln p(\theta_i|\Theta)] = -\int_{\theta_i} p(\theta_i|\Theta, G_i)\ln p(\theta_i|\Theta)\, d\theta \tag{14}$$

In practice, topic score is used to detect anomaly groups while the likelihood score is used to detect the point-based groups.

6. Experiments

6.1. Experiments on synthetic data

6.1.1. Generate the synthetic data

To verify the performance of CHGM, two kinds of synthetic data sets are generated for comparative experiments. In the first kind (denoted by ind-data), topics are independent with each other, and data point features are also independent with each other. In the second kind (denoted by cor-data), topics are correlated with each other and data point features are also correlated with each other.

The ind-data sets are generated as follow. More than 5000 data points for non-anomalous groups are randomly sampled from a Gaussian Mixture Model (GMM) with three components (K=3), and then, they are random divided into 50 groups (M=50). Two genres (T=2) of non-anomalous groups are defined by two topic distributions. The two genres for groups are given sampled from a two-dimensional GMM. The parameters of our used GMM are set as follow: the means of three components are (−1.7, −1), (1.7, −1) and (0,2), respectively; their covariance matrixes are same as $0.2 \times \mathbf{I}_2$, where $\mathbf{I}_2$ denotes the 2×2 identity matrix; the mixture weights of three components have two genres: (0.33, 0.64, 0.03) and (0.33, 0.03, 0.64); the probability(π) of chosen each genre is (0.48, 0.52).

The cor-data sets are generated by Algorithm 2. In CHGM, we still take three components, namely the number of three topics. Compared with ind-data, the logistic normal distribution are used to generate topic distribution(θ). About logistic normal distribution, the mean value(μ) of the topics is (2,2,3) and covariance matrix (Σ) is (3,0.94,0.94;0.94,3,0.94;0.94,0.94,3). Two genres of topic distribution (θ) belonging to non-anomalous groups are generated. The means of the three components still are (-1.7,-1), (1.7,-1) and (0, 2) and covariance [5] is (0.2, 0.14; 0.2, 0.14).

6.1.2. Experiment results and their analysis

We randomly generate four cor-data sets and one ind-data set, their weights for topics are shown in Table 1. Each data set has 50 groups, and the data in the first 3 groups have been modified by injecting anomaly. The successful detection model should get the lowest abnormal score in the first 3 groups.

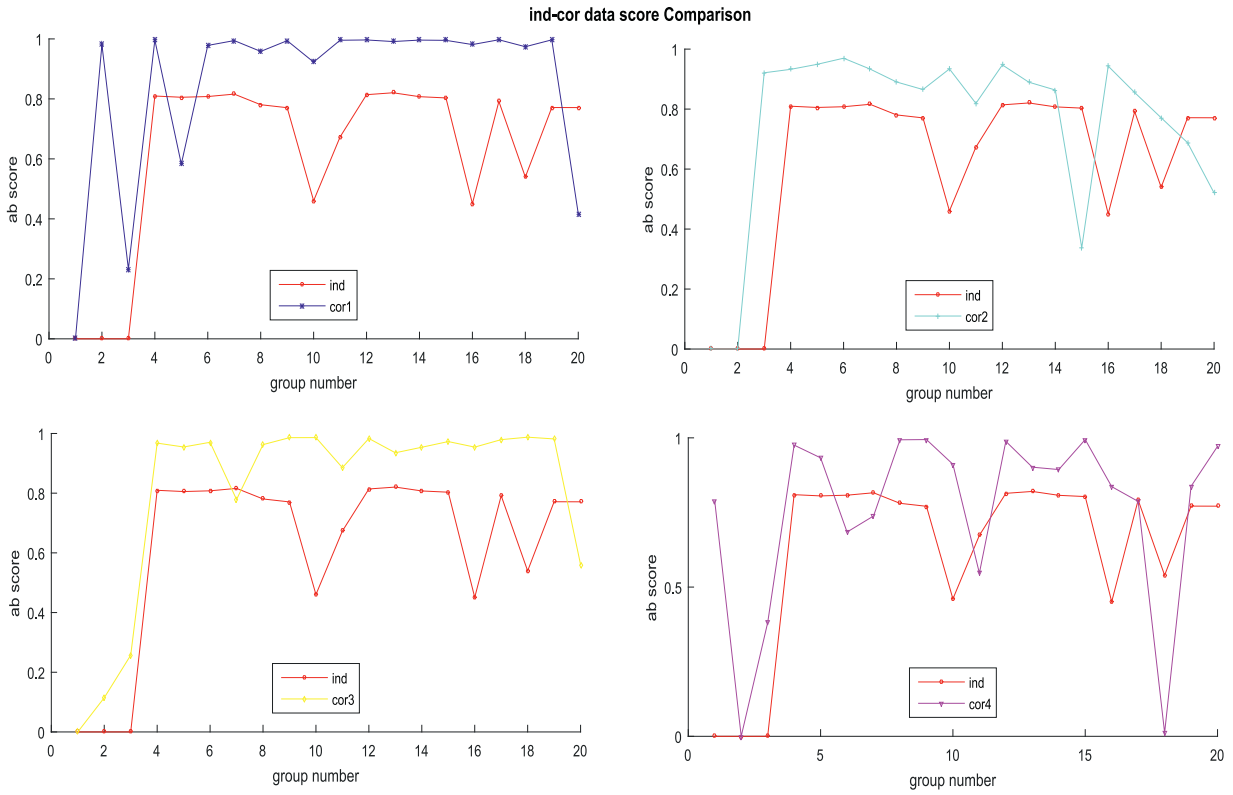

Fig. 4. MGMM for cor-data and ind-data. In order to clearly compare the effect of MGMM method on related data and independent data, we respectively compare one same group of independent data (colored in red) and different four groups of related data (colored in blue, green, yellow and Magenta), and the results are shown in the four subgraphs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Using MGMM [38],we detect anomaly groups in two kinds data(ind-data and cor-data) described above. Fig. 4 shows the experimental results,which is a line chart of anomaly scores obtained by using MGMM model. We conducted four groups comparative experiments. The anomaly scores of the first 20 groups of data in the data sets are shown in Fig. 4, which includes four line charts representing the score comparison of four different sets of correlated data with the same set of independent data respectively. The ind-data(denoted by red line in Fig. 4) acquires significant low scores in the first 3 groups,so they can be inferred as abnormal groups properly. Meanwhile, it can be seen that almost all the anomaly scores of cor-data are larger than ind-data. Moreover, the first 3 groups acquires higher score, so we cannot detect the anomaly groups completely.

From the above experiments, we can see that MGMM cannot effectively detect anomaly groups in correlated data. We will try to detect abnormal groups by our model in the correlated data. We randomly generate five sets of correlated data and inject anomalies into the first 3 groups in the same way. Fig. 5 shows our model's anomaly scores for each group on these five sets of data sets. Obviously, our model detected anomaly groups successfully.

6.2. Experiments on astronomical data

In this section, we compared our model with MGMM,OCSVM,OCSMM and SMDD on real data: The Sloan Digital Sky Survey (SDSS DR8) project. SDSS is a major multi-spectral imaging and spectroscopic redshift survey. Data release 8 (DR8) includes all photometric observations taken with the SDSS imaging camera, covering 14,555 square degrees on the sky (just over 35% of the full sky). The data set contains about 7×10^5 galaxies, each with 4000-dimensional spectral characteristics. We cluster these galaxies according to their spatial distances [38] and produce 505 clusters of galaxies, each containing 10 to 50 galaxies. 505 clusters correspond to 505 galaxy groups to be anomaly detected. Data preprocessing can reduce the dimension of data and make calculation easier. First, we use downsampling to reduce the dimension to 500 dimensions. However, the 500-dimensional data is still not easy for our next work. Finally, we use PCA to reduce the dimension to 4 dimensions and retain 85% of the data information. The boxplot of data features is shown in Fig. 6. Abscissa represents four dimensions of data and longitudinal coordinate represents the values of each data point on these four dimensions. Although there are many salient points in the box diagram, we still haven't done such operation as logarithm for visually showing the range of values of each dimension feature.

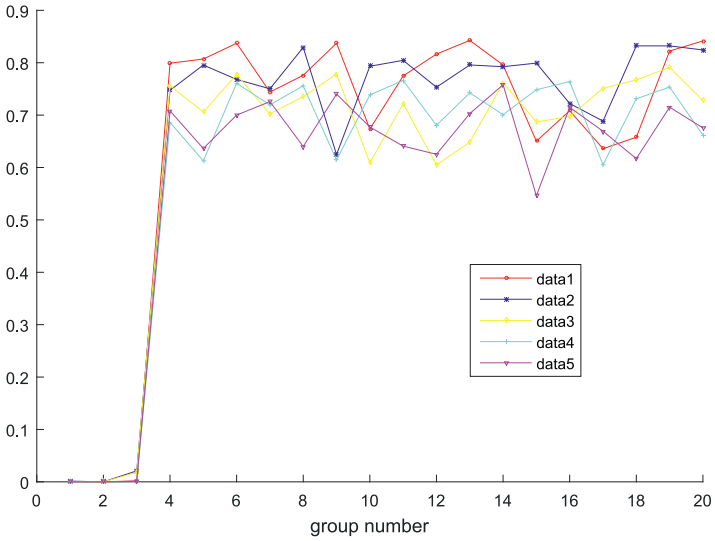

Fig. 5. CHGM for cor-data. Each color corresponds to a set of correlated data. We compare five correlated data sets and acquire good detection effect. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

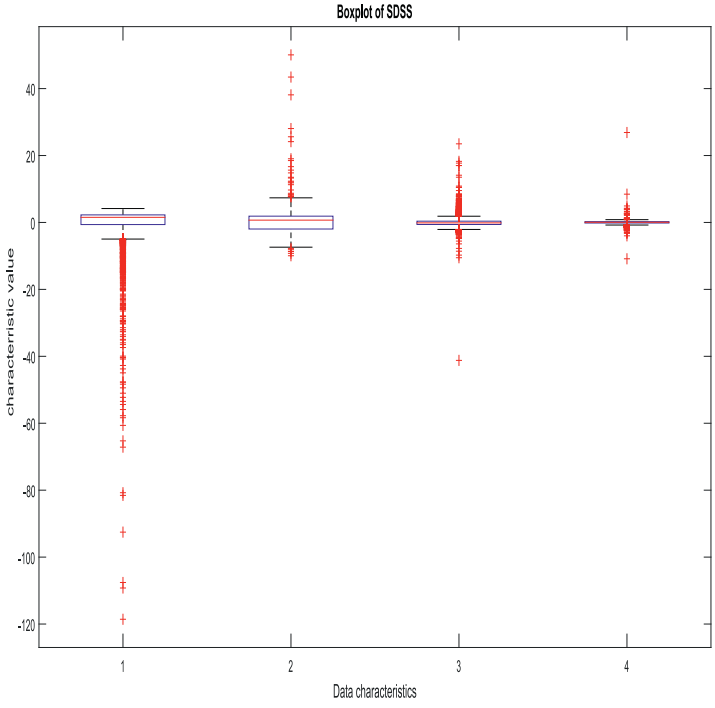

Fig. 6. Boxplot of characteristic data value on SDSS. After preprocessing, we get the feature values of four dimensions.

Since SDSS data set is not labeled, we injected anomalies into 50 groups of 505 data groups, totaling 555 groups of data to participate in the experiment. We compare the existing methods(MGMM,OCSVM,OCSMM,SMDD) with our model proposed in this paper, measuring the average accuracy(AP) and area under the receiver operating characteristic curve(AUC) of anomaly group detection. We carried out 50 runs and adapted GA to train the datas to get reliable results, the box diagram of the performance indicators is shown in the Fig. 7. AUC performances indicate that our model tends to give the anomalies high scores. Further, the AP of our model is much higher than other methods, showing that our model is easier to detect the top anomalies. The MGMM method has also achieved good results. In comparison, the probabilistic model method achieves better performance in anomaly group detection.

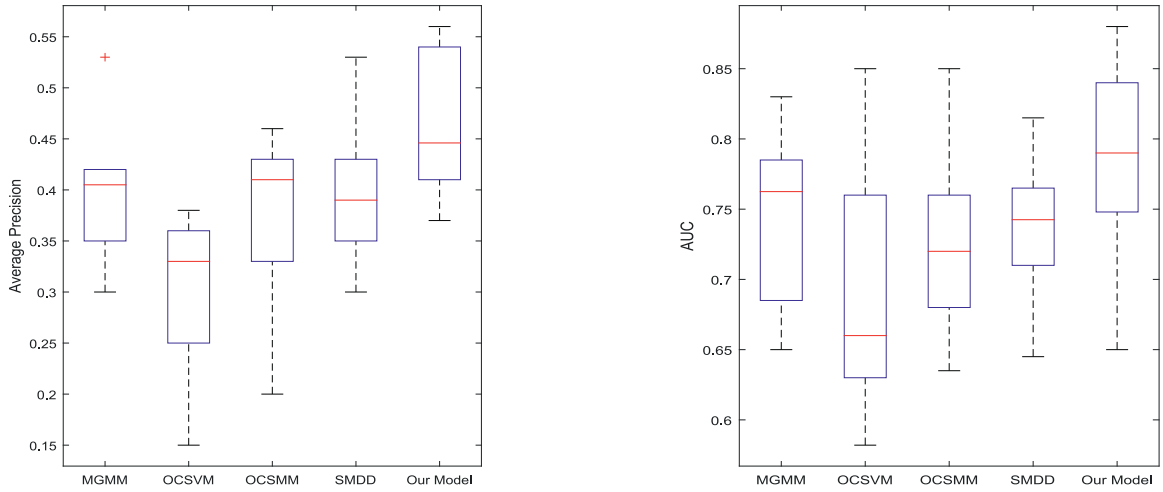

Fig. 7. (Left)The average precision(AP) of different group anomaly detection algorithms on the SDSS data set.(Right)The area under the ROC curve(AUC) of different group anomaly detection algorithms on the SDSS data set.

7. Conclusion

This paper presents a more comprehensive hierarchical generative model to detect anomaly groups. Since the other generative models, such as MGMM and LDA, cannot model the correlation among the topics shared by groups, a logistic normal function is used to solve this issue. Unlike LDA, the proposed model cannot explicitly estimate its parameters, the variational method and GA is employed to calculate them. Some experiments on synthetic data and astronomical data have been carried out on CHGM and its compared model. For synthetic data sets, CHGM exhibits overwhelming merit compared with MGMM in terms of accuracy rate. For astronomical data set, CHGM is superior to the compared methods in terms of Average Precision (AP) and AUC. There are still serval issues to be overcome for our model. The first one is that it is time-consuming to estimate the parameters of CHGM because the prior distribution does not conjugate with the post distribution, it is necessary to find new methods to reduce its complexity. The second one is that we only verify the performance of CHGM on astronomical data, some deep experiments should be performed the other data sets from real applications, such as social networks, video media.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Wanjuan Song: Writing - original draft, Visualization, Investigation, Software, Validation. **Wenyong Dong:** Writing - original draft, Conceptualization, Methodology. **Lanlan Kang:** Writing - review & editing.

References

[1] M. Ahmed, A.N. Mahmood, M.J. Maher, Heart Disease Diagnosis Using Co-Clustering, Scalable Information Systems, Springer International Publishing, 2014.
[2] J. Atchison, S.M. Shen, Logistic-normal distributions:some properties and uses, Biometrika 67 (2) (1980) 261–272.
[3] D.M. Blei, A.Y. Ng, M.I. Jordan, J. Lafferty, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[4] Jie Cao, Juxiang Luo, Xiaoxu Li, Image annotation probabilistic topic model fusing class information, Comput. Eng. Appl. 53(10) (2017) 187–192.
[5] R. Chalapathy, E. Toth, S. Chawla, Group anomaly detection using deep generative models, ECML PKDD 2018, 2018.
[6] Z. Chen, C.K. Yeo, B.S. Lee, C.T. Lau, Y. Jin, Evolutionary multi-objective optimization based ensemble autoencoders for image outlier detection, Neurocomputing 309 (2018) 192–200.
[7] Mi Chen, Yaohua Yi, Deren Li, Qin Qianqing, Application of projection pursuit based on dynamical evolutionary algorithm to anomaly target detection in hyperspectral images, Geom. Inf. Sci. Wuhan Univ. 31 (1) (2006) 55–58.
[8] J. Dutta, B. Banerjee, C.K. Reddy, Rods: rarity based outlier detection in a sparse coding framework, IEEE Trans. Knowl. Data Eng. 28 (2) (2016) 483–495.
[9] W. Fei, S. Chawla, D. Surian, Latent outlier detection and the low precision problem, Acm Sigkdd Workshop on Outlier Detection and Description, 2013.
[10] W. Feng, H. Zhang, K. Li, Z. Lin, J. Yang, X. Shen, A hybrid particle swarm optimization algorithm using adaptive learning strategy, Inf. Sci. 436 (2018) 162–177.
[11] J.J. Flores, A. Antolino, J.M. Garcia, Evolving HMMs for network anomaly detection learning through evolutionary computation, in: Sixth International Conference on Networking and Services IEEE, 2010.

[12] J. Fujiao, S. Yanfeng, G. Junbin, H. Yongli, Y. Baocai, Image outlier detection and feature extraction via l1-norm-based 2D probabilistic PCA, IEEE Trans. Image Process. 24 (12) (2015) 4834.
[13] J. Guevara, S.E. Canu, R. Hirata, Support measure data description, in: International Conference on Knowledge Discovery and Data Mining, 2015.
[14] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor, Development and validation of a genetic algorithm for flexible docking, J. Mol. Biol. 267 (3) (1997) 0–748.
[15] M. Killam, Data mining: concepts and techniques. the morgan Kaufmann series in data management systems, Antimicrob. Agents Chemother. 59 (3) (2015) 1435–1440.
[16] B.J.D. Lafferty, A correlated topic model of science, Ann. Appl. Stat. 1(1) (2007) 17–35.
[17] F.F. Li, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.
[18] K. Li, Y. Chen, W. Li, J. He, Y. Xue, Improved gene expression programming to solve the inverse problem for ordinary differential equations, Swarm Evol. Comput. 38 (2018) 231–239.
[19] K. Li, H. Wang, F. Tang, W. Li, Y. Lu, A mobile node localization algorithm based on the angle self-adjustment model for wireless sensor networks, Int. J. Pattern Recognit. Artif. Intell. 33 (2) (2018) 1958004.
[20] K. Li, Zhuozhi, Performance analyses of differential evolution algorithm based on dynamic fitness landscape, Int. J. Cogn. Inf. Nat. Intell. (IJCINI) 13 (1) (2019) 36–61.
[21] J.W. Liu, L.I. Hai-En, J.J. Zhou, X.L. Luo, D.O. Automation, A survey on the representation theory of probabilistic graphical models, Acta Electron. Sin. 44 (2016) 1219–1226.
[22] S. Liu, Q. Qiang, S. Wang, Heterogeneous anomaly detection in social diffusion with discriminative feature discovery, Inf. Sci. 439-440 (2018) 1–18.
[23] Y. Liu, W. Dong, D. Gong, L. Zhang, Q. Shi, Deblurring natural image using super-gaussian fields, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 452–468.
[24] Y. Liu, W. Dong, M. Zhou, Frame-based variational Bayesian learning for independent or dependent source separation, IEEE Trans. Neural Netw. Learn. Syst. 29 (10) (2018) 4983–4996.
[25] W. Lu, I. Traore, Unsupervised anomaly detection using an evolutionary extension of k-means algorithm, Int. J. Inf. Comput. Secur. 2 (2) (2001) 200–201.
[26] K. Muandet, B. Schoelkopf, One-class support measure machines for group anomaly detection, in: Twenty-Ninth Conference on Uncertainty in Artificial Intelligence., 2013, pp. 449–458.
[27] E.H.M. Pena, L.F. Carvalho, S. Barbon Jr., J.J.P.C. Rodrigues, M.L. Proença Jr., Anomaly detection using the correlational paraconsistent machine with digital signatures of network segment, Inf. Sci. 420 (2017). S0020025517309131.
[28] M. Radovanovic, A. Nanopoulos, M. Ivanovic, Reverse nearest neighbors in unsupervised distance-based outlier detection, IEEE Trans. Knowl. Data Eng. 27 (5) (2015) 1369–1382.
[29] H. Ren, Z. Ye, Z. Li, Anomaly detection based on a dynamic Markov model, Inf. Sci. 411 (2017) 52–65.
[30] Yi Ren, Siqiing Yin, Songyang Li, Scene and place recognition using improved LDA topic model, Comput. Eng. Des. 38 (2) (2017) 506–510.
[31] I.B. Samira Douzi, B. ElOuahidi, Hybrid approach for intrusion detection using fuzzy association rules, in: 2018 2nd Cyber Security in Networking Conference (CSNet), 2018.
[32] I.M. Stephanakis, I.P. Chochliouros, E. Sfakianakis, S.N. Shirazi, D. Hutchison, Hybrid self-organizing feature map (SOM) for anomaly detection in cloud infrastructures using granular clustering based upon value-difference metrics, Inf. Sci. 494 (2019) 247–277.
[33] F. Wang, Y. Li, H. Zhang, T. Hu, X.-L. Shen, An adaptive weight vector guided evolutionary algorithm for preference-based multi-objective optimization, Swarm Evol. Comput. 49 (2019) 220–233.
[34] L.D. Wang, B.G. Wei, J. Yuan, Document clustering based on probabilistic topic model, Acta Electron. Sin. 40 (11) (2012) 2346–2350.
[35] P. Wei, F. Qin, F. Wan, Y. Zhu, J. Jiao, Q. Ye, Correlated topic vector for scene classification, IEEE Trans. Image Process. PP (99) (2017). 1–1.
[36] M. Xi, H. Mo, S. Zhao, J. Li, Application of anomaly detection for detecting anomalous records of terroris attacks, in: IEEE International Conference on Cloud Computing and Big Data Analysis, 2017.
[37] L. Xiong, B. Pczos, J. Schneider, Group anomaly detection using flexible genre models, Adv. Neural Inf. Process. Syst. (2011) 1071–1079.
[38] L. Xiong, B. Pczos, J.G. Schneider, A. Connolly, J. Vanderplas, Hierarchical probabilistic models for group anomaly detection, J. Mach. Learn. Res. 15 (2011) 789–797.
[39] S. Yang, K. Li, L. Wei, W. Chen, C. Yan, Dynamic fitness landscape analysis on differential evolution algorithm, in: International Conference on Bio-inspired Computing: Theories and Applications, 2016.
[40] R. Yu, X. He, L. Yan, Glad: group anomaly detection in social media analysis, in: ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, 2014.
[41] J. Zeng, R. Qin, W. Tang, An extended negative selection algorithm for unknown malware detection, J. Comput. Theor. Nanosci. 13 (6) (2016) 4010–4017.
[42] H.Y Zhang, L.W. Wang, Y.X. Chen, Research progress of probabilistic graphical models: A survey, J. Softw. 24 (11) (2013) 2476.