THEORY AND APPLICATIONS OF SOFT COMPUTING METHODS

OPE-HCA: an optimal probabilistic estimation approach for hierarchical clustering algorithm

Jiancong Fan[1,2,3]

Received: 26 February 2015 / Accepted: 21 July 2015 / Published online: 5 August 2015

Abstract The Survival of the Fittest is a principle which selects the superior and eliminates the inferior in the nature. This principle has been used in many fields, especially in optimization problem-solving. Clustering in data mining community endeavors to discover unknown representations or patterns hidden in datasets. Hierarchical clustering algorithm (HCA) is a method of cluster analysis which searches the optimal distribution of clusters by a hierarchical structure. Strategies for hierarchical clustering generally have two types: agglomerative with a bottom-up procedure and divisive with a top-down procedure. However, most of the clustering approaches have two disadvantages: the use of distance-based measurement and the difficulty of the clusters integration. In this paper, we propose an optimal probabilistic estimation (OPE) approach by exploiting the Survival of the Fittest principle. We devise a hierarchical clustering algorithm (HCA) based on OPE, also called OPE-HCA. The OPE-HCA combines optimization with probability and agglomerative HCA. Experimental results show that the OPE-HCA has the ability of searching and discovering patterns at different description levels and can also obtain better performance than many clustering algorithms according to NMI and clustering accuracy measures.

Keywords Clustering · Hierarchical clustering algorithm · Data mining · Probabilistic estimation

1 Introduction

The phrase "Survival of the Fittest" is originated from evolutionary theory as a way of describing the natural selection mechanism. Generally, it refers that the probability of survivors is high if the survivors are fit for the natural environment. So it is more commonly used today to refer to a supposed greater probability that "fit" as opposed to "unfit" individuals will survive some context.

Clustering is a general task to be solved in data analysis and mining. The clusters obtained by various clustering algorithms differ significantly in their tasks and objectives. So far, however, it is still a hard work to predict what constitutes a cluster hidden in dataset and how to efficiently discover them with high accuracy. One of the main reasons is that clustering is an unsupervised analysis process. It is unknown how many clusters there exist and what the names of clusters are. But this problem occurs in most practical applications, such as web data mining and big data analysis, because it is difficult to foresee exactly the hidden patterns in black box or the event that has not occurred. There have emerged rich clustering strategies and algorithms attempting to solve the blindness in clustering process. However, most of them are specialized algorithms that one algorithm is only suitable to solve one particular dataset.

Among the numerous clustering methods, hierarchical clustering [1, 2], also called connectivity-based clustering,

✉ Jiancong Fan
fanjiancong@sdust.edu.cn

[1] State Key Laboratory of Mining Disaster Prevention and Control Co-founded by Shandong Province and the Ministry of Science and Technology, Shandong University of Science and Technology, Qingdao 266590, China

[2] College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China

[3] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

is one of the most popular strategies. In hierarchical clustering, agglomerative method occupies a prominent position because of its inductive process. But hierarchical clustering has some flaws. For example, the intermediate results are difficult to be reevaluated once the clustering starts to construct the dendrogram for data points because the points that have been grouped into a node cannot move to other branches dynamically. This causes that an incorrect assignment for a point made early in clustering process cannot be corrected. This deficiency has driven the development of other clustering techniques borrowed from distribution density of data points, mathematical models, and others in the artificial intelligence community.

In this paper, we study how to overcome the shortcoming discussed above under the condition of preserving the advantages of agglomerative clustering. We develop a hierarchical clustering algorithm (HCA) based on optimal probabilistic estimation (OPE), also called OPE-HCA, which combines the dendrogram and the optimal probability of survivors. The basic idea of our proposed OPE-HCA algorithm is to select individuals with higher probability from probabilistic distribution as optimal individuals. These optimal individuals are combined as one population which can generate a parent node. Thus, the OPE-HCA algorithm is able to produce different kinds of cluster distributions at different levels of tree, which can offset the incorrectness of agglomerative process.

The rest of the paper is organized as follows. The next section overviews the clustering algorithms based on hierarchical idea and probabilistic estimation over the years, especially in recent years. Section 3 briefly describes the concepts and advantages of the optimal probabilistic estimation (OPE) and estimation of distribution algorithms (EDAs) because we apply the idea of EDAs for the development of OPE. We present the OPE based on EDAs in Sect. 4.1. The OPE-HCA algorithm is given and analyzed in Sect. 4.2. And Sect. 5 gives detailed experimental results with twelve real datasets. Finally, Sect. 6 concludes the paper.

2 Related work

There have been numerous clustering algorithms used in a wide range of research and application areas [3]. Among those algorithms, the partitional and hierarchical clustering methods are applied widely due to their simplicity, practicality, and ease of implementation. The idea of partitional clustering method firstly determines a set of initial seeds, and each seed represents a cluster. All seeds are then improved iteratively by the designed criterion functions. The typical representatives of partitional clustering methods include K-means clustering and its variations [4–6], K-medoids clustering [7, 8], and fuzzy c-means clustering [9, 10]. Although K-means-type clustering has inherent limitations, such as subjective determination of the number of clusters and local convergence, this type of algorithms is applied widely in many data domains due to their simplicity and understandability and also has good performances in many cases. Hierarchical clustering method is to construct dendrogram using tree data structure. Two ways of constructing such a tree are bottom-up or top-down methods. The bottom-up clustering method is an inductive process and more natural than the top-down, so the bottom-up clustering strategy is more widely applied [2, 11]. In addition to the above two kinds of clustering methods, probability-based clustering method has caused the great concern of many researchers in recent years [12–21]. Probability-based clustering method mainly includes model-based strategy [12–16] and evolutionary probabilistic modeling strategy [17–21].

Probabilistic estimation approach has been widely applied in data mining and machine learning community. Yan et al. [22] presented a stochastic optimization algorithm based on the idea of the gradient method which incorporated probabilistic estimation technique. Sánchez et al. [23] extended the probabilistic estimation of stochastic context-free grammars (SCFGs) by proving the property of consistency for all SCFGs without restrictions. Apte et al. [24] used the probabilistic estimation predictive-modeling data mining kernel and discovered risk-characterization rules by analyzing large and noisy datasets. Ferri et al. [25] investigated several issues in order to improve the performance of probabilistic estimation trees, the quality of which is evaluated by the 1-versus-1 multiclass extension of the area under the ROC curve (AUC) measure. Jaulin [26] transformed probabilistic estimation problems into a set of estimation problems by assuming that some rare events will never happen to study the nature of probabilistic estimation. Choi et al. [27] proposed a multiple-criteria decision-making approach based on probabilistic estimation technique, called maximum A posteriori (MAP), to analyze users' physiological status. Han et al. [28] presented two probabilistic estimation techniques to identify the most likely number of frequency-hopping transmitters and compare their performances. Jiang et al. [29] investigated the class probability-estimation performance of Tree Augmented Naive Bayes (TAN) using conditional log likelihood and presented an algorithm to improve its class probability-estimation performance by the spanning TAN classifiers. Pimentel et al. [30] presented a probabilistic estimation approach using a Gaussian process framework which estimated the respiratory rate from the different sources of modulation. Duchi et al. [31] studied the estimation of probability discrete and continuous distributions and given a sharp mini-max rates of convergence for estimation in locally private settings. Azad et al. [32] applied a color probabilistic estimation technique to detect face skin in human face perception.

In probabilistic modeling-based hierarchical agglomerative clustering domain, Friedman [33] described a hierarchical agglomerative clustering method based on probabilistic modeling, which used an adaptive probabilistic modeling approach to adjusting probabilistic models to adapt distinctions in the process of learning different variances for different cells. Segal and Koller [34] described a general probabilistic clustering framework called probabilistic abstraction hierarchies, which can cluster data into a hierarchical structure. In clustering algorithms based on evolutionary probabilistic modeling, Fan et al. [20, 35] proposed clustering algorithm and semi-supervised clustering algorithm based on EDAs.

The previous work on clustering based on hierarchical approaches has their own flaws. To our knowledge, no work exists on clustering based on OPE with the virtues of hierarchal approaches. In this paper, we give a probability-estimation-based agglomerative clustering algorithm based on our proposed OPE technique. This technique borrows the idea from estimation of distribution algorithms (EDAs). And the next section introduces the idea of EDAs and our proposed idea of OPE based on EDAs.

3 Brief review of estimation of distribution algorithms (EDAs) and the optimal probabilistic estimation (OPE)

Estimation of probabilistic model is one of the key ideas of estimation of distribution algorithms (EDAs) [36] which are a class of evolutionary computation methods without random crossover and mutation operators. EDAs combine probabilistic modeling with evolutionary computation to select optimal individuals with high probability instead of using crossover and mutation operators. EDAs are also called probabilistic model-building genetic algorithm. In EDAs, each position of the population in generation G can be computed based upon the population in generation G-1. In a population at generation G, the probability of value j appearing in position i in a solution vector v can be computed as follows:

$$P(i,j) = P(v_i = j) = \frac{\sum_{v \in (G-1) \wedge v_i = j} \text{Evaluate Vector}\,(v)}{\sum_{v \in (G-1)} \text{Evaluate Vector}\,(v)} \tag{1}$$

The aim of EDAs is to construct probability vectors and use vectors with high probability values to represent populations. The approach need not design genetic operators any more. Instead, the individuals in a population are sampled from a probability distribution estimated by a database containing the individuals of the previous generation.

The idea of the OPE proposed in this paper comes from EDAs. EDAs have two key ideas, estimation and sampling. Estimation is to build the probability distribution model in the solution space. Sampling is to choose the best local solutions according to the estimated probability distribution model in order to produce another better solution space. The OPE focuses on estimating the best solution sampling from the ones formed by merging the subsolutions having maximum probabilities among all subsolutions. Assume that $C = \{c_1, c_2, \ldots, c_m\}$ is the set of all possible categories of dataset D, $\{D_j \mid D_j \subseteq D, j = 1, 2, \ldots, n\}$, and D_j is the cluster of jth instance in some view V_k. The estimation model of maximum probability is derived from:

$$\begin{aligned} E_{\text{best}} &= \arg\max P(i,\ j) \\ &= \arg\max P(v_i = j) \\ &= \arg\max P(D_j \in c_i) \\ &= \mathop{\arg\max}_{\substack{1 \le i \le m \\ 1 \le j \le n}} P(c_i | D_j, D_j \subseteq V_k, \quad st., \quad k = 1,\ 2, \ldots) \end{aligned}$$

Let V_{k1} and V_{k2} are two arbitrary views corresponding to D_{j1} and D_{j2}, the subsets of D_j. The estimation of maximum probability model for V_k in their higher layer is given by:

$$\begin{aligned} E_{D_{j1}} &= \arg\max P(c_{i1} | D_{j1}, D_{j1} \subseteq V_{k1}) \\ E_{D_{j2}} &= \arg\max P(c_{i2} | D_{j2}, D_{j2} \subseteq V_{k2}) \\ E_{\text{best}} &= E_{D_{j1}} \infty E_{D_{j2}} \\ &= \arg\max P(c_{i1} \cup c_{i2} | D_{j1} \cup D_{j2},\ D_{j1} \cup D_{j2} \subseteq V_k) \end{aligned}$$

We adopt the above idea to estimate the optimal clustering probability to obtain the most probably correct clusters. Generally speaking, it is difficult to guess which clusters should be classified into the same clusters in addition to those distance-based approaches. However, distance-based approach is good at recognizing those globular clusters. Probability-based approach is an alternative technique to find the similar clusters in any positions in space because it need not calculate distance but only consider the frequency. OPE thinks that those clusters with high probabilities are the ones belonging to the same class.

4 Hierarchical clustering algorithm based on optimal probabilistic estimation

In this section, we will introduce our proposed approach OPE-HCA. OPE-HCA algorithm includes two key procedures. One is to use distance-based computation to cluster similar data instances in the bottom layer. The other key procedure is to deal with clusters in each hierarchy by OPE. Our proposed OPE is given in Sect. 4.1. Section 4.2 presents the OPE-HCA algorithm in detail.

4.1 Optimal probabilistic estimation (OPE)

OPE can deal with non-deterministic problems by maximum probabilistic modeling. The basic idea of OPE is to select samples with higher probability as optimal individuals from probabilistic distribution. Assume that dataset D consists of n ($n > 1$) subsets denoted by sub_1, sub_2, …, sub_n. Each of the subsets is called an individual, the occurrence probability of which is denoted by $p_1 = P(\text{sub}_1)$, $p_2 = P(\text{sub}_2)$,…, $p_n = P(\text{sub}_n)$, respectively, where $P(\text{sub}_i)$ is given by

$$P(\text{sub}_i) = \frac{|\text{sub}_i|}{|\text{sub}_1| + |\text{sub}_2| + \cdots + |\text{sub}_n|}$$

Assume that the maximum probability is $p_i = \max_{1 \le j \le n}\{p_j\}$ which corresponds to the occurrence probability of individual sub_i. Let *st*. denotes the fitness constraint. Under *st*., when the probability of the union of sub_i and sub_j ($i \neq j$) is greater than the union of sub_i and another sub_k ($k \neq j$), as formula (2) shows,

$$\begin{aligned}&P(\text{sub}_i \cup \text{sub}_j | st., i \neq j) > \\ &P(\text{sub}_i \cup \text{sub}_k | k = 1, 2, \ldots, n \quad \text{and} \quad k \neq i, k \neq j)\end{aligned} \tag{2}$$

we can obtain a new individual sub_{ij}, the union of sub_i and sub_j. And the occurrence probability of sub_{ij} is denoted by p_{ij}. If sub_i with the maximum probability value cannot satisfy Eq. (2), we need to take the subset with the second largest probability value to continue the above process. Each union of the individuals leads to the update of data distribution in D, and each iteration can generate a distribution view of D. Equation (2) is called optimal probabilistic estimation (OPE) criteria. In order to introduce the computation algorithm based on OPE, we give Definition 1.

Definition 1 Clustering view. A set D with n data points is partitioned into $\Bbbk$ clusters (also called subset), C_1, C_2, …, C_k, where $|C_i| \ge 1$ ($i = 1, 2,\ldots, k$) and $C_i \cap C_j = \phi$, then $\{C_1, C_2, \ldots, C_k\}$ is called a clustering view of D, denoted by $\mathbb{S}$. When $|C_i| = 1$ ($i = 1, 2,\ldots, k$),that is, $n = k$,$\{C_1, C_2, \ldots, C_k\}$ is called source view of D, denoted by $\mathbb{S}_0$.

We give the optimal probabilistic estimation algorithm in Fig. 1.

In the OPE algorithm described in Fig. 1, $\mathbb{D}^{(1)}$ and λ are two key parameters. $\mathbb{D}^{(1)}$ is the source view, and λ denotes the threshold for permitting two clusters to be merged into one, which is defined as Definition 2 in Sect. 4.2.

4.2 OPE-based hierarchical clustering algorithm

The OPE-HCA is designed based on OPE procedure. We firstly introduce Definition 2.

```
OPE ($D^{(1)}$, $\lambda$) Algorithm
Step 1  Give a population with $t$ individuals $C_1, C_2, \ldots, C_t$, and
        $\{C_1, C_2, \ldots, C_t\}$ is a clustering view of data set $D$
        denoted by $Pop^{(0)} = \{C_1, C_2, \ldots, C_t\}$;
Step 2  Compute the occurrence probability of $C_1, C_2, \ldots, C_t$ in
        $Pop^{(0)}$, $p_1 = P(C_1)$, $p_2 = P(C_2)$, ..., $p_t = P(C_t)$, where
            $p_i = P(C_i) = \frac{|C_i|}{|\mathbf{D}|}$, $i = 1, 2, \ldots, t$.
        and $p_1 + p_2 + \ldots + p_t = 1$;
Step 3  Sort over all individuals $\{C_i \mid i = 1, 2, \ldots, t\}$ in $Pop^{(0)}$ by
        their occurrence probabilities $p_i$ ($i = 1, 2, \ldots, t$)
        descending, and then obtain a sorted set $D^{(1)}$;
Step 4  Set $k = 1$;
Step 5  while (true) {
            Select individuals in turn from $D^{(k)}$, that is,
            firstly select the ones with higher probability;
            Given criterion $\Theta$, if there exists an individual
            $C_j$, for arbitrary integer $u$, $u \neq i$ and $u \neq j$,
            when $C_{ij} = C_i \cup C_j$ ($i \neq j$), the following formula
            holds,
                $P(C_{ij} \mid \Theta) \ge P(C_u \mid \Theta, u \neq i \text{ and } u \neq j)$,
            and $C_i$ and $C_j$ satisfy
                $|\text{mean}(C_i) - \text{mean}(C_j)| \le \lambda$,
            $C_i$ and $C_j$ are merged into one set to obtain a new
            population $Pop^{(k)}$,
                $Pop^{(k)} = \{C_{ij}\} \cup (Pop^{(k-1)} - \{C_i, C_j\})$.
            if (there exist elements in $D^{(k)}$ which are not
            traversed yet) continue;
            Sort over the individuals in $Pop^{(k)}$ descending by
            their probability to obtain sorted population
            denoted by $D^{(k+1)}$, which is a new clustering
            view with less clusters;
                $S_{(k+1)} \leftarrow \mathbf{D}^{(k+1)}$;
            $k = k + 1$;
        } until (only one element remains in $D^{(k)}$, or the
        termination conditions are satisfied.)
```

Fig. 1 OPE algorithm

Definition 2 Minimum merging distance between clusters (MinMD). For arbitrary two clusters C_i and C_j, their mean values are $c_{\text{mean}}^{(i)}$ and $c_{\text{mean}}^{(j)}$, respectively. For a given distance threshold λ, if Eq. (3) can be satisfied, C_i and C_j are merged into one cluster. The threshold λ is called minimum merging distance.

$$\left|c_{\text{mean}}^{(i)} - c_{\text{mean}}^{(j)}\right| \le \lambda \tag{3}$$

The MinMD defined in Definition 2 is used to decide whether two clusters can be merged. The value of MinMD is assigned as the actual parameter and transferred to OPE procedure in OPE-HCA algorithm as shown in Fig. 2.

In OPE-HCA clustering algorithm, the criterion Θ is needed to unite clusters. We give two definitions as follows.

```
OPE-HCA algorithm
Input:
    Data set $D=\{D_1, D_2, \ldots, D_n\}$, and $D_i=(d_{i1}, d_{i2}, \ldots, d_{im})$ is
    $m$-dimensional tuple;
    Distance parameter $\theta$.
Procedure:
    (1) Compute the distances between data by following
    formula;
    Distance($D_i$, $D_j$)=
        $\sqrt{(d_{i1}-d_{j1})^2+(d_{i2}-d_{j2})^2+\cdots+(d_{im}-d_{jm})^2}$ , $i \neq j$;
    (2) Merge $D_i$ and $D_j$ into one cluster if satisfying
        Distance($D_i$, $D_j$)$\leq\theta$, thereby generate $k$ clusters
        $C_1$, $C_2$, …, $C_k$ to obtain the source clustering view
        $S_1$;
    (3) Assign the value of MinMD to $\lambda$;
    (4) Call algorithm OPE ($S_1$, $\lambda$);
Output:
    The clustering views on the condition of different
    parameters $\theta$ and $\lambda$.
```

Fig. 2 OPE-HCA clustering algorithm

Definition 3 Reachability from set $\mathbb{Z}^j$ to another set $\mathbb{Z}^i$. Assume $z_{\text{mean}}^{(j)}$ is the mean value of $\mathbb{Z}^j$ and $z_{mean}^{(i)}$ is the mean value of $\mathbb{Z}^i$. Given a small positive number ε, if Eq. (4) holds, the set $\mathbb{Z}^j$ can reach the set $\mathbb{Z}^i$ for parameter ε.

$$\left|z_{\text{mean}}^{(i)} - z_{\text{mean}}^{(j)}\right| \leq \varepsilon \tag{4}$$

Definition 4 First reachability of the sets. Assume that there are three sets, $\mathbb{Z}^i$, $\mathbb{Z}^j$, and $\mathbb{Z}^k$. $z_{\text{mean}}^{(j)}$ is the mean value of $\mathbb{Z}^j$, $z_{\text{mean}}^{(i)}$ is the mean value of $\mathbb{Z}^i$, and $z_{\text{mean}}^{(k)}$ is the mean value of $\mathbb{Z}^k$. Given a small positive number ε, compute the following equations:

$$\begin{aligned} \left\|z_{\text{mean}}^{(j)} - z_{\text{mean}}^{(i)}\right\| &= \varepsilon_1 \\ \left\|z_{\text{mean}}^{(j)} - z_{\text{mean}}^{(k)}\right\| &= \varepsilon_2 \end{aligned} \tag{5}$$

If $\varepsilon_1 < \varepsilon_2 \leq \varepsilon$, the set $\mathbb{Z}^j$ first reaches the set $\mathbb{Z}^i$.

The criterion Θ in this paper is defined as follows: The set that can merge C_j is the one that C_j firstly reaches.

In addition to initialization and parameter setting in OPE-HCA algorithm, there are also two important procedures: (1) cluster the nearest data instances in source view by distance-based measurement to generate the first abstract view; (2) merge clusters by OPE from the first abstract view to the follow-up.

The distance-based measures can be put into direct relationship to the datasets by a lot of formulations according to many kinds of data types and formats. We adopt Euclidean distance to cluster data instances because it is reportedly faster than most other means of determining correlation and is a fair measure of how similar ratings are for specific preferences or items. Whatever distance-based measure we use, distance threshold θ is the key for OPE-HCA algorithm and can even influence the clustering results. The experiments with different values of parameter θ will be carried out in Sect. 5. The second key procedure is how to generate views at different abstract levels. The procedure is OPE algorithm, which has been presented in Sect. 4.1, and is called in OPE-HCA algorithm.

The general solution structure is illustrated as Fig. 3 after executing OPE-HCA algorithm.

Figure 3 is a bottom-up hierarchical structure. In Fig. 3, q clustering views are generated when OPE-HCA algorithm finishes normally. View$_1$ is the source clustering view. In real data clustering, each sub_i ($i = 1, 2, \ldots, n$) in view$_1$ is a data instance. Distance parameter θ is only applied to view$_1$ to merge similar data instances and generate view$_2$. From view$_2$ to view$_q$, the parameter λ is used to merge nearest subsets (clusters) in each view$_i$ ($i = 2, \ldots, q$). A view is stable when no subsets can be merged under condition of λ and OPE. The top of the general solution structure is the last output of OPE-HCA algorithm, which denotes that all clusters cannot be merged and the algorithm is terminated.

5 Experimental results

In this section, we give experiments on many datasets by our OPE-HCA algorithm and other classical clustering algorithms. Section 5.1 explains the experimental settings including experimental datasets, compared algorithms, and evaluation metrics. Section 5.2 presents the sensitivity

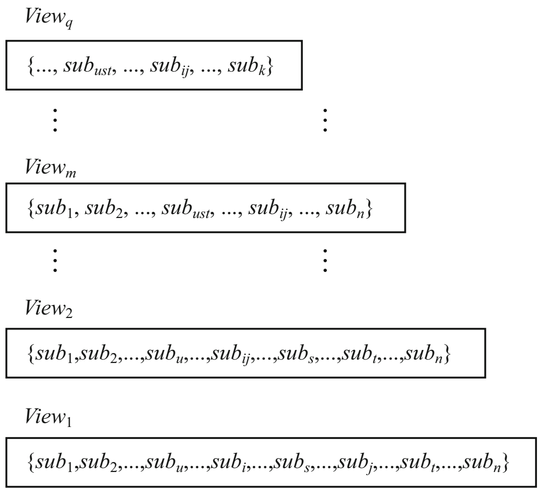

View$_q$
{..., sub_{ust}, ..., sub_{ij}, ..., sub_k}
View$_m$
{sub_1, sub_2, ..., sub_{ust}, ..., sub_{ij}, ..., sub_n}
View$_2$
{sub_1,sub_2,...,sub_u,...,sub_{ij},...,sub_s,...,sub_t,...,sub_n}
View$_1$
{sub_1,sub_2,...,sub_u,...,sub_i,...,sub_s,...,sub_j,...,sub_t,...,sub_n}

Fig. 3 General solution structure after executing OPE-HCA algorithm

analysis of the parameters used in OPE-HCA algorithm. Section 5.3 presents the detailed experimental results.

5.1 Experimental setting

5.1.1 Datasets

The OPE-HCA algorithm we proposed has been tested with twelve datasets taken from UCI machine learning repository [37]. The important statistics of twelve datasets are summarized in Table 1.

The datasets used in this paper listed in Table 1 are all real and multivariate data instances with class label attributes.

5.1.2 Compared algorithms

In order to test our clustering algorithm OPE-HCA and other well-known clustering algorithm, including K-means [6], fuzzy c-means (FCM) [10], BIRCH [11], and Bayesian hierarchical clustering (BHC) [15], the class label attributes in the datasets are completely deleted. The reason why we choose K-means, FCM, BIRCH, and BHC as the comparing algorithms with our algorithm is that our proposed algorithm combines hierarchical idea, distance-based method with probabilistic computation. The K-means clustering is based on distance measurement. FCM is based on Euclidean distance function and associated with fuzzy mathematics. BIRCH is a classical hierarchical clustering algorithm with good performances. BHC is a probabilistic clustering method based on Bayesian model.

Table 1 Information of the tested benchmark datasets

Dataset	Size	# Attribute	# Cluster
Iris	150	4	3
Seeds	210	7	3
Yeast	1484	8	10
Glass	214	10	7
Ecoli	336	8	8
Haberman’s survival (HS)	306	3	2
Balance scale	625	4	3
Transfusion	748	5	2
Contraceptive method choice (CMC)	1473	9	3
Page-blocks	5473	10	5
User knowledge modeling (UKM)	403	5	4
Wine	178	13	3

5.1.3 Evaluation metrics

1. Clustering accuracy;
2. Normalized Mutual Information (NMI).

The above metrics will be explained in Sect. 5.3.

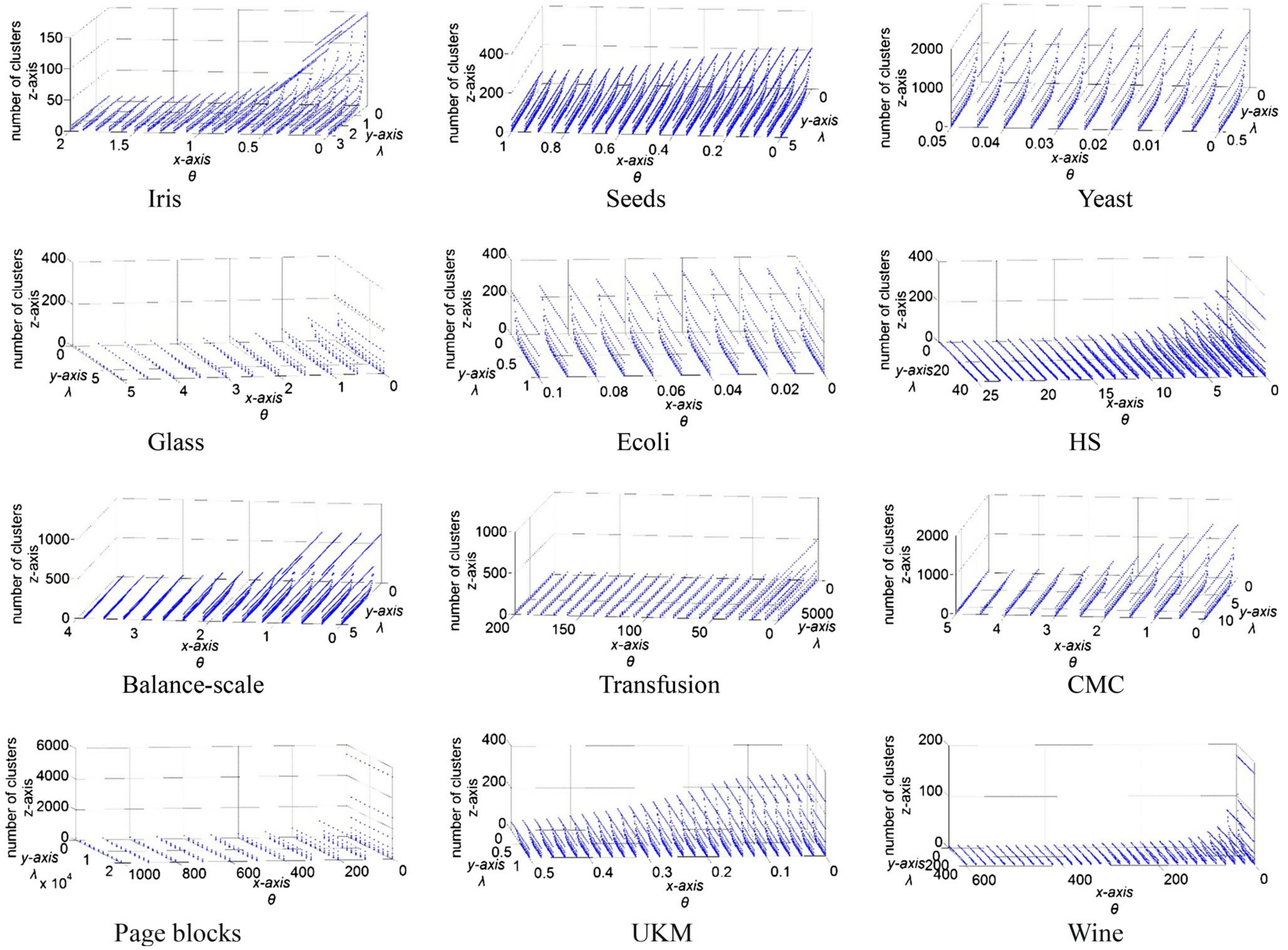
5.2 Sensitivity analysis

In our experiments, two parameters are used to analyze the performance of OPE-HCA algorithm: the parameter θ and the parameter λ. The accuracy and the Normalized Mutual Information (NMI) [38] are applied to evaluate the performance of our algorithm OPE-HCA and other clustering algorithms.

The first step of OPE-HCA algorithm is to cluster data instances by threshold θ to generate clustering view. The greater the value of θ, the less the number of clusters. It is due to the dependence of the change of similarity degree between data instances over the size of θ. The greater value of θ can make farther-away data clustered into the same class. The parameter λ is applied to merge clusters starting from the first abstract view. Similarly, the greater value of λ causes that more clusters can be merged into the same more abstract cluster. Figure 4 shows the variations of the number of clusters for the twelve benchmark datasets, respectively, as the parameter θ and λ changes. From Fig. 4, it is easily observed that OPE-HCA algorithm is able to converge to the minimum number of clusters with the increasing value of θ and λ.

In Fig. 4, each point in x axis, y axis, and z axis denotes the value of θ, λ, and the number of clusters in each view, respectively. The range of θ is from zero to the difference value of the maximum and minimum of dataset. The range of λ takes the mean values of all clusters in each view according to Definitions 3 and 4. The values on z axis are generated by OPE-HCA algorithm. From Fig. 4, the number of clusters in a certain view is more when the values of θ and λ are smaller. When the values of θ and λ increase, the number of clusters in a view gets closer to the actual cluster number of dataset. The results of the number of clusters in a view obtained from OPE-HCA algorithm and the true number of clusters for all datasets are shown in Table 2. It is easy to see that the number of clusters obtained by our algorithm almost equals to the true number for most of the tested datasets.

The number of hierarchies generated by OPE-HCA algorithm increases with increasing MinMD (see Definition 2) λ for the same parameter values of θ. The reason is that the values of λ is smaller, the number of reachable individuals from a certain individual is less according to Definition 3, which causes less individuals to be able to be merged to generate consequent views; on the contrary, the



Iris

Seeds

Yeast

Glass

Ecoli

HS

Balance-scale

Transfusion

CMC

Page blocks

UKM

Wine

Fig. 4 Illustrations of the number of clusters during executing OPE-HCA algorithm with the changes in θ and λ for all datasets, where x axis is the value of θ, y axis is the value of λ, and z axis is the number of clusters along the x axis and y axis. The number of clusters converges to a stable value with the increase in θ and λ values for all datasets

larger values of λ cause more individuals to be able to be merged to generate more consequent views. Under the condition of reasonable value of λ, we can obtain a view which is close or equal to the exact clusters and the number of clusters. The results show that OPE-HCA algorithm can be convergent to a certain view. With same value of λ, various values of θ have little influences on the number of generated views. It is due to the usage of parameter θ only at the first step of OPE-HCA algorithm to generate the source view. However, θ is a very important basic parameter which can influence the clustering performances. If the value of θ is too small, the number of individuals in source view will be too big, which results in consuming extra time at consequent steps; if the value of θ is too big, some data instances far from others are clustered into the same class, which may cause the higher error rate.

5.3 Comparison of OPE-HCA and other algorithms

OPE-HCA algorithm and other compared algorithms described in Sect. 5.1 are implemented, and the results in different evaluation metrics are compared in this section.

5.3.1 Comparison of clustering accuracy results on classical algorithms

Table 3 shows the comparison of OPE-HCA algorithm with K-means, BIRCH, FCM, and BHC algorithms in clustering accuracy. From Table 3, we can see that our proposed OPE-HCA algorithm outperforms the baselines on seven datasets and only performs marginally worse than the best results on the rest five datasets according to accuracy metric.

Table 2 Comparison of the optimum number of clusters obtained by our proposed OPE-HCA with the true number of clusters for every datasets

Datasets	True # cluster	# Cluster obtained from OPE-HCA
Iris	3	3
Seeds	3	3
Yeast	10	10
Glass	7	7
Ecoli	8	8
HS	2	2
Balance scale	3	3
Transfusion	2	2
CMC	3	3
Page-blocks	5	4
UKM	4	4
Wine	3	3

Since different datasets have different value ranges, the parameters θ and λ take different values from their corresponding ranges

Table 3 Comparison of clustering accuracy of OPE-HCA algorithm with the best results of other four clustering algorithms

Dataset	Accuracy (%)				
	Algorithms				
	k-means	BIRCH	FCM	BHC	OPE-HCA
Iris	88.67	90.48	89.33	89.33	**94.00**
Seeds	89.05	90.95	89.52	88.57	**91.43**
Yeast	51.18	41.45	42.60	N/A	**52.34**
Glass	57.94	49.53	**63.08**	55.43	62.15
Ecoli	77.38	**80.65**	78.87	76.44	77.08
HS	73.53	73.53	73.54	72.88	**73.86**
Balance scale	66.88	66.24	71.20	69.63	**71.84**
Transfusion	76.20	76.74	76.21	N/A	**76.87**
CMC	44.81	43.58	**45.62**	N/A	43.86
Page-blocks	89.88	**89.97**	**89.97**	N/A	89.95
UKM	57.36	58.53	53.88	55.44	**63.95**
Wine	**70.22**	65.59	68.54	61.17	65.73

The values that are better than others in corresponding performance indices are in bold

The computational method of accuracy Acc is

$$\mathrm{Acc} = 1 - \mathrm{Err}$$

where Err is given by

$$\mathrm{Err} = \frac{\sum \mathrm{count}\{d_i | d_i \in Y \wedge d_j \in Y,\ d_i \in X \wedge d_j \in Y,\ i \neq j\}}{\mathrm{count}(D)} \tag{6}$$

Equation (6) means that the sum of all the number of data instances belonging to set X but clustered into cluster Y for all data instances is divided by the size of the whole dataset.

In Eq. (6), the parameter d_i and d_j denote arbitrary data instances, X and Y denote the sets of data instances which d_i and d_j belong to, respectively, and Y denotes a cluster generated by OPE-HCA algorithm. The function *count(S)* is used to compute the number of elements in set S. The accuracy of OPE-HCA algorithm over all datasets is obtained when θ and λ take the values listed in Table 4, respectively.

5.3.2 Comparison of NMI results on classical algorithms

We also take another important measurement of clustering performance NMI [38] to compare OPE-HCA with the four classical algorithms. NMI is calculated by:

$$\mathrm{NMI}(X, Y) = \frac{I(X, Y)}{\sqrt{H(X)H(Y)}} \tag{7}$$

where X and Y denote the random variables, $I(X, Y)$ denotes the mutual information between X and Y, and $H(X)$ denotes the entropy of X. $I(X, Y)$, $H(X)$, and $H(Y)$ are given by Eqs. (8), (9), and (10), respectively,

$$I(X, Y) = \sum_{h=1}^{k^{(a)}} \sum_{l=1}^{k^{(b)}} n_{h,l} \log\left(\frac{n \cdot n_{h,l}}{n_h^{(a)} n_l^{(b)}}\right) \tag{8}$$

$$H(X) = \sum_{h=1}^{k^{(a)}} n_n^{(a)} \log \frac{n_h^{(a)}}{n} \tag{9}$$

$$H(Y) = \sum_{l=1}^{k^{(b)}} n_l^{(b)} \log \frac{n_l^{(b)}}{n} \tag{10}$$

Table 4 Values of parameter θ and λ for the best results of OPE-HCA algorithm

Dataset	[θ (OPE-HCA), λ (OPE)]
Iris	[0.6, 2.2]
Seeds	[0.5, 4.3]
Yeast	[0.035, 0.36]
Glass	[1.5, 6.5]
Ecoli	[0.09, 0.4]
HS	[19.8, 30]
Balance scale	[3.3, 2.7]
Transfusion	[30, 3600]
CMC	[4.5, 9.5]
Page-blocks	[10,000, 40,000]
UKM	[0.362, 0.55]
Wine	[560, 280]

where $n_h^{(a)}$ is the number of data instances in cluster h according to their label a, and $n_l^{(b)}$ is the number of data instances in cluster l according to their label b. And $n_{h,\ l}$ denote the number of instances that are in cluster h according to a as well as in group l according to b. Table 5 illustrates the compared results of OPE-HCA algorithm with other four classical algorithms using NMI measures. As it can be seen, our proposed algorithm outperforms most of the baselines over the datasets.

5.3.3 Comparison of NMI results on "affinity propagation"

The affinity propagation (AP) method is one of the state-of-the-art methods proposed recently [39]. AP takes as input measures of similarity between pairs of data points and has much lower error than other methods. In this section, we give the comparison results of our OPE-HCA method with AP method. Table 6 gives the results of OPE-HCA and AP in NMI measure.

Although AP has higher NMI values on three dataset, Yeast, Ecoli, and UKM, OPE-HCA outperforms AP on most of the twelve datasets in NMI. Also, AP is not able to converge to the correct number of clusters on Seeds data and Balance scale data.

Table 5 Comparison of OPE-HCA algorithm with the best results of other clustering algorithms using NMI measure

Dataset	NMI (%)				
	Algorithms				
	k-means	BIRCH	FCM	BHC	OPE-HCA
Iris	84.22	84.78	84.69	82.45	**89.80**
Seeds	82.25	**85.03**	81.31	81.55	83.76
Yeast	46.87	47.46	39.80	N/A	**54.32**
Glass	58.77	62.24	53.40	59.60	**64.35**
Ecoli	67.06	**81.97**	73.88	73.52	78.89
HS	52.32	67.88	52.24	64.13	**73.16**
Balance scale	48.17	46.98	49.18	45.46	**61.94**
Transfusion	59.35	71.82	58.00	N/A	**73.31**
CMC	41.20	43.81	41.23	N/A	**48.14**
Page-blocks	64.58	77.76	64.84	N/A	**78.40**
UKM	57.77	58.89	51.91	57.12	**64.49**
Wine	65.13	62.42	64.42	64.59	**70.56**

The values that are better than others in corresponding performance indices are in bold

Table 6 Comparison of OPE-HCA algorithm with the state-of-the-art clustering method in NMI

Dataset	NMI (%) Algorithms OPE-HCA	AP
Iris	**89.80**	77.38
Seeds	**83.76**	N/A
Yeast	54.32	**97.94**
Glass	**64.35**	49.66
Ecoli	78.89	**92.40**
HS	**73.16**	1.34
Balance scale	**61.94**	N/A
Transfusion	**73.31**	71.65
CMC	**48.14**	31.62
Page-blocks	**78.40**	41.43
UKM	64.49	**93.97**
Wine	**70.56**	44.00

The values that are better than others in corresponding performance indices are in bold

6 Conclusions

Data clustering is an important research and application area. It is also one of the most promising research and application trends in data mining. The clustering methods are data-driven. Different algorithms are constrained with different data types, data structure, and data domains. Most data type is numeric, such as the data used in our experiment, and the patterns or structures hidden in dataset occur with certain probabilities. Therefore, using probability-based methods has significant advantages to cluster them. Moreover, unsupervised clustering is a mining technology without prior knowledge. Therefore, estimation approach may be a good selection for it. Based on the above analysis, OPE is an ideal technology for clustering. OPE clustering based on EDAs can randomly select data as individuals to construct initial population. The probability distribution of population is computed to estimate the distribution of dataset. The optimal individuals in population are selected by the designed fitness function. Then, the new individuals that combine with the optimal ones to form the next generation are chosen according to the patterns of the optimal individuals.

In this paper, OPE idea-based data clustering algorithm OPE-HCA is designed, which obtains better performance and results. OPE-based clustering algorithm is proven to have unique advantages. The accuracy and NMI of the maximum probability-based OPE-HCA and the non-

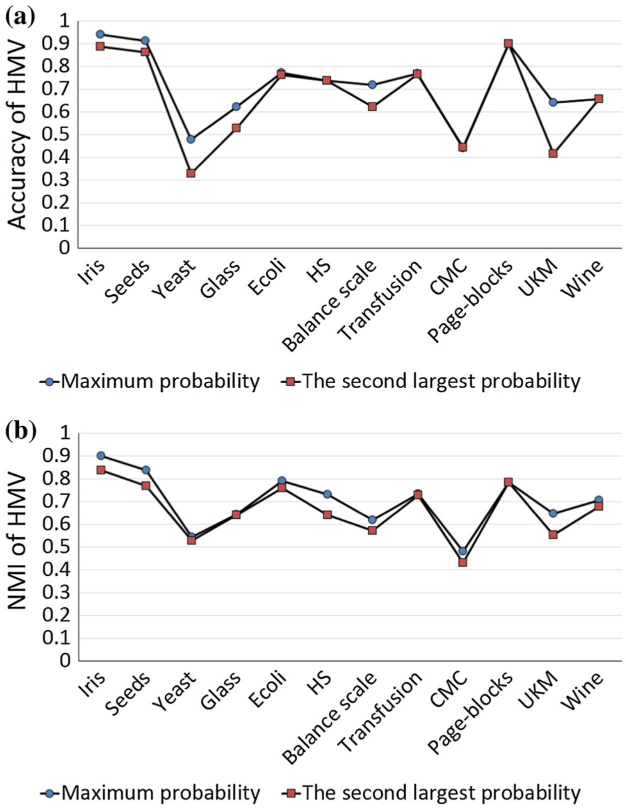


Fig. 5 Illustration of **a** accuracy and **b** NMI of the maximum probability-based OPE-HCA and the non-maximum probability-based OPE-HCA

maximum probability-based OPE-HCA are illustrated in Fig. 5a, b. From Fig. 5, we can see that maximum probability-based OPE-HCA outperforms the baselines of non-maximum probability-based OPE-HCA.

There is no doubt that the range of data is wide and diverse. In this paper, we have processed the pattern clustering problem of numeric data. Text data and other complex data clustering problems are also needed to solve, which we will study in the future.

Acknowledgments The author wishes to thank the anonymous reviewers and the JEO Assistant for their constructive comments and suggestions. The author thank the students of the laboratory Zheng Feng, Wenhua Liu, Yuhao Cai, and Tianyi Liang for participating in the experiment. This paper is supported by National Natural Science Foundation of China under Grant 61203305 and Shandong Provincial Natural Science Foundation of China under Grant ZR2012FM003.

References

1. Tan P-N, Steinbach M, Kumar V (2005) Introduction to data mining. Pearson Addison Wesley, London
2. Han J, Kamber M, Pei J (2011) Data mining: concepts and techniques, 3rd edn. Morgan Kaufmann, San Francisco
3. Aggarwal CC, Reddy CK (eds) (2013) Data clustering: algorithms and applications. CRC Press, Boca Raton, FL
4. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J (eds) Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol I, statistics, 281–297
5. Arthur D, Vassilvitskii S (2007) k-means++: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, pp 1027–1035
6. Nazeer KAA, Sebastian MP (2010) Clustering biological data using enhanced k-means algorithm. Electronic Engineering and Computing Technology. Springer, Berlin, pp 433–442
7. Kaufman L, Rousseeuw P (1990) Finding Groups in data: an introduction to cluster analysis. Wiley, New York
8. Park HS, Jun CH (2009) A simple and fast algorithm for K-medoids clustering. Expert Syst Appl 36(2):3336–3341
9. Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy c-means clustering algorithm. Comput Geosci 10(2–3):191–203
10. Pal NR, Pal K, Keller JM et al (2005) A possibilistic fuzzy c-means clustering algorithm. IEEE Trans Fuzzy Syst 13(4):517–530
11. Zhang T, Ramakrishnan R, Livny M (1997) BIRCH: a new data clustering algorithm and its applications. Data Min Knowl Discov 1(2):141–182
12. P. Smyth. Probabilistic model-based clustering of multivariate and sequential data. Proceedings of the Seventh International Workshop on AI and Statistics, San Francisco, CA: Morgan Kaufman, 1999: 299-304
13. Cadez IV, Gaffney S, Smyth P (2000) A general probabilistic framework for clustering individuals and objects. In: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2000, pp 140–149
14. Basu S, Bilenko M, Mooney RJ (2004) A probabilistic framework for semi-supervised clustering. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2004, pp 59–68
15. Heller KA, Ghahramani Z (2005) Bayesian hierarchical clustering. In: Proceedings of the 22nd international conference on machine learning, Bonn, Germany, 2005, pp 297–304
16. Papapetrou O, Siberski W, Fuhr N (2012) Decentralized probabilistic text clustering. IEEE Trans Knowl Data Eng 24(10):1848–1861
17. Boudjeloud-Assala L (2012) Visual interactive evolutionary algorithm for high dimensional outlier detection and data clustering problems. Int J Bio-Inspir Comput 4(1):6–13
18. Larrañaga P, Lozano JA (eds) (2002) Estimation of distribution algorithms: a new tool for evolutionary computation. Kluwer Academic Publishers, Boston
19. Furey E, Curran K, McKevitt P (2012) HABITS: a Bayesian filter approach to indoor tracking and location. Int J Bio-Inspir Comput 4(2):79–88
20. Fan J, Liang Y, Xu Q, Jia R, Cui Z (2011) EDA-USL: unsupervised clustering algorithm based on estimation of distribution algorithm. Int J Wirel Mob Comput 5(1):88–97
21. Fan J, Feng Z, Liu W et al (2014) Predicting yeast protein localization sites by a new clustering algorithm based on weighted feature ensemble. J Comput Theor Nanosci 11(6):1563–1568
22. Yan D, Mukai H (1993) Optimization algorithm with probabilistic estimation. J Optim Theory Appl 79(2):345–371
23. Sánchez JA, Benedí JM (1997) Consistency of stochastic context-free grammars from probabilistic estimation based on growth transformations. IEEE Trans Pattern Anal Mach Intell 19(9):1052–1055

24. Apte C, Grossman E, Pednault EP, Rosen BK, Tipu FA, White B (1999) Probabilistic estimation-based data mining for discovering insurance risks. IEEE Intell Syst 14(6):49–58
25. Ferri C, Flach PA, Hernández-Orallo J (2003) Improving the AUC of probabilistic estimation trees. In: Machine learning: ECML 2003, pp 121–132. Springer, Berlin
26. Jaulin L (2010) Probabilistic set-membership approach for robust regression. J Stat Theory Pract 4(1):155–167
27. Choi A, Woo W (2011) Multiple-criteria decision-making based on probabilistic estimation with contextual information for physiological signal monitoring. Int J Inf Technol Decis Mak 10(1):109–120
28. Han Y, Wen J, Cabric D, Villasenor JD (2011) Probabilistic estimation of the number of frequency-hopping transmitters. IEEE Trans Wirel Commun 10(10):3232–3240
29. Jiang L, Cai Z, Wang D, Zhang H (2012) Improving Tree augmented Naive Bayes for class probability estimation. Knowl Based Syst 26(2):239–245
30. Pimentel MA, Clifton DA, Clifton L, Tarassenko L (2013) Probabilistic estimation of respiratory rate using Gaussian processes. In: 2013 35th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp 2902–2905
31. Duchi J, Wainwright MJ, Jordan MI (2013) Local privacy and minimax bounds: sharp rates for probability estimation. In: Advances in neural information processing systems, the 27th annual conference on neural information processing systems (NIPS 2013), Lake Tahoe, Nevada, pp 1529–1537
32. Azad R, Davami F (2014) A robust and adaptable method for face detection based on color probabilistic estimation technique. arXiv preprint arXiv:1407.6318
33. Friedman N (2003) Pcluster: probabilistic agglomerative clustering of gene expression profiles. Technical Report 2003-80, Hebrew University
34. Segal E, Koller D (2002) Probabilistic hierarchical clustering for biological data. In: Proceedings of the sixth annual international conference on computational biology, ACM, pp 273–280
35. Fan J, Xu Q, Liang Y (2012) A novel classification learning framework based on estimation of distribution algorithms. Int J Comput Sci Math 3(4):353–366
36. Hauschild M, Pelikan M (2011) An introduction and survey of estimation of distribution algorithms. Swarm Evolut Comput 1(3):111–128
37. Newman DJ, Hettich S, Blake CL, Merz CJ (1998) UCI repository of machine learning databases. University of California, Department of Information and Computer Science, Irvine, CA. http://www.ics.uci.edu/~mlearn/MLRepository.html
38. Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York
39. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315:972–976