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

Leo Vattoly
Department of Computer Science
Central University of Kerala
Kasaragod, Kerala, India
leo.vattoly@ieee.org

V Kumar
Assistant Professor, Department of Computer Science
Central University of Kerala
Kasaragod, Kerala, India
kumar.sanvi@gmail.com

***Abstract*—Hierarchical clustering algorithm (HCA) is on of the best method using in data mining. It proposes two different types of method to discover the hidden patterns in a given data set. Agglomerate and Divisive methods are the methods proposed by HCA. However, clustering algorithms are facing two different drawbacks: the use of distance based approach and the difficulty while integrating two local clusters. We can easily optimize any clustering algorithm by reducing the drawbacks. In this paper we are introducing a Genetics Algorithm (GA) based approach to integrate the local clusters,and optimize HCA. Introducing "The survival of the fittest" technique, to optimize the Agglomerate HCA algorithm. Here we are including GA using probabilistic method, exploring the data set and finding the fittest local cluster using Estimation of Distribution Algorithm (EDA), and generating the next generation using Optimal Probabilistic Estimation (OPE). OPE-HCA showing better performance, also increased ability to searching and discovering the hidden patterns.**

***Index Terms*—Clustering, Hierarchical clustering algorithm, Data mining, Probabilistic estimation**

I. Introduction

Machine learning has made a great leap in all the field, it conquering all the R&D areas and enhancing the capability of the machine to an extended level. It took the machine to the next dimension, act like a human being or replace the human effort. Nowadays machines can predict or suggest according to the data trained by the practitioner, this practice is called supervised learning. While training the machine the practitioner facing some challenges, such as making Knowledge Discovery in Databases (KDD) from a set of raw data. It can be avoided by the unsupervised machine learning mechanism. In unsupervised learning technique, the machine learns from hidden patterns, which obtained from a set of unlabelled datasets. The term cluster has a great role in unsupervised learning, especially to catch the hidden patterns from the datasets [1] [2].

HCA is one of the most popular plan of action in connectivity based clustering method [3], [4]. The agglomerative hierarchical clustering algorithm has some defects,for example its very difficult to be reevaluated once the clustering has started, because once a node integrated with a group can't move dynamically. Which leads the entire clustering accuracy to be reduced.

In this paper, we study how to overcome the flaws discussed above under the condition of keeping the assets of agglomerative clustering. We develop a HCA based on a Optimal Probabilistic Estimation (OPE), also called OPE-HCA, which combines the dendrogram and the optimal probability of survivors. The core idea behind our proposed OPE-HCA algorithm is to select the individual with higher probability from the probability distribution as a survivor to the next generation. This algorithm can produce a more better and different cluster distribution at different levels of trees.

The rest of the paper is organized as follows. The next section is briefly explain the clustering algorithms based on hierarchical idea and probabilistic estimation in recent years. Section 3 describing the Optimal Probabilistic estimation (OPE) and Estimation of distribution algorithm (EDAs), because we apply the idea of EDA for the development of OPE. We formulating the algorithms of OPE and OPE-HCA in section 4. And Section 5 explaining the experimental settings, dataset and results with twelve real datasets. Finally section 6 concludes the paper.

II. Related Work

Agglomerative and divisive methods are the two types of hierarchical clustering methods. It creates dendrogram using tree data structure. Bottom - up(agglomerative) and top-down(divisive) are the methods to create the tree data structure. The bottom-up process is an natural and inductive process. when compared with top-down. So, agglomerative clustering method is widely used [6], [7]. Although, probability based methods mainly includes, model based strategy [8] - [12]

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.

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

Genetic algorithm will take place by the estimation of distribution algorithms (EDAs), using its key idea estimation of probabilistic model [13]. Instead of random mutation and cross over EDA prefer probability as a key to determine the next generation. Simply combining the optimal individual which is having high probability. In EDAs, each position of the next generation G can be computed based upon the previous 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:

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

Estimation and sampling are the two basic ideas of EDA. Estimation is to build the probability distribution from the dataset and, sampling is to choose the best local solution according to the estimated probability distribution model in order to produce the next generation. Mutation and crossover could be full filled using these two operations, also it avoid the random choosing and create a semi-supervised learning model. The OPE focuses on estimating the best solution sampling from the ones formed by merging the sub solutions having maximum probabilities among all sub solutions. OPE cluster the local points which having higher probability, accordingly it creates the clusters.

IV. Hierarchical clustering algorithm based on optimal probabilistic estimation

In this section we discuss about the proposed algorithm OPE-HCA. OPE-HCA is consist of two major procedures. One is to use the distance based calculation, to produce a cluster with similar data instances in the bottom layer. And the second is to deal with clusters in each hierarchy by OPE. The following sub sections A and B are discussing about the OPE and OPE-HCA algorithm in detail.

A. Optimal probabilistic estimation (OPE)

Non-deterministic problems can be skillfully deal with OPE by maximum probabilistic modelling.The key idea of OPE is to determine probability and select the high probability items to the next generation from probabilistic distribution. Let's consider that dataset D consist 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 $p_1 = P(sub_i)$ 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 $P_i = max_i \leq j \leq n\{P_j\}$ which corresponds to the occurrence prob-ability 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 $(k \neq j)$

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

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*.

B. OPE-based hierarchical clustering algorithm

The OPE-HCA is designed based on OPE procedure.

OPE ($\mathbf{D}^{(1)}$, λ) 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 $\mathbf{D}$ 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 $Pop^{(0)}$, $p_1 = P(C_1), p_2 = P(C_2), \ldots, p_t = P(C_t)$, 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 $\mathbf{D}^{(1)}$;

Step 4 Set $k = 1$;

Step 5 while (true) {

Select individuals in turn from $\mathbf{D}^{(k)}$, that is, firstly select the ones with higher probability;

Given criterion Θ, 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,

$$P(C_{ij} \mid \Theta) \geq P(C_u \mid \Theta, u \neq i \; and \; u \neq j),$$

and C_i and C_j satisfy

$$|\text{mean}(C_i) - \text{mean}(C_j)| \leq \lambda,$$

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 $\mathbf{D}^{(k)}$ which are not traversed yet) continue;

Sort over the individuals in $Pop^{(k)}$ descending by their probability to obtain sorted population denoted by $\mathbf{D}^{(k+1)}$, which is a new clustering view with less clusters;

$$\mathbf{S}_{(k+1)} \leftarrow \mathbf{D}^{(k+1)};$$

$k = k + 1$;

} until (only one element remains in $\mathbf{D}^{(k)}$, or the termination conditions are satisfied.)

Figure 1 OPE algorithm

Definition 1 Minimum merging distance between clusters (MinMD). For arbitrary two clusters C_i and C_j, their mean values are C_i(mean) and C_j(mean), 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.

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

The MinMD defined in Definition 1 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.

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 θ.

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 $C_1, C_2, \ldots, C_k$ to obtain the source clustering view S_1;

(3) Assign the value of MinMD to λ;

(4) Call algorithm OPE (S_1, λ);

Output:

The clustering views on the condition of different parameters θ and λ.

Figure 2 OPE - HCA algorithm

In OPE-HCA algorithm it create the first level cluster using the euclidean distance formula. This algorithm merge two local clusters, if the given value θ is less than the distance between two local clusters. Once a generation S_1 is created then the algorithm OPE will be called with parameters S_1 and λ (MinMD). OPE algorithm will recursively call either the items in the dataset is 1 or probability distribution contains the same values, when one of the condition became true OPE algorithm return the clustered dataset.

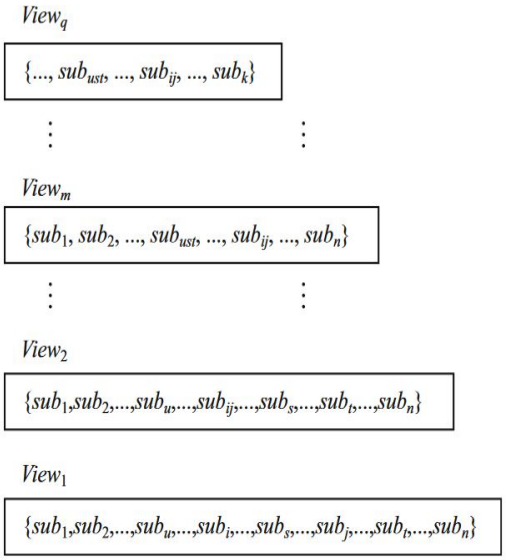

Figure 3 General solution structure after executing OPE-HCA algorithm

V. Experimental Setting and Results

In this section we are applying OPE-HCA algorithm and other classical clustering algorithms with twelve different public data sets. Section A explains about the experimental settings, including data sets, other classical algorithms to compare with OPE-HCA. Section B deals with the sensitivity analysis of the parameters used in OPE-HCA algorithm. Detailed results and comparisons are presented in Section C.

A. Experimental Settings

Dataset:

We are testing twelve datasets with proposed OPE-HCA. Datasets are taken from UCI machine learning repository [14]. The important details about the datasets are mentioned in the Table 1.

Compared algorithms:

The well-known clusters K - means [15], fuzzy c-means (FCM), BIRCH [16] and Bayesian hierarchical clustering (BHC) [17] are comparing with OPE-HCA in order to find the accuracy. The idea behind choosing these algorithms, K-means using the distance based algorithm in our proposed algorithm initially creating a view (cluster) using the Euclidean distance method. In other side FCM is implemented using the Euclidean and fuzzy method to cluster the dataset. Also, BHC is a probabilistic clustering method based on Bayesian model. OPE-HCA following the probabilistic method to evolve the next generation.

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

Table 1 Information of the tested benchmark datasets

B. Sensitivity analysis

1. Clustering accuracy

The above metric will discuss in section D.

C. Sensitivity analysis

The proposed algorithm OPE-HCA performed according to the following parameters θ and λ. The accuracy is used to evaluate the performance of our algorithm and other clustering algorithms.

The first step in OPE-HCA is to develop a cluster using euclidean distance method. Two local points will merge when the distance between the local points are less than the given threshold value (θ). From view 1 next generation will be created using the MinMD

value (λ). Equation 3 is defining the condition for merging two or more local clusters using the MinMD parameter value. The results show that OPE-HCA algorithm can be convergent to a definite view. With same value of λ, various values of θ have little influence on the generated views. That is because θ only using to create first level of view. Anyway, θ is more influencing while the generation of source view, which may influence the whole output of the algorithm. Table 2 listed the θ and λ values used in different datasets.

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]

Table 2 Values of parameter θ and λ for the best results of OPE-HCA algorithm

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

Table 3 Comparison of the optimum number of clusters obtained by our proposed OPE-HCA with the true number of clusters for every datasets

D. Comparison of OPE-HCA and other algorithms

In this section we are comparing OPE-HCA algorithm with different classical clustering algorithms. Table 3 showing the output of our proposed clustering algorithm, with different datasets. Then we are comparing this algorithm with other classical clustering algorithms mentioned in Section A.

Table 4 shows the comparison of OPE-HCA algorithm with K-means, BIRCH, FCM, and BHC algorithms in clustering accuracy. 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.

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

Table 4 Comparison of clustering accuracy of OPE-HCA algorithm with the best results of other four clustering algorithms

The computational method of accuracy Acc is

Acc = 1 - Err

where Err is given by,

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

The above equation 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.

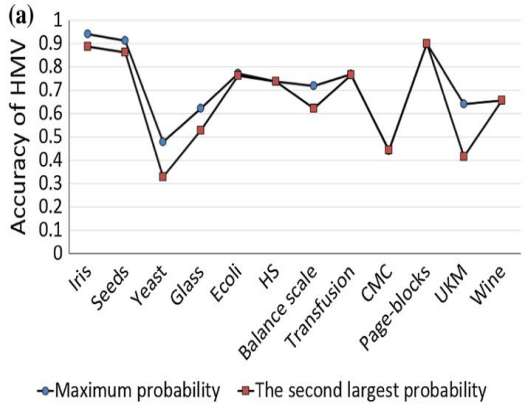

Figure 4 Illustration of accuracy of the maximum probability-based OPE-HCA and the non-maximum probability-based OPE-HCA

VI. Conclusion

Clustering is an unsupervised learning methodology, we don't have prior knowledge about a set of raw data. Then, we obtain a hidden pattern or information from the dataset. Most of the time dataset consist of numeric data, which would be more easy to cluster and obtain the hidden information. We have several algorithms are there to cluster raw data, with out any prior knowledge. Using probabilistic method is more useful, because the hidden patterns or structures in a dataset occur certain probabilities. Although, we don't have any prior knowledge about the raw data, so estimation approach is also useful to select the individual

to cluster the data points. In GA we are using random selection but using the EDAs we can obtain a new generation, using the probability distribution of the population. The probability distribution of population is computed to estimate the distribution of dataset. Then, the new individual will be selected based upon the probability distribution.

In this paper we have proposed an optimized clustering method of hierarchical clustering algorithm called OPE-HCA. This algorithm obtains better performance and results. OPE based clustering have unique advantages. The accuracy of the maximum probability-based OPE-HCA and non-maximum probability based OPE-HCA is illustrated in Figure 4, we can see that maximum probability -based OPE-HCA well performing compared with non-maximum probability-based OPE-HCA.

In future we would like to study and luster not only numeric data, but also text and other complex datasets.

ACKNOWLEDGMENT

I would like to thank V Kumar, Assistant professor, Department of Computer Science, Kerala Central University, Kasaragod, India, for supporting this minor project.

REFERENCES

[1] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to data mining. Pearson Education India, 2016.

[2] Mucherino, Antonio, Petraq Papajorgji, and Panos M. Pardalos. Data mining in agriculture. Vol. 34. Springer Science & Business Media, 2009.

[3] Tan P-N, Steinbach M, Kumar V (2005) Introduction to datamining. Pearson Addison Wesley, London

[4] Han J, Kamber M, Pei J (2011) Data mining: concepts andtechniques, 3rd edn. Morgan Kaufmann, San Francisco

[5] Han J, Kamber M, Pei J (2011) Data mining: concepts andtechniques, 3rd edn. Morgan Kaufmann, San Francisco

[6] Papapetrou O, Siberski W, Fuhr N (2012) Decentralized proba-bilistic text clustering. IEEE Trans Knowl Data Eng24(10):1848–1861

[7] P. Smyth. Probabilistic model-based clustering of multivariateand sequential data. Proceedings of the Seventh InternationalWorkshop on AI and Statistics, San Francisco, CA: MorganKaufman, 1999: 299-304.

[8] Cadez IV, Gaffney S, Smyth P (2000) A general probabilisticframework for clustering individuals and objects. In: Proceedingsof the sixth ACM SIGKDD international conference on knowl-edge discovery and data mining, ACM, 2000, pp 140–149.

[9] Basu S, Bilenko M, Mooney RJ (2004) A probabilistic framework for semi-supervised clustering. In: Proceedings of the tenthACM SIGKDD international conference on knowledge discoveryand data mining, ACM, 2004, pp 59–68.

[10] Heller KA, Ghahramani Z (2005) Bayesian hierarchical cluster-ing. In: Proceedings of the 22nd international conference onmachine learning, Bonn, Germany, 2005, pp 297–304.

[11] Zhang T, Ramakrishnan R, Livny M (1997) BIRCH: a new dataclustering algorithm and its applications. Data Min Knowl Discov1(2):141–182.

[12] P. Smyth. Probabilistic model-based clustering of multivariateand sequential data. Proceedings of the Seventh InternationalWorkshop on AI and Statistics, San Francisco, CA: MorganKaufman, 1999: 299-304.

[13] Hauschild M, Pelikan M (2011) An introduction and survey ofestimation of distribution algorithms. Swarm Evolut Comput1(3):111–128

[14] Newman DJ, Hettich S, Blake CL, Merz CJ (1998) UCI reposi-tory of machine learning databases. University of California,Department of Information and Computer Science, Irvine, CA.http://www.ics.uci.edu/*mlearn/MLRepository.html

[15] Nazeer KAA, Sebastian MP (2010) Clustering biological datausing enhanced k-means algorithm. Electronic Engineering andComputing Technology. Springer, Berlin, pp 433–442

[16] Zhang T, Ramakrishnan R, Livny M (1997) BIRCH: a new dataclustering algorithm and its applications. Data Min Knowl Discov1(2):141–182.

[17] Heller KA, Ghahramani Z (2005) Bayesian hierarchical cluster-ing. In: Proceedings of the 22nd international conference onmachine learning, Bonn, Germany, 2005, pp 297–304