Expert Systems With Applications 67 (2017) 12–18

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An improved overlapping k-means clustering method for medical applications

Sina Khanmohammadi[a,*], Naiier Adibeig[b], Samaneh Shanehbandy[b]

[a] Department of Systems Science and Industrial Engineering, The State University of New York at Binghamton, 4400 Vestal Parkway East, Binghamton, NY-13902, USA

[b] Department of Computer Engineering, Islamic Azad University at Shabestar, Shabestar-Dizajkhalil Road, Shabestar, East Azerbaijan-5381637181, Iran

ARTICLE INFO

Article history:
Received 25 May 2016
Revised 25 August 2016
Accepted 14 September 2016
Available online 17 September 2016

Keywords:
Overlapping clustering
Overlapping k-means
K-harmonic means
FBCubed
Data Mining
Medical informatics

ABSTRACT

Data clustering has been proven to be an effective method for discovering structure in medical datasets. The majority of clustering algorithms produce exclusive clusters meaning that each sample can belong to one cluster only. However, most real-world medical datasets have inherently overlapping information, which could be best explained by overlapping clustering methods that allow one sample belong to more than one cluster. One of the simplest and most efficient overlapping clustering methods is known as overlapping k-means (OKM), which is an extension of the traditional k-means algorithm. Being an extension of the k-means algorithm, the OKM method also suffers from sensitivity to the initial cluster centroids. In this paper, we propose a hybrid method that combines k-harmonic means and overlapping k-means algorithms (KHM-OKM) to overcome this limitation. The main idea behind KHM-OKM method is to use the output of KHM method to initialize the cluster centers of OKM method. We have tested the proposed method using FBCubed metric, which has been shown to be the most effective measure to evaluate overlapping clustering algorithms regarding homogeneity, completeness, rag bag, and cluster size-quantity tradeoff. According to results from ten publicly available medical datasets, the KHM-OKM algorithm outperforms the original OKM algorithm and can be used as an efficient method for clustering medical datasets.

1. Introduction

One of the most primitive human actions is grouping objects into different catagroies based on their similarity. In the data mining domain, this task is known as clustering and is one of the most important and useful concepts for analyzing large quantities of data. More formally, clustering can be defined as finding heterogeneous groups of data using some dissimilarity criterion. Clustering has been employed in many applications from wireless sensor networks (Abbasi & Younis, 2007; Liu, 2012) to medicine (Andreopoulos, An, Wang, & Schroeder, 2009; Kalyani, 2012; Xu, Wunsch et al., 2010). Some of the applications of clustering algorithms in medical domain include medical diagnosis (NITHYA, Duraiswamy, & Gomathy, 2013; Paul & Hoque, 2010; Vogt & Nagel, 1992), biological data analysis (Li & Zhu, 2013; Nugent & Meila, 2010; Wiwie, Baumbach, & Röttger, 2015; Yasodha & Mohanraj, 2011), medical image segmentation (Ma, Tavares, Renato, & Jorge, 2009; Naik & Shah, 2014), patient database management (Narmadha, alias Balamurugan, Sundar, & Priya, 2016; da Veiga, 1996), and hospital resource management (Dilts, Khamalah, & Plotkin, 1995).

Most clustering algorithms identify mutually exclusive groups of data, where objects are not allowed to belong to more than one cluster (exclusive clustering). However, the inherited pattern of most real-world datasets cannot be fully explained using such hard constraint. This is especially true in a medical domain where various diseases are characterized by complex overlapping symptoms such as “poor muscle tone” in Hypothyroidism and Pompe diseases (Howell et al., 2006; Kishnani et al., 2006); or more than one disorders co-occur (comorbidity) such as diabetes and hypertension (Bretzel, 2007; Long & Dagogo-Jack, 2011). Hence, overlapping clustering methods have become exceedingly popular since they enable identifying clusters where one object can belong to more than one cluster. The overlapping clustering should not be confused with soft clustering methods where we allow objects to belong to one cluster with some degree of membership [0,1] rather than a crisp membership value of 0,1 (Jain, Murty, & Flynn, 1999). The overlapping clustering methods can have either soft or hard clustering membership values.

* Corresponding author.
E-mail addresses: skhanmo1@binghamton.edu (S. Khanmohammadi), adibeig.n@gmail.com (N. Adibeig), samane.shanebandy@gmail.com (S. Shanehbandy).

http://dx.doi.org/10.1016/j.eswa.2016.09.025
0957-4174/

Several overlapping clustering algorithms have been proposed in literature (NCir, Cleuziou, & Essoussi, 2015). However, most of them suffer from high computational complexity. One of the methods that provide satisfactory results using less computational power is an extension of the k-means algorithm called overlapping k-means method (OKM) (Cleuziou, 2007; 2008). However, similar to k-means algorithm this approach is sensitive to the initialization meaning that the results are heavily dependent on the randomly chosen initial cluster centroids. For the k-means algorithm, the problem of sensitivity to initial points is solved using a method called k-harmonic means algorithm, which integrates the harmonic average to the objective function of the k-means algorithm (Zhang, Hsu, & Dayal, 1999). Recognizing the limitations of the OKM method and the opportunity provided by KHM method, in this paper we propose a hybrid KHM-OKM approach to improve the overall performance of the OKM method by reducing its sensitivity to initialization of cluster centroids. The main idea is to use the KHM method to identify the initial points for the OKM approach and obtain the overlapped clusters based on the identified points. We tested the performance of proposed KHM-OKM method using ten public medical datasets from UCI (2016), and the results show that the proposed hybrid method improves the overall performance of the original OKM method regarding FBCubed metric. Furthermore, we analyzed the value of OKM's objective function at a certain random point, and the results suggest that initializing OKM method using KHM method improves the speed of convergence to the optimal objective function value in the OKM method.

In summary, in this study we investigate the hypothesis that using a more systematic initialization approach instead of the random initialization in OKM overlapping clustering algorithm will improve its overall performance. In this regards, the specific research objectives include:

- Propose an alternative approach to the random initialization of the cluster centers in the OKM algorithm.
- Analyze and compare the properties of the OKM algorithm (such as convergence speed) with random and nonrandom initialization.
- Understand the effects of dataset size on the performance of the OKM algorithm.

The remainder of the paper is organized as follows. Section 2 provides a background of clustering algorithms and some of the recent advances in overlapping clustering methods. In Section 3, a full description of the proposed hybrid overlapping clustering algorithm is presented. Section 4 includes the details of the experimental results, and the paper is concluded in Section 5.

2. Background

In this section, a brief overview of the overlapping clustering algorithms is provided. First, a brief discussion of different classes of clustering algorithms is provided followed by a more in-depth background of overlapping clustering methods.

2.1. Overview of clustering methods

Clustering algorithms in literature have been categorized according to different criteria such as the type of input data, proximity measure (similarity measure), nature of generated cluster, membership function style, and clustering strategy (Andreopoulos et al., 2009; Berkhin, 2006; Grira, Crucianu, & Boujemaa, 2004; Jain et al., 1999; NCir et al., 2015; Soni & Ganatra, 2012; Wong, 2015; Xu, Wunsch et al., 2005; 2010). In terms of input data, the clustering methods are divided into three groups of numerical, categorical, and mixed clustering. With regards to proximity measure, various similarity measures have been defined for each of the input data types. For example, the k-means clustering algorithm uses the Euclidean distance also known as the L2 norm for calculating the similarity between various data points (a complete review of similarity measures is available in Xu et al. (2010)). Regarding generated clusters, the clustering algorithms fall into two categories of Exclusive (Non-overlapping) and overlapping methods. As mentioned earlier, in exclusive clustering the data points can only belong to one of the identified disjoint clusters whereas in overlapping clustering method the data points can belong to more than one clusters. The type of generated clusters should not be confused with the membership function style, where we have hard (crisp) and soft (fuzzy) clustering. In hard clustering methods, one data point either belongs or doesn't belong to a cluster (binary membership function), whereas in soft clustering methods one data point can belong to a cluster with some degree of membership between zero and one.

Considering the strategy used to cluster different data points, the clustering methods are categorized into partitioning, hierarchical, density-based, model-based, graph-based, and grid-based approaches. In partitioning clustering methods, the dataset is decomposed into a set of joint (in overlapping methods) or disjoint (in exclusive methods) clusters by optimizing some similarity criterion. Instead of checking all the possible decompositions and selecting the best one (which is computationally expensive), an iterative approach is usually used to identify the best clustering structure according to the selected evaluation criterion. In hierarchical clustering as the name suggest, a hierarchy of clusters is identified using either agglomerative (bottom-up) or divisive (top-down) approaches. In agglomerative approach, we consider each data point as one cluster and iterative merge them according to some criterion, whereas in the divisive approach we consider the whole dataset as one cluster and iteratively split them into multiple clusters according to some evaluation criterion. Density-based clustering methods rely on the local density of clusters and aim at identifying neighborhoods which are separated by low-density subspaces. Model-based approaches assume that the dataset can be modeled using some mathematical, physical, or more commonly probabilistic model. In probabilistic clustering algorithms, the goal is to explain the distribution of the data points using a parametric distribution function. However, not all model-based methods are probabilistic, gravitational clustering methods assume each data point as a particle and use a physical model (such as Newton's laws) to simulate the dynamics of the dataset and cluster the data points accordingly (Sylos Labini, 2008). More recently, nature inspired herd clustering algorithm are introduced, where the goal is to mimic the herd behavior seen in real-world to cluster datasets into multiple groups (Wong, Peng, Li, & Chan, 2014). Grid-based approaches map each data points to a cell of a grid structure and use the obtained grid structure to cluster data points into various clusters. Similarly, graph-based approaches map data points into a network structure and identify the clustering based on the relationship of different objects without knowing the actual representation of the objects (Becker, 2005). One specific type of graph-based clustering known as correlation clustering aims at identifying clusters without specifying the number of clusters in advance. Table 1 provides an overview of current clustering methods proposed in literate.

2.2. Overlapping clustering methods

As mentioned previously, overlapping clustering methods allow data points to belong to more than one cluster (Fig. 1). Among overlapping clustering algorithm, partitioning methods are more popular mainly because of their simplicity and effectiveness on large datasets. One of the most common partitional

Table 1
Overview of different clustering algorithms according to the clustering strategy.

Clustering Strategy:	Clustering Method:
Partitioning	ALS Family, BCM, Bisecting k-Means, CLARANS, CLARA, CLICK, COOLCAT, ECM, Farthest First Travel, Forgy, Fuzzy c-Means (FCM), Fuzzy k-Modes, GA-Based, k-Means, k-Medians, k-Medoids(PAM), k-Modes, k-Prototypes, Kernel k-Means, KOKM, KOKMΦ, LDAOC, OKM, OKMED, Paramterized R-OKM, PCL, Possibilistic c-Means (PCM), PSO-Based, Squeezer, x-Means, WOKM
Hierarchical	Agglomoreative Family, AGNES, BIRCH, CHEMELEON, CURE, DIANA, LIMBO, MONA, Power Graphs, Pyramide, ROCK, Spectral
Density-Based	CACTUS, CLOPE, DBCLASD, DBSCAN Family, DENCLUE, MULIC, OPTICS, PROCLUS, SNN, STIRR, WaveCluster
Model-Based	ART Family, AutoClass, BILCOM, COBWEB, EM, ENCLUS, Fuzzy ART, IOMM, LVQ-Based, MMM, MOC, ORCLUS, OSOM, SNOB, SOFM, SOM Neural Nets, SVM Family
Graph-Based	DClusteR, DCS, ICSD, ISC, MCL, MCODE, OClusteR, OCC, RNSC, SPC, Star, STC
Grid-Based	BANG, CLIQUE, MAFIA, OPTIGRID, STING

*It should be noted that the clustering categories provided here are not mutually exclusive.

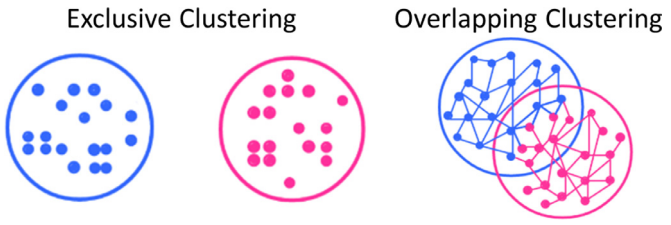

Fig. 1. Graphical representation of overlapping clustering.

overlapping clustering algorithms is the overlapping k-means algorithm (OKM), which is an extension of the k-means algorithm to generate overlapping clusters (Cleuziou, 2007; 2008). Some of the recent extensions of the OKM method include overlapping k-medoid (OKMED), weighted overlapping k-means (WOKM), kernel overlapping k-means (KOKM/KOKMΦ), and parametrized OKM methods. The OKMED method aggregates the data around the cluster representatives (medoids) and is basically an extension of the k-medoids method to identify overlapping clusters (Cleuziou, 2010). The WOKM method is a generalization of the weighted k-means method and assigns a weight to take into account the importance of each data point for clustering (Cleuziou, 2010). The kernel overlapping k-means method takes a different approach and aims at identifying nonlinear separation of datasets by using a kernel function. The first method (KOKM) achieves this by kernelization of the Euclidean distance employed in the traditional k-means algorithm (BenNCir, Essoussi, & Bertrand, 2010) whereas the second approach (KOKMΦ) performs all the clustering steps in the high dimensional space (BenNCir & Essoussi, 2012). Finally, the parametrized OKM method adds flexibility to the OKM method by allowing the regulation of the amount of overlap using a parameter (α). For more detailed review of overlapping clustering algorithms including the extensions of OKM algorithm readers can refer to Aroche-Villarruel, Carrasco-Ochoa, Martínez-Trinidad, Olvera-López, and Pérez-Suárez (2014); NCir et al. (2015). One of the limitations of the discussed methods is the fact that they are essentially extensions of the k-means method, which is very sensitive to the initialization process. The sensitivity of the traditional k-means method to initial points is already resolved by the k-harmonic means algorithm (KHM) (Zhang, 2000; Zhang et al., 1999). Hence, here we propose a hybrid KHM-OKM algorithm to address the sensitivity of OKM method to initial points.

Table 2
Mathematical notations.

Symbol	Definition
X	Data
n	Total number of samples
h, i	Sample indexes
k	Total number of clusters
j	Cluster index
Q	Objective function
π	Set of clusters
Z	Centroid of clusters
ϕ	Image of x (center of clusters that x belongs to)
F	Total number of features
v	Feature index (dimension of x)
δi	Total number of clusters that x_i belongs to ($\lvert\pi(x_i)\rvert$)
m	The membership function in KHM algorithm
w	Weight of each data point in KHM algorithm
C	Class labels
$D(x)$	Set of datapoints that share at least one cluster with x
$H(x)$	Set of datapoints that share at least one class with x

3. Methodology

The full description of each component of the proposed hybrid HKM-OKM algorithm is provided in the following sections. A complete list of notations used in the methodology section is provided in Table 2.

3.1. OKM Method

The K-means algorithm aims at clustering $X = \{\vec{x_1}, \ldots \vec{x_n}\}$ into k clusters by minimizing the following objective function (Aroche-Villarruel et al., 2014).

$$Q(\pi) = \sum_{j=1}^{k} \sum_{\vec{x_i} \in \pi_j} ||\vec{x_i} - \vec{z_j}||^2, \tag{1}$$

where, $\vec{x_i}$ is a v-dimensional vector of observations, $\pi = \{\pi_1, \ldots, \pi_k\}$ is the set of k clusters ($\pi_i \cap \pi_j = \emptyset$), and $Z = \{\vec{z_1}, \ldots, \vec{z_k}\}$ is the set of cluster centroids. The OKM approach relaxes the objective function used in k-means to allow overlapping clusters. More specifically, removing the $\pi_i \cap \pi_j = \emptyset$ constraint enables the algorithm to identify overlapping clusters (Eq. 2).

$$Q'(\pi) = \sum_{i=1}^{n} ||\vec{x_i} - \phi(\vec{x_i})||^2 \tag{2}$$

The $\phi(\vec{x_i})$ is the representation of $\vec{x_i}$ (barycenter of clusters that $\vec{x_i}$ belongs to), and $\phi(\vec{x_i})$ is calculated as Aroche-Villarruel et al. (2014):

$$\phi\left(\vec{x_i}\right) = \frac{\sum_{\vec{z_j} \in \pi\left(\vec{x_i}\right)} \vec{z_j}}{\left|\pi\left(\vec{x_i}\right)\right|}, \tag{3}$$

Here, the centroid $\vec{z_j} \in \pi(\vec{x_i})$, where $\pi(\vec{x_i})$ is the list of all clusters that $\vec{x_i}$ belongs to is updated using the following equation (Aroche-Villarruel et al., 2014):

$$\vec{z_j} = \frac{1}{\sum_{\vec{x_i} \in \pi_j} \frac{1}{\delta_i^2}} \sum_{\vec{x_i} \in \pi_j} \frac{1}{\delta_i^2} \cdot \left(\delta_i \times \vec{x_i} - \sum_{\vec{z_j} \in \pi(\vec{x_i})/\vec{z_i}} \vec{z_j} \right), \tag{4}$$

where δ_i corresponds to the total number of clusters that $\vec{x_i}$ belongs to (in this case $\delta_i = |\pi(\vec{x_i})|$).

3.2. KHM Method

The OKM method is sensitive to outliers, which is inherited from the k-means algorithm that is based on. More specifically, when datasets include outliers they might be selected as the initial cluster centroids in the k-means algorithm, and this could cause the k-means method not to converge to the best cluster formation (Tan, Steinbach, Kumar et al., 2006). The KHM algorithm, proposed by Zhang et al. (1999) and Zhang (2000), resolves this problem by minimizing the harmonic average of all data points from the cluster centers. The harmonic average gives a weight to each data point based on its proximity to each center. This weight is considered as the importance of each point in identifying the clusters in the dataset. In other words, the KHM algorithm introduces a bias (using the weight) to shift the cluster centers to the data points that are more important according to some criterion. Similar to k-means algorithm, the KHM method can be formulated as an optimization problem where the objective is to minimize:

$$Q''(\pi) = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{||\vec{x_i} - \vec{z_j}||^p}}, \tag{5}$$

where p is a free parameter (typically $p \geq 2$) and the expression inside the summation $\left(\frac{k}{\sum_{j=1}^{k} \frac{1}{||\vec{x_i} - \vec{z_j}||^p}} \right)$ is the harmonic mean. In order to calculate the harmonic mean, we first need to calculate the centroid of the clusters ($\vec{z_j}$) using the following equation:

$$\vec{z_j} = \frac{\sum_{i=1}^{n} m(\vec{z_j}|\vec{x_i}) w(\vec{x_i}) \vec{x_i}}{\sum_{i=1}^{n} m(\vec{z_j}|\vec{x_i}) w(\vec{x_i})} \tag{6}$$

Here, the $m(\vec{z_j}|\vec{x_i})$ is the membership of data point $\vec{x_i}$ to centroid of the cluster j calculated by Eq. (7), and $w(\vec{x_i})$ is the weight associated with data point $\vec{x_i}$ calculated by Eq. (8). After the objective function converged to some optimal value, each data point is assigned to the cluster that yields maximum membership function value.

$$m(\vec{z_j}|\vec{x_i}) = \frac{||\vec{x_i} - \vec{z_j}||^{-p-2}}{\sum_{j=1}^{k} ||\vec{x_i} - \vec{z_j}||^{-p-2}} \tag{7}$$

$$w(\vec{x_i}) = \frac{\sum_{j=1}^{k} ||\vec{x_i} - \vec{z_j}||^{-p-2}}{(\sum_{j=1}^{k} ||\vec{x_i} - \vec{z_j}||^{-p})^2} \tag{8}$$

3.3. Hybrid KHM-OKM method

In this paper, the KHM-OKM is proposed to improve the performance of OKM method and reduce its sensitivity to initialization. The proposed OKM-KHM method includes four main steps as:

1. Find the center of clusters using the KHM method (Eq. 6).
2. Initialize the OKM method based on the results of previous step.
3. Use the OKM method to obtain the final cluster centers (Eq. 4).
4. Identify clusters that each data point belongs to using Eq. (9).

The final clusters for each data point are calculated using the Eq. (9), where $\vec{x_i}$ and $\vec{z_j}$ represent the current data point and cluster center, respectfully (Fig. 2).

$$m(\vec{z_j}|\vec{x_i}) = \frac{||\vec{x_i} - \vec{z_j}||^2}{\sum_{j=1}^{k} ||\vec{x_i} - \vec{z_j}||^2} \tag{9}$$

4. Results and discussion

In this section the detailed experimental results on ten publicly available medical datastest are provided. First, we provide the description of the sample datasets which were retrieved from UCI repository (UCI, 2016). Next, we discuss the details of the employed evaluation criteria (FBCubed metric) followed by the detailed results of applying the proposed KHM-OKM algorithm to sample datasets.

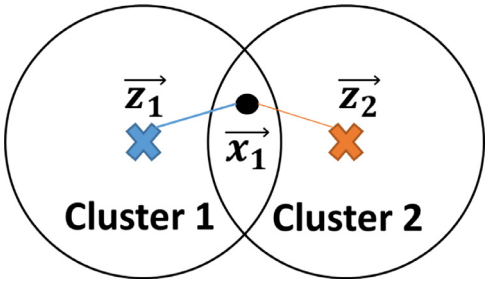

Fig. 2. Distance of datapoint $\vec{x_1}$ from cluster centers $\vec{z_1}$ and $\vec{z_2}$.

4.1. Dataset

A total of ten medical datasets (summarized in Table 3) from UCI repository was used in this study (UCI, 2016). These datasets include both continuous and categorical features. Considering the total number of features (F) in each dataset, we have categorized them into five groups of very small, small, medium, large, and very large.

4.2. Evaluation metrics

Evaluating clustering algorithms is more challenging than the classification algorithms considering the unsupervised nature of the clustering methods (i.e. the absence of ground truth information). Hence, there is a significant amount of research devoted to validation metrics for the clustering algorithms (Amigó, Gonzalo, Artiles, & Verdejo, 2009; Halkidi, Batistakis, & Vazirgiannis, 2001). Among these metrics, the recently proposed FBCubed metric has been proven to provide more accurate information about the quality of identified clusters when we have access to the true class labels (Amigó et al., 2009; Aroche-Villarruel et al., 2014). Hence, in this study, we have employed the FBCubed metric to compare our proposed method to the original OKM method. The FBCubed is an extrinsic metric that evaluates the degree of conformity between clusters and their class labels based on formal mathematical constraints (Amigó et al., 2009). The FBCubed measure is calculated as:

$$FBCubed = \frac{2(\frac{1}{n}\sum_i^n BCP(\vec{x_i}))(\frac{1}{n}\sum_i^n BCR(\vec{x_i}))}{(\frac{1}{n}\sum_i^n BCP(\vec{x_i})) + (\frac{1}{n}\sum_i^n BCR(\vec{x_i}))}, \tag{10}$$

where BCP and BCR measures correspond to BCubed Precision and BCubed Recall, respectively. The BCubed precision for overlapping clustering methods is defined as Amigó et al. (2009):

$$BCP(\vec{x_i}, \vec{x_h}) = \frac{\sum_{\vec{x_h} \in D(\vec{x_i})} \frac{\min(|\pi(\vec{x_i}) \cap \pi(\vec{x_h})|, |C(\vec{x_i}) \cap C(\vec{x_h})|)}{|\pi(\vec{x_i}) \cap \pi(\vec{x_h})|}}{|D(\vec{x_i})|}, \tag{11}$$

where $\pi(\vec{x_i})$ is the set of clusters associated with datapoint $\vec{x_i}$, $C(\vec{x_i})$ is the class labels associated with datapoint $\vec{x_i}$, and $D(\vec{x_i})$ is the set of datapoints that share at least one cluster with $\vec{x_i}$. Similarly, the BCubed recall for overlapping clustering methods is defined as Amigó et al. (2009):

$$BCR(\vec{x_i}, \vec{x_h}) = \frac{\sum_{\vec{x_h} \in H(\vec{x_i})} \frac{\min(|\pi(\vec{x_i}) \cap \pi(\vec{x_h})|, |C(\vec{x_i}) \cap C(\vec{x_h})|)}{|C(\vec{x_i}) \cap C(\vec{x_h})|}}{|H(\vec{x_i})|} \tag{12}$$

Here, $H(\vec{x_i})$ is the set of datapoints that share at least one class with $\vec{x_i}$. The FBCubed definition is general and works even if there

Table 3
Details of tested datasets.

Category:	Dataset:	# of Samples (n):	# of Features (F):
Very Small ($0 < F < 10$)	Liver disorder	345	6
	Breast cancer wisconsin (Original)	699	9
Small ($10 \leq F < 19$)	Indian liver patients	683	10
	Heart disease (Original)	303	13
	Heart disease (Statlog)	270	13
	Hepatitis	155	19
Medium ($19 \leq F < 34$)	Parkinsons	195	22
	Breast cancer wisconsin (Diagnostic)	569	30
Large ($34 \leq F < 50$)	Dermatology	366	34
Very Large ($F \geq 50$)	Lung cancer	32	56

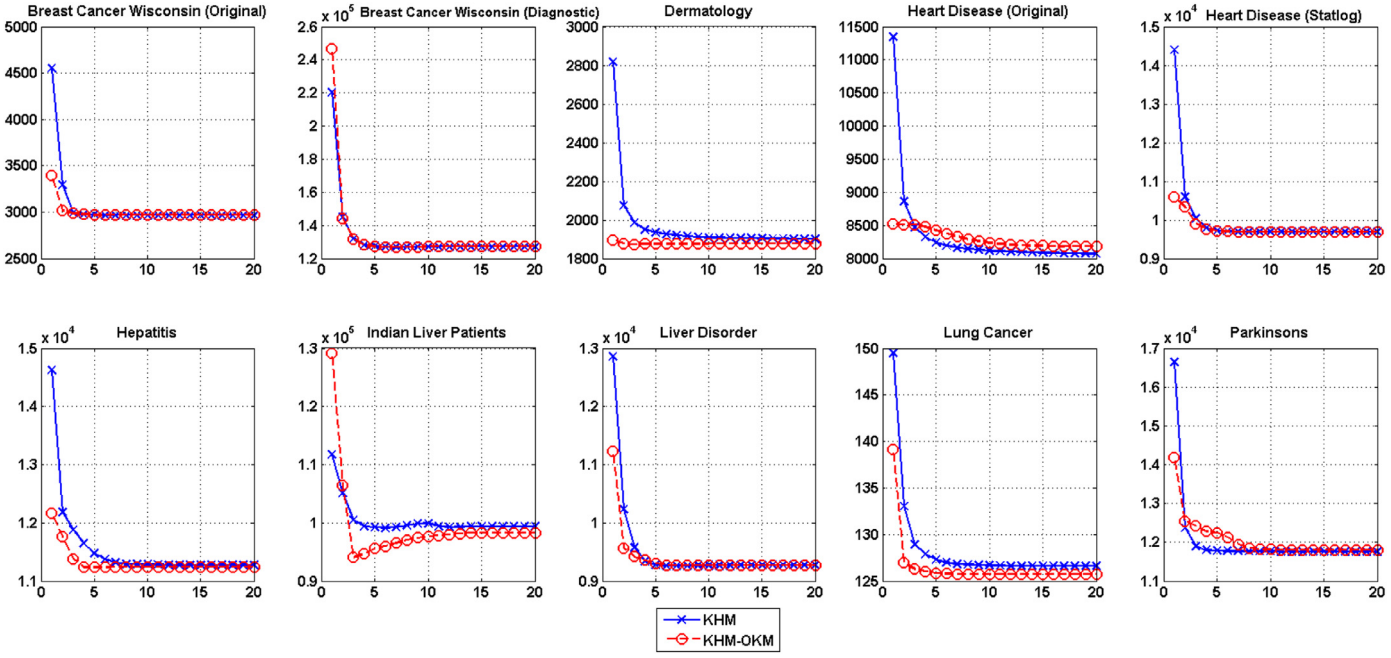

Fig. 3. Objective function values in 20 iterations.

are more than one classes assigned to each datapoint. Intuitively, the FBCubed measures the likelihood of two datapoints with same class labels being in the same cluster. In addition to evaluating the performance of proposed method using FBCubed criterion, we have also used the value of the objective function to compare the quality of obtained clusters. This evaluation method is justified since the OKM method identifies the clusters by minimizing the objective function in Eq. (2). In order to minimize the bias during comparison of objective function values, all the reported results are average of 20 replications (runs).

4.3. KHM-OKM Performance

The detailed results based on FBCubed metric are provided in Table 4. According to the results, in five datasets the proposed method performed better than the original OKM method. In three datasets (Breast Cancer Wisconsin (Original), Breast Cancer Wisconsin (Diagnostic), and Heart Disease (Statlog)) the results were identical, which might be an indication of lack of outliers in these datasets. The reason that the overall improvement over the regular OKM method might seem to be small is the fact that we allowed both methods to run until their converge. Hence, we also compared the objective function value of the KHM-OKM method to OKM method over 20 iterations to gain more insights about the performance of KHM-OKM approach. According to Fig. 3 except two datasets (Heart Disease (Original) and Parkinsons) the proposed method provided better or comparable results to the original OKM method (the KHM-OKM plot was below or similar to the OKM plot). Furthermore, we can see that in six of the datasets the KHM-OKM method convergence faster to the optimal value when compared to the OKM approach. Finally, the effects of systematic initialization of cluster centers are evident in the first iteration, since in all cases except two datasets the objective function value of the OKM-KHM method is much less than the corresponding value in OKM method. This is yet another reflection of the effectiveness of the KHM method in initializing the cluster centers of OKM algorithm.

5. Conclusion

The recently proposed overlapping k-means algorithm is one of the simplest and most effective methods for identifying overlapping clusters. However, the OKM method is sensitive to the randomly selected initial cluster centroids. Hence, in this study, we addressed this limitation by proposing a hybrid KHM-OKM algorithm, where the initial points are selected according to the results of KHM clustering algorithm. Experimental results using ten publicly available medical datasets show that the proposed hybrid method provides better or comparable results compared to the original OKM algorithm when considering the FBCubed performance criterion. Furthermore, we have demonstrated the effectiveness of the systematic initialization of OKM algorithm by

Table 4
Detailed FBCubed results.

Data Category	Datasets	OKM			KHM-OKM			
		precision	Recall	FBcubed	precision	Recall	FBcubed	*p*-value (K-S test)
Very small	Liver disorder	0.5068 ± 0.0002	0.8489 ± 0.0015	0.6347 ± 0.0006	0.5069 ± 0.0003	0.8504 ± 0.0021	0.6352 ± 0.0008	0.059 *
	Breast cancer wisconsin (Original)	0.8471 ± 0.0000	0.9846 ± 0.0000	0.9107 ± 0.0000	0.8471 ± 0.0000	0.9846 ± 0.0000	0.9107 ± 0.0000	1.000
Small	Indian liver patients	0.5927 ± 0.0001	0.9229 ± 0.0135	0.7218 ± 0.0041	0.5927 ± 0.0001	0.9176 ± 0.0074	0.7202 ± 0.0023	0.965
	Heart disease (Original)	0.3630 ± 0.0052	0.4895 ± 0.0147	0.4168 ± 0.0056	0.3595 ± 0.0003	0.5138 ± 0.0011	0.4230 ± 0.0002	0.000 ***
	Heart disease (Statlog)	0.4957 ± 0.0000	0.7733 ± 0.0000	0.6041 ± 0.0000	0.4957 ± 0.0000	0.7733 ± 0.0000	0.6041 ± 0.0000	1.000
	Hepatitis	0.6610 ± 0.0020	0.8400 ± 0.0074	0.7398 ± 0.0038	0.6619 ± 0.0000	0.8661 ± 0.0000	0.7504 ± 0.0000	0.000 ***
Medium	Parkinsons	0.6301 ± 0.0020	0.8664 ± 0.0074	0.7295 ± 0.0038	0.6285 ± 0.0032	0.8727 ± 0.0170	0.7306 ± 0.0039	0.497
	Breast cancer wisconsin (Diagnostic)	0.6740 ± 0.0000	0.9380 ± 0.0000	0.7844 ± 0.0000	0.6740 ± 0.0000	0.9380 ± 0.0000	0.7844 ± 0.0000	1.000
Large	Dermatology	0.2249 ± 0.0111	0.4181 ± 0.0201	0.2924 ± 0.0127	0.2205 ± 0.0000	0.3848 ± 0.0000	0.2803 ± 0.0000	0.003 ***
Very large	Lung cancer	0.4315 ± 0.0357	0.7365 ± 0.0723	0.5432 ± 0.0425	0.4645 ± 0.0206	0.6563 ± 0.6620	0.5433 ± 0.0349	0.000 ***
	Average	0.5427 ± 0.0058	0.7818 ± 0.0147	0.6377 ± 0.0073	0.5451 ± 0.0024	0.7758 ± 0.0094	0.6382 ± 0.0042	

* p-value < 0.1; ** p-value < 0.05; *** p-value < 0.01.

comparing the objective function values at the first iteration of the OKM algorithm. According to results (except two datasets), the objective function value of the KHM-OKM algorithm was better than original OKM algorithm with random initialization. Even though we have only focused on the medical datasets in this study, the applications of the KHM-OKM method is not limited to medical domain, and it could be applied to any other domain where the overlapping clustering methods are useful.

Despite some encouraging results, there are some limitations that need to be addressed in our future work. First of all, most of the public datasets including the ones used in this study are preprocessed and therefore the performance of the proposed method could not be fully understood. Experimental data from real-world scenarios should be used to further analyze the performance of the proposed KHM-OKM approach. Next, even though the FBCubed metric has been shown to be the most effective performance criteria for analyzing clustering algorithms, it still depends on the class label information. Hence, further studies are required to define a performance metric that could efficiently capture the performance of clustering algorithms without relying on the class label information. Finally, even though we addressed one of the limitations of OKM algorithm (sensitivity to initial cluster centers), the OKM method has several other limitations that need to be addressed for an effective overlapping clustering method. One of the other major limitations of the OKM method is its reliance to the Euclidean distance. Using pairwise Euclidean distance for capturing the similarity of data points ignores the global distance variation in the dataset. Some methods such as OKM-σ have been proposed to address this limitation (NCir & Essoussi, 2013), which should be considered for a possible integration with the KHM-OKM method.

A possible future direction of this research is integrating more efficient metaheuristic optimization algorithms such as genetic algorithm to the proposed hybrid framework instead of the iterative approach used in the KHM-OKM method to overcome the local minima problem that is common among iterative optimization methods. Another future direction is to integrate correlation clustering method with the KHM-OKM method to make the KHM-OKM method a non-parametric approach that doesn't require the prior information about the number of clusters. Finally, the plausibility of directly using the harmonic means for overlapping clustering can be considered.

References

Abbasi, A. A., & Younis, M. (2007). A survey on clustering algorithms for wireless sensor networks. *Computer Communications, 30*(14), 2826–2841.

Amigó, E., Gonzalo, J., Artiles, J., & Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval, 12*(4), 461–486.

Andreopoulos, B., An, A., Wang, X., & Schroeder, M. (2009). A roadmap of clustering algorithms: Finding a match for a biomedical application. *Briefings in Bioinformatics, 10*(3), 297–314.

Aroche-Villarruel, A. A., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F., Olvera-López, J. A., & Pérez-Suárez, A. (2014). Study of overlapping clustering algorithms based on k-means through FBCubed metric. In *Pattern recognition* (pp. 112–121). Springer.

Becker, H. (2005). A survey of correlation clustering. *Advanced Topics in Computational Learning Theory*, 1–10.

BenNCir, C., & Essoussi, N. (2012). Overlapping patterns recognition with linear and non-linear separations using positive definite kernels. *International Journal of Computer Applications, 56*(9).

BenNCir, C., Essoussi, N., & Bertrand, P. (2010). Kernel overlapping k-means for clustering in feature space. In *International conference on knowledge discovery and information retrieval (KDIR)* (pp. 250–256).

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25–71). Springer.

Bretzel, R. G. (2007). Comorbidity of diabetes mellitus and hypertension in the clinical setting: A review of prevalence, pathophysiology, and treatment perspectives. *Clinical Therapeutics, 29*, S35–S43.

Cleuziou, G. (2007). A generalization of k-means for overlapping clustering. *Rapport Technique*, 54.

Cleuziou, G. (2008). An extended version of the k-means method for overlapping clustering. In *Pattern recognition, 2008. ICPR 2008. 19th international conference on* (pp. 1–4). IEEE.
Cleuziou, G. (2010). Two variants of the OKM for overlapping clustering. In *Advances in knowledge discovery and management* (pp. 149–166). Springer.
Dilts, D., Khamalah, J., & Plotkin, A. (1995). Using cluster analysis for medical resource decision making. *Medical Decision Making, 15*(4), 333–346.
Grira, N., Crucianu, M., & Boujemaa, N. (2004). Unsupervised and semi-supervised clustering: A brief survey. *A Review of Machine Learning Techniques for Processing Multimedia Content, 1*, 9–16.
Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems, 17*(2), 107–145.
Howell, R. R., Byrne, B., Darras, B. T., Kishnani, P., Nicolino, M., & van der Ploeg, A. (2006). Diagnostic challenges for pompe disease: An under-recognized cause of floppy baby syndrome. *Genetics in Medicine, 8*(5), 289–296.
Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR), 31*(3), 264–323.
Kalyani, P. (2012). Approaches to partition medical data using clustering algorithms. *International Journal of Computer Applications, 49*(23).
Kishnani, P. S., Steiner, R. D., Bali, D., Berger, K., Byrne, B. J., Case, L. E., et al. (2006). Pompe disease diagnosis and management guideline. *Genetics in Medicine, 8*(5), 267–288.
Li, X., & Zhu, F. (2013). On clustering algorithms for biological data. *Engineering, 5*(10), 549.
Liu, X. (2012). A survey on clustering routing protocols in wireless sensor networks. *Sensors, 12*(8), 11113–11153.
Long, A. N., & Dagogo-Jack, S. (2011). Comorbidities of diabetes and hypertension: Mechanisms and approach to target organ protection. *The Journal of Clinical Hypertension, 13*(4), 244–251.
Ma, Z., Tavares, J. M. R., Renato, M., & Jorge, N. (2009). A review on the current segmentation algorithms for medical images. In *IMACAPP* (pp. 135–140).
Naik, D., & Shah, P. (2014). A review on image segmentation clustering algorithms. *International Journal of Computer Science and Information Security, 5*(3), 328993.
Narmadha, D., alias Balamurugan, A., Sundar, G. N., & Priya, S. J. (2016). Survey of clustering algorithms for categorization of patient records in healthcare. *Indian Journal of Science and Technology, 9*(8).
NITHYA, N., Duraiswamy, K., & Gomathy, P. (2013). A survey on clustering techniques in medical diagnosis. *International Journal of Computer Science Trends and Technology (IJCST), 1*(2), 17–23.
Nugent, R., & Meila, M. (2010). An overview of clustering applied to molecular biology. *Statistical Methods in Molecular Biology*, 369–404.
NCir, C.-E. B., Cleuziou, G., & Essoussi, N. (2015). Overview of overlapping partitional clustering methods. In *Partitional clustering algorithms* (pp. 245–275). Springer.
NCir, C.-E. B., & Essoussi, N. (2013). Non-disjoint cluster analysis with non-uniform density. In *Mining intelligence and knowledge exploration* (pp. 100–111). Springer.
Paul, R., & Hoque, A. S. M. L. (2010). Clustering medical data to predict the likelihood of diseases. In *Digital information management (ICDIM), 2010 fifth international conference on* (pp. 44–49). IEEE.
Soni, N., & Ganatra, A. (2012). Categorization of several clustering algorithms from different perspective: A review. *International Journal of Advanced Research in Computer Science and Software Engineering, 2*(8), 63–68.
Sylos Labini, F. (2008). Gravitational clustering: An overview. arXiv preprint arXiv: 0806.2560.
Tan, P.-N., Steinbach, M., Kumar, V., et al. (2006). *Introduction to data mining*: Vol. 1. Pearson Addison Wesley Boston.
UCI (2016). UCI repository. Accessed: 2016-02-01. http://archive.ics.uci.edu/ml/.
da Veiga, F. A. (1996). Structure discovery in medical databases: A conceptual clustering approach. *Artificial Intelligence in Medicine, 8*(5), 473–491.
Vogt, W., & Nagel, D. (1992). Cluster analysis in diagnosis.. *Clinical Chemistry, 38*(2), 182–198.
Wiwie, C., Baumbach, J., & Röttger, R. (2015). Comparing the performance of biomedical clustering methods. *Nature Methods, 12*(11), 1033–1038.
Wong, K.-C. (2015). A short survey on data clustering algorithms. In *2015 second international conference on soft computing and machine intelligence (ISCMI)* (pp. 64–68). IEEE.
Wong, K.-C., Peng, C., Li, Y., & Chan, T.-M. (2014). Herd clustering: A synergistic data clustering approach using collective intelligence. *Applied Soft Computing, 23*, 61–75.
Xu, R., Wunsch, D., et al. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on, 16*(3), 645–678.
Xu, R., Wunsch, D. C., et al. (2010). Clustering algorithms in biomedical research: A review. *Biomedical Engineering, IEEE Reviews in, 3*, 120–154.
Yasodha, M., & Mohanraj, M. (2011). Clustering algorithms for biological data-a survey approach. *Data Mining and Knowledge Engineering, 3*(3), 148–152.
Zhang, B. (2000). Generalized k-harmonic means. *Technical Report*. Hewlett-Packard Laboratoris.
Zhang, B., Hsu, M., & Dayal, U. (1999). K-harmonic means-a data clustering algorithm. *Technical Report HPL-1999-124*. Hewlett-Packard Labs.