Applied Soft Computing 35 (2015) 199–213

Contents lists available at ScienceDirect

Applied Soft Computing

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

# Robust fuzzy clustering algorithms in analyzing high-dimensional cancer databases

S.R. Kannan[a,*], R. Devi[a], S. Ramathilagam[b], T.-P. Hong[c], A. Ravikumar[a]

[a] Pondicherry University (A Central University of India), India
[b] Periyar Govt. College, Cuddalore, India
[c] National University of Kaohsiung, Taiwan

ARTICLE INFO

*Article history:*
Received 25 August 2013
Received in revised form 14 May 2015
Accepted 21 May 2015
Available online 25 June 2015

*Keywords:*
Fuzzy C-means
Kernel distances
Uncertain objects
Cancer databases

ABSTRACT

Due to uncertainty value of objects in microarray gene expression high dimensional cancer database, finding available subtypes of cancers is considered as challenging task. Researchers have invented mathematical assisted clustering techniques in clustering relevant gene expression of cancer subtypes, but the techniques have failed to provide proper outcome results with less error. Hence, it is an essential one in finding efficient computational clustering techniques to cluster the high dimensional gene expression cancer database for perfect diagnosis of cancer subtypes. This paper presents robust clustering techniques to identify perfect similarity between the uncertain objects of high dimensional cancer database. In order to obtain the robust clustering techniques, this paper incorporates both membership functions of fuzzy c-means and possibilistic c-means. In addition, this paper presents prototype initialization algorithm to avoid random initialization of initial prototypes. Benchmarks datasets were used to show the effectiveness of the proposed methods. The proposed methods were successfully implemented with microarray high dimensional gene expression cancer databases to separate available subtypes of cancer regions. The clustering accuracies of proposed and existed clustering methods indicate that the proposed methods are superior to the existed methods.



## 1. Introduction

The goal of this paper is to cluster the available subtypes of cancers in high dimensional microarray gene expression cancer database. The technology of microarray high dimensional database has significant impact on cancer research [14,40,44] in finding different types of tissues. A microarray gene expression database consisting of genes and tissue samples is typically organized in a 2D matrix. Each element in the 2D matrix gives the expression level of the gene for the tissue sample. The original gene expression database obtained from a scanning process which is contaminated by noise, and missing values. A major problem in microarray cancer database analysis is the large number of dimensions, and therefore the intensity between the tissue samples is almost similar [23]. Hence, researchers have introduced many clustering techniques to capture the available subgroups of genes in microarray high dimensional database [13,27,33,40,41]. But the prediction of subgroups in cancer database has not reached the expected percentage of confidence. Recently the unsupervised [9,17,19,36,38,39], and supervised [6,12,18] clustering techniques play an important role in clustering functionally related genes together for predicting the unlabelled gene expression about their classes of tissues. However the existing methods were not robust in finding subgroups of genes in high dimensional microarray cancer database with similar intensity tissue samples [21]. Very recently fuzzy set based fuzzy clustering [3,15,20,22,29,31] has been implemented to obtain appropriate subtypes of cancer classes in cancer database. The fuzzy set based fuzzy clustering allows gradual memberships to the data points to place an object in all clusters. The memberships offer a much finer degree of detail of the data model to cluster it into several groups and memberships can also express how ambiguously or definitely a data point should belong to a cluster [11]. Even though there are lots of benefits using fuzzy c-means algorithms, it has considerable drawbacks such as the result of clustering process deteriorates while uncertainty exists in the high dimensional medical database [25,46]. Our previous works in [22,31] have worked well in clustering low dimensional databases, but the methods are unsuitable for analyzing microarray gene expression high dimensional database. Therefore this paper attempts to provide robust fuzzy clustering techniques to capture the similar gene expression

* Corresponding author. Tel.: +91-(0)413-2654703(office).
*E-mail address:* srkannan.pu@gmail.com (S.R. Kannan).

http://dx.doi.org/10.1016/j.asoc.2015.05.035
1568-4946/

of cancer subtypes from high dimensional cancer database. This paper tries to obtain the robust kernel based fuzzy clustering algorithms in the combination of both fuzzy membership function and typicality of possibilistic c-means. Possibilistic approach has been shown to be advantageous in noisy environments, the algorithm helps to find valid clusters, and in finding a robust estimate of the cluster prototypes. Typicality-based fuzzy memberships automatically reduce the effect of noise points, and improve the accuracy of the results considerably. In order to obtain high degree of memberships for the data points that are equidistant from the prototype of the clusters, this paper obtained the possibilistic c-means based objective function of fuzzy c-means. The performance of obtaining membership to the noisy object is improved by relaxing the membership constraints using the typicality of the possibilistic c-means [28,45]. To overcome the undesirable effects of similar gene expression in updating reliable prototypes the penalized constraints of typicality is used with the proposed algorithms. Here the typicality values are constrained and the sum of the overall data points of typicalities to a cluster is equal to one. The proposed objective functions are enhanced by introducing new kernel induced distance called hyper tangent kernel Bray Curtis distance to evaluate the relations between cluster prototypes and data objects. Tangent kernel induced distance of proposed clustering techniques overcomes the difficulties in clustering the uncertain objects [37]. The neighborhood information of this paper effectively finds the difference between cluster prototypes and data points. The random selection of initial prototypes of fuzzy c-means leads more number of iteration to converge the termination condition [24,34], therefore this paper presents a mathematical prototype initialization method to reduce the number of iterations.

The rest of the paper is organized as follows. In Section 2, this paper gives terminology of clustering techniques and workflow of the paper. Section 3 contains the proposed algorithms. Section 4 presents prototypes initialization method. Section 5 gives the method for clustering accuracy. The experimental results on Synthetic Dataset, Checkerboard Dataset, Wine dataset, IRIS Dataset, and High Dimensional Cancer Databases of the proposed clustering methods are reported in Section 6. Section 7 provides conclusion of this paper.

## 2. Terminology of clustering techniques and workflow

### 2.1. Stipulation of fuzzy C-means

Consider the data which contains N objects, and p attributes. Hence we have $N \times p$ dimension of the dataset. Let the data set $G$, which contains n data points say $x_1, x_{2,\ldots,}x_n$. Assume we have to find $c$ clusters in $G$, where $2 \le c \le N$. In crisp clustering, the goal would be to partition $G$ into the disjoint non-empty partitions $G = \sum_{k=1}^{C} G_k$, and $k \in \{1, 2, 3, \ldots, c\}$. The objective function is given by $J_{cm}(G, V) = \sum_{i=1}^{n}\sum_{k=1}^{c} D(x_i, v_k)$, where $v_k$ is a prototypes of $k$th cluster. In fuzzy clustering, the goal would be to find the partition matrix $U$. The partition matrix is a real $N \times c$ matrix that defines membership degrees for each feature vector. $u$ is defined by $u \in R^{NxC} = [u_{ik}]$, where $u_{ik} \in [0, 1]$, $\forall i, k$. The objective function of fuzzy c-means [4] is as

$$J_{fcm}(G, V) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^{m} D(x_i, v_k), (m > 1) \tag{1}$$

In fuzzy clustering the results of a given clustering method is dependent on the similarity measure. The smaller value of distance between two objects represents the larger similarity between objects; conversely, the larger value of distance between the objects represents the dissimilarity of the objects. The fuzzy clustering method uses the concept of prototype or center from members of cluster. Thus, the clustering problem becomes finding a set of $c$ prototypes. The challenging to researchers in clustering is no single capable clustering technique for identifying clusters in all of real world problems, because of complicated structure of dataset and various noises in the dataset.

### 2.2. Kernel induced fuzzy C-means

Recently kernel function is considered as an effective similarity measure in clustering methods and kernel based clustering techniques have been successfully implemented with many real life applications [7]. The common ground of Kernel based clustering is to map the input data element into a feature space with higher dimension via a nonlinear transformation. Generally, the kernel function is defined in term of inner product space as

$$H(x, y) = \left\langle \varphi(x), \varphi(y) \right\rangle = \varphi(x)^T \varphi(y) \tag{2}$$

where $\varphi : x \mapsto \varphi(x)$ is a linear transformation and $x \in X$. Here, $\varphi(x)$ is considered as higher dimensional feature space. The function $H(x, v)$ is called a kernel function, and assumed as known. Now the kernel induced distance function can be expanded using inner product space as:

$$\begin{aligned} \left\| \phi(x_n) - \phi(v_k) \right\|^2 &= \left\langle \phi(x_n) - \phi(v_k), \phi(x_n) - \phi(v_k) \right\rangle \\ \left\| \phi(x_n) - \phi(v_k) \right\|^2 &= \left\langle \phi(x_n), \phi(x_n) \right\rangle + \left\langle \phi(v_k), \phi(v_k) \right\rangle - 2\left\langle \phi(x_n), \phi(v_k) \right\rangle \end{aligned} \tag{3}$$

Therefore we have the new kernel induced distance function from Eq. (3) as:

$$\left\| \phi(x_n) - \phi(v_k) \right\|^2 = H(x_n, x_n) + H(v_k, v_k) - 2H(x_n, v_k),$$

where $H(x_n, v_k) = \left\langle \phi(x_n), \phi(v_k) \right\rangle$

If $H(x_n, x_n) = 1$ and $H(v_k, v_k) = 1$, then the distance function can be rewritten [45] as

$$\left\| \phi(x_n) - \phi(v_k) \right\|^2 = 2(1 - H(x_n, v_k)) \tag{4}$$

The above distance function is known as hyper tangent based distance measure.

The derivation of the prototypes depends on the specific selection of the kernel function. If we consider the tangent kernel, then $H(v_k, v_k) = 1$ ($k = 1,\ldots, c$) and the objective function of the Fuzzy C-Means[FCM] can be expressed as

$$J_{\mathrm{FCM}}(U, V) = 2\sum_{k=1}^{n}\sum_{i=1}^{c} (u_{ik}^{m})(1 - H(x_n, v_k))$$

### 2.3. Workflow of the paper

The aim of this paper is to find robust fuzzy clustering techniques to find subtypes of cancers in high dimensional microarray cancer databases. To reduce the iterations of clustering algorithms in finding available subtypes of cancers the prototype initialization method is introduced. The proposed algorithms are implemented with artificial datasets and real benchmark datasets for evaluating the performance of the algorithms. The clustering validity methods are used to evaluate the clustering accuracy of the proposed algorithms. The working procedure in clustering cancer database into subtypes is stipulated in Fig. 1.