Fuzzy Kernel Based Effective Clustering Techniques in Analyzing Heterogeneous Databases

S R Kannan¹, M Siva¹, R Devi², S Ramathilagam³ and Mark Last⁴
¹Pondicherry University (A Central University of India), Puducherry, India
²Pachaiyappa’s College for Men, Chennai, India
³Periyar Govt. Arts College, Cuddalore, Tamil Nadu, India
⁴Ben-Gurion University of the Negev, Beersheba, Israel

Correspondence to: srkannan.mat@pondiuni.edu.in

Abstract. The aim of this paper is to introduce an effective fuzzy clustering technique based kernel function to find appropriate subgroups in heterogeneous databases. This paper introduces the effective fuzzy clustering that incorporates weighted bias field information, kernel distance, possibilistic memberships and fuzzy memberships into memberships equation and prototype equation. The effectiveness and efficiency of the proposed clustering techniques have been shown through the experimental results on benchmark heterogeneous databases.

Keywords: Clustering, Fuzzy C-Means, Kernel Distance, heterogeneous databases

1. Introduction
The main aim of this paper is to analyze the heterogeneous databases which can be capable to analyze liver cancer databases [3], [11], [12]. Research is focused on analyzing the available partition in heterogeneous databases using clustering techniques [2], [4], [6], [7], [13], [21], [32], [34]. Fuzzy clustering techniques are of considerable benefits because they could retain more information from the heterogeneous databases than other clustering techniques [10], [14], [15], [17], [22], [29], [31]. But the existed methods are not robust and received the low accuracy in finding subclasses [25], [27], [28]. Therefore this paper tries to introduce effective fuzzy clustering techniques by incorporating the fuzzy membership function, typicality of possibilistic c-means, weighted bias field information, and kernel distance functions into the objective function of fuzzy c-means.

The weighted kernel induced distance converts the original pattern space into the higher dimensional feature space to have reliable membership for an object in the noise region. This paper shows the effectiveness of the proposed clustering technique in clustering the available classes through benchmark heterogeneous database. The rest of this paper is organized as follows. This paper reviews the standard fuzzy c-means in section 2. The kernel based clustering methods are discussed in section 3. Section 4 proposes the proposed method. The experimental results on bench mark heterogeneous databases are reported in section 5. Section 6 provides the conclusion of this paper.
2. Standard Fuzzy C-Means (SFCM)

The Standard Fuzzy C-Means algorithm (SFCM) was first proposed by Dunn in [9] and then it has been extended by Bezdek [1]. It is an iterative approach, where the membership values and the cluster centers are updated in each step. The FCM algorithm involves in partitioning the \( n \) elements \( X = \{x_1, x_2, \ldots, x_n\} \) into \( c \) clusters. Given a finite set of data, the algorithm returns a list of \( c \) cluster centers \( V = \{v_1, v_2, \ldots, v_c\} \) and a partition matrix \( U = \{u_{ij}\}, u_{ij} \in [0,1], i = 1,2,\ldots,n \text{ and } j = 1,2,\ldots,c \), where each element \( u_{ij} \) gives the degree to which element \( x_i \) belongs to cluster \( v_j \). The SFCM algorithm is a probabilistic clustering algorithm, which means that the sum of the membership degrees for each data item equals

\[
\sum_{j=1}^{c} u_{ij} = 1 \forall i \in \{1,2,\ldots,n\}
\]

and the condition for the partitioning is as:

\[
\sum_{i=1}^{n} u_{ij} > 0 \forall j \in \{1,2,\ldots,c\}
\]

The main objective of fuzzy c-means algorithm is to minimize the following objective function:

\[
J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^m \|x_i - v_j\|^2
\]

Where \( \|x_i - v_j\| \) is the Euclidean distance between the object \( x_i \) and cluster center \( v_j \) & \( m > 1 \) is the degree of fuzzifier lies in \((1,\infty)\). Minimization of the objective function is achieved by iteratively optimizing in terms of \( U \) and \( V \). The cluster centers and the memberships are computed as follows:

\[
v_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m}
\]

\[
u_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{m-1}}
\]

The FCM algorithm consists of the following steps:

- Fix the number of prototypes and then select initial prototypes randomly
- Use Equation (5) for obtaining membership partition matrix
- Update the prototypes using Equation (4).
- Repeat Steps (4)–(5) until the termination criterion is satisfied.

3. Kernel Based Distance

In recent years kernel function is used widely in analyzing the available information in high dimensional medical [5, 24, 30] databases by mapping the data objects into a higher dimensional feature space from lower dimension. The kernel function connects the gap between the linearity and
the non-linearity to any algorithm that can be expressed in terms of dot product of two vectors. For the given dataset, a non linear mapping \( \phi : X^p \rightarrow F \) exists, that transforms the \( p \)-dimensional data from \( X^p \) into a high or infinite dimensional inner product feature space \( F \). A clustering method is performed in this space \( F \). To compute the inner products, the kernel inner product space is defined as \( K(x, y) = \phi(x)^T \phi(y) \). By applying the kernel trick the kernel based distance is defined as

\[
\|\phi(x) - \phi(y)\|^2 = 2(1 - K(x, y))
\]  

3.1 Kernelized fuzzy C-means algorithm (KFCM)

Kernel function is used to convert the data object from lower dimension to higher dimensional spaces which can effective to separate the data objects into better structure. Therefore objective function of SFCM has been modified using kernel function and a new kernelized objective function of FCM algorithm is discussed as:

\[
J(U, V) = 2 \sum_{i=1}^{c} \sum_{k=1}^{c} (u_{ik}^m) \|\phi(x_k) - \phi(v_i)\|^2
\]  

The parameter \( m \) is called weighting parameter which control the fuzziness of memberships in resulting classification. The \( x_k \) is the partitioning dataset and \( v_i \) is a center of cluster. The partitioning matrix \([u_{ik}]\) will satisfy the following conditions:

\[
\sum_{k=1}^{c} u_{ik} > 0 \text{ for all } i
\]

\[
\sum_{i=1}^{n} u_{ik} = 1 \text{ for all } k
\]

3.2 Kernel Possiblyilistic Fuzzy C-Means (KPFCM)

In order to reduce the sensitivity of FCM algorithm in the region of noise and outliers, kernel based fuzzy possibilistic c-means (KPFCM) clustering algorithm was introduced [33]. The KPFCM relax the condition of membership function and the objective function of KPFCM is

\[
J(\tau, U, V) = 2 \sum_{i=1}^{c} \sum_{k=1}^{c} \left( u_{ik}^m + \tau_{ik}^n \right) \|\phi(x_k) - \phi(v_i)\|^2
\]

Where \( \|\phi(x_k) - \phi(v_i)\|^2 = \langle \phi(x_k), \phi(x_k) \rangle + \langle \phi(v_i), \phi(v_i) \rangle - 2\langle \phi(x_k), \phi(v_i) \rangle \)

4. Proposed Method

4.1 Weighted Kernel Induced Fuzzy Possiblyilistic C-Means (WFPCM)

To evaluate the relationship between cluster prototypes and high dimensional Liver medical data objects, this subsection proposes effective clustering methods by enhancing the objective function of proposed methods using new kernel induced distance, fuzziness weighting exponent, the typicality
The proposed objective function of weighted fuzzy possibilistic c-means is given by

$$J_{WFPCM}(U,V) = 2\sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik}^m + \tau_{ik}^\eta) \left(1 - T(x_k - w_k b_k, v_j)\right)$$

(10)

Where $T(x_k - w_k b_k, v_j) = 1 - \|T(x_k - w_k b_k, v_j)\|^2$, and $T$ represents weighted kernel induced distance. The proposed objective function satisfies the following conditions:

$$0 \leq u_{ik} \leq 1, \text{ for } 1 \leq i \leq c, 1 \leq k \leq n$$

$$\sum_{i=1}^{c} u_{ik} = 1, k = 1, \ldots, n$$

(11)

$$\sum_{k=1}^{n} \tau_{ik} = 1, i = 1, \ldots, c$$

The parameters $m$ & $\eta$ are weighting exponent and the parameters try to reduce the vagueness in classifying the objects for the suitable classes. The weighting exponents compute the amount of fuzziness in the resulting classification in order to obtain proper center of cluster from the database which has similar gene expression. Using Lagrangian multiplier, the equation (10) can be written:

$$L_{WFPCM}(U,V) = 2\sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik}^m + \tau_{ik}^\eta) \left(1 - T(x_k - w_k b_k, v_j)\right) - \lambda_k \left(\sum_{i=1}^{c} u_{ik} - 1\right) - \delta_i \left(\sum_{k=1}^{n} \tau_{ik} - 1\right)$$

(12)

Where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_c), \delta = (\delta_1, \delta_2, \ldots, \delta_c)$ and $b_k$ is weighted Bias field information. The above objective function is minimized subject to the following constraints:

$$\sum_{i=1}^{c} u_{ik} = 1 \ \& \ \sum_{k=1}^{n} \tau_{ik} = 1$$

(13)

4.1.1 Membership & Typicality

By minimizing the objective function in (12) this subsection gets the following equation of membership function $u_{ik}$:

$$u_{ik} = \left(\frac{\lambda_k}{m}\right)^{\frac{1}{m-1}} \left(\frac{1}{1 - T(x_k - w_k b_k, v_j)}\right)^{\frac{1}{m-1}}$$

Using the fuzzy membership constraint, we have the following generalized $u_{ik}$:
The objective function in (12) is minimized subject to the constraint of typicality \(\tau_{ik}\) and we obtain the following typicality membership function:

\[
\tau_{ik} = \left(\frac{\delta_i}{2\eta}\right)^{-\frac{1}{\eta-1}} \left(\frac{1}{1-T(x_k-w_kb_k,v_i)}\right)^{\frac{1}{\eta-1}}
\]

Using the constraint of typicality \(\tau_{ik}\), we have the generalized typicality \(\tau_{ik}\):

\[
\tau_{ik} = \left(\frac{1}{1-T(x_k-w_kb_k,v_i)}\right)^{\frac{1}{\eta-1}}
\]

4.1.2 Cluster Center

The objective function in \(J_{WFPCM}(U,V)=2\sum_{i=1}^{n}(u_{ik}^m + \tau_{ik}^\eta)(1-T(x_k-w_kb_k,v_i))\) is minimized with respect to \(v_i\), to obtain the equation for updating the cluster center or prototypes of WFPCM. The generalized cluster center equation is as:

\[
v_j = \frac{\sum_{k=1}^{n}(u_{ik}^m + \tau_{ik}^\eta)(x_k-w_kb_k)}{\sum_{k=1}^{n}(u_{ik}^m + \tau_{ik}^\eta)}
\]

4.1.3 Weighted Bias field information

The objective function \(J_{WFPCM}(U,V)=2\sum_{i=1}^{n}\sum_{k=1}^{c}(u_{ik}^m + \tau_{ik}^\eta)(1-T(x_k-w_kb_k,v_i))\) is minimized with respect to \(b_k\) and this subsection is obtained the following weighted bias field information:

\[
b_k = \frac{1}{w_k} \left( x_k - \frac{\sum_{i=1}^{c}(u_{ik}^m + \tau_{ik}^\eta)v_i}{\sum_{i=1}^{c}(u_{ik}^m + \tau_{ik}^\eta)} \right)
\]
The clustering steps of WFPCM Algorithm are summarized as:
- Fix the number of cluster
- Initialize the prototypes using Prototype initialization method
- Calculate the membership using (14)
- Update the prototypes using (16)
- Repeat Step (14) and (16) until the algorithm reaches the terminal value.

5. Experimental results with Heterogeneous Benchmark Databases

In order to investigate the effects of the proposed algorithms, this subsection uses Banana dataset (Data Dimension: 5300x2, No. of Classes: 2), Heart dataset (Data Dimension: 462x9, No. of Classes: 2), and Thyroid Disease dataset (Data Dimension: 462x9, No. of Classes: 2) [29] during the experimental works. Table 1 & 2 list the clustering results on Banana dataset, Heart dataset, and Thyroid Disease dataset using the algorithms SFCM [1], FPCM [29], KFCM [30], KPFCM [33], and the proposed methods. As shown in Table 1, the best clustering accuracy, running time, and number of iterations was obtained using proposed algorithm. The error matrix [8], [20] in Table 2 provides the accuracy between reference classes and the obtained classes of Banana dataset, Heart dataset, and Thyroid disease dataset during the experimental study. The Table 2 affords the superiority of the proposed methods through the accuracy between reference classes and the obtained classes.

Table 1: Comparison of Iteration Count, Running Time and clustering accuracy

|                | SFCM | FPCM | KFCM | KPFCM | WFPCM |
|----------------|------|------|------|-------|-------|
| No. of Iterations | 18   | 16   | 15   | 15    | 5     |
| Accuracy        | 74%  | 80.81% | 84.23% | 85.65% | 99.17% |

Heart Dataset

|                | SW Cluster1 | SW Cluster2 | ASW | Accuracy | Running Time | No. of Iterations |
|----------------|-------------|-------------|-----|----------|--------------|-------------------|
| SFCM           | 0.75        | 0.75        | 0.75| 75%      | 7 Seconds    | 15                |
| FPCM           | 0.81        | 0.81        | 0.81| 81%      | 7 Seconds    | 15                |
| KFCM           | 0.82        | 0.82        | 0.82| 81%      | 7 Seconds    | 12                |
| KPFCM          | 0.83        | 0.85        | 0.83| 84%      | 4 Seconds    | 7                 |
| WFPCM          | 0.98        | 0.98        | 0.98| 98%      | 2 Seconds    | 5                 |
Thyroid Dataset

|                | SW Cluster1 | SW Cluster2 | ASW | Accuracy | Running Time | No. of Iterations |
|----------------|-------------|-------------|-----|----------|--------------|-------------------|
| SFCM           | 0.74        | 0.74        | 0.75| 74%      | 7 Seconds    | 17                |
| FPCM           | 0.76        | 0.76        | 0.76| 76%      | 8 Seconds    | 16                |
| KFCM           | 0.84        | 0.84        | 0.84| 84%      | 5 Seconds    | 11                |
| KPFCM          | 0.85        | 0.85        | 0.85| 85%      | 5 Seconds    | 10                |
| WFPCM          | 0.93        | 0.94        | 0.94| 94%      | 2 Seconds    | 6                 |

Table 2. Error Matrix

|                | SFCM | FPCM | KFCM | KPFCM | WFPCM |
|----------------|------|------|------|-------|-------|
| Banana dataset | 76%  | 82%  | 84%  | 85%   | 97%   |
| Heart dataset  | 71%  | 81%  | 84%  | 85%   | 97%   |
| Thyroid dataset| 76%  | 82%  | 86%  | 92%   | 96%   |

The results on Banana dataset, Heart dataset, and Thyroid Disease dataset are shown that the proposed method is effective in determining the available subtypes in heterogeneous datasets.

6. Conclusion
The analysis of subclasses in heterogeneous database through effective fuzzy clustering techniques has been done, and shown the proposed methods are robust in finding the subclasses. The superiority of the proposed methods has been shown through cluster validation, error matrix, running time, number of iterations and well separated clusters in clustering the heterogeneous database.

7. Acknowledgement
This work was financially supported by DST India and MOST Israel.

References
[1] Bezdek J C 1981 *Pattern Recognition with Fuzzy Objective Function Algorithms* Plenum Press New York
[2] Bing Liu, Chunru Wan, and Wang L P 2006 An efficient semi-unsupervised gene selection method via spectral biclustering *IEEE Transactions on Nano-Bioscience* 5 110-14
[3] Burden of non-communicable disease in India report by Cameron Institute 2010
[4] Carlos Alzate, Johan A K and Suykens 2011 Sparse kernel spectral clustering models for large-scale data analysis *Neurocomputing* 74 1382-90
[5] Chen S C and Zhang D Q 2004 Robust image segmentation using FCM with spatial constraints
based on new kernel-induced distance measure IEEE Trans. Syst. Man Cybern. 34 1907-16
[6] Chih-Hsuan Wang 2009 Outlier identification and market segmentation using kernel-based clustering techniques Expert Systems with Applications 36 3744-50
[7] Ching-Hao Lai 2006 Oncogenes and Subtypes of Diffuse Large B-Cell Lymphoma Discoveries from Microarray Database JCIS Atlantis Press
[8] Congalton R G and Green K 1992 Assessing the Accuracy of Remotely Sensed Data: Principles and Practices Lewis Publishers USA
[9] Dunn J C 1974 A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well Separated Clusters Journ. Cybern 95-104
[10] Francesco Masulli et al 1999 A fuzzy clustering based segmentation system as support to diagnosis in medical imaging Artificial Intelligence in Medicine 16 129-47
[11] Häme Yand Pollari M 2012 Semi-automatic liver tumor segmentation with hidden Markov measure field model and non-parametric distribution estimation Med Image Anal. 16 140-9
[12] Heimann T, Van Ginneken B, Styner MA, Arzhaeva Y, Aurich Vand Bauer C et al. 2009 Comparison and evaluation of methods for liver segmentation from CT datasets IEEE Trans Med Imaging 28 1251-65
[13] Hu Yang and Nicolino J Pizzi 2004 Biomedical Data Classification Using Hierarchical Clustering Proc IEEE Canadian Conf Elect Comput. Eng, Niagara Falls 4 1861-64
[14] Hongmei Zhang and Guidong Yu 2016 A Novel Clustering and Mining Algorithm for High Dimensional Data based on Uncertainty Criteria and Fuzzy Mathematics Rev. Téc. Ing. Univ. Zulia 39 1-11
[15] Roland Winkler, Frank Klawonn and Rudolf Kruse 2011 Fuzzy C-Means in High Dimensional Spaces International Journal of Fuzzy System Applications 1 1-16
[16] Rong-Ho Lin 2009 An Intelligent model for liver disease diagnosis Artificial Intelligence in Medical, 47 53-62
[17] Jezewski M 2011 An application of modified fuzzy clustering to medical data classification Journal of Medical Informatics and Technologies 17 51-7
[18] Kannan S R, Devi R, Ramathilagam S, Hong T P and Ravikumar A 2015 Robust fuzzy clustering algorithms in analyzing high-dimensional cancer databases Applied Soft Computing 35 199-213
[19] Liang Bai and Jiye Liang 2011 An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data Knowledge-Based Systems, 24 785-95
[20] Lunetta R S et al. 1991 Remote sensing and geographic information system data integration: error sources and research issues Photogramm. Eng. Remote Sens. 57 677-87
[21] Mohamad Amin Bakhshali 2017 Segmentation and enhancement of brain MR images using Fuzzy clustering based on information theory Soft Computing 21 22 6633-40
[22] Niladri Shekhar Mishra, Susmita Ghosh and Ashish Ghosh 2012 Fuzzy clustering algorithms incorporating local information for change detection in remotely sensed images Applied Soft Computing 12 2683-92
[23] Rousseeuw PJ 1987 Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis Journal of Computational and Applied Mathematics 20 53-65
[24] ShanmugaPriya S and Valarmathi A 2017 Efficient fuzzy c-means based multilevel image segmentation for brain tumor detection in MR images Design Automation for Embedded Systems 1-13
[25] Smet, Frank De, Mathys, Janick, Marchal, Kathleen, Thijs, Gert, Moor, Bart De and Moreau and Yves 2002 Adaptive quality-based clustering of gene expression profiles Bioinformatics 18 735-46
[26] Soumadip Ghosh et al. 2014 A novel Neuro-fuzzy classification technique for data mining *Egyptian Informatics Journal* **15** 129-47

[27] Tamayo P et al. 1999 Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation *Proceedings of the National Academy of Sciences of the United States of America* **96** 2907

[28] Tavazoie S, Hughes J D, Campbell M J, Cho R J and Church G M 1999 Systematic determination of genetic network architecture *Nature Genetics* **22** 281-5

[29] Vanisri D and Loganathan C 2011 An Efficient Fuzzy Possibilistic C-Means with Penalized and Compensated Constraints *Global Journal of Computer Science and Technology* **11**

[30] Zanaty E A, Aljahdali S and Debnath, N. A 2009 Kernelized fuzzy c-means algorithm for automatic magnetic resonance image segmentation *J. Comput. Methods Sci. Eng.* **9** 123-36

[31] Xiangyu Chang, Qingnan Wang, Yuewen Liu, and Yu Wang 2017 Sparse Regularization in Fuzzy c-Means for High-Dimensional Data Clustering *IEEE Transactions on knowledge and Data Engineering* **47** 2616-27

[32] Zheng Y, Jeon B and Xu D et al. 2015 Image segmentation by generalized hierarchical fuzzy c-means algorithm *J Intell Fuzzy System* **28** 961-73

[33] Zhao Xiao-qiang and Zhou Jin-hu 2015 Improved Kernel Possibilistic Fuzzy Clustering Algorithm Based on Invasive Weed Optimization *J. Shanghai Jiaotong Univ. (Sci.)* **20** 164-70

[34] Zhao Xuemei, Yu Li and Quanhua Zhao 2018 A Fuzzy Clustering Approach for complex Color Image Segmentation Based on Gaussian Model with Interactions between Color Planes and Mixture Gaussian Model *International Journal of Fuzzy Systems* **20** 309-17