Kernel Induced Possibilistic Unsupervised Clustering Techniques in Analyzing Breast Cancer Database

S.R. Kannan	extsuperscript{1}, M. Siva	extsuperscript{2}, R. Devi	extsuperscript{3}, Mark Last	extsuperscript{4}, Ramathilagam	extsuperscript{5}

1 Department of Mathematics, Pondicherry University, India
2 Department of Mathematics, Pachaiyappa’s College for men, Chennai, India
3 Ben-Gurion University of the Negev, Israel
4 Department of Mathematics, Periyar Govt. Arts College, Tamilnadu, India
5 Corresponding Author: srkannaniiitm@mail.com

DOI: https://doi.org/10.26438/ijcse/v7si14.9398 | Available online at: www.ijcseonline.org

Abstract—The challenge in medical breast cancer database is to differentiate the sub types of cancers in the data. Analyzing the medical breast cancer database is most important one in identifying cancer types which cause deaths. Therefore in order to analyze the types of diseases in cancer databases this paper develops fuzzy set based unsupervised effective clustering technique and implements it with breast cancer database to divide it into available subtypes. This paper introduces the objective function of unsupervised effective proposed clustering technique by incorporating kernel induced distance, kernel functions, and possibilistic memberships. Through the experimental part of this paper the efficiency of proposed method is proved.

Keywords—Clustering, Fuzzy C-Means, Kernel Distance, Breast Cancer Data

I. INTRODUCTION

The main aim of this paper is to analyze the high dimensional Breast cancer database into the available subtypes of cancers. Breast cancer is one of the main leading causes of death among women since the last decades, the breast cancer is curable cancer types if it can be identified early[10]. Through the report of US for woman cancer accessed in September 2009, the death rate of breast cancer is higher than any other cancer, approximately 40,480 deaths among 182,460 diagnosed breast cancer cases. Early recognition of the types either cancerous or non-cancerous can help in the diagnosis of the disease for woman and it can help strongly to enhance the expectancy of survival. As per the World Health Organization the early diagnosis of the types of cancerous can reduce one-third of the cancer deaths[7]. High dimensional gene expression breast cancer database is considered as a best technique in analyzing the types of cancers [1]. Due to missing attributes and overlapping of objects, analyzing the types in high dimensional gene expression cancer database is considered as difficult task. Handling the missing attributes in gene expression databases with improper techniques can easily lead to biased outcome. Therefore design of an effective diagnosis model is an important issue in breast cancer data for finding available types of cancers. Researchers have introduced clustering based algorithms to analyze the available subtypes of cancers in breast database [2, 3, 5, 6, 11]. Clustering is an important and powerful tool in analyzing the large dimension of the databases in various data analyzing process[12, 26, 27] and it is capable of recognizing the unknown patterns in high dimensional database [5, 6, 8, 9]. The unsupervised fuzzy clustering technique is performed well in high dimensional medical databases for analyzing the available subtypes of diseases [11, 13, 14, 16, 17, 18]. The existed fuzzy clustering techniques are receiving low accuracy in analyzing high dimensional databases with heavy noise. Hence this paper introduces 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 proposed objective function finds successfully the relations between the centers and the objects in the breast cancer database which has missing attributes. The kernel induced distance of the proposed objective functions transforms the original lower dimensional pattern space into the higher dimensional feature space in order to obtain reliable membership to the object in the. The paper is organized as follows. Section 2 proposes the proposed method. Section 3 presents the experimental results on Breast cancer databases. Section 4 concludes the paper.

II. PROPOSED METHOD

A. Exponential Kernel Induced Fuzzy Possibilistic C-Means (EFPFCM)

To evaluate correct feature of data substructure in clustering the high dimensional breast database, this subsection...
introduces the effective clustering technique using possibilistic memberships together fuzzy clustering with exponential kernel induced distance. The proposed objective function assigns strong membership to place the object into more appropriate cluster. The objective function of EFPCM is defined by

\[ J_{EPCM}(U,V) = 2 \sum_{i=1}^{n} \sum_{k=1}^{m} \left( u_{ik}^{m} + \tau_{ik}^{m} \right) \left( 1 - E(x_i, v_j) \right) + \sum_{k=1}^{m} \left( \sum_{i=1}^{n} \frac{u_{ik}^{m}}{|k_{-i} - \alpha_{i}|} \right) \ln \alpha_{i} \]

where \( E(x_i, v_j) = \exp \left( -\frac{\|x_i - v_j\|^2}{\sigma^2} \right) \) and \( \alpha_{i} \) is the regularized parameter.

**B. Membership & Typicality**

The objective function (1) is minimized with respect to \( u_{ik} \) using the necessary condition of Lagrangian method and the following generalized membership equation is obtained:

\[
u_{ik} = \left( \frac{2}{m} \right) ^{\frac{1}{p+1}} \left( \frac{1}{2 \left( 1 - E(x_i, v_j) \right) + \frac{1}{|k_{-i} - \alpha_{i}|} \ln \alpha_{i}} \right) ^{\frac{1}{p+1}}
\]

Using the fuzzy membership constraint \( \sum_{k=1}^{n} u_{ik} = 1 \), the generalized membership equation is modified as:

\[
u_{ik} = \left( \frac{2}{m} \right) ^{\frac{1}{p+1}} \left( \frac{1}{2 \left( 1 - E(x_i, v_j) \right) + \frac{1}{|k_{-i} - \alpha_{i}|} \ln \alpha_{i}} \right) ^{\frac{1}{p+1}}
\]

The objective function (1) is minimized with respect to \( \tau_{ik} \) using the necessary condition of Lagrangian method to obtain the following generalized membership of typicality:

\[
u_{ik} = \left( \frac{2}{m} \right) ^{\frac{1}{p+1}} \left( \frac{1}{(1 - T(x_i, v_j))} \right) ^{\frac{1}{p+1}}
\]

Using the constraint of typicality \( \sum_{k=1}^{n} \tau_{ik} = 1 \), we have

\[
u_{ik} = \left( \frac{2}{m} \right) ^{\frac{1}{p+1}} \left( \frac{1}{(1 - T(x_i, v_j))} \right) ^{\frac{1}{p+1}}
\]

This subsection implements the proposed method on GSE841 breast cancer database in order to evaluate the performance of the proposed method. This subsection used the version of dataset GSE841 sample size 22574x7 with two cancer types. The algorithms involved in this subsection are executed with HP Z800 INTEL Xeon HEX (6) Dual Core Processor workstation and the results of proposed algorithm at the existed algorithms SFCM [4], and FPCM [25], are used to compare the results with the proposed algorithm in analyzing the breast cancer data into two available subgroups. The Breast dataset in Figure. 1 consists of 22574x7 samples.
This subsection illustrates the results of algorithms in Figures 2-14 in analyzing the available subgroups in the GSE841 Breast dataset. The resulted available two classes of Breast dataset using SFCM is given in Figure 2 and the separated classes are given in Figure 3 (a-b). The obtained memberships to decide the objects for class 1 and class 2 are plotted in Figure 4 and the separated memberships for class 1 and class 2 are given in Figure 5 (a-b). Further the successive numbers of objects in cluster 1 and cluster 2 are shown explicitly in Figure 6. From the results given Figures 4-5, it is cleared that the SFCM has obtained not strong memberships to represent the objects for the particular class, due to Euclidean distance based weak objective function of the algorithms. The results of FPCM are illustrated in Figure 7-10. Figure 7 gives the available shape of two classes in GSE841 breast database and Figure 8 (a-b) provides the separated class 1 and class 2 from the database. Figure 9 depicts the obtained memberships of objects for two classes of Breast cancer database and Figure 10 illustrate the memberships for separated classes 1 & 2. It is observed that the FPCM fails to allot a stronger membership grade for assigning the object into the available classes, due to non-kernalized distance function with the objective function of the FPCM. The results of proposed method are illustrated in Figure 11-14. Figure 11 gives the available shape of two classes in GSE841 breast database and Figure 12 (a-b) provides the separated class 1 and class 2 from the database. Figure 13 depicts the obtained memberships of objects for two classes of Breast cancer database and Figure 14 illustrate the memberships for separated classes 1 & 2. It is observed that the proposed method assigns the objects into the available classes with strong membership grades, due to the kernalized distance function with the effective objective function of the proposed method.
Figure 7. Available two clusters in Breast Cancer by FPCM

Figure 8. (a) Cluster 1 by FPCM (b) Cluster 2 by FPCM

Figure 9. (a) Memberships of cluster 1 & cluster 2 by FPCM

Figure 10. Memberships of cluster 1 and cluster 2 by FPCM

Figure 11. Available two clusters in Breast Cancer by EFPCM

Figure 12. (a) Cluster 1 by EFPCM (b) Cluster 2 by EFPCM

Figure 13. Memberships of cluster 1 & cluster 2 by EFPCM

Figure 14. Memberships of cluster 1 and cluster 2 by EFPCM
The results in Table 1 are cleared that the proposed methods on Breast database are established best accuracy comparing with the existed methods. The performance evaluation of the proposed methods has been shown using the research decisive factor Silhouette accuracy [20]. Silhouette accuracy is an important factor to assess the effect of classes received by the proposed clustering algorithms. The silhouette value differs from -1 and 1 with well-clustered interpretation includes values near 1 and weakly clustered interpretation includes values near -1.

Table 1: Comparison of Clustering Accuracy

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Accuracy</th>
<th>Running Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFCM</td>
<td>78%</td>
<td>6Minutes</td>
</tr>
<tr>
<td>FPCM</td>
<td>84.5%</td>
<td>4Minutes</td>
</tr>
<tr>
<td>EFPCM</td>
<td>91%</td>
<td>1.2Minutes</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

The analysis of subclasses in Breast medical database through effective fuzzy clustering techniques have been done, and shown the proposed methods are robust in finding the subclasses. The proposed fuzzy clustering methods with kernel induced distance, exponential functions, possibilistic memberships and fuzzy memberships have provided strong prototypes, and strong membership to classify the objects for the available subclasses in Breast cancer database. The superiority of the proposed methods has been shown through cluster validation, running time and well separated clusters in clustering Breast medical database.

Acknowledgement

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

REFERENCES

Authors Profile

S.R. Kannan received his Ph.D. from Indian Institute of Technology (www.iitm.ac.in), Madras, India and PDF at DISI (www.disi.unige.it), University of Genova, Genova, Italy. Recently he has received post doctoral fellowship from department of Electrical Engineering, National Cheng Kung University (web.ncku.edu.tw), Taiwan. Presently, Dr. S.R. Kannan is working as Professor in Department of Mathematics, Pondicherry University (A Central University of India), India. He had been awarded a grant in the framework of a joint agreement between the Direzione Generale per la Cooperazione allo Sviluppo of the Italian Ministry of Foreign Affairs and the ICTP Programme for Training and Research in Italian Laboratories (www.ictp.it). He had been invited by Director General, National Agriculture Research Center, Tsukuba, Japan, for joint research work on remote sensing data to estimate rice yield (http://narc.naro.affrc.go.jp/narc-e/index.html). He has received two major research grants from UGC India, major research grant from CSIR India, bilateral research grants from DST India & NSC Taiwan for joint collaborative research project Indo Taiwan, and bilateral research grants from DST India & MOST Israel for joint collaborative research project Indo Israel.

M. Siva is currently pursuing his Ph.D. in Pondicherry University and he is working under the guidance of Dr. S. R. Kannan, Professor, Pondicherry University.

R. Devi is currently working as Assistant Professor in Mathematics, Pachaiyappa’s College for men, Chennai, India.

Mark Last is a Full Professor at the Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Israel and the Head of the Data Science Research Center at Ben-Gurion University. Prof. Last obtained his Ph.D. degree from Tel Aviv University, Israel in 2000. Prior to starting his appointment at Ben-Gurion University in March 2001, Mark Last was a Visiting Assistant Professor at the Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA (1999-2001). Between the years 2009-2012, Prof. Last has served as the Head of the Software Engineering Program at Ben-Gurion University. He has published over 200 peer-reviewed papers and 11 books on data mining, text mining, and cyber security. Prof. Last is a Senior Member of the IEEE Computer Society and a Professional Member of the Association for Computing Machinery (ACM). He currently serves as an Associate Editor of IEEE Transactions on Cybernetics and an Editorial Board Member of Data Mining and Knowledge Discovery. From 2007 to 2016, he has served as an Associate Editor of Pattern Analysis and Applications. His main research interests are focused on data mining, cross-lingual text mining, soft computing, cyber intelligence, and medical informatics.

S Ramathilagam is currently working as Assistant Professor in Mathematics, Pachaiyappa’s College for men, Chennai, India.