IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-14, May 2019 E-ISSN: 2347-2693

Kernel Induced Possiblistic Unsupervised Clustering Techniques in Analyzing Breast Cancer Database

S.R. Kannan[1*], M. Siva[2], R. Devi[3], Mark Last[4], Ramathilagam[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

[*]*Corresponding Author: srkannaniitm@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 (EFPCM)

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

$$J_{EFPCM}(U,V)=2\sum_{i=1}^{c}\sum_{k=1}^{n}\left(u_{ik}^{m}+\tau_{ik}^{\eta}\right)\left(1-E\left(x_k,v_i\right)\right)+\sum_{i=1}^{c}\sum_{k=1}^{n}\frac{u_{ik}^{m}}{\left|x_k-\alpha_i\right|}\ln\alpha_i \quad (1)$$

where $E(x_k,v_i)=\exp\left(-\frac{\left\|x_k-v_i\right\|}{\sigma^2}\right)$ and α_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:

$$u_{ik}=\left(\frac{\lambda_k}{m}\right)^{\frac{1}{m-1}}\left(\frac{1}{\left(2\left(1-T\left(x_k,v_i\right)\right)+\frac{\gamma}{2}\frac{1}{\left|x_k-\alpha_i\right|}\ln\alpha_i\right)}\right)^{\frac{1}{m-1}}$$

Using the fuzzy membership constraint $\sum_{i=1}^{c}u_{ik}=1$, the generalized membership equation is modified as:

$$u_{ik}=\frac{\left(\frac{1}{\left(2\left(1-E\left(x_k,v_i\right)\right)+\frac{1}{\left|x_k-\alpha_i\right|}\ln\alpha_i\right)}\right)^{\frac{1}{m-1}}}{\sum_{j=1}^{c}\left(\frac{1}{\left(2\left(1-E\left(x_k,v_j\right)\right)+\frac{1}{\left|x_k-\alpha_j\right|}\ln\alpha_j\right)}\right)^{\frac{1}{m-1}}} \quad (2)$$

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

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

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

$$\tau_{ik}=\frac{\left(\frac{1}{\left(1-E\left(x_k,v_i\right)\right)}\right)^{\frac{1}{\eta-1}}}{\sum_{l=1}^{n}\left(\frac{1}{\left(1-E\left(x_l,v_i\right)\right)}\right)^{\frac{1}{\eta-1}}} \quad (3)$$

C. Cluster Center

Minimizing the objective function , the generalized center v_i is obtained. The cluster center is given by

$$v_i^{t}=\frac{\sum_{k=1}^{n}\left(u_{ik}^{m}+\tau_{ik}^{\eta}\right)\exp\left(\frac{-\left\|x_k-v_i^{t-1}\right\|}{\sigma^2}\right)x_k}{\sum_{k=1}^{n}\left(u_{ik}^{m}+\tau_{ik}^{\eta}\right)\exp\left(\frac{-\left\|x_k-v_i^{t-1}\right\|}{\sigma^2}\right)} \quad (4)$$

where t represents the iteration count.

The clustering steps of EFPCM Algorithm are summarized as:

- Fix the number of cluster
- Calculate the membership using(2)
- Calculate the typicality using(3)
- Update the prototypes using (4)
- Repeat (2), (3) and (4) until the algorithm reaches the termination value

III. EXPERIMENTAL RESULTS ON BREAST CANCER DATABASE

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 samples 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. Breast cancer data analysis is considered as a best technique in analyzing the subtypes of breast cancer. Clustering the GSE841 breast cancer into cancerous and non-cancerous is considered as difficult task due to high similarity between the objects and missing attributes of databases. 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.

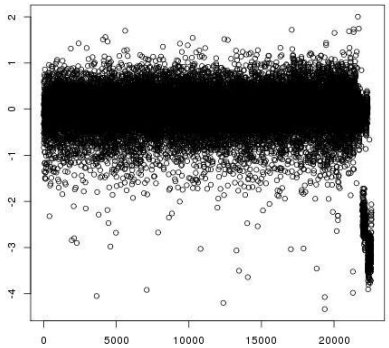


Figure. 1 Breast Cancer Database

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 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.

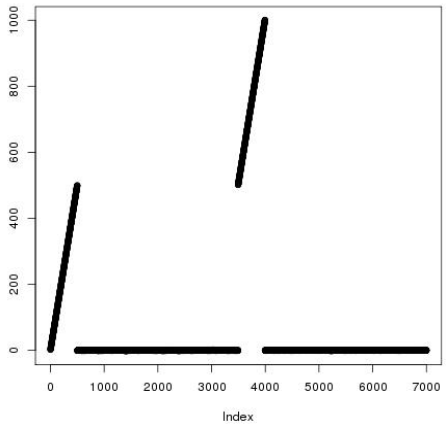
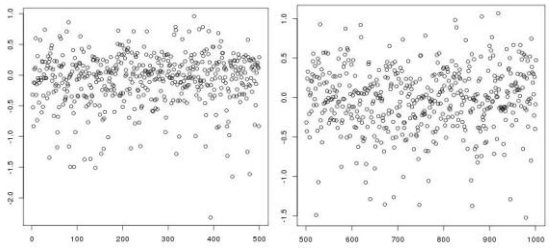


Figure2. Available two clusters in Breast Cancer by SFCM



(a) (b)

Figure.3 (a) Cluster 1 by SFCM (b) Cluster 2 by SFCM

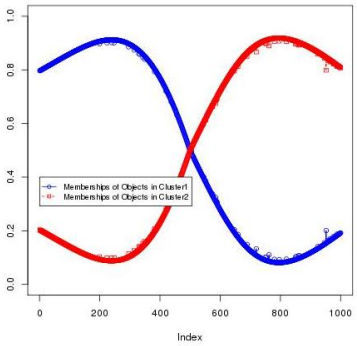
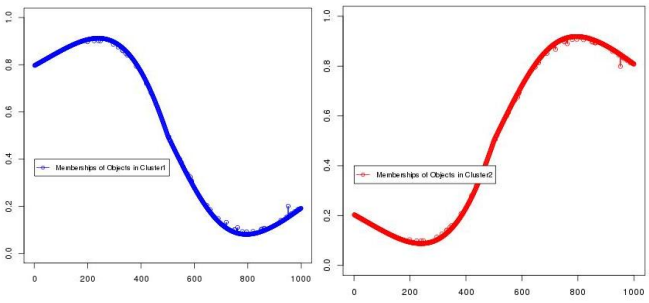


Figure 4. Memberships of cluster 1 and cluster 2 by SFCM



(a) (b)

Figure 5. Memberships of cluster 1and cluster 2 by SFCM

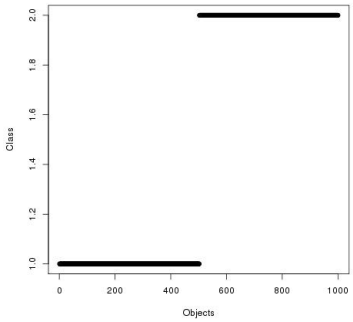


Figure 6. Objects in Cluster 1and Cluster2 by SFCM

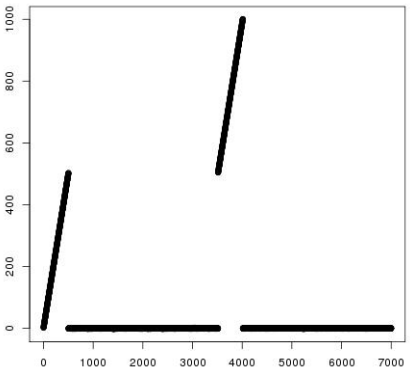


Figure 7. Available two clusters in Breast Cancer by FPCM

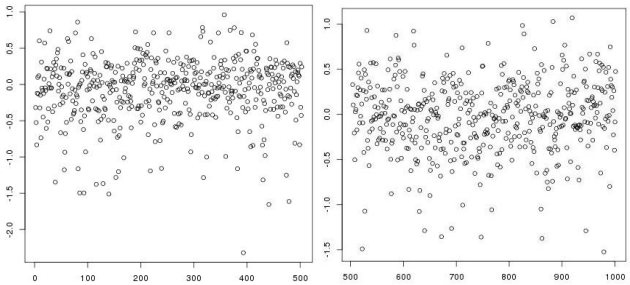


Figure 8. (a) Cluster 1 by FPCM (b) Cluster 2 by FPCM

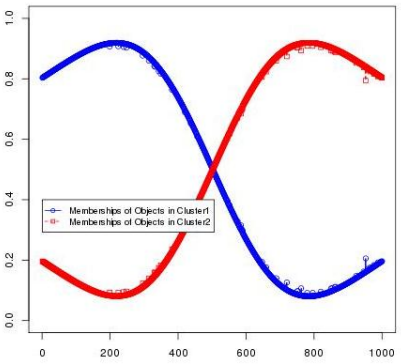
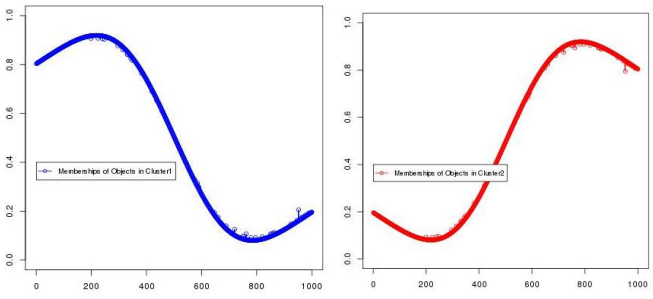


Figure 9. (a) Memberships of cluster 1 & cluster 2 by FPCM



(a) (b)

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

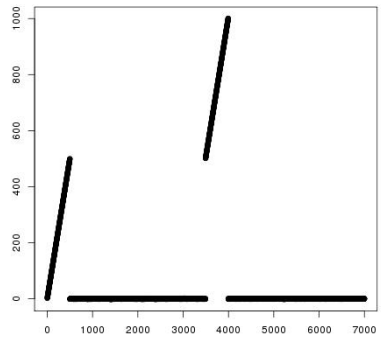


Figure 11. Available two clusters in Breast Cancer by EFPCM

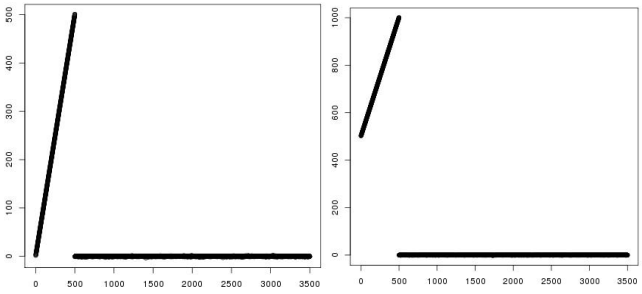


Figure 12. (a) Cluster 1 by EFPCM (b) Cluster 2 by EFPCM

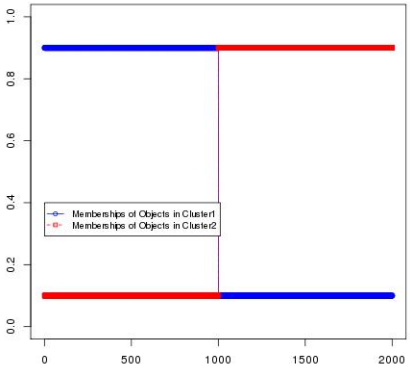
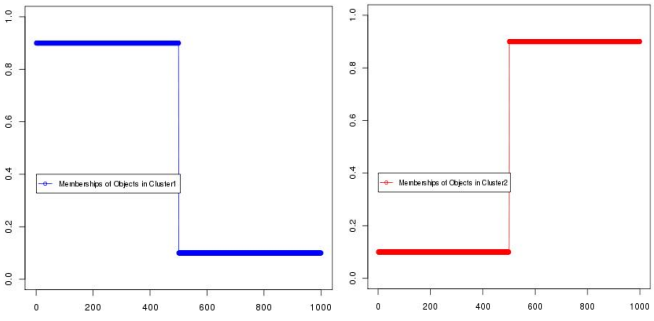


Figure 13. Memberships of cluster 1 & cluster 2 by EFPCM



(a) (b)

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

Algorithms	Accuracy	Running Time
SFCM	78%	6Minutes
FPCM	84.5%	4Minutes
EFPCM	91%	1.2Minutes

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

[1] Akay, Mehmet Fatih, *"Support vector machines combined with feature selection for breast cancer diagnosis"* , Expert systems with applications, Vol.36, Issue 2, PP.3240-3247, 2009.

[2] Antony, S. Julian Savari, "*Detected Breast Cancer on Mammographic Image Classification Using Fuzzy C-Means Algorithm*", International Journal of Innovations in Engineering and Technology, Vol.36, Issue 2, 2014.

[3] Basha, S. Saheb, K. Satya Prasad, *"Automatic detection of breast cancer mass in mammograms using morphological operators and Fuzzy C-Means clustering",* Journal of Theoretical & Applied Information Technology, Vol.**5,** Issue.**6**, 2009

[4] Bezdek J.C, *"Pattern Recognition with Fuzzy Objective Function Algorithms"*, Plenum Press, New York, 1981.

[5] Bing Liu, Chunru Wan, L.P. Wang, "*An efficient semi-unsupervised gene selection method via spectral biclustering"*, IEEE Transactions on Nano-Bioscience, Vol.**5,** Issue.**2**, pp.110-114, 2006.

[6] Carlos Alzate, Johan A.K. Suykens, "*Sparse kernel spectral clustering models for large scale dataanalysis*", Neurocomputing, Vol.**74,** Issue.**9**, pp.1382-1390, 2011.

[7] Chen, Hui-Ling, *"A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis"*, Expert Systems with Applications, Vol.**38,** Issue.**7**, pp. 9014-9022, 2011.

[8] Chih-Hsuan Wang, *"Outlier identification and market segmentation using kernel-based clustering techniques"*, Expert Systems with Applications Vol.**36,** Issue.**2**, pp. 3744-3750, 2009.

[9] Ching-Hao Lai et al., "*Oncogenes and Subtypes of Diffuse Large B-Cell Lymphoma Discoveries from Microarray Database"*, JCIS, Atlantis Press, *2006.*

[10] Etehad Tavakol, M., Saeed Sadri, E. Y. K. Ng, *"Application of K- and fuzzy c-means for color segmentation of thermal infrared breast images"*, Journal of medical systems, Vol.**34,** Issue.1, pp. 35-42, 2010.

[11] Francesco Masulli, Schenone A, *"A fuzzy clustering based segmentation system as support to diagnosis in medical imaging"*, Artificial Intelligence in Medicine, Vol.**16,** Issue.**2**, pp.129-147, 1999.

[12] Frank Klawonn, *"What Can Fuzzy Cluster Analysis Contribute to Clustering of High-Dimensional Data?"*, International workshop on Fuzzy Logic and Applications, Springer, Cham,pp.1-47, 2013.

[13] Hongmei Zhang, Guidong Yu, *"A Novel Clustering and Mining Algorithm for High Dimensional Data based on Uncertainty Criteria and Fuzzy Mathematics"*, Rev. Téc. Ing. Univ. Zulia, Vol.39**,** Issue.**2**, pp.1-11, 2016.

[14] Hu Yang, Nicolino J. Pizzi, *"Biomedical Data Classification Using Hierarchical Clustering"*, Proc IEEE Canadian Conf Elect Comput. Eng, Niagara Falls, Vol.*4*, pp.1861-1864, 2004.

[15] Jezewski M, *"An application of modified fuzzy clustering to medical data classification"*, Journal of Medical Informatics and Technologies, Vol.17, pp.51-57, 2011.

[16] Klifa, C., et al. *"Quantification of breast tissue index from MR data using fuzzy clustering"*, Engineering in medicine and biology society, IEEE, Vol.3, pp.1667-70, 2004.

[17] Liu, Xiaoming, et al. *"Microcalcification detection in full-field digital mammograms with PFCM clustering and weighted SVM-based method"*, EURASIP Journal on Advances in Signal Processing, Vol.1**,** Issue.**73**, 2015.

[18] Muhic Indira, *"Fuzzy analysis of breast cancer disease using fuzzy c-means and pattern recognition"*, Southeast Europe Journal of Soft Computing, Vol.39**,** Issue.**2**, 2013.

[19] Roland Winkler, Frank Klawonn, Rudolf Kruse, *"Fuzzy C-Means in High Dimensional Spaces, International Journal of Fuzzy System Applications"*, Vol.1**,** Issue.**1**, pp.1-16, 2011.

[20] Rousseeuw PJ, *"Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis"*, Journal of Computational and Applied Mathematics, vol.20, pp.53-65, 1987.

[21] Şahan, S., Polat, K., Kodaz, H., & Güneş, S, "A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis", Computers in Biology and Medicine, Vol.*37, Issue.*3, pp.415-423, 2007.

[22] Senthilkumar, B., and G. Umamaheswari, *"Combination of novel enhancement technique and fuzzy c means clustering technique in breast cancer detection"*, Biomedical Research, Vol.24, Issue.2, pp.252-256, 2013.

[23] Soria, Daniele et al., *"A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients"*, Computers in biology and medicine,Vol. 40, Issue.3, pp.318-330, 2010.

[24] Singh Nalini et al., *"GUI Based Automatic Breast Cancer Mass and Calcification Detection in Mammogram Images using K-means and Fuzzy C-means Methods"*, International Journal of Machine Learning and Computing, Vol.2, Issue.1, pp.7-12, 2012.

[25] D.Vanisri, C.Loganathan, *"An Efficient Fuzzy Possibilistic C-Means with Penalized and Compensated Constraints"*, Global

Journal of Computer Science and Technology, Vol.11, Issue.3, pp.15-22, 2011.
[26] Vivona Letizia, et al., *"Fuzzy technique for micro calcifications clustering in digital mammograms"*, BMC medical imaging, *Vol.*14, Issue.1, 23, 2014.
[27] Zheng, Bichen, Sang Won Yoon, Sarah S. Lam, *"Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms"*, Expert Systems with Applications, Vol.41, Issue.4, pp.1476-1482, 2014.

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.

