International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Medical Data Mining Using Efficient QPSO-FCM Clustering & Hybrid SVM-Decision Tree Classification Techniques

[1]T. Thamaraiselvan, [2]Dr. K. Saravanan

Submitted: 06/02/2024 **Revised**: 14/03/2024 **Accepted**: 20/03/2024

Abstract: Conventional medical or health care services are rapidly shifting to the internet with the rise of Internet of Health (IoH) era and have been generating a significant measure of health data related to medicine, medical infrastructure, doctors, patients, and so on. The health care services benefit from the effective analyses of these IoH results. Data mining and information discovery is a recent, fundamental research area which has significant applications in medicine, education, science, engineering and industry. It is a method of calculating and determining useful information from a large data set. The goal of data mining is to create, analyze, and apply simple induction processes that make it easier to extract useful knowledge and information from unstructured data. An effective clustering technique aids in partitioning a dataset into many groups, with the similarity in each group being higher than the similarity between groups. In this paper, the Fuzzy C-Means Clustering algorithm is combined with the Quantum-behaved Particle Swarm Optimization (QPSO). The QPSO algorithm's global search capacity helps to prevent local optima stagnation, whereas FCM's soft clustering method helps to divide the data on the basis of membership probabilities. Data classification is a crucial technique for extracting useful data. In this paper, a hybrid classification method is proposed that aims to combine the benefits of both decision trees and support vector machine (SVM) to produce better classification results. The proposed approach reduces the training dataset for SVM classification by using decision tree algorithm and it produces faster results with higher accuracy rates.

Key words: Internet of Health (IoH), FCM, QPSO, SVM and Decision tree.

I. INTRODUCTION

With the growing popularity of information technology and the progressive adoption of digital software in medical and health care sectors, numerous medical departments have collected a significant amount of historical data (For e.g., health care solutions, past information about diseases, patient details and so on) which serves as a primary source of IoH data. In general, the past IoH records contain useful informations, particularly for medical or health departments, like expert medical solution, patient's past diseases, etc. Mining and analyzing the past IoH data records will help doctors make more scientific and practical diagnoses and treatment decisions. The method of finding information from data, also known as data processing, is the process of extricating data from IoH dataset. In this process, either mathematical or non-mathematical methods have been used to achieve data to information mining conversion [1]-[4]. Data mining technique has emerged as a promising research field aimed at extracting secret predictive information, uncovering information from educational datasets and analyzing intrinsic data structures. Clustering, Classification, neural networks, and relationship mining are some of the approaches used [5]. For commercial and scientific applications, the size of database is extremely large, having datasets with some thousands to millions of documents. Therefore, clustering is a of data mining process that requires algorithms to find the distribution of data in the underlying data storage. Cluster formation is based on the maximization of similarities between patterns belonging to different clusters [6].

The K-means algorithm is a basic unsupervised learning approach used for solving the clustering issues. Due to its quick and easy implementation quality, K-means is commonly used in clustering. The k-mean algorithm divides the dataset into k clusters and the key concept is assigning a centroid for one and all cluster. Different centroid position yields distinct clustering result while defining the centroid. However, it is failed due to its inability to handle data in the form of strings or character [7], [8]. The fuzzy c-means (FCM) algorithm is a standard prototype-based solution and it is the most commonly used algorithm. In contrast to hard clustering, the FCM algorithm divides cluster partitioning into the degree to which data points belong to a specific class.

[1]*Research Scholar, Department of Computer Science PRIST UNIVERSITY*
selvanthamarai84@gmail.com
[2]*Professor, Department of Computer Science PRIST UNIVERSITY*
ks_tnj@yahoo.co.in

FCM, on the other hand, makes the number of the degrees of each entity in each category equal to1, which often results in meaningless output for datasets with outliers and noise [9]. In order to rectify the issues with FCM, an efficient hybrid SC-FCM method is employed. To generate centroid candidates, the method first employs Subtractive Clustering. Subtractive Clustering was chosen because of its capability to produce an output with a low computational cost. Anyhow, no tests of validity of cluster have ever been performed. The measurement is necessary to find the validity of SCFCM clustering method [10]. Therefore, S^3FCM has been chosen to overcome the drawbacks of SC-FCM. The S^3FCM safely examine the labelled samples using a combination of unsupervised clustering and semi supervised clustering (SSC). However, the efficiency of S^3FCM highly dependent on the graph quality [11] which has been considered as a serious drawback. Therefore, the Fuzzy C-Means Clustering algorithm is combined with the Quantum-behaved Particle Swarm Optimization (QPSO) is proposed in this paper to achieve better results.

For the intelligentization of medical knowledge, classification of medical health big data is crucial. The KNN (K-Nearest Neighbor) classification algorithm is commonly used in many fields due to its simplicity. The performance of KNN classification algorithm is significantly reduced when the size of the sample is high and the function attributes are large [12]. On the other hand, Support vector machine (SVM) is a powerful data classification tool that has been used to solve scientific and engineering problems including, time series prediction, text classification, fault diagnosis and MAC protocol identification. Though it has several advantages, it has serious limitation as it causes data privacy risk to both server and user data [13]. To overcome the limitations of SVM, ANN-based classification algorithm is used, which builds a model by learning from the training dataset but it is difficult to train the parameters [14]. Therefore, a hybrid classification method is proposed that aims to combine the benefits of both decision trees and SVM to produce better classification results. Association rule mining is one of the most popular and well-considered features of data mining. It not only offers a well-organized method of finding patterns and recognizing the model, but it also verifies existing laws, assisting in the creation of new rules [15].

Thus this paper presents the detailed study of data mining process with effective QPSO-FCM clustering algorithm along with a brief review of hybrid SVM-decision tree based classification technique.

II. PROPOSED CONTROL SYSTEM

Different data mining models exist, and they differ depending on the application domain. However, it is divided into two categories and are known as Predictive and Descriptive Model. The following are some relevant data mining activities in the medical and healthcare domains.

- Association
- Classification
- Clustering
- Trend analysis
- Regression
- Summarization

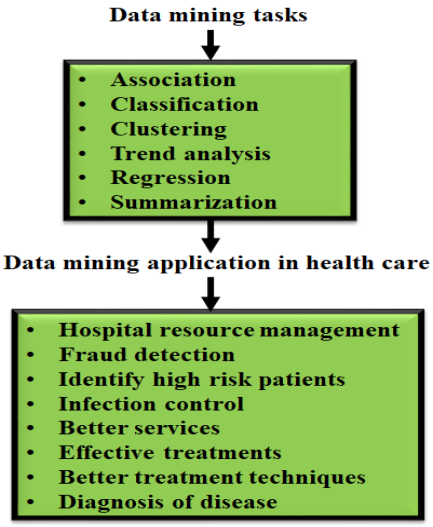


Figure 1 Application of data mining in medical service

Data mining (DM) is a technique for extracting new instructions, designs and details from vast amounts of sales data in transactional and interpersonal catalogues. It is a new technology that combines artificial intelligence, data retrieval, machine learning, and high-performance computing. It gives crucial and useful information to decision maker results in immeasurable economic

benefits. Figure 2 portrays the process of discovering knowledge in data mining.

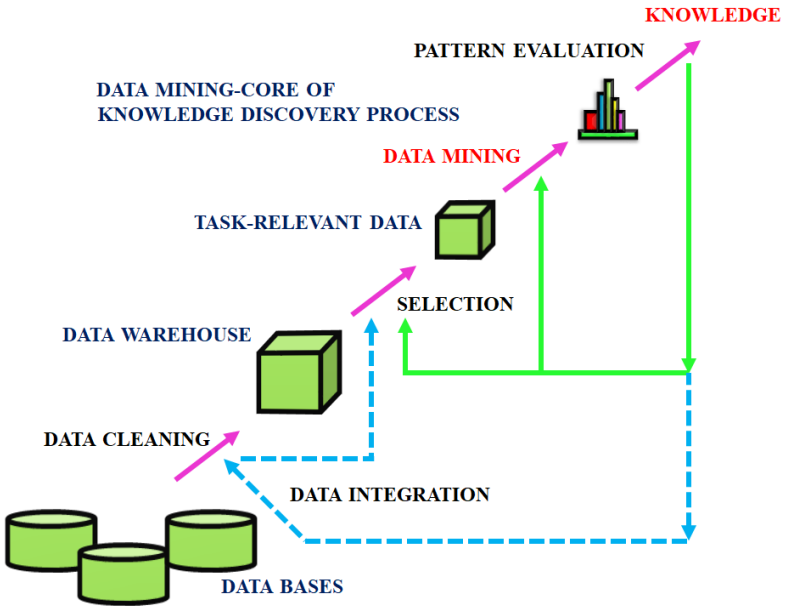


Figure 2 Representation of Data mining process in knowledge discovery

Cluster analysis, outlier analysis, classification and prediction, association and correlation analysis are all functions of data mining. The universal search power of the QPSO approach helps in avoiding the local optima stagnation whereas the soft clustering approach of the FCM algorithm helps in partitioning the data on the basis of membership probabilities. Data classification is the most important area in data mining. Data mining, machine learning, bioinformatics, medical science and statistics are all depend on classification. The proposed method uses a two-class training dataset. The decision tree technique reduces the complexity in training the SVM by reducing the dataset size.

III. MODELING PROPOSED CONTROL SCHEME

A. Clustering

The distinction between grouping and clustering is subtle. Clustering is an unsupervised learning approach whereas classification is a supervised learning method. The information of the class levelled is established in classification, but the information of the class levelled is unknown in clustering. Related data is grouped together in one cluster, while dissimilar data is grouped together in another. For partitioning the data, clustering requires very little or no details. The disadvantage of clustering is that we must first classify the clusters before assigning a new instance to them. Therefore, clustering algorithm is preferred to overcome the disadvantages of clustering approach.

B. FCM clustering Algorithm

The aim of FCM is to partition N datasets into C clusters. Therefore, it group the similar objects in the dataset $D = \{D_1, D_2, D_3 \dots \dots . D_N\}$ into C clusters $1 < C < N$. Let Q be the center of the cluster $Q = \{Q_1, Q_2, Q_3 \dots \dots . Q_C\}$. Every data point corresponds to a cluster along with randomly assigned centroids. Thus the membership function (MF) μ_{ij} is given as,

$$\mu_{ij} = \frac{1}{\sum_{r=1}^{C} \left(\frac{d_{ij}}{d_{rj}}\right)^{\frac{a}{m-1}}} \tag{1}$$

$$d_{ij} = \|x_i - y_j\|$$

$$d_{ij} = \|x_r - y_j\|$$

Where, d_{ij} denotes the distance among the i-th center and j-th data point and the distance among the r-th center and j-th data point is represented by d_{rj} and the fuzzifier is given as $m\epsilon[1, \infty)$. To compute centroids, FCM uses iterative gradient descent which can be expressed as,

$$x_i = \frac{\sum_{j=1}^{N} \mu_{ij}^{m} y_j}{\sum_{j=1}^{N} \mu_{ij}^{m}} \tag{2}$$

The sum of membership weighted Euclidean distances is the objective function that has been minimized by FCM and is given as,

$$\varphi = \sum_{i=1}^{C} \sum_{j=1}^{N} \mu_{ij}^{m} \left(\|x_i - y_j\|\right)^2 \tag{3}$$

The eqn (1) and (2) has been periodically calculated, after attaining the preset convergence criteria, the FCM can be terminated. In a multidimensional fitness landscape, like many other gradient descent algorithms,

the FCM has a local optima issue. A stochastic optimization technique can be used to prevent this.

C. QPSO clustering approach

The PSO particle updated position by utilizing personal best ($pbest$) and global best ($gbest$) position. The particle position and velocity after each iteration is give as,

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \quad (4)$$

$$v_{ij}(t+1) = \omega v_{ij}(t) + C_1 r_1(t)\left(p_{ij}(t) - x_{ij}(t)\right) + C_2 r_2(t)\left(p_{gj}(t) - x_{ij}(t)\right) \quad (5)$$

Where, C_1 and C_2 are the acceleration constants, r_1 and r_2 are the random numbers $(0< r < 1), p_{ij}(t)$ and $p_{gj}(t)$ represents the $pbest$ and $gbest$ positions and ω denotes the inertia of the i-th particle. The $pbest$ update uses a greedy update scheme with a cost reduction objective which can be expressed as follows,

$$f\left(x_i(t+1)\right) < f\left(p_i(t)\right) \Rightarrow p_i(t+1) = x_i(t+1)$$

Else, $p_i(t+1) = p_i(t)$ (6)

Where, f represents the cost. The global best (p_g) is the lowest cost bearing portion of the past collection of personal bests p_i of a specific particle. Anyhow, the PSO has low convergence rate and it has been rectified by choosing QPSO. The following equations define the particle's state update equation.

$$mbest_j = \frac{1}{N}\sum_{i=1}^{N} p_{ij} \quad (7)$$

$$\phi_{ij} = \theta_{ij} + (1-\theta)\, p_{gj} \quad (8)$$

$$x_{ij} = \phi_{ij} + \beta\left|mbest_j - x_{ij}(t)\right| \ln\left(\frac{1}{q}\right) \ \forall k \geq 0.5 = \phi_{ij} - \beta\left|mbest_j - x_{ij}(t)\right| \ln\left(\frac{1}{q}\right) \ \forall k < 0.5 \quad (9)$$

Where, $mbest$ denotes the mean value of $pbest$ in all dimension, ϕ_{ij} represents the particle local attractor and β is the contraction expansion coefficient and it can be give as follows,

$$\beta = (1-0.1)\left(\frac{iteration_{max} - iteration_{current}}{iteration_{max}}\right) + 0.1 \quad (10)$$

D. Hybridized FCM – QPSO Clustering

In FCM-QPSO clustering, one and all particle belongs to the C cluster is a D dimensional candidate solution and it can be expressed as,

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1D} \\ \vdots & \ddots & \vdots \\ x_{C1} & \cdots & x_{CD} \end{pmatrix} \quad (11)$$

Personal and global best positions are calculated after the random initialization of the particle population. Following that, the values of membership parameters are calculated and each particle assign a cost. The QPSO technique reduces the particle cost by repeatedly computing the mean best location (eqn 7), updating the candidate cluster center solution X, cost function and membership values. If the global best does not change and the QPSO-FCM is stagnant, or if the desired number of iterations is reached, the algorithm is terminated. The instability issue in a local minima in a multidimensional search space is greatly reduced to a greater extent by utilizing stochastic QPSO technique and non-differentiable parameter value handling ability in the FCM algorithmic framework. The algorithm of the FCM PSO algorithm is given as follows,

FCM-QPSO Algorithm

1. **for** each particle x_i
2. Position initialization
3. **end for**
4. Estimate the membership function values using eqn(1)
5. Set $pbest$ and $qbest$ evaluate the cost factor using eqn (3)
6. **do**
7. Calculate mean best position by using eqn (7)
8. **for** each particle x_i
9. **for** each dimension j
10. calculate ϕ_{ij} value using eqn (8)
11. **if** $k \geq 0.5$
12. update the value of x_{ij} using eqn (9) with '+'
13. **else** update the value of x_{ij} using eqn (9) with '-'
14. **end if**
15. **end for**
16. Cost estimation using eqn (3) and set $pbest$ and $gbest$
17. **end for**
18. **while** max iteration or not achieving the convergence criterion

E. Hybrid SVM-Decision Tree Classification

As the dataset is huge, the complexity of the training process and the necessary storage memory for saving these data have been increased. Therefore, it is important to provide a model that reduces the uncertainty. Thus the

hybrid classification algorithm is adopted in order to optimize the accuracy of the data classifier.

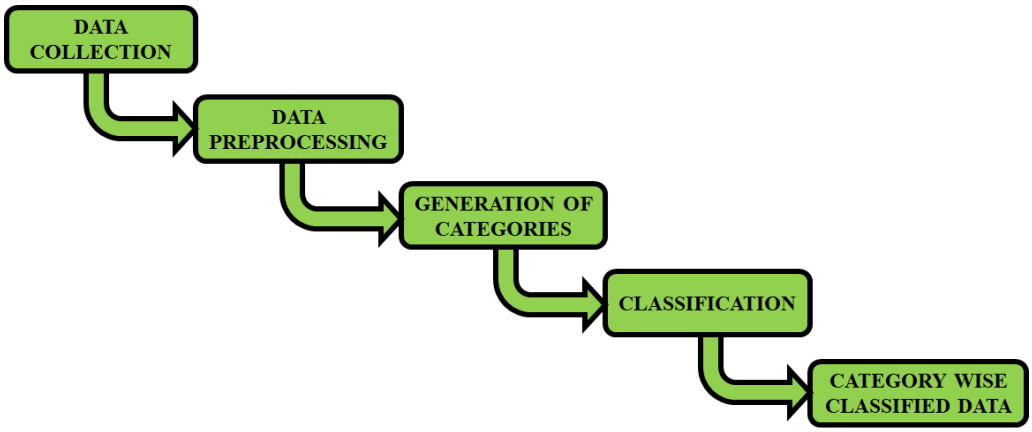


Figure 3 Steps involved in hybrid classification algorithm

In order to produce an effective integrated classification procedure, the proposed model combines SVM and Decision tree algorithm. This hybrid approach entails randomly dividing all data into two groups: experimental and test data with a ratio proportion of 70 to 30. The experimental data is then given to the standard SVM which estimates the performance. The collected coefficients are used to reclassify the data using SVM. Thus, the estimated class is named as new target. Following that, support vectors corresponding to the approximate class are used to measure the distance between individual data, and their average rate is computed. The flow chart of combined SVM and decision tree algorithm is shown in Figure 4.

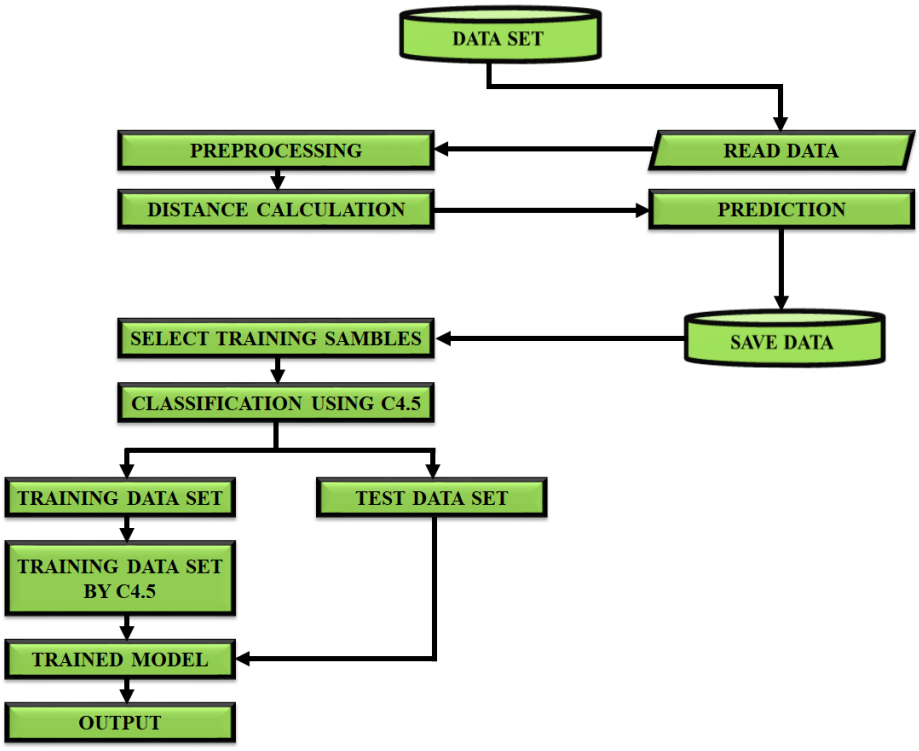


Figure 4 Flow diagram of hybrid SVM and decision tree classification algorithm

The following are the step involved in the hybrid SVM-Decision tree classification process

Step 1: Read the data

Step 2: Data preprocessing (includes eliminating the deviated data and normalization)

Step 3: Use SVM to analyze a dataset

Step 4: Estimating the distance between sample data and related support vectors for each class

Step 5: Together with the actual label of one and all sample class, the predicted label and obtained distance as data membership in one class is determined.

Step 6: Step 5 yields results that are saved in a new dataset.

Step 7: Using the new dataset, a decision tree is used to classify the training results.

Step 8: Data testing

Step 9: Estimate the result of this hybrid approach.

The predicted class and estimated distance value for each experimental data set, as well as the actual data class, are loaded into the decision tree classifier to recalculate the results. During the test process, the preceding steps are repeated in order to obtain individual test data using the SVM model. After that, it is classified and the new objective is the estimated class for one and all set of test data. The distance among test data from support vectors belonging to the target is averaged as the next move and the resultant value is used as the second parameter. In order to evaluate the class of test data, two obtained features (distance and estimated class) are incorporated into the previously acquired decision tree classifier.

IV. RESULTS & DISCUSSIONS

To allow global observation in the initial stages and development in the later stages, the training parameters C_1 and C_2 have been initialized and inertia weight value in PSO is linearly reduced over the course of iteration and the contraction-expansion parameter has been varied accordingly. The study included four well-defined original datasets from the UCI Machine Learning Repository are given in table 1,

Table 1 Attribute Descriptions for Datasets

S.no.	Primary Description	Minimum level	Maximum level	Diabetes above function level
1	BLOOD PRESSURE	155	380	T2 diabetes
2	BMI	20.5	88.6	Normal
3	GLUCOSE	130	260	T2 diabetes level-1
4	INSULIN LEVEL	110	300	Pre-diabetes

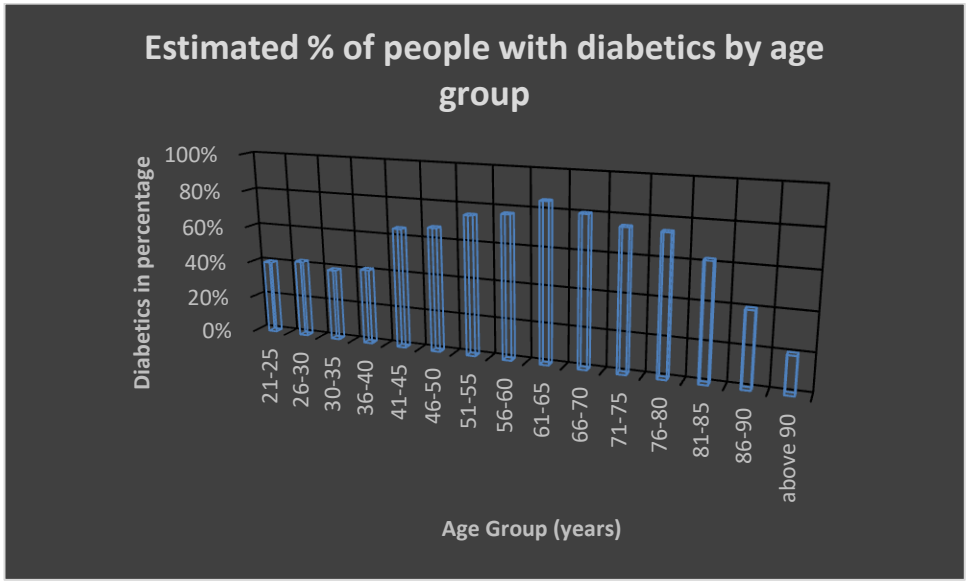
EFFICIENCY ANALYSIS

Effectiveness is a measurable concept measured by the percentage of useful yield that shows up in the reports and reveals the results. Feasibility is a more straightforward definition of having the ability to achieve a proper result that can be expressed without difficulty, but it rarely necessitates more stressed mathematics than enlargement. The clustering performance and accuracy of SCFCM, S^3FCM and QPSO-FCM have been compared through the Table 2.

Table 2 Comparison of Dataset Attributes

S. No	Primary Description	Data Partitioning			Membership Relation			Overlapping reduction		
		SC-FCM	S^3FCM	QPSO-FCM	SC-FCM	S^3FCM	QPSO-FCM	SC-FCM	S^3FCM	QPSO-FCM
1	Blood pressure	87%	88%	89%	78%	82%	84%	89%	90%	92%
2	BMI	86%	87%	89%	76%	80%	85%	78%	82%	86%
3	Glucose	74%	78%	80%	79%	84%	86%	73%	77%	79%
4	Insulin level	65%	67%	69%	70%	74%	78%	76%	78%	81%
5	Skin thickness level	77%	87%	92%	22%	20%	18%	47%	57%	63%
6	Diabetes Pedigree Function	70%	65%	63%%	35%	31%	28%	45%	50%	57%

The performance indicators such as data partitioning, membership relation and overlapping reduction are computed from the results achieved in table 2 and it is clearly observed that the QPSO-FCM gives the impressive result. In comparison to conventional SCFCM and S^3FCM implementations, the improvements in clustering accuracy obtained by FCM QPSO is sufficiently high.



Graph 1 Diabetes-causing factors based on age group

This graph depicts a comparison of diabetes-causing factors based on age. It shows that the percentage of diabetics is higher among the people with age group of 61-65 and despite the fact that diabetes affects all older people and those with multiple comorbidities are often omitted.

V. CONCLUSION

On a variety of datasets, this paper compares and contrasts the accuracy of three different clustering approaches such as SCFCM, S^3FCM and FCM when hybridized with the quantum based fully-connected topology of PSO version. FCM-QPSO combines FCM's fuzzy membership rules with QPSO's assured convergence ability thereby preventing the local optima instability in the aspect of multidimensional fitness. For competitive clustering techniques like SCFCM and S^3FCM, the performance indices of data mining with clustering technique such as data partitioning, membership relation, overlapping reduction, and accuracy have been analyzed. When compared to the others, experimental findings show that QPSO-FCM gives superior result. The experimental result shows that the hybrid SVM-Decision tree classification approach decreases the training time while improving the accuracy. It also collects the data pattern and offers enough information to achieve a successful result.

REFERENCES

[1] Hanqing Sun;Zheng Liu;Guizhi Wang;Weimin Lian;Jun Ma, Year: 2019, "Intelligent Analysis of Medical Big Data Based on Deep Learning", IEEE Access, Vol: 7, pp: 142022 - 142037.

[2] Rong Jiang;Mingyue Shi;Wei Zhou, Year: 2019, "A Privacy Security Risk Analysis Method for Medical Big Data in Urban Computing", IEEE Access, Vol: 7, pp: 143841 - 143854.

[3] Mao Ye;Hangzhou Zhang;Li Li, Year: 2019, "Research on Data Mining Application of Orthopedic Rehabilitation Information for Smart Medical", IEEE Access, Vol: 7, pp: 177137 - 177147.

[4] Qingguo Zhang;Bizhen Lian;Ping Cao;Yong Sang;Wanli Huang;Lianyong Qi, Year: 2020, "Multi-Source Medical Data Integration and Mining for Healthcare Services", IEEE Access, Vol: 8, pp: 165010 - 165017.

[5] Samina Kausar;Xu Huahu;Iftikhar Hussain;Zhu Wenhao;Misha Zahid, Year: 2018, "Integration of Data Mining Clustering Approach in the Personalized E-Learning System", IEEE Access, Vol: 6, pp: 72724 - 72734.

[6] Fanyu Bu;Chengsheng Hu;Qingchen Zhang;Changchuan Bai;Laurence T. Yang;Thar Baker, Year: 2021, "A Cloud-Edge-Aided Incremental High-Order Possibilistic c-Means Algorithm for Medical Data Clustering", IEEE Transactions on Fuzzy Systems, Vol: 29, Issue: 1, pp: 148 - 155.

[7] Liang Li;Jia Wang;Xuetao Li, Year: 2020, "Efficiency Analysis of Machine Learning Intelligent Investment Based on K-Means

Algorithm", IEEE Access, Vol: 8, pp: 147463 - 147470.
[8] Natália Maria Puggina Bianchesi;Estevão Luiz Romão;Marina Fernandes B. P. Lopes;Pedro Paulo Balestrassi;Anderson Paulo De Paiva, Year: 2019, "A Design of Experiments Comparative Study on Clustering Methods", IEEE Access, Vol: 7, pp: 167726 - 167738.
[9] L. F. Zhu;J. S. Wang;H. Y. Wang, Year: 2019, "A Novel Clustering Validity Function of FCM Clustering Algorithm", IEEE Access, Vol: 7, pp: 152289 - 152315.
[10] Victor Utomo;Dhendra Marutho, Year: 2018, "Measuring Hybrid SC-FCM Clustering with Cluster Validity Index", 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI).
[11] Haitao Gan, Year: 2019, "Safe Semi-Supervised Fuzzy C -Means Clustering", IEEE Access, Vol: 7, pp: 95659 - 95664.
[12] Wenchao Xing;Yilin Bei, Year: 2020, "Medical Health Big Data Classification Based on KNN Classification Algorithm", IEEE Access, Vol: 8, pp: 28808 - 28819.
[13] Xingxin Li;Youwen Zhu;Jian Wang;Zhe Liu;Yining Liu;Mingwu Zhang, Year: 2018, "On the Soundness and Security of Privacy-Preserving SVM for Outsourcing Data Classification", IEEE Transactions on Dependable and Secure Computing, Vol: 15, Issue: 5, pp: 906 - 912.
[14] Alan J. X. Guo;Fei Zhu, Year: 2019, "Spectral-Spatial Feature Extraction and Classification by ANN Supervised With Center Loss in Hyperspectral Imagery", IEEE Transactions on Geoscience and Remote Sensing, Vol: 57, Issue: 3, pp: 906 - 912.
[15] Ahmed M. Khedr;Zaher Al Aghbari;Amal Al Ali;Mariam Eljamil, Year: 2021, "An Efficient Association Rule Mining From Distributed Medical Databases for Predicting Heart Diseases", IEEE Access, Vol: 9, pp: 15320 - 15333.