DOI: 10.21917/ijsc.2021.0357

AN EFFICIENCT RULE MINING MODEL USING SAFE SEMI SUPERVISED FUZZY C MEANS AND ANN TECHNIQUES

T. Thamaraiselvan and K. Saravanan

Department of Computer Science, Ponnaiyah Ramajayam Institute of Science and Technology University, India

Abstract

We are living in the era known as information era where numerous sectors especially health sectors are put to handle tremendous amount of information. To reduce this burden, an efficient data mining technology has been employed which is a successful evolving technique that has a big future for helping businesses and concentrates on the most valuable data in their data warehouses. Thus, this paper presents the review of Safe Semi Supervised Fuzzy C Means (S^3FCM) clustering algorithm which is the key aspect of this work aids in achieving the goal. It has been utilized to cluster and classify the clear and relevant global data formats that many other clustering approaches unable to handle. By limiting the subsequent predictions obtained by unsupervised clustering, incorrectly labelled samples are thoroughly investigated. Meanwhile, the other labelled sample's predictions are equivalent to the assigned labels. As a result, the labelled samples are safely examined using a combination of unsupervised clustering and Semi-Supervised Clustering (SSC). Therefore, it has been clear that S^3FCM yields better result compared to other techniques. The Artificial Neural Network (ANN) based classification algorithm is used, which learns from the training dataset to construct a model. This model helps in the classification of new objects.

Keywords:

Data Mining, Clustering, S^3FCM, Classification, ANN

1. INTRODUCTION

The gradual growth in the information technology paves way for the implementation of digital software in medical and health care fields, numerous agencies and medical departments have accrued a large amount of historical data such as medical records of patient, medical solutions, details regarding the diseases, management details, etc., gives rise to huge Internet of Health (IoH) data [1]. Medical information has been conveyed in variety of ways, such as images, symbols and languages. To obtain useful information from these large and complex datasets, we must conduct data analysis. The process of retrieving information from dataset is known as data analysis. The goal is to raise the level of data mining, increase the access speed of medical data resulting in smart medical system and beneficiary services to people. Mining and reviewing historical data records will greatly assist doctors in making scientific and fair diagnosis and treatment decisions [2]. However, choosing the wrong assessment criterion for mining trends will result in lower profits. Adopting acceptable evaluation criteria such as frequency, utility or taxonomy, on the other hand, can result in more actionable derived patterns. As a consequence, a pattern mining approach that allows for tradeoffs between different criteria is needed [3]. Adopting clustering algorithm eliminates the burden occurs while retrieving the appropriate data from the complex historical data, in other words it makes the data mining process more efficient. It eventually organizes the *n* elements of the database in to *k* clusters by adopting the data from each case's variables. The precision of the clusters however, is determined by the adopted clustering technique. There are numerous clustering algorithms are in practice and here we have overviewed the efficiency and limitations of cluster ensemble technique, multi-relational clustering algorithm, k-means algorithm, hierarchical clustering algorithm, Fuzzy C-Means (FCM) algorithm and SCFCM algorithm [4]-[5].

Firstly, the ensemble clustering algorithm has been utilized to perform effective clustering as it is an innovative technique adopted to minimize the time complexity of selecting k-mean values and also it enhances the system's robustness. However, the major pitfall with these algorithms is as it is not helpful for large scale applications due to scalability issue [6]. By considering the limitations of ensemble cluster approach, k-means clustering technique has been adopted as it groups the input data in to corresponding cluster using the nearest mean criterion in k-means clustering and it is vividly used to construct the group of basis function which improves the convergence rate. However, there are some limitations to this convenience as it not be able to handle data in the form of string or characters [7]. Thus, the k-means clustering technique has been replaced with hierarchical clustering (HC) which is one of the most widely used clustering techniques because of its willingness to aggregate data structures in an informative and coherent manner. It seeks for and combines the two nearest groups until all objects are grouped in to a single cluster. A linkage method has been defined in the context to determine the distance between two clusters. However, some drawbacks are enlisted as it requires a greater number of iterations to execute HC and each iteration of HC estimate and upload the pairwise distance among all intermediary cluster. As a result, the exact algorithm of HC is bound to have time and space complication [8]. To overcome the issues with HC, Fuzzy C-Means clustering technique has been adopted. The iterative nature of clustering algorithm increases the time required for computation, but this partially offset by the smaller size of the clustered dataset. Because of this approach even small classes are characterized by a fixed number of rules. The force that repels prototype from each other need to be as strong as possible to make sure the largest possible variety of prototypes. The Fuzzy C-Means clustering technique derive this force from a probabilistic constraint and it is improved by lowering the weighting exponent value to 1. But the greater reduction in the weighting exponent value causes the method to become unstable [9]. Therefore, in this work Safe Semi Supervised Fuzzy C means clustering algorithm is proposed to achieve the goal.

Classification is a method of labelling unknown data by assigning a task that defines and distinguishes data classes. When it comes to deal with large amount of data, classification techniques help tremendously. To predict categorical class names, classification is used which assigns a class label to newly available data [10], [11]. For the intelligence of medical knowledge, classification of medical health big data is critical.

The K- Nearest Neighbor (KNN) classification algorithm is a commonly used algorithm for data classification. It is used in many fields because of its simplicity but the efficiency of KNN classification algorithm is significantly reduced with greater sample size and large function attributes [12], [13]. In order to overcome the drawbacks of KNN classification algorithm, Support Vector Machine (SVM) classification technique is employed. The dimensionality reduction data is classified using the SVM algorithm. It is not affected by dimension and is only bound to the support vector of the classification margin. In text and image classification, the SVM algorithm is commonly used. As the data often contains redundant and useless features that makes the computation more complex and lower the classification accuracy [14], [15]. Therefore, the Artificial Neural Network (ANN) based data classification technique is proposed in this paper as it eliminates distortive and useless data.

Diagnosis is a difficult process with a wide range of errors that leads to unreliable end-results. Therefore, the relational association rule mining method is utilized to determine the likelihood of disease based on a variety of variables and symptoms. This interface is also expanded by using learning strategies to introduce new symptoms and define correlations between the new signs and the diseases they refer to. The safe semi supervised fuzzy C means clustering technique adopted in data mining process and the ANN classifier used for effective classification; the result of which are clearly discussed in this paper.

2. PROPOSED CONTROL SYSTEM

Data mining is the process of extracting most required knowledge or data from large amounts of data from the past. Using data mining tools, it is much easier to sift through the databases and uncover the previously stored hidden information. Since the data mining is supported by three technologies that are likely to be sufficient: Massive data collection, Multiprocessor computers and data mining algorithms. Medical databases are expanding at faster rates therefore cost-effective computational data mining techniques has been adopted to make it simpler. The Fig.1 portrays the data mining process in discovering information.

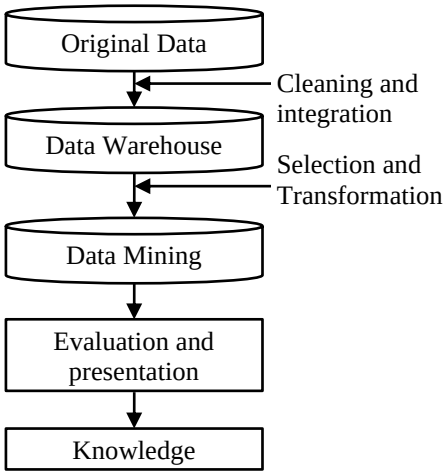

Fig.1. Representation of Data mining process in knowledge discovery

The data mining concept have been in progress for decades in fields including statistics, machine learning and artificial intelligence. These methods are applicable for modern data warehouse environment as it is assisted by high performance database engines and extensive data integration efforts. The data mining process includes clustering, classification and prediction, association analysis and forecasting. The grouping of data is known as clustering and in this paper an efficient clustering technique called safe semi supervised FCM clustering approach is employed. The ANN based classification algorithm is adopted which constructs a model by learning from the training dataset. This model aids in categorizing new objects. Association rule mining concept is adopted which is appropriate for non-numeric, categorical data and necessitates a little more than simple counting. It aims in detecting common trends, similarities, and associations in datasets stored in a variety of databases, like relational databases, transactional databases, and other data warehouses.

3. MODELING PROPOSED CONTROL SCHEME USING S^{3}FCM CLUSTERING

The conventional semi supervised clustering (SSC) methods simply assumes that data labelling is often beneficial to clustering efficiency. However, due to the ignorance and exhaustion of experts, some samples may be obtained with incorrect labels as a result of the selection process. These incorrect labels affect the performance of SSC. As a result, it is necessary to thoroughly investigate the information contained in the labelled samples using unsupervised testing process in order to minimize the risk. Therefore, an unsupervised output-dependent control parameter has been modeled based on this concept to achieve a safe exploration of the risk labelled samples. In the beginning FCM has been executed on X by ignoring the labels and it is done before partitioning the dataset into c clusters. Thus, the mapping algorithm has been used to map the estimated cluster labels to the equivalent given ones because the cluster labels generated by FCM are frequently inconsistent with the given ones. Therefore, the mathematical expression of permuted partition matrix as per the relationship among the given ones and the cluster labels is given as,

$$\hat{U} = \left[\hat{U}_{ik}\right]_{c\times n} \tag{1}$$

The objective function of S^{3}FCM is given as,

$$J_{sa} = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^2 d_{ik}^2 + \lambda_1 \sum_{k=1}^{n}\sum_{i=1}^{c}\left(u_{ik}^2 - f_{ik}b_k\right)d_{ik}^2 + \lambda_2 \sum_{k=1}^{n}\sum_{i=1}^{c}\left(u_{ik} - \hat{u}_{ik}b_k\right)^2 d_{ik}^2$$

$$\sum_{i=1}^{c} u_{ik} = 1,\ \forall k=1,2,\ldots,n$$

$$0\le u_{ik}\le 1,\ \forall k=1,2,\ldots,n \tag{2}$$

where λ_1 and λ_2 are the control parameters. The last two terms, in particular, limits the prediction of SSC to the given labels as well as the FCM's predictions, respectively. Therefore, with Eq.(2), the objective of safe discovery of labelled sample is achieved.

We use the Lagrangian multiplier method to get the value of u_{ik} when v_i is fixed. The Lagrangian function is given as follows,

$$L = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^2 d_{ik}^2 + \lambda_1 \sum_{k=1}^{n}\sum_{i=1}^{c} \left(u_{ik}^2 - f_{ik} b_k\right) d_{ik}^2 + \lambda_2 \sum_{k=1}^{n}\sum_{i=1}^{c} \left(u_{ik} - \hat{u}_{ik} b_k\right)^2 d_{ik}^2 - \gamma\left(\sum_{i=1}^{c} u_{ik} - 1,\right) \quad (3)$$

Therefore, by setting the derivative to 0, we can get the following equation in derivative form

$$2u_{ik} d_{ik}^2 + 2\lambda_1 \left(u_{ik}^2 - f_{ik} b_k\right) d_{ik}^2 + 2\lambda_2 \left(u_{ik} - \hat{u}_{ik} b_k\right)^2 d_{ik}^2 - \gamma = 0 \quad (4)$$

The value of u_{ik} is given as,

$$u_{ik} = \frac{1}{1+\lambda_1+\lambda_2}\left(\frac{1+\lambda_1+\lambda_2-\sum_{j=1}^{c}\Delta_{jk}}{d_{ik}^2 / d_{jk}^2} + \Delta_{ik}\right) \quad (5)$$

$$\therefore \ \Delta_{ik} = \lambda_1 f_{ik} b_k + \lambda_2 \hat{u}_{ik} b_k$$

When u_{ik} is fixed, v_i can be expressed according to the equation $d_{ik}=\|x_k-v_i\|_2$. Therefore, J_{sa} with respect to v_i is given as,

$$\frac{\partial j_{sa}}{\partial v_i} = 2\sum_{k=1}^{n} u_{ik}^2 \left(v_i - x_k\right) + 2\lambda_1 \sum_{k=1}^{n} \left(u_{ik} - f_{ik} b_k\right)^2 \left(v_i - x_k\right) + 2\lambda_2 \sum_{k=1}^{n} \left(u_{ik} - \hat{u}_{ik} b_k\right)^2 \left(v_i - x_k\right) \quad (6)$$

The maximum possible solution of U and V is obtained after attaining the convergence criteria, $\left|J_{sa}^{(t)} - J_{sa}^{(t-1)}\right| < \eta$

where t denotes the number of iterations and η represents the predefined threshold value.

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^2 x_k + \lambda_1 \sum_{k=1}^{n} \left(u_{ik} - f_{ik} b_k\right)^2 x_k + \lambda_2 \sum_{k=1}^{n} \left(u_{ik} - \hat{u}_{ik} b_k\right)^2 x_k}{\sum_{k=1}^{n} u_{ik}^2 + \lambda_1 \sum_{k=1}^{n} \left(u_{ik} - f_{ik} b_k\right)^2 + \lambda_2 \sum_{k=1}^{n} \left(u_{ik} - \hat{u}_{ik} b_k\right)^2} \quad (7)$$

3.1 S^{3}FCM ALGORITHM

Input: The first I samples in dataset $X=[x_1,x_2,..x_n]$ are numbered while the rest are unlabeled. The relative labels of the labeled samples are given as $Y=[y_1,y_2,..y_n]^T$, λ_1,λ_2 and η.

Output: The center V and the Partition matrix U are the outputs

Step 1: To get the cluster result $\hat{U}$, FCM has been executed on the entire dataset;

Step 2: By evaluating the mean value of the labeled sample in one and all cluster, the cluster centers $V^{(0)}$ has been initialized;

Step 3: **For** t=1:Max_{iter} **do**

a. Update $u_{ik}^{(t)}$;

b. Update $v_i^{(t)}$;

c. Compute $J_{sa}^{(t)}$;

d. **If** $\left|J_{sa}^{(t)} - J_{sa}^{(t-1)}\right| < \eta$ **then**

i.**return** U and V.

e. end if

Step 4: **End for**

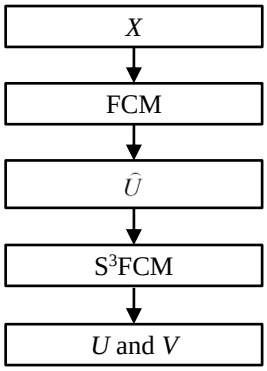

Fig.2. Flowchart of S^{3}FCM clustering algorithm

3.2 CLASSIFICATION

The main aim of classification is to predict the accurate target class from the data cluster. Classification has been done by two steps: In the first step, a framework is designed on the basis of data available for training and then the same design is used to classify an unknown tuple in to class label in the second step.

Step 1: Framework Construction

Step 2: Unknown tuple modelling

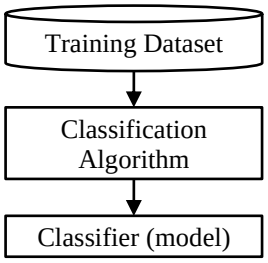

Fig.3. Construction of framework

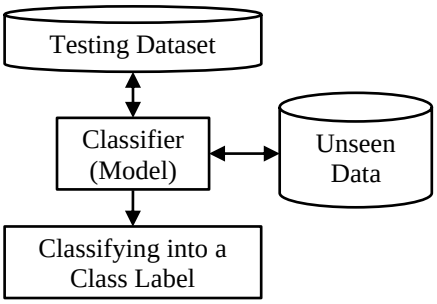

Fig.4. Unknown tuple model

3.3 ANN CLASSIFIER

The gradient decent approach is used in the Neural Network (NN), which is based on the concept of biological nervous system comprising many interconnected processing components. These processing components are named as neurons. Set of rules have been retrieved from the trained NN to enhance the connectivity of the learned network. Therefore, neural network use neurons, which are standardized processing elements to solve such specific problems.

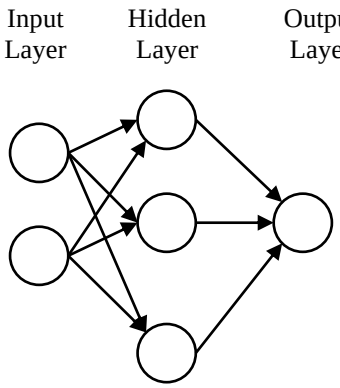

Fig.5. ANN as a classifier

The ANN used for classification and prediction. In order to mitigate the error, the ANN adjusts its framework and weights. During the learning process, weights are adjusted based on knowledge that streams internally and externally across the network. The problem with multiclass ANN is solved by adopting multilayer feed forward technique in which neurons are used in the output layer rather utilizing a single neuron.

4. RESULTS AND DISCUSSIONS

In order to attain more effective clustering, consider the crime datasets to see how the data is grouped and separated after and before clustering. The data are categorized using the ANN classification method, and our proposed scheme S^3FCM demonstrates the efficacy of clustering. The classification process begins with the dataset which is defined as a collection of data. A typical dataset represents the contents available in a single database table or a matrix of statistical information here one and all column of the table denotes a specific variable and every row denotes a specific member of the dataset. Furthermore, datasets are not limited to numbers and text; they may also contain collections of photographs or videos.

Table.1. Attribute Description of Datasets

Primary Description	Min Level	Max Level	Diabetes above Function Level
Blood pressure	155	380	T2 diabetes
BMI	20.5	88.6	Normal
Glucose	130	260	T2 diabetes Level 1
Insulin level	110	300	Prediabetes

4.1 EFFICIENCY ANALYSIS

The overall efficiency of the data mining process is verified by implementing the classification and clustering techniques and the selected inputs are analyzed by measuring the strength of the association among the variables.

It demonstrates that the S^3FCM regularization method is feasible and capable of achieving the objective of secure exploration of risky labelled samples. Furthermore, the two regularization parameters λ_1 and λ_2 in Eq.(2) have significant effects on clustering efficiency. As a result, it is important to verify the behavior patterns with various values of λ_1 and λ_2. Comparatively, S^3FCM gives better result than Semi Conquer Fuzzy C Means (SCFCM) clustering algorithm which is shown in Table.2.

Table.2. Comparison of Dataset Attributes

Primary Description	Data Partitioning		Membership Relation		Overlapping Reduction	
	SCFCM	S^3FCM	SCFCM	S^3FCM	SCFCM	S^3FCM
Blood Pressure	87%	88%	78%	82%	89%	90%
BMI	86%	87%	76%	80%	78%	82%
Glucose	74%	78%	79%	84%	73%	77%
Insulin Level	65%	67%	70%	74%	76%	78%
Skin Thickness Level	77%	87%	22%	20%	47%	57%
Diabetes Pedigree Function	70%	65%	35%	31%	45%	50%

When all of these concepts are compared, the proposed S^3FCM strategy has solved all of the challenges in clustering both implicit and explicit data using rule mining. Typically, datasets are larger in size and can be imported for clustering; however, in this case, the datasets are up to 1GB in size and are imported by 100mb per clustering, as shown in Fig.6.

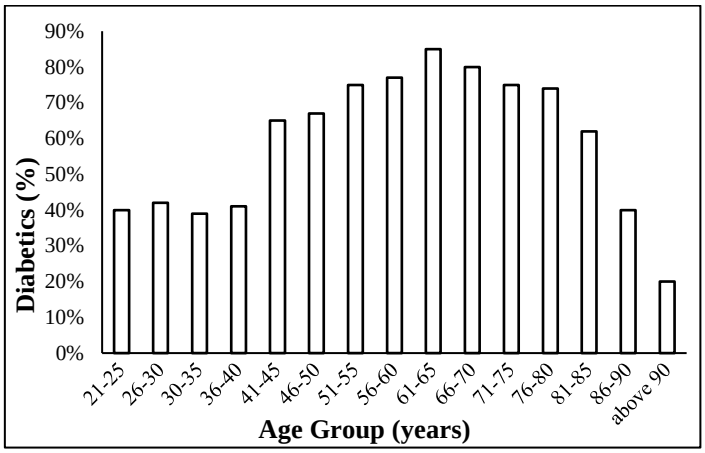

Fig.6. Percentage Estimation of diabetics among people with different age group

5. CONCLUSION

Medical care services and scientific disaster diagnosis benefit from effective analysis of huge amount of data. Data mining technique aids in searching for desired medical records in a time efficient and highly confidential manner and it supports the health workers to provide better medical and health care to patients. The proposed S^3FCM clustering algorithm helps in improving the overall efficiency of the data mining process by efficiently reducing the risk of labelled samples as shown in the reported result. ANN classifier is used to solve classification issue which is a machine learning, complex adaptive technique that alter its internal structure depending on the information it receives. When

comparing the efficiency of the proposed S^3FCM clustering method to that of other clustering models, it is obvious that S^3FCM clustering method has the higher efficiency.

REFERENCES

[1] Qingguo Zhang, Bizhen Lian and Ping Cao, "Multi-Source Medical Data Integration and Mining for Healthcare Services", *IEEE Access*, Vol. 8, pp. 165010-165017, 2020.

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

[3] Chun Hao Chen, Ji Syuan He and Tzung Pei Hong, "Post-Analysis Framework for Mining Actionable Patterns Using Clustering and Genetic Algorithms", *IEEE Access*, Vol. 7, pp. 108101-108115, 2019.

[4] Quang Thinh Bui, Bay Vo and Vaclav Snasel, "SFCM: A Fuzzy Clustering Algorithm of Extracting the Shape Information of Data", *IEEE Transactions on Fuzzy Systems*, Vol. 29, No. 1, pp. 75-89, 2021.

[5] Natalia Maria Puggina Bianchesi, Estevao Luiz Romao and Anderson Paulo De Paiva, "A Design of Experiments Comparative Study on Clustering Methods", *IEEE Access*, Vol. 7, No. 1, pp. 167726-167738, 2019.

[6] Dong Huang, Chang-Dong Wang, Jian-Sheng Wu and Jian-Huang Lai, "Ultra-Scalable Spectral Clustering and Ensemble Clustering", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 32, No. 6, pp. 1212-1226, 2020.

[7] Dal Jae Yun, In Il Jung, Haewon Jung and Hoon Kang, "Improvement in Computation Time of the Finite Multipole Method by using K-Means Clustering", *IEEE Antennas and Wireless Propagation Letters*, Vol. 18, No. 9, pp. 1814-1817, 2019.

[8] Yongkweon Jeon, Jaeyoon Yoo, Jongsun Lee and Sungroh Yoon, "NC-Link: A New Linkage Method for Efficient Hierarchical Clustering of Large-Scale Data", *IEEE Access*, Vol. 5, pp. 5594-5608, 2017.

[9] Jacek M. Leski, Robert Czabanski and Michal Jezewski, "NC-Link: Fuzzy Ordered c-Means Clustering and Least Angle Regression for Fuzzy Rule-Based Classifier: Study for Imbalanced Data", *IEEE Access*, Vol. 28, No. 11, pp. 2799-2813, 2020.

[10] Ali Arshad, Saman Riaz and Licheng Jiao, "Semi-Supervised Deep Fuzzy C-Mean Clustering for Imbalanced Multi-Class Classification", *IEEE Access*, Vol. 7, pp. 28100-28112, 2019.

[11] Gabriele Cavallaro, Morris Riedel and Matthias Richerzhage, "On Understanding Big Data Impacts in Remotely Sensed Image Classification using Support Vector Machine Methods", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 8, No. 10, pp. 4634-4646, 2015.

[12] Peng Wan, Yafeng Zhan and Weiwei Jiang, "Study on the Satellite Telemetry Data Classification Based on Self-Learning", *IEEE Access*, Vol. 8, pp. 2656-2669, 2020.

[13] Wenchao Xing and Yilin Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm", *IEEE Journal of Selected Topics in Signal Processing*, Vol. 8, pp. 28808-28819, 2020.

[14] Anne Pavy and Brian Rigling, "SV-Means: A Fast SVM-Based Level Set Estimator for Phase-Modulated Radar Waveform Classification", *IEEE Access*, Vol. 12, No. 1, pp. 1544-1556, 2018.

[15] T. Thamaraiselvan, "A Pulse Based Automated Diagnostic System", *Proceedings of IEEE 3rd International Conference on Electronics Computer Technology*, pp. 305-308, 2011.

[16] T. Thamaraiselvan and K. Saravanan, "A Survey on Hybrid Item Dependencies in Association Rule Mining", *The International Journal of Analytical and Experimental Modal Analysis*, Vol, 11, No. 12, pp.1244-1249, 2019.

[17] T.Thamaraiselvan, "An Efficient Clustering on Hybrid Item Dependency using SCFCM and SVM Techniques", *Design Engineering*, Vol. 7, No. 7, pp. 2275-2286, 2021.

[18] Weiping Ding, "SVM-Based Feature Selection for Differential Space Fusion and Its Application to Diabetic Fundus Image Classification", *IEEE Access*, Vol. 7, No. 1, pp. 149493-149502, 2019.