MEDICAL DATA MINING PROCESS USING EFFICIENT CLUSTERING AND CLASSIFICATION APPROACHES

V.D. Ambeth Kumar

Professor,Department of Computer Science and Engineering,Panimalar Engineering College,Chennai,Tamil Nadu, India. E-mail:dr.vdambethkumar@gmail.com

Abstract- Data mining technology is used by a variety of companies due to its numerous benefits. As the health care is playing a major role in today's world, the significance of data mining is gradually raised in the health care sectors. Various new technologies are developed to analyze patient's physical conditions and detect symptoms of various diseases. It involves a large amount of data, including a patient's previous medical records, test history, and even personal information. Since there is a large number of data associated with medical systems, an efficient method for extracting relevant information from the database is needed. One of the best solutions for this is data mining which alleviate the pressure in handling massive amount of medical data. This paper presents a review on data mining applications in healthcare as well as some recent mining techniques like clustering and classification approaches applied in this area. As a result, to enhance the data mining operation, an efficient clustering algorithm is used. Thus the Safe Semi Supervised Fuzzy C Means (S^3FCM) clustering algorithm is presented in this research work. Also, an effective hybrid classification algorithm is employed that attempts to combine the advantages of both decision trees and support vector machines (SVM) to achieve accurate classification result.

Keywords: Data mining, health care, medical data, S^3FCM clustering, hybrid SVM-decision tree classification.

1 INTRODUCTION

The growth of information technology has resulted in large number of databases and massive amount of data in the field of medicine[1]. The systematic accumulation of medical data from a single patient or a group of patients gives rise to these massive medical data [2]. Medical records, particularly historical diagnosis records, medical prescriptions, health examination photographs and laboratory test rescords are the most common medical data formats. For a particular disease, doctors examine the disease by referring the past medical data and detect the disease in its early stage to prevent the disease before it manifests, reliably estimate disease progression, and identify the patients in high risk [3]. Thus the Medical and Healthcare Management group has recently shown a lot of interest in intelligently extracting information from their data warehouse in order to enhance healthcare quality while also lowering costs [4]. However, to obtain valuable knowledge from these massive and complex datasets, data mining has been performed. The process of determining/extracting information from the data is known as data mining which helps in improving information extraction and health care quality[5]-[7].

Medical data analysis is a difficult task because of its complex health data structure, numerous missing documents, and noise. Therefore an efficient clustering technique is employed to enhance the data mining process. Data clustering is an important multivariate data analysis and design technique. This method divides data into groups with a significant level of natural correlation between members of the same group and a lower level of natural correlation between members of different groups. Clusters are a term used to describe such groups. Cluster analysis has also been widely applied in a variety of fields, including information retrieval, data processing, knowledge discovery, artificial intelligence, and applications in science, healthcare, psychology, finance, and a variety of other fields. Nowadays, this theory is gaining popularity, with over 100 algorithms reviewed and the most commonly used algorithms are, the ensemble, Gaussian mixtures models k-means, fuzzy C-means and hierarchical clustering [8], [9].

Generally, the ensemble clustering algorithm has been utilized to perform successful clustering because it is a novel technique for reducing the time complexity of selecting k-mean values while also improving the system's robustness. The main disadvantage of these algorithms is that they are not suitable for large-scale applications due to scalability issues [10], [11]. In order to overcome the drawbacks of the ensemble cluster method, the k-means clustering technique is employed, which groups the input data into corresponding clusters using the nearest mean criterion in k-means clustering. It builds the category of basis functions, which increases the convergence rate. However, this convenience has some drawbacks, such as the inability to handle data in the form of strings or characters [12]-[14]. As a result, the k-means clustering algorithm has been succeeded by hierarchical clustering (HC), which has been considered as a popular clustering techniques due to its ability to integrate data structures in a meaningful and coherent way. It looks for and merges the two closest groups before all of the objects are clustered together in a single

cluster. In order to determine the distance between two clusters, a linkage method has been established. However, there are some disadvantages, such as the fact that it takes more iterations to run HC, and each iteration of HC estimates and uploads the pairwise distance between all intermediate clusters. As a consequence, the specific HC clustering method is known to be complicated in terms of time and space [15], [16]. The Fuzzy C-Means clustering technique is utilized to solve the problems in HC clustering. The time needed for computation is high beacause of the iterative nature of the FCM, but this is partly compensated by the dimensionally reduced clustered dataset. In this clustering concept, even a small groups have been defined by certain laws. The force that repels prototypes from each other must be as intense as possible to make sure the highest expected variety of prototypes. This force is derived from a probabilistic restriction by the Fuzzy C-Means clustering process, which is strengthened by lowering the weighting exponent value to 1. However, when the weighting exponent value is reduced, the process becomes unstable [17]-[19]. As a result, this paper proposes the Safe Semi Supervised Fuzzy C Means (S^3FCM) clustering algorithm, which is a key aspect of this work.

During classification, the data is mapped into predefined groups. Since the classifier is built using learning data their corresponding classes that have been already known, then the classification process is known as supervised learning. The algorithm is trained on dataset examples in this task, it attempts to classify each collection of data into the appropriate class. When it comes to classification, the aim is usually to figure out what distinguishes classes from one another. As a result, when a dataset with no label is loaded into the process, the classification algorithm will identify, which class the series belongs to. Several classification algorithms are in practice to enhance the classification accuracy[20].

Many researchers are interested in KNN(K-Nearest Neighbor) classification methods because of their applications in pattern classification, remote sensing, image processing, bioinformatics, and other fields. But, the enlarged sample size and the large feature attributes significantly reduces the efficiency of the KNN classification algorithm [21]. Thus the decision algorithm has been chosen to overcome the drawbacks of KNN classification technique. Under ambiguity, the decision tree is used as a model for sequential decision issues. It supports in explaining the decisions that has been taken, the situations that occur, and the outcomes that are relevant to one and all incidents and decisions. However, the performance and accuracy of the decision tree classifier is significantly less [22]. To overcome this limitation data classification based on random forest forest algorithm has been employed as it provides excellent classification accuracy. This algorithm includes bootstrap aggregating or bagging technique. In bagging, a huge number of classifiers are trained on the random subset of the training dataset, the training dataset is classified using all classifier's predominance voting. Boosting works similarly to bagging, but adds weights to one and all classifier on the basis of their efficiency over the training dataset. Eventhough it has several advantages, it suffers due to poor performance [23], [24]. As a result, a hybrid classification approach is proposed, which seeks to incorporate the advantages of both decision trees and Support Vector Machine (SVM) to improve classification results.

2 RELATED WORKS

Mingchen Feng et al [25] presented the concept of big data analytics (BDA) for analysing and recognising patterns, relationships, and trends in massive amounts of data. We apply BDA to crime data in this article, with data mining for visualisation and pattern prediction. A variety of cutting-edge deep learning and data mining methods have been employed. A variety of cutting-edge deep learning and data mining approaches are employed. By considering crime data, some fascinating facts and trends are discovered by numerical analysis and visualisation. The Keras stateful LSTM and Prophet concept outperforms NN models in terms of prediction, with three years of training data being considered to be the optimum size. These encouraging results will help law enforcement agencies and police departments better identify crime concerns.

Sunil Kumar et al [26] presented the data mining and data analysis are strategies for analysing data and extracting secret knowledge. Since big data is complex and large in volume, conventional methods to evaluation and extraction do not fit well. Data clustering is a popular data mining technique that groups data into clusters and makes it possible to obtain knowledge from these clusters. Current clustering approaches, such as hierarchical, and k-means are inefficient because the consistency of the clusters they generate is harmed. As a result, a highly efficient and scalable clustering algorithm is needed. Have proposed hybrid clustering as a novel clustering algorithm to address the shortcomings of current clustering algorithms.

Gaspard Harerimana et al [27] illustrated the concept that because of the abundance of data, the health sector has received increased attention, and a growing number of studies aimed at using information to improve healthcare have been performed. Innovative data from sources such as genomic and social network service data have been used to create customised medication programmes. As a result, medical data have heen collected in several forms from multiple sites like

images, contexts, graphs, tables and their existence obstruct the research. In this research, we look at the main problems, strategies, data sources and innovations in the area of clinical information systems, as well as future research directions.

Weidong Liu et al [28] analyzed the multi-modal specific healthcare data feature representation training material, and various feature training images for disease risk assessment were proposed in order to extract knowledge from primary healthcare data and create smart application-related problems. The disease risk management programme uses convolutional neural network text processing techniques. The deep learning approach is used to describe medical data features. The model's simplicity is realised by using the same approach for studying and extracting different disease attributes. A CNN for health data function development and smart detection is built using a basis experimental data sample preprocessing, which includes power frequency denoising and convolution regularisation. Several simulation and experimental results were conducted on the basis of it to investigate the impact of the convolution kernel and learning rate choice.

Xuyang Yan et al [29] presented the concept that features of real-time data are uncertain, so conventional data clustering approach rely on previous data or predetermined variables. Furthermore, the user-defined threshold has been utilized to mitigate the effects of noises and outliers, which have a major impact on clustering efficiency. The latest stream clustering approaches face another big challenge: cluster overlap. Their real-time applications are severely limited by these constraints. The author suggest a two-phase stream clustering algorithm on the basis of fitness proportionate sharing. When prior information is unavailable, it manages data streaming and the clustering issue has been converted into a multimodal optimization issue.

Bharath K. Samanthula et al [30] With the recent rise in development of cloud computing, users can now redistributing their results, in encrypted format and also the data mining techniques to the cloud. Current privacy-preserving classification methods aren't applicable because the data in the cloud is encrypted. The classification issue over encrypted data is the subject of this paper. We introduce a stable KNN classification technique for handling the data in encrypted form in the cloud, in particular. The suggested protocol safeguards data secrecy, preserves the user's query privacy, and conceals data accessing trends. To our knowledge, this is the first time a stable KNN classification has been developed over encrypted data using the semi-honest design. We also use a real-world dataset to experimentally test the efficacy of our proposed scheme with various parameter settings.

3 PROPOSED WORK

Data mining is the process of discovering most important information or data from the massive amounts of historical data. It is much easier to filter the desired data from the databases and discover previously stored special information using data mining techniques. Data preprocessing and data mining are the two pats of the data mining process. The process flow diagram of data mining with data preprocessing, feature extraction, clustering and classification model is shown in Figure 1.

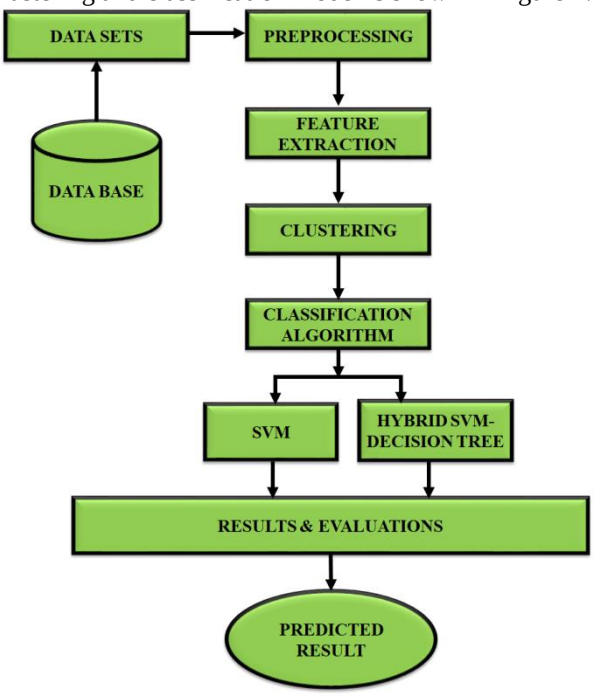

Figure 1 Data mining process with effective clustering and classification models

Missing values, noise and uncertainities are the common problems with raw medical data. Hence, the raw data has to be preprocessed. Data preprocessing is an important step used in the data mining process, since it describes the preparation and modification of initial dataset. The data preprocessing method includes data cleansing, integration, transformation and reduction.

Data cleaning: The phase noise data and the data that is unrelated to the research are excluded from the collected data. This process of removing irrelevant noise data data from the dataset is called data cleaning.

Data Integration: Multiple data sources are merged into a single source during the data integration process and it is worth assigning here that the data sources are often heterogeneous.

Data Transformation: The data transformation process, also known as data consolidation, involves in

transforming the previously selected data into the data formats that are fit for mining process.

Data Reduction: It is the process of converting the experimentally or empirically derived alphabetical or numerical digital data into a corrected, sorted and simplified form.

After data preprocessining, feature extraction has been done to optimize the precision of the model by selecting more important features. There are a variety of data mining models, and they vary depending on the application domain. It is, however, divided into two types: Predictive and Descriptive Models. The techniques involved in data mining tasks are summarization, association, clustering, classification, trend analysis and regression. Clustering is the process of grouping similar data, and in this paper, an effective clustering technique known as the safe semi-supervised FCM clustering method is used. For data classification, an efficient hybrid classification approach is utilized, which seeks to incorporate the advantages of both decision trees and SVM to improve the classification results.

3.1 Data Preprocessing

The Normalization technique is necessary to preserve the wide variance in prediction and forecasting and to bring their values close to each other. Therefore, in this work, the Min-Max normalization technique is adopted for data preprocessing stage. Normalization is a data preprocessing technique that applies a linear transformation to the original data set. Thus it maintains the relation between the original data in the dataset. It is also a simple technique that allows you to fit data into a pre-defined boundary. Thus the formula is given as follows,

$$Min - Max\ Normalization = \frac{Original\ value - Min\ value}{Max\ value - Min\ value}$$

3.2 Feature Extraction

Feature extraction is a technique of selecting the attribute which contribute much in information exploration and thus enhance the performance of the model. It also helps in reducing the time complexity of the computation process. Feature extraction has been used in data mining to optimize the precision of the model by selecting more important features. In this work, SIFT algorithm is adopted for feature extraction. Using SIFT algorithm, the distinct feature invariance of the medical images are extracted without relying on key-point extraction, grey space features or colour. The key-points, also known as descriptors, are found after the features are extracted, and these descriptors are independent of alteration. The SIFT algorithm employs the methods of detecting scale space extreme, correct localization of key-points, assignment orientation, and local image definition. The flowchart of SIFT algorithm is portrayed in Figure 2.

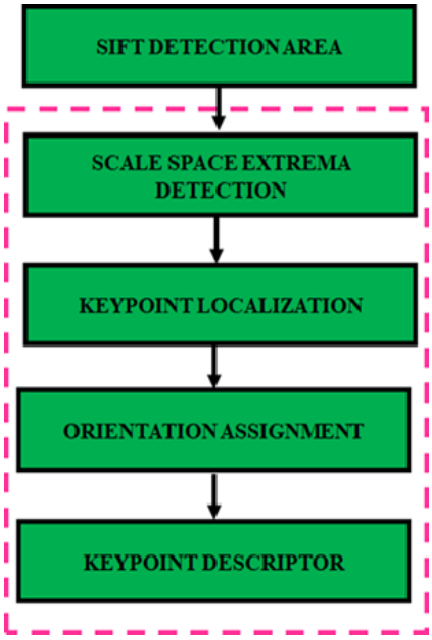

Figure 2 Flow diagram of SIFT Algorithm

3.3 S^3FCM Clustering

Clustering is the process of identifying objects that belong to similar groups. Here, Safe Semi Supervised Fuzzy C-Means (S^3FCM) Clustering algorithm is used to investigate the wrongly labelled samples by constraining the corresponding predictions to those obtained by unsupervised clustering. Similarly, the estimates of other labelled samples have to be close to the specified labels. As a result, the labelled samples can be safely explored using a combination of unsupervised clustering and SSC. To ensure a sustainable discovery of the risk labelled samples, an unsupervised output-dependent control parameter has been modelled based on this principle. Until partitioning the dataset into c clusters, FCM was performed on X without taking into account the labels. Since this cluster labels created by FCM are often inconsistent with the given ones, the mapping algorithm was used to map the estimated cluster labels to the equivalent given ones. As a result, the mathematical expression for the permuted partition matrix based on the relationship between the given ones and the cluster labels is:

$$\hat{Y} = [\widehat{Y_{ik}}]_{c \times n} \quad (1)$$

$$Q_{sa} = \sum_{k=1}^{n}\sum_{i=1}^{c} y_{ik}^2\, d_{ik}^2 + \lambda_1 \sum_{k=1}^{n}\sum_{i=1}^{c}(y_{ik}^2 - f_{ik}b_k)^2 d_{ik}^2 + \lambda_2 \sum_{k=1}^{n}\sum_{i=1}^{c}(y_{ik} - \widehat{y_{ik}}b_k)^2 d_{ik}^2 \quad (2)$$

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

$$0 \le y_{ik} \le 1, \ \forall k = 1, \ldots, n$$

Where, Q_{sa} is the objective function. The control parameters λ_1 and λ_2 are used here. The last two terms, in particular, restrict SSC's prediction to the given labels as well as the FCM's predictions. As a result, Equation 2 achieves the goal of a secure discovery of a labelled sample.

When v_i is set, we use the Lagrangian multiplier technique to analyze the value of y_{ik}. The following is the Lagrangian function:

$$L = \sum_{k=1}^{n}\sum_{i=1}^{c} y_{ik}^2\, d_{ik}^2 \lambda_1 \sum_{k=1}^{n}\sum_{i=1}^{c}(y_{ik}^2 - f_{ik}b_k)^2 d_{ik}^2 + \lambda_2 \sum_{k=1}^{n}\sum_{i=1}^{c}(y_{ik} - \widehat{y_{ik}}b_k)^2 d_{ik}^2 - \gamma(\sum_{i=1}^{c}(y_{ik} - 1) \quad (3)$$

As a result, by setting the derivative to 0, we can obtain the derivative form of the following equation.

$$2y_{ik}d_{ik}^2 + 2\lambda_1(y_{ik}^2 - f_{ik}b_k)d_{ik}^2 + 2\lambda_2(y_{ik} - \widehat{y_{ik}}b_k)d_{ik}^2 - \gamma = 0 \quad (4)$$

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

$$\Delta_{ik} = \lambda_1 f_{ik}b_k + \lambda_2\widehat{y_{ik}}b_k$$

When the value of y_{ik} is kept constant, then the v_i can be expressed on the basis of equation $d_{ik} = \|x_k - v_i\|_2$ Therefore, the value of Q_{sa} on the basis of v_i can be expressed as follows,

$$\frac{\partial q_{sa}}{\partial v_i} = 2\sum_{k=1}^{n} y_{ik}^2(v_i - x_k) + 2\lambda_1 \sum_{k=1}^{n}(y_{ik} - f_{ik}b_k)^2 (v_i - x_k) + 2\lambda_2 \sum_{k=1}^{n}(y_{ik} - \widehat{y_{ik}}b_k)^2 (v_i - x_k) \quad (6)$$

After meeting the convergence requirements, the maximum possible solution of Y and V is found.

$$\left|Q_{sa}^{(t)} - Q_{sa}^{(t-1)}\right| < \eta$$

The number of iterations is denoted by t, the predefined threshold value is denoted by.

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

S^3FCM Algorithm

Input: In dataset X= $[x_1, x_2, .. x_n]$, the first I samples are labelled, while the rest are unlabeled. $Y = [y_1, y_2, .. y_n]^T, \lambda_1, \lambda_2, and\ \eta$ are the relative labels for the labelled samples.

Output: The outputs are the V present at the center and the Partition matrix Y.

1) FCM was run on the entire dataset to obtain the cluster result ($\hat{Y}$).
2) The cluster centres $V^{(c)}$have been initialised by evaluating the mean value of the labelled sample in one and all clusters.
3) **For** $t = 1 : M\ axiter$ **do**
4) Update $y_{ik}^{(t)}$;
5) Update $v_i^{(t)}$;
6) Compute $Q_{sa}^{(t)}$;
7) **if** $\left|Q_{sa}^{(t)} - Q_{sa}^{(t-1)}\right| < \eta$ **then**
8) **return** Y and V
9) **end if**
10) **end for**

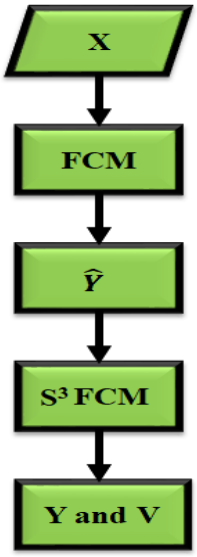

Figure 3 S^3FCM clustering Flowchart

3.4 Hybridized SVM-Decision Tree Classifier

The primary goal of classification is to accurately predict the target class from a data cluster. Since the dataset is so large, the training process has become more complicated, and the amount of storage memory required to save the data is increased. As a result, providing a model that reduces the uncertainty is crucial. Therefore, the hybrid classification algorithm is used to increase the data classifier's accuracy.

The developed scheme incorporates SVM and the Decision Tree algorithm to create an efficient integrated classification process. This hybrid method involves splitting all data into two categories at random: experimental and test data, with a 70:30 ratio. After that, the experimental data is fed into a standard SVM, which

calculates the results. SVM is used to relabel the data using the obtained coefficients. As a result, the estimated class is called "new goal." The distance among individual data is then measured using support vectors belonging to the estimated class, and the average rate is calculated. Figure 4 depicts the flow chart of the combined SVM and decision tree algorithm.

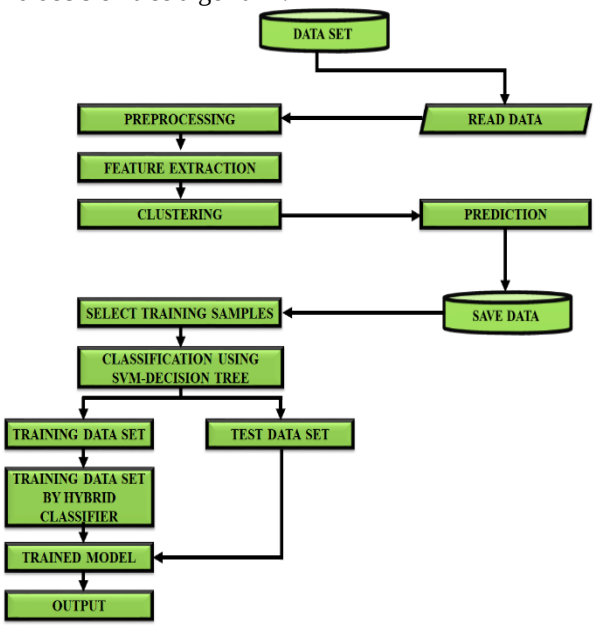

Figure 4 Hybrid SVM and decision tree algorithm flowchart

Intially read the data and it has to be preprocessed in order to remove the erroneous data, and then the data have to be normalized. The clustering approach is employed which calculates the distance between sample data. The SVM is utilized to perform dataset estimation and for each class and calculate related support vectors. The predicted label and obtained distance as data membership in one class are calculated in conjunction with the actual label of one and all sample classes. The obtained result is stored in the new dataset. Then this new dataset is taken by the decision tree which classifies the training results.

The actual data class, as well as the expected class for each observational data set, are put into the decision tree classifier to reconfigure the results. The preceding steps are replicated during the evaluation phase in order to achieve individual test data using the SVM model. The data is then categorised, and the new goal is to determine the approximate class for each collection of test data. As the next step, the test data from target support vectors is summed, and the resultant value is used as the second parameter. Two obtained features are integrated into the previously acquired decision tree classifier in order to determine the class of test data.

4 RESULTS & DISCUSSION

Consider the medical dataset attribute to see how the data is clustered and classified to achieve more accurate information. The liver patient dataset (https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset) we have considered in this experiment has been obtained from the online UCI machine learning repositories and these UCI dataset are freely available for research purpose in the field of medicine. The clustering algorithm S^3FCM which provides an excellent clustering result is employed in this research work. Also, the hybrid classification approach on the basis of previously defined events have been implemented and tested. The pairwise comparisons on data from different patients are conducted. For every pair of individuals, this pairwise comparison resulted in a confusion matrix. Eventually, as per the consultation with the expert, the individuals with outliers are classified using efficient classification technique. Thus the data are classified using the hybrid SVM-decision tree classification technique in this work. The entire process is verified by using MATLAB Simulink. The medical dataset representing the list of indian liver patient is considered and the expert criteria and technique comparison by the confusion matrix is given in Table1.

Table 1 Confusion Matrix of Proposed Hybrid method

		Actual Result	
		Other patient	**Liver patient**
Predicted Result	**Other patient**	42	4
	Liver patient	5	95

Table 2 Confusion Matrix of SVM method

		Actual Result	
		Other patient	**Liver patient**
Predicted Result	**Other patient**	36	10
	Liver patient	8	92

This dataset contains 416 liver patient records and the record of 167 patient with some other diseases.The dataset has collected from test samples in North east Andhra Pradesh. The class label ‘is patient' has been used

to classify the dataset into classes (liver patient or not). There are 441 male patient records and 142 female patient records in this dataset. Any patient who is over the age of 89 is referred to as being "90". In this dataset, 80% of data are used for training and 20% of data are used for testing. However, the confusion matrix clearly shows that the accuracy obtained with the proposed classification model is high as it exactly define the number of liver patient and other patient data available in the dataset. Thus the value of 'true positive' and 'false negative' is high which clearly shows that the classification accuracy of the proposed hybrid approach is high.

Table 3 Comparison of Result with Liver patient dataset

Indicator	Proposed method	Exisitng method
Precision	90.0	87.0
Recall	92.5	90.2
Specificity	98.6	95.1
Accuracy	98.2	95.7
Processing time (sec)	40	30

Table 3 reveals that the proposed hybrid classification model gives better performance result compared to the existing SVM approach in terms of precision, accuracy, recall, specificity and processing time. Thus, the performance indices determines that the overall efficiency of the hybrid approach is good. Here, the hybrid classification approach is tested on smaller dataset for our understanding and the behavior of the proposed model vary when applied to larger dataset. However, the proposed classification approach has higher accuracy for all datasets. The Graph 1 depicts the comparison result of performance indicators of existing and proposed classification model with liver patient dataset.

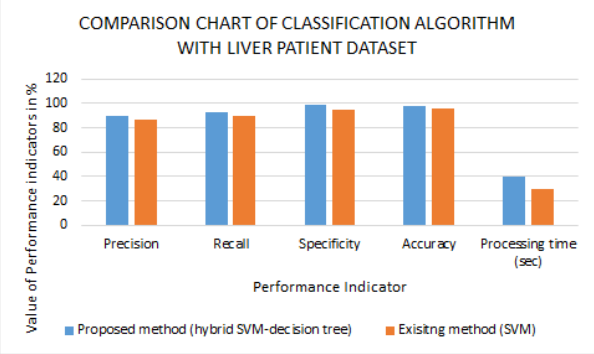

Figure 5 SVM and SVM-Decision tree classification algorithm comparison chart with liver patient dataset

5 CONCLUSION

Data mining is an iterative method in which the mining process is optimised and new data is incorporated to produce more effective results. Data mining satisfies the need for effective, adaptable and scalable data analysis. The suggested S^3FCM clustering algorithm aids in improving the overall efficiency of the data mining process by effectively grouping the similar data and reducing the risk of labelled samples. The proposed classification algorithm is verified in this paper which reduces the size of training dataset. Also, the hybrid classification approach has easy implementation, simple concept and it is faster than conventional SVM training approach. It also collects the data pattern and offers enough information to achieve a successful result. Experimental result shows that the proposed method is scalable for broad data classification while maintaining high classification accuracy and effectiveness.

References

[1] Xueyuan Gong, Liansheng Liu, Simon Fong, Qiwen Xu, Tingxi Wen, Zhihua Liu, "Comparative Research of Swarm Intelligence Clustering Algorithms for Analyzing Medical Data", IEEE Access, Vol. 7, pp. 137560 – 137569, 2019.

[2] Jianqiang Li, Xiyue Tan, Xi Xu, Fei Wang, "Efficient Mining Template of Predictive Temporal Clinical Event Patterns From Patient Electronic Medical Records", IEEE Journal of Biomedical and Health Informatics, Vol. 23, no. 5, pp. 2138 – 2147, 2019.

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

[4] Altug Akay, Andrei Dragomir, Björn-Erik Erlandsson, "Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care", IEEE Journal of Biomedical and Health Informatics, Vol. 19, no. 1, pp. 210 – 218, 2015.

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

[6] Ying Yang, Tao Chen, "Analysis and Visualization Implementation of Medical Big Data Resource Sharing Mechanism Based on Deep Learning", IEEE Access, Vol. 7, pp. 156077 – 156088, 2019.

[7] Ahmed Banimustafa, Nigel Hardy, "A Scientific Knowledge Discovery and Data Mining Process Model for Metabolomics", IEEE Access, Vol. 8, pp. 209964 – 210005, 2020.

[8] Foued Saâdaoui, Pierre R. Bertrand, Gil Boudet, Karine Rouffiac, Frédéric Dutheil, Alain Chamoux, "A Dimensionally Reduced Clustering Methodology for Heterogeneous Occupational Medicine Data Mining", IEEE Transactions on NanoBioscience, Vol. 14, no. 7, pp. 707 – 715, 2015.

[9] Yan Yang, Hao Wang, "Multi-view clustering: A survey", IEEE Transactions on NanoBioscience, Vol. 1, no. 2, pp. 83 – 107, 2018.

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

[11] Dong Huang, Chang-Dong Wang, Jian-Huang Lai, "Locally Weighted Ensemble Clustering", IEEE Transactions on Cybernetics, Vol. 48, no. 5, pp. 1460 – 1473, 2018.

[12] Dal-Jae Yun, In Il Jung, Haewon Jung, Hoon Kang, Woo-Yong Yang, In Yong Park, "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.

[13] Siwei Wang, Miaomiao Li, Ning Hu, En Zhu, Jingtao Hu, Xinwang Liu, Jianping Yin, "K-Means Clustering With Incomplete Data", IEEE Access, Vol. 7, pp. 69162 – 69171, 2019.

[14] Junjie Wu, Hongfu Liu, Hui Xiong, Jie Cao, Jian Chen, "K-Means-Based Consensus Clustering: A Unified View", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, no. 1, pp. 155 – 169, 2015.

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

[16] Weiyu Huang, Alejandro Ribeiro, "Hierarchical Clustering Given Confidence Intervals of Metric Distances", IEEE Transactions on Signal Processing, Vol. 66, no. 10, pp. 2600 – 2615, 2018.

[17] Jacek M, Leski, Robert Czabański, Michal Jezewski, Janusz Jezewski, "Fuzzy Ordered c-Means Clustering and Least Angle Regression for Fuzzy Rule-Based Classifier: Study for Imbalanced Data", IEEE Transactions on Fuzzy Systems, Vol. 28, no. 11, pp. 2799 - 2813.

[18] Tien-Loc Le, 2019, "Fuzzy C-Means Clustering Interval Type-2 Cerebellar Model Articulation Neural Network for Medical Data Classification", IEEE Access, Vol. 7, pp. 20967 – 20973, 2019.

[19] Dan Zhu, Yue Li, Chao Zhang, "Automatic Time Picking for Microseismic Data Based on a Fuzzy C-Means Clustering Algorithm", IEEE Geoscience and Remote Sensing Letters, Vol. 13, no. 12, pp. 1900 – 1904, 2016.

[20] Mohammed Ali, Ali Alqahtani, Mark W, Jones, Xianghua Xie, "Clustering and Classification for Time Series Data in Visual Analytics: A Survey", IEEE Access, Vol. 7, pp. 181314 – 181338, 2019.

[21] Zhiwen Yu, Hantao Chen, Jiming Liu, Jane You, Hareton Leung, Guoqiang Han, "Hybrid k -Nearest Neighbor Classifier", IEEE Transactions on Cybernetics, Vol. 46, no. 6, pp. 1263 – 1275, 2016.

[22] Edson Farias de Oliveira, Maria Emília de Lima Tostes, Carlos Alberto Oliveira de Freitas, Jandecy Cabral Leite, "Voltage THD Analysis Using Knowledge Discovery in Databases With a Decision Tree Classifier", IEEE Access, Vol. 6, no. 6, pp. 1177 – 1188, 2018.

[23] Hongyan Cui, Yazhou Wang, Guangsheng Li, Yongcan Huang, Yong Hu, "Exploration of Cervical Myelopathy Location From Somatosensory Evoked Potentials Using Random Forests Classification", IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 27, no. 11, pp. 2254 – 2262, 2019.

[24] Jiangtao Ma, Yaqiong Qiao, Guangwu Hu, Yongzhong Huang, Arun Kumar Sangaiah, Chaoqin Zhang, Yanjun Wang, Rui Zhang, "De-Anonymizing Social Networks With Random Forest Classifier", IEEE Access, Vol. 6, pp. 10139 – 10150, 2017.

[25] Mingchen Feng, Jiangbin Zheng, Jinchang Ren, Amir Hussain, Xiuxiu Li, Yue Xi, Qiaoyuan Liu, "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data", IEEE Access, Vol. 7, pp. 106111 – 106123, 2019.

[26] Sunil Kumar, Maninder Singh, "A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem", Big Data Mining and Analytics, Vol. 2, no. 4, pp. 240 – 247, 2019.

[27] Gaspard Harerimana, Beakcheol Jang, Jong Wook Kim, Hung Kook Park, "Health Big Data Analytics: A Technology Survey", IEEE Access, Vol. 6, pp. 65661 – 65678, 2018.

[28] Weidong Liu, Caixia Qin, Kun Gao, Heng Li, Zuen Qin, Yafei Cao, Wen Si, "Research on Medical Data Feature Extraction and Intelligent Recognition Technology Based on Convolutional Neural Network", IEEE Access, Vol. 7, pp. 150157 – 150167, 2019.

[29] Xuyang Yan, Mohammad Razeghi-Jahromi, Abdollah Homaifar, Berat A. Erol, Abenezer Girma, Edward Tunstel, "A Novel Streaming Data

Clustering Algorithm Based on Fitness Proportionate Sharing", IEEE Access, Vol. 7, pp. 184985 – 185000, 2019.

[30] Bharath K, Samanthula, Yousef Elmehdwi, Wei Jiang, "k-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, no. 5, pp. 1261 – 1273, 2015.