Health Informatics Journal

Regular Article

Health Informatics Journal
2020, Vol. 26(1) 539–562

DOI: 10.1177/1460458219839629
journals.sagepub.com/home/jhi
SAGE

A novel optimized initial cluster center and enhanced objective function: Medical diagnosis through classification

Binay Subedi, Abeer Alsadoon and PWC Prasad
Charles Sturt University, Australia

Omar Hisham Alsadoon
Al Iraqia University, Baghdad, Iraq

Sami Haddad
Greater Western Sydney Area Health Services, Central Coast Area Health, Australia

Ahmad Alrubaie
University of New South Wales, Sydney, Australia

Abstract

Medical diagnosis through classification is often critical as the medical datasets are multilabel in nature, that is, a patient may have more than one health condition: high blood pressure, obesity, and diabetes. The aim of this article is to improve the accuracy and performance of multilabel classification using multilabel feature selection and improved overlapping clustering method. The proposed system consists of Optimized Initial Cluster Centers and Enhanced Objective Function technique to reduce the number of iterations in the clustering process thereby improving the clustering performance and to improve the clustering accuracy which will result in improving the accuracy and performance of multilabel classification. Ratios of clustering distance to class distance and execution time are used as the evaluation metric for accuracy and total execution time is used as the evaluation metric for performance. Based on the different combination with the number of labels, attributes, instances, and number of clusters, different values of accuracy and performance are obtained. The results on all 10 datasets show that the proposed technique is superior to the current technique. Furthermore, on average, the proposed technique has improved the classification accuracy by 5%–7%. Furthermore, the performance of new technique is improved by decreasing the processing time by 0.5–1 s on average. The proposed system targets on improving the accuracy and performance of the multilabel classification for medical diagnosis, which consists of multilabel feature selection and enhanced overlapping clustering technique. This study provides an acceptable range of accuracy with improved processing time, which assists the doctors in medical diagnosis (high blood pressure, obesity, and diabetes) of patients.

Corresponding author:
Abeer Alsadoon, School of Computing and Mathematics, Charles Sturt University, Sydney, NSW 2010, Australia.
Email: aalsadoon@studygroup.com

Keywords
feature extraction, fuzzy C-means, knowledge discovery, multilabel classification, overlapping clustering

Introduction

Medical diagnosis to a patient using multilabel classification technique is the emerging area of study within the field of medical science as a patient may have more than one health condition: high blood pressure, obesity, and diabetes.[1] In the past, doctors used manual and traditional methods for medical diagnosis to a patient. All the medical reports, including blood report, urine report, and so on, were collected, and the diagnosis was done manually.[2] A few years back, the diagnosis was automated with multilabel classification based on hard division clustering or nonoverlapping clustering, which can assign the patient to only one class of disease. However, a patient may have multiple diseases at the same time. To overcome the limitations of traditional procedure, the multilabel classification used for medical diagnosis is based on overlapping clustering technique where a patient can be assigned to multiple classes of disease. The latest such technology is multilabel classification based on fuzzy C-means (FCM).[3] FCM is an overlapping clustering technique that allows an instance to belong to more than one cluster in multilabel datasets. Medical datasets are multilabel in nature, and multilabel classification with FCM will be advantageous for medical diagnosis purpose.

FCM clustering technique provides a significant advantage for the medical field in terms of diagnosing critical diseases at the same time; for example, a patient may have high blood pressure, obesity, and diabetes at the same time. FCM-based multilabel classification has been successfully implemented at the field of medical science, bioinformatics, text, and image classification. However, it is still the subject of research as the FCM is sensitive to the initial cluster centers[4] and the FCM's objective function uses magnitude-based Euclidean distance, to calculate similarity between the data, which is sensitive to outliers and noise and may not cover overall similarity between the data; thus, the FCM still has limitations in similarity metric, high number of iterations, low accuracy, and high processing time. The best technique can be defined as the one which can enhance the FCM in terms of accuracy, lower iterations, improve similarity measures, and produce high performance.

Current studies of FCM-based multilabel classification in medical diagnosis and bioinformatics use a range of techniques and algorithms to improve accuracy and performance of clustering and classification results. The maximum generated result for accuracy is 0.1983 in terms of hamming loss (which is the ratio of the number of erroneous labels to the total number of labels), and performance is 0.6745 in terms of the micro F1 score.[3] Such greater hamming loss value (low accuracy) may result in failure of the medical diagnosis. Furthermore, the low value of micro F1 score may delay the process of diagnosis.

The purpose of this article is to increase the accuracy and performance of overlapping clustering and multilabel classification techniques. The FCM algorithm may increase the processing time through higher numbers of iterations as it is sensitive to initial cluster centers. In addition, the Euclidean distance used in FCM may not capture the overall similarity between data points, which may result in low accuracy. This research proposes an Improved FCM with Optimized Initial Cluster Centers and Enhanced Objective Function (OICCaEOF) technique to decrease the number of iterations thereby increasing performance and to improving accuracy.

Literature review

The literature reviews are divided into three different sections based on feature extraction, clustering, and classification. And, the last section consists of state of art solution.

Feature extraction

Pacheco et al.[5] investigated the performance of un-supervised feature selection approach in terms of classification accuracy. This is done using Attribute Clustering Algorithm using Rough Set (ACARS) theory that leads the execution time of 0.06211 s with a classification accuracy of 89% for LSVT dataset having 140 attributes. This is an improvement over the current solution. In Hierarchical UnSupervised Feature Selection (HUSFS) with an execution time of 14.8699 s, ACARS uses hard clustering technique (k-means) which is unsuitable for multilabel datasets because it can assign an object to a single cluster. Therefore, it is unsuitable for the proposed multilabel classification system.

Saha et al.[6] investigated the effectiveness, in terms of Minkowski Score (MS), of a feature selection–based clustering technique for classification accuracy in the bio informatics area. This is done using multi-objective clustering as well as feature selecting technique called FeaClusMOO technique that led to MS of 0.31 in cancer dataset having 683 records and 9 features. While this is an improvement over the classical k-means clustering with MS values of 0.37, as the lower score is better,[7] FeaClusMOO includes a lot of computations, and its processing time is significantly high. It would, therefore, seem that the combination of techniques does not offer further possibilities for improvement in the proposed system.

Karimi and Farrokhnia[8] investigated the predictive performance in terms of classification accuracy in cancer detection. This is done using the combination of genetic algorithm and linear discriminant analysis (GA-LDA) based on dimension reduction technique that lead to the predictive accuracy of 94% in leukemia dataset. While this is an improvement over the current solutions—K-Nearest Neighbor (K-NN) and Classification and Regression Trees (CART) with a predictive accuracy of 85%—GA-LDA based on dimension reduction technique uses principal component analysis (PCA) without providing an optimal way to select threshold variance in extracting useful principal components, which may result in loss of important information.[9] Therefore, this technique does not provide any assistance to the proposed multilabel classification system.

Das et al.[10] investigated the performance of feature selection in terms of classification accuracy. This is done using hybrid feature selection algorithm using graph-based Unified Feature Association Map (U-FAM) technique that led to an average accuracy of 70% in 17 different datasets. While this is an improvement over the current solutions—correlation-based feature selection (CFS) and minimum redundancy maximum relevance (MRMR)[11] with an average accuracy of 60% and 61%, respectively—U-FAM is based on graph-theoretic principles of maximal independent set and minimal vertex cover to derive a feature subset which is NP-complete in nature and difficult to select. Therefore, it is unsuitable for the proposed multilabel classification system.

Peng and Liu[3] investigated the accuracy in terms of hamming loss and performance in terms of F1 score of multilabel feature selection in multilabel datasets. This is done using multilabel feature selection method called mutual information (MI) technique that led to hamming loss value (less is better) of 0.1986 and F1 score value (more is better) of 0.6650 in emotion dataset. This is an improvement over the current feature selection method called max-dependence criterion with hamming loss value of 0.2034 and F1 score value of 0.6532. Many studies have revealed that MI-based feature selection techniques are effective because the MI can handle different types of attributes, can measure nonlinear relations between variables, and does not make any assumptions.[12] This is, therefore, of significance for the proposed multilabel classification system.

Clustering

Hu and Pan[13] investigated the appropriate way for clustering of data from Really Simple Syndication (RSS). This is done using the hierarchical clustering (HC) method that led to two-dimensional (2D) format description of 25 RSS demonstrating the distance between different RSS. This is an improvement over the current solution using dendrogram where relationships among the fed are

not clear. Although results are reproducible in HC, the time complexity is quadratic, that is, $O(n^2)$.[14] Therefore, it is unsuitable for the proposed multilabel classification system.

Javadi et al.[15] investigated the performance in terms of correlation coefficient and accuracy in terms of Student t-value of clustering in groundwater vulnerability assessment. This is done using k-means clustering algorithm that led to the correlation coefficient of 61%, 54%, and 58%, respectively, for nitrate, chloride, and total dissolved solids (TDS), and Student t-value of 0.001, 0.002, and 0.002, respectively, for nitrate, chloride, and TDS. While this is an improvement over the current solution DRASTIC[16] with the correlation coefficients of 46%, 32%, and 53%, respectively, for nitrate, chloride, and TDS, and Student t-value of 0.002, 0.005, and 0.002, respectively, for nitrate, chloride, and TDS, k-means is non-overlapping clustering technique and is not suitable for multilabel datasets. Therefore, such non-overlapping clustering technique does not provide any assistance to the proposed multilabel classification system.

Shakya and Makwana[17] investigated the accuracy of clustering results in terms of classification accuracy in intrusion detection. This is done using k-means++ algorithm that led to an average accuracy of 96.922% in Knowledge Discovery and Data Mining (KDD) cup dataset. While this is an improvement over the current solution using a combination of simple k-means clustering and Support Vector Machine Classification (KMSVM) technique with an average accuracy of 84.062%, k-means++ is just an extension of k-means and does not support overlapping clustering.[18] It would, therefore, seem that the combination of algorithms does not offer further possibilities for improvements in the proposed multilabel classification system.

Hamdi et al.[19] investigated the accuracy in terms of error rate of clustering results in a medical field. This is done using the Fuzzy Ant System (F-ASClass) algorithm that combines the Ant System (AS) algorithm[20] and the FCM clustering algorithm that led to error rate of 0.476 for thyroid dataset. While this is an improvement over the current solutions' k-means and FCM with an error rate of 0.774 and 0.847, respectively, F-ASClass performance is still unacceptably low because of the high computational complexity. Therefore, it is unsuitable for the proposed multilabel classification system.

Peng and Liu[3] investigated the accuracy in terms of hamming loss and performance in terms of F1 score of overlapping clustering in multilabel datasets. This is done using overlapping clustering algorithm called FCM that lead to hamming loss value (less is better) of 0.1899 and F1 score value (more is better) of 0.6486 in yeast dataset. While this is an improvement over the non-overlapping clustering algorithm k-means with hamming loss value of 0.1994 and F1 score value of 0.6332, FCM is sensitive to initial cluster centers, noise, and outliers, which may affect accuracy and performance of clustering results as the initial cluster centers are selected randomly.[4] As this solution provides overlapping clustering technique, it would be significant to the proposed multilabel classification system.

Yanli and Jizhu[4] investigated the performance and effectiveness of clustering results based on cluster centers obtained in cab mobility area. This is done using density and grid-based clustering technique[21] to optimize the initial cluster center of FCM that led to a stable cluster center of (37.63, −122.40) in five different runs for first cluster. This is an improvement over the current overlapping clustering algorithm FCM with varying cluster centers of (37.62, −122.40), (37.63, −122.39), (37.63, −122.40), (37.64, −122.41), and (37.65, −122.40) in five different runs for first cluster. The experimental results show that the proposed technique optimizes the initial cluster centers of FCM and greatly reduces the number of iterations required for FCM, thereby enhancing the performance and improving the accuracy of FCM clustering results. This is, therefore, of significance for the proposed multilabel classification system.

Classification

Sheydaei et al.[22,23] investigated the classification accuracy of multilabel classification in classifying documents. This is done using Bit-priori Association Classification Algorithm (BACA) that led

to the classification accuracy of 83% for cultural class. While this is an improvement over the current solution—K-NN classifier with a classification accuracy of 80.77%—BACA classification accuracy is still critical because important association rules can be missed as there is no optimized process for selecting a threshold for rules.[24] Therefore, this technique does not provide any assistance to the proposed multilabel classification system.

Barak and Gelbard[25] investigated the classification accuracy in terms of Harmonic Clustering Score (HCS) in the datasets where the target attributes are known in advance. This is done using the combination of Classification by Clustering (CBC) and Bounded Rationality Theory (BRT) that led to the HCS values of 99.1% for Acute Inflammations datasets with two classes. While this is an improvement over the current solution Decision Tree[26] with HCS value of 92.1%, CBC and BRT have some limitations that do not support real-time applications, testing is not done for excessively large datasets, and the threshold of saliency requires manual input. Therefore, it is not suitable for the proposed multilabel classification system.

Chen et al.[27] investigated the classification accuracy of objects that are on the border of two clusters in bioinformatics area. This is done using semi-supervised classification based on clustering adjusted similarity (SSC-CAS) classification technique that led to the classification accuracy of 96% for breast dataset. While this is an improvement over the current solution Linear Neighborhood Propagation (LNP)[28] classification method with a classification accuracy of 85.83%, SSC-CAS can only assign the object to one class, which is unsuitable for multilabel datasets. Therefore, this technique does not provide any assistance to the proposed multilabel classification system.

Zhun-Ga et al.[29] investigated the classification accuracy in terms of error rate of objects that is in the border of two classes in bio informatics area. This is done using Harmonic Clustering Score (HCS) which combines K-NN and self-organizing map (SOM) that lead to an error rate of 2.51% for breast dataset. While this is an improvement over the current solution Belief-based K-Nearest Neighbor (BK-NN)[30] classification method with a classification accuracy of 2.85%, HCS can only assign an object to one class, which is unsuitable for multilabel datasets. Therefore, this technique does not provide any assistance to the proposed multilabel classification system.

Peng and Liu[3] investigated the accuracy in terms of hamming loss and performance in terms of F1 score of multilabel classification in multilabel datasets. This is done using Overlapping Clustering Based Multilabel Classification (OCBMLC) which combines FCM clustering and Multilabel Classification based K-Nearest Neighbor (ML-KNN) classifier that lead to hamming loss value (less is better) of 0.1983 and F1 score value (more is better) of 0.6745 in emotion dataset. This is an improvement over the current classifier using a combination of hard clustering technique k-means and ML-KNN with hamming loss value of 0.2021 and F1 score value of 0.6562. OCBMLC can classify a single instance into multiple classes as it uses overlapping clustering algorithm. This is, therefore, of significance for the proposed multilabel classification system. See the Comparative analysis Table in Appendix 1.

State of art

This part describes the overall features and techniques presented in the multilabel classification method proposed by Peng and Liu.[3] The recognized features of this method are highlighted in blue (Figure 1) and limitations are highlighted in red (Figure 1). The method proposed by Peng and Liu[3] provided a way in which multilabel classification can be performed through overlapping clustering. This method provides a feature selection process based on MI and greedy selection (GS). This method consists of two main stages (Figure 1) which are feature selection and overlapping clustering (Table 1).

Feature selection stage. In this stage, the original dataset, for example, yeast dataset, is evaluated, and a set of useful and important features are extracted based on interaction information between

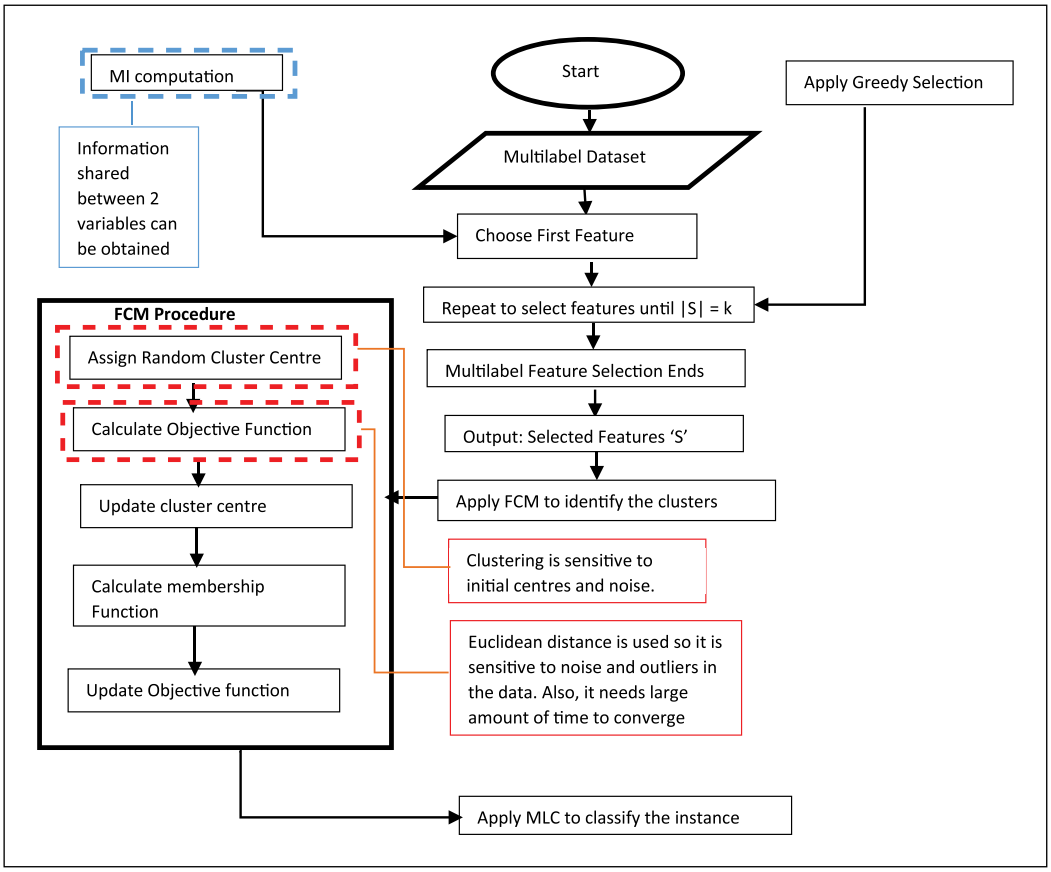

Figure 1. (a) The workflow of multilabel classification using FCM clustering[3] and (b) the advantages (in blue) and limitations (in red) of the current multilabel classification technique using FCM.

attributes and labels as shown in Figure 1. MI is calculated between the feature set (original set of features in dataset) and the output class set (class labels in the dataset) which determines the amount of information shared by the particular feature and particular class labels. The feature that maximizes the MI is selected. GS is used to select the required number of features.

Overlapping clustering phase. In this stage, FCM clustering technique is used for overlapping clustering purpose. FCM starts by assigning the random points as initial cluster centers. The objective function is calculated using the Euclidian distance (ED) which gives the similarity between cluster center and the data point. FCM is based on minimizing the objective function which is optimized iteratively by updating the membership function and cluster centers. The process stops when the difference between two consecutive objective functions is less than the termination criteria (€) which are defined in advance.

However, there are some limitations in the stages of FCM. In FCM clustering, the initial cluster centers are selected by assigning random points as initial cluster centers.[4] Thus, FCM clustering is sensitive to initial centers and noise. Mathematically, this requires more iterations to get the optimized centers, and the computation time may increase. Also, in calculating the objective function,

Table 1. Pseudo code for state of art fuzzy C-means (FCM) algorithm.

Input: Data set $X = \{x_1, x_2, \ldots, x_n\}$, the number of clusters c and termination criteria €.
Output: cluster center set V.

1. BEGIN
2. Assign random points as cluster centers.
3. At k-step: update the centers vectors C(k) = [cj] with U(k) using equation (4).
4. Update U(k), U(k + 1) membership function using equation (2).
5. Update objective function using equation (3).
6. If $||U(k+1) - U(k)|| <$ € then STOP; otherwise return to step 2.
7. END

ED is used, which is sensitive to noise and outliers in the data.[31] Thus, it affects in the accuracy of the clustering result.

In multilabel classification, one object may belong to multiple classes. However, algorithms based on non-overlapping clustering like k-means do not consider such situations. In contrast, overlapping clustering-based methods consider this situation of assigning each object to more than one class when they handle datasets. Therefore, overlapping clustering method like FCM works well for the clusters where one object may belong to multiple classes (Figure 2).

For multilabel classification, the classification performance and accuracy depend highly on the feature selection process which helps in the refinement of the data and reduction of computational cost during clustering and classification steps. For feature selection processes, MI method is used. MI based feature selection methods are effective and efficient because the MI can handle different types of attributes, does not make any assumptions, and can measure nonlinear relations between variables. MI is the amount of information shared by two variables and is defined in equation (1)[3]

$$I(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log\left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right) \tag{1}$$

where $X \in \{x_1, x_2, \ldots x_n\}$ and $Y \in \{y_1, y_2, \ldots y_m\}$ are discrete variables, I(X, Y) is the MI between two discrete variables X and Y, i and j are two positive integer variables, n is the total number of data points in X, m is the total number of data points in Y, $p(x_i)$ is the probability of X, $p(y_j)$ is the probability of Y, and $p(x_i, y_j)$ is the joint probability of X and Y.

After the multilabel feature selection, overlapping clustering is performed through FCM. Unlike hard clustering (nonoverlapping clustering), here data points are assigned the membership to each cluster center; as a result, probability of belonginess of each data points to all the cluster centers can be obtained. The FCM membership function is defined in equation (2)[3]

$$u_{i,j} = \sum_{t=1}^{c}\left[\left(\frac{\|x_j - v_i\|}{\|x_j - v_t\|}\right)^{\frac{2}{(m-1)}}\right]^{-1} \tag{2}$$

where $u_{i,j}$ is the membership value of the jth object and ith cluster, t is positive integer variable, c is the number of clusters, x_j is an object/data point, V_i is the cluster center of the ith cluster, and m is any real number greater than 1.

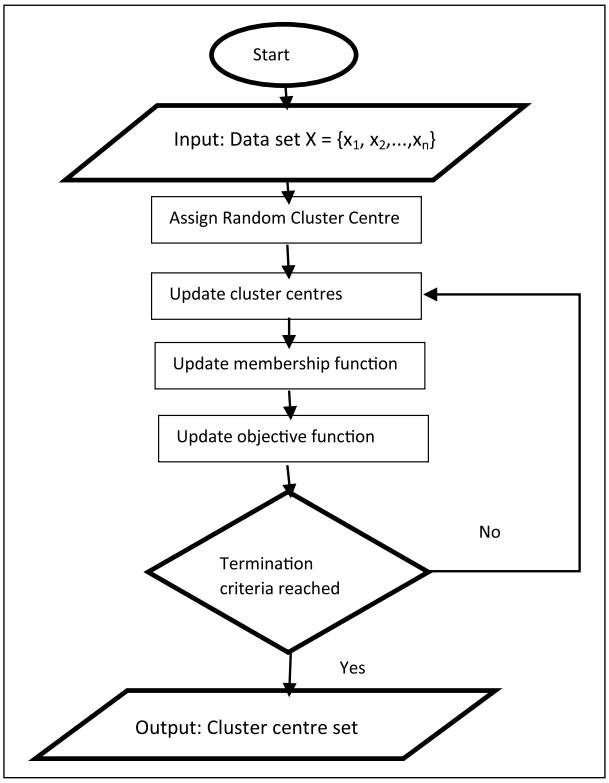

Figure 2. The logical flow of the FCM clustering algorithm used by the current best multilabel classification technique.

The objective function is calculated as shown in equation (3)[3]

$$J_m = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \tag{3}$$

where J_m is an objective function, i and j are two positive integer variables, n is the total number of data points, c is the total number of clusters, u_{ij} is the membership value of the jth object and ith cluster, m is any real number greater than 1, xi is the ith of d-dimensional measured data, *cj* is the d-dimension center of the cluster, and || || is any Euclidean norm expressing the similarity between any measured data and the center.

The cluster center is updated as equation (4)[3]

$$v_i = \frac{\sum_{k=1}^{n} \left(u_{ik}\right)^m x_k}{\sum_{k=1}^{n} \left(u_{ik}\right)^m} \tag{4}$$

where v_i is the ith cluster center, k is positive integer variable, n is total number of data points, u_{ik} is the membership value of kth object and ith cluster, m is any real number greater than 1, and X_k is the object/data point in the cluster.

The experimental results show that the proposed model (OCBMLC) can increase predictive performance compared to a non-overlapping clustering framework like k-means.

But still, it is still possible to increase the predictive performance and accuracy by optimizing the initial cluster center and enhancing the objective function.

Proposed solution

Various existing feature selection, clustering, and classification algorithms and techniques have been reviewed in depth for this research. Based on these existing solutions, it is found that accuracy and performance of feature selection, clustering, and classification approaches are the prime factors. The multilabel classification based on feature selection and overlapping clustering proposed by Peng and Liu[3] is used as the foundation for the proposed solution (Table 2). However, there are some limitations in this solution as it follows the classical FCM algorithm as mentioned in the previous section. Furthermore, a range of other solutions were analyzed, and a solution proposed by Yanli and Jizhu[4] was found, which provides a way to optimize the initial cluster center in FCM algorithm. This article proposes a way to overcome the shortcomings of FCM clustering. Yanli and Jizhu[4] is proposing an approach for initial center optimization method based on density and grid to avoid the sensitivity of FCM to initial centers.

Here, the entire data are partitioned using the grid-based method. The cell center (CC) is given by equation (5)[4]

$$CC = \min_{x_i \in cell} \sum_{j=1}^{n} \left\| x_i - x_j \right\| \tag{5}$$

where CC is cell center, x_i and $x_j \in$ cell, n is the number of data in the cell, i and j are two positive integer variables, and $\| \ \|$ is the Euclidean norm between x_i and x_j.

The similarity (SL) between the cell is given by equation (6)[4]

$$SL\left(cell_i, cell_j\right) = \frac{\left\| CC_i - CC_j \right\|}{\max \left\| CC_i - CC_j \right\|} \tag{6}$$

where $SL(cell_i, cell_j)$ is the similarity between $cell_i$ and $cell_j$, CC_i is the cell center of $cell_i$, and CC_j is the cell center of $cell_j$.

Based on the similarity, two cells with maximum similarity are merged until the number of cells is equal to the number of clusters. At this point, the center of the cells becomes the center of the clusters for FCM. Hence, the initial center for FCM is optimized.

Also, the Euclidean distance in FCM objective function can be replaced by cosine similarity (CS) metric[31] to improve the accuracy and decrease the iterations of FCM using the equation (7)[31]

$$CS\left(x,y\right) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} = \frac{x.y}{\left\| x \right\| \left\| y \right\|} \tag{7}$$

Table 2. Pseudo code of proposed solution (multilabel classification using feature selection and modified fuzzy C-means (FCM)).

Input: Data set $X = \{x_1, x_2, \ldots, x_n\}$, number of features k and the number of clusters c
Output: Feature set S and cluster centers V.

1. Initialize $S = \{\}$
2. Computation of the MI with the output class set
For $\forall x_i \in X, \forall l_i \in L$ compute $\Sigma_{lk \in L} MI(x_i; l_k)$
3. (Choice of the first feature) Find the feature x that maximizes $\Sigma_{lk \in L} MI(x_i; L)$;
set $X \leftarrow X \setminus \{x_i\}$; set $S \leftarrow \{x_i\}$
4. Repeat until $|S| = k$
5. Apply grid-based method to partition S into different "m" cells.
6. For each cell do:
7. Calculate cell center according to equation (5).
8. Calculate similarity between two arbitrary adjacent cells according to equation (6).
9. While $m > c$ do:
10. Merge two cells which have maximum similarity.
11. Calculate similarity between new cell and other adjacent cells.
12. End
13. Calculate the center of final cells and assign initial centers to c clusters.
14. Calculate enhanced objective function according to equation (10).
15. While termination criteria is not reached do:
16. Update cluster center

$$v_i = \frac{\sum_{j=1}^{n} \left(w_{ij}\right)^m x_j}{\sum_{j=1}^{n} \left(w_{ij}\right)^m}$$

17. Update membership function

$$u_{i,j} = \sum_{t=1}^{c} \left[\left(\frac{\left\| x_j - v_i \right\|}{\left\| x_j - v_t \right\|} \right)^{2/(m-1)} \right]^{-1}$$

18. Update objective function
19. End
20. Classify the instance based on multilabel classification model.

where CS(x, y) is the cosine similarity between two vectors x and y, i and j are two positive integer variables, and x_i and y_i are components of vector x and y, respectively.

Proposed equation

An enhanced Fuzzy C-means (EFCM) is proposed that follows the useful feature (feature extraction) of state of art given in Figure 1 and enhances the FCM clustering with initial center optimization to improve the performance and accuracy of overlapping clustering, which is the recognizing feature from Yanli and Jizhu[4] solution. Furthermore, the objective function used in FCM is modified with cosine similarity metric, which overcomes the drawback of ED used in FCM thereby increasing the accuracy of clustering and decreasing the number of iterations required in FCM.[31] The proposed technique will improve the accuracy by decreasing the hamming loss value to

0.14~0.16 compared to 0.1983 in the state of art. Furthermore, the performance will be improved by increasing the micro F1 value to 0.71~0.75 compared to 0.6745 in the state of art.

The initial cluster center (ICC) can be optimized using the equation (8)[4]

$$ICC=\min_{x_i \in cluster} \sum_{k=1}^{n} \left\| x_i - x_k \right\| \tag{8}$$

where ICC is initial cluster center, x_i and $x_k \in$ cluster, n is the number of data in the cluster, and $\|\cdot\|$ is the Euclidean norm between x_i and x_k.

The enhanced objective function is given by equation (9)

$$EJ_m = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} \frac{x_i c_j}{\left\| x_i \right\| \left\| c_j \right\|} \tag{9}$$

where EJ_m is enhanced objective function, i and j are two positive integer variables, n is the total number of data points, c is the total number of clusters, u_{ij} is the membership value of the jth object and ith cluster, m is any real number greater than 1, *xi* is the ith of d-dimensional measured data, and *cj* is the d-dimension center of the cluster.

Area of improvements

The part of which the state of art is enhanced is the optimization of initial center, and the enhancement of the object function used is FCM. To optimize the initial cluster center, we use the grid and density-based clustering technique used in the solution proposed by Peng and Liu.[3,4] The solution proposed by Peng and Liu[3] assigns the cluster centers randomly; thus, the result is sensitive to noise and also the number of iteration will be increased. Furthermore, next improvement in the proposed technique is the enhancement of objective function using cosine similarity metric. This will help in increasing the accuracy of clustering and minimization of iterations.

The proposed multilabel classification process, including FCM with OICCaEOF techniques, consists of two main stages, which are feature selection and overlapping clustering. The feature selection step uses MI and GS approach to extract the important features that contain most of the information needed to generate clustering and classification results. The overlapping clustering uses enhanced FCM to calculate the required cluster centers where the number of clusters are known in advance. The overall components of the proposed multilabel classification, including FCM with OICCaEOF technique, are demonstrated in Figure 3.

For feature selection, MI technique has been chosen followed by the GS method. Many researchers have shown that MI-based feature selection methods are effective and efficient because the MI can handle various types of attributes, does not make any assumptions, and can measure nonlinear relations between variables.[12]

The first contribution in the proposed technique includes the optimized initial cluster center (OICC) in FCM, which helps in reducing the number of iterations of the FCM and enhancing clustering accuracy. With the classical FCM clustering, clustering result is sensitive to initial cluster centers and noise as initial centers were selected randomly. This will reduce the accuracy of the clustering result and also increase the number of iterations in FCM.

Also, another contribution in the proposed solution is to use cosine similarity metric instead of Euclidean distance metric. The Euclidean distance is sensitive to noise and outliers in the data and

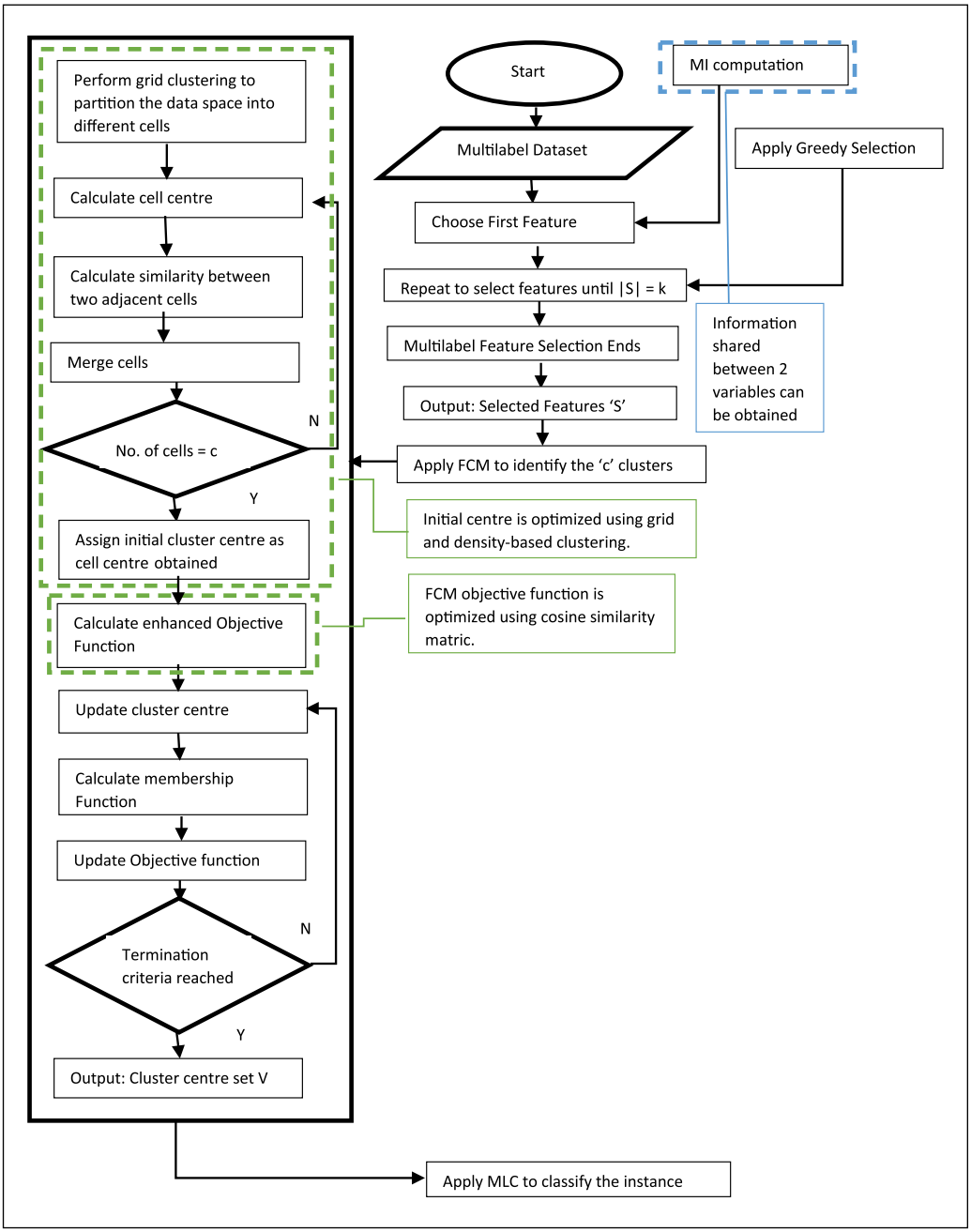

Figure 3. (a) The workflow of multilabel classification using proposed OICCaEOF technique and (b) the inherited features of current best technique (in blue) and features enhanced by OICCaEOF (in green).

takes more time to converge. Thus, the enhanced objective function (EOF) technique of FCM using cosine similarity helps in improving the performance and accuracy as it is based on orientation (angles) rather than magnitude. The advantage of angel-based metric is that it is bound on the

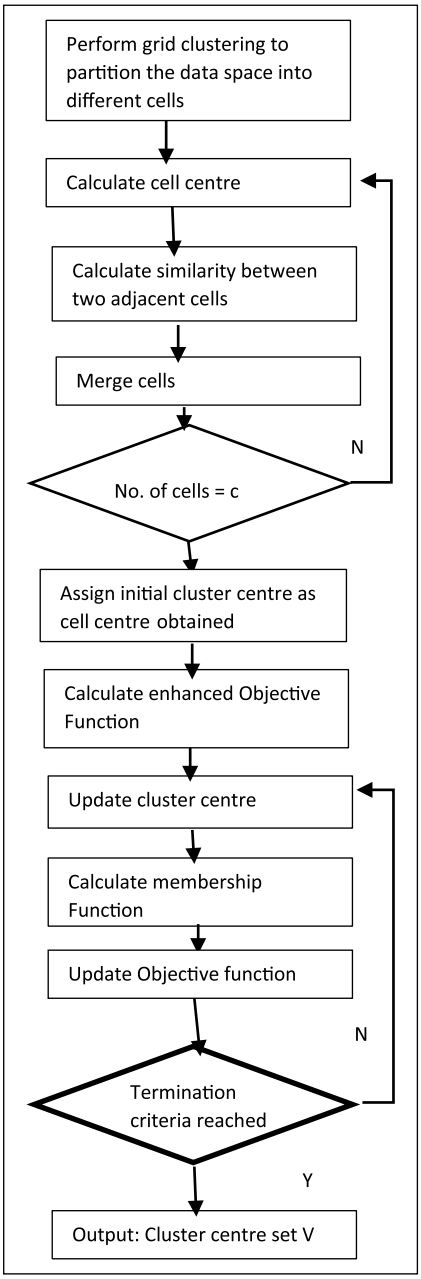

Figure 4. The logical flow of the enhanced FCM clustering algorithm used in proposed OICCaEOF technique.

interval [–1,1]. Furthermore, the distribution of correlation between random vectors becomes narrowly focused around zero as the dimensionality grows in high dimensional dataset. So, the significance of small correlation increases with growing dimensionality.

The proposed OICCaEOF technique is the enhancement of the classical FCM (Figure 4). The enhancement is done by adding one step to optimize the initial cluster center selection process and to modify the objective function. In FCM, the objective function is calculated using Euclidean distance; thus, it takes magnitude into account, but it is not sufficient to determine the similarity between two data in a high dimensional data space. Therefore, orientation of data needs to be significantly considered. Based on this fact, we are proposing an EOF that uses cosine similarity metric. This will increase the accuracy of the clustering result and also help to reduce the number of iterations in FCM thereby enhancing performance.

And, the enhanced objective function is given by equation (10)

$$EJ_m = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m} \frac{x_i c_j}{\|x_i\| \|c_j\|} \tag{10}$$

where EJ_m is enhanced objective function, i and j are two positive integer variables, n is the total number of data points, c is the total number of clusters, u_{ij} is the membership value of the jth object and ith cluster, m is any real number greater than 1, *xi* is the ith of d-dimensional measured data, and *cj* is the d-dimension center of the cluster.

The main limitation of state of art[3] is that it assigns the initial cluster centers to the random points which make the cluster result sensitive to the initial selected points and effects in overall accuracy and performance. The reason of that is when Peng and Liu's solution[3] was run for the different times, different clustering results were produced as different initial cluster centers were assigned in each run. The proposed method will solve this issue by providing the OICC technique based on grid and density clustering. Another limitation of the Peng's solution is that it uses ED to measure the relationship between the data points, but it is magnitude-based not orientation-based, which is unsuitable to capture overall relationship in high dimension dataset. This issue is solved by using orientation-based cosine similarity metric. This will increase the accuracy and performance of the clustering result.

Why OICCaEOF?

The state of art solution used classical FCM, which produced a different result each time the algorithm is run. This occurred because the initial cluster centers were selected randomly. This problem can be solved by providing an effective way to select the OICC in FCM. The proposed OICC technique follows this concept and uses grid and density-based technique to optimize the initial cluster centers. Thus, the accuracy and performance of a clustering result will be enhanced.

The Euclidean distance is magnitude-based and is sensitive to the noise and outlier in the data. It is also unsuitable to cover overall relationship in the high dimensional dataset. Thus, orientation-based cosine similarity can help to solve this issue. The proposed EOF technique replaces the Euclidean distance with cosine similarity metric to obtain the overall similarity between objects in high dimensional datasets. The cosine similarity metric can improve the performance and accuracy of FCM.

Results

Python 3.6 was used in the implementation of proposed technique using 10 samples of multilabel datasets from the different biological and medical domain that requires multilabel classification. The datasets that are taken as samples to test have different label cardinality (number of records

Table 3. Implementation sample of proposed technique.

Dataset	Number of labels	Initial number of attributes	Selected number of attributes	Number of clusters	Classification accuracy
Diabetes	24	33	12	2	89%
				3	87%
				4	83%
				5	88%

Table 4. Comparison of accuracy and processing time between state of art and proposed technique for datasets having labels less than 25.

S.N.	Sample datasets	Labels	Attributes	Instances	Number of clusters	State of art		Proposed	
						Classification accuracy	Processing time (s)	Classification accuracy	Processing time (s)
1	Diabetes	24	33	1364	2	0.77	0.33	0.89	0.29
					3	0.81	0.39	0.87	0.33
					4	0.72	0.43	0.83	0.37
					5	0.69	0.43	0.88	0.38
					Average	0.75	0.40	0.87	0.34
2	Emotion	6	72	593	2	0.88	0.31	0.88	0.29
					3	0.92	0.33	0.95	0.29
					4	0.83	0.33	0.90	0.28
					5	0.87	0.35	0.89	0.31
					Average	0.88	0.33	0.90	0.29
3	Iris	8	77	653	2	0.98	0.25	0.98	0.17
					3	0.79	0.19	0.85	0.22
					4	0.76	0.18	0.80	0.20
					5	0.73	0.26	0.79	0.20
					Average	0.81	0.22	0.85	0.20
4	Yeast	14	103	2417	2	0.88	0.22	0.89	0.19
					3	0.83	0.27	0.88	0.23
					4	0.81	0.27	0.85	0.21
					5	0.81	0.25	0.83	0.23
					Average	0.83	0.25	0.86	0.21
5	Scene	6	294	2407	2	0.89	0.29	0.93	0.24
					3	0.92	0.29	0.93	0.26
					4	0.93	0.28	0.95	0.25
					5	0.91	0.29	0.93	0.25
					Average	0.91	0.29	0.94	0.25

belonging to a particular label) that varies from 2 to 5. Comprehensive samples have been taken including various combinations of labels, attributes, and records (instances). The datasets have been taken from the open source Java library. They are freely available in Mulan library for multi-label learning.[32] The feature extraction and overlapping clustering are performed using Python by importing sklearn python library. The output obtained from overlapping clustering (FCM) is used to perform the multilabel classification. The accuracy of multilabel classification and total processing time has been tested, which is shown in Tables 4 and 5.

Table 5. Comparison of accuracy and processing time between state of art and proposed technique for datasets having labels more than 25.

S.N.	Sample datasets	Labels	Attributes	Instances	Number of clusters	State of art		Proposed	
						Classification accuracy	Processing time (s)	Classification accuracy	Processing time (s)
1	Breast cancer	27	73	1393	2	0.76	0.24	0.83	0.24
					3	0.73	0.30	0.81	0.28
					4	0.65	0.28	0.79	0.28
					5	0.61	0.32	0.80	0.31
					Average	0.69	0.28	0.81	0.27
2	Medical	45	1449	978	2	0.83	0.44	0.86	0.38
					3	0.87	0.39	0.87	0.38
					4	0.78	0.42	0.82	0.34
					5	0.75	0.44	0.80	0.40
					Average	0.81	0.42	0.84	0.38
3	Genbase	27	1186	662	2	0.96	0.29	0.98	0.26
					3	0.92	0.32	0.96	0.25
					4	0.89	0.32	0.93	0.28
					5	0.89	0.31	0.91	0.30
					Average	0.91	0.31	0.94	0.27
4	Corel	374	499	5000	2	0.87	0.46	0.92	0.41
					3	0.91	0.45	0.96	0.39
					4	0.88	0.43	0.92	0.39
					5	0.86	0.45	0.93	0.40
					Average	0.88	0.45	0.93	0.40
5	Eron	53	1001	1702	2	0.91	0.33	0.92	0.31
					3	0.93	0.35	0.95	0.33
					4	0.93	0.33	0.95	0.30
					5	0.92	0.33	0.94	0.31
					Average	0.92	0.34	0.94	0.31

The result is compared with the classification accuracy of the multilabel datasets affected by overlapping clustering stage for the medical diagnosis of diseases that can occur at the same time (e.g. diabetes and high blood pressure) as shown in Table 3.

Samples were compared from the current best technique and the proposed technique. With the help of graphs and the data reports that are shown in Figures 5 and 6, results were compared from the current best technique and the proposed technique. The results from multilabel dataset samples are reviewed in Tables 4 and 5. All samples on the table contain the results obtained when the number of clusters are set at different values ranging from 2 to 5 during the overlapping clustering stage. The results are divided according to the overlapping clustering phase to see the impact on the number of clusters on accuracy of multilabel classification. Here, the results from the sample are presented in terms of accuracy and processing time. Accuracy is calculated in terms of the ratio of cluster distance to class distance, where class distance is the actual mean of distance calculated if the data points are clustered accurately by class and cluster distance is the mean distance calculated based upon the membership value assigned by the overlapping clustering technique to each data point. Comprehensive test has been conducted for

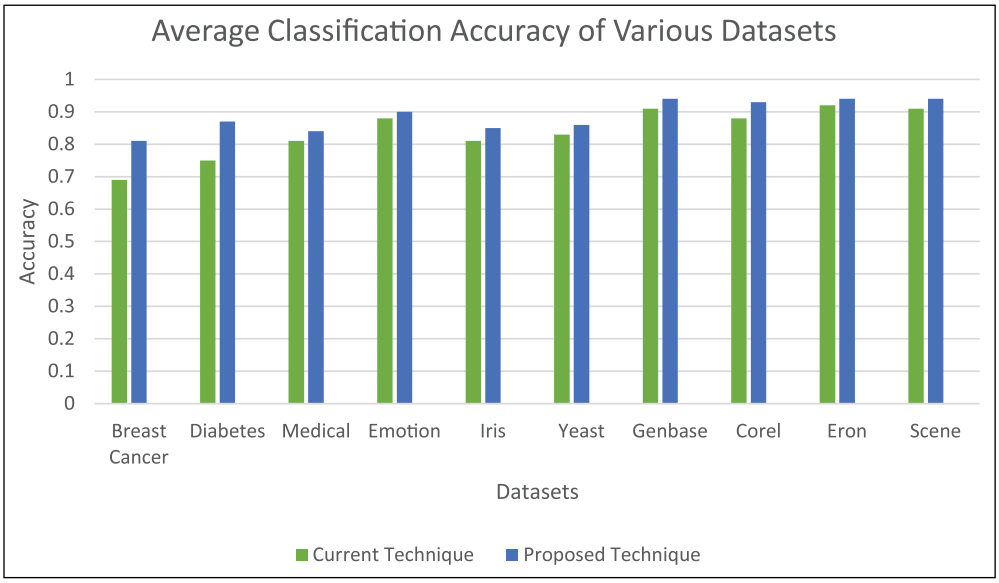

Figure 5. The percent accuracy of multilabel classification for each of 10 multilabel datasets using the current FCM clustering technique (green) versus the proposed OICCaEOF technique (blue).

10 tests; each test has four cases based upon the number of clusters. The accuracy result has been calculated by taking the average result of each test case. Then, the result has been calculated by taking the average for all test cases in 10 datasets.

These results were compared during different stages of multilabel learning in the multilabel feature extraction, overlapping clustering, and multilabel classification as mention above (Table 6). The proposed solution has improved the accuracy of multilabel classification by using the orientation-based cosine similarity metric and also has reduced the processing time by decreasing the number of iterations required to perform overlapping clustering. This will speed up the medical diagnosis process.

Discussion

Results show the difference in accuracy and processing time between the current and the proposed solution with respect to the classification result. For the creation of the multilabel classification system, a range of techniques have been implemented, but continuously refined by the desire for accuracy and lower processing time. The proposed technique improves classification accuracy by 5%~7% in average along with the decrease in the number of iteration required for FCM, thereby reducing the processing time by 0.5~1 s. The accuracy and processing time were accessed by importing sklearn python library. To access the accuracy, the distance from each data point at its cluster center is calculated. For the multilabel data, first, the actual mean of distance is calculated if the data points are clustered accurately by class. And, second, the mean distance is calculated based upon the membership value assigned by the clustering algorithm to each data point. Finally, the ratio of cluster distance to class distance is calculated to obtain the actual classification accuracy. The degree of improvement in processing time is quantified by running the state of art and proposed algorithms and the duration of running each algorithm.

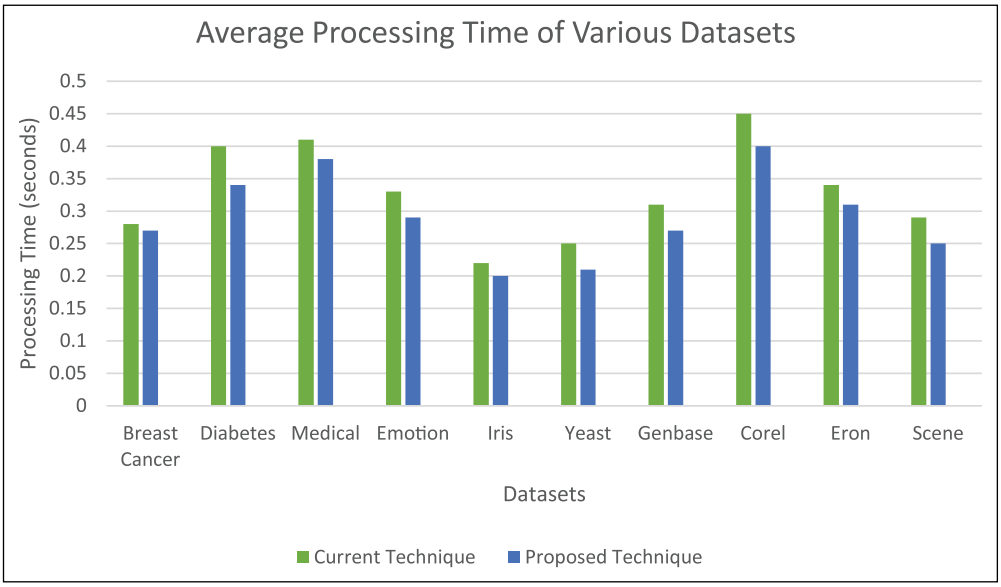

Figure 6. The processing time in seconds of multilabel classification for each of 10 multilabel datasets using the current FCM clustering technique (green) versus the proposed OICCaEOF technique (blue).

Table 6. Comparative results of state of art and proposed technique.

	State of art technique (FCM)	Proposed technique (OICCaEOF)
Applied area	General multilabel classification	Multilabel classification for medical diagnosis
Features	Performs overlapping clustering and assigns each data points membership value to each cluster centers. Euclidian distance is used to calculate similarity between datapoints and cluster centers.	Performs initial cluster center optimization to reduce the number of iterations required for overlapping clustering. Assigns each data points membership value to each cluster centers. Cosine similarity metric is used to calculate similarity between data points and cluster centers.
Equation	a. Initial cluster center, $V = \{v_1, v_2, v_3, \ldots\ldots\ldots v_i\}$ b. Objective Function, $J_m = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m} \left\|x_i - c_j\right\|^2$	a. Optimized initial cluster center, $Ov_i = \min_{x_i \in \text{cluster}} \sum_{j=1}^{n} \left\|x_k - x_j\right\|$ Enhanced Objective Function, $EJ_m = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m} \frac{x_i c_j}{\left\|x_i\right\| \left\|c_j\right\|}$

FCM: fuzzy C-means; OICCaEOF: Optimized Initial Cluster Centers and Enhanced Objective Function.

The proposed system has been tested in Python and has shown to reduce the processing time and increase accuracy. The initial center optimization technique helped to reduce the number of iterations required for FCM clustering process, which as a result decreased the processing time. And, the use of cosine similarity metric helped to capture the orientational

similarity between the data points which as a result improved the accuracy of clustering and classification process. With the overlapping clustering technique, FCM, the membership of each data points to various cluster centers can be obtained, which means if three clusters represent diabetes, high blood pressure, and obesity, then the record of a patient is assigned membership value to each disease. This information can be used by the multilabel classification technique to diagnose the disease of a patient. In conclusion, the combination of techniques greatly improved the multilabel classification process which will provide great assistance in the medical diagnosis of diseases.

Conclusion

Medical datasets are multilabel in nature as same patients may have high blood pressure, obesity, and diabetes at the same time. For the multilabel classification of medical datasets with high accuracy and performance overlapping clustering, technique is essential. FCM, example of overlapping clustering, is capable of assigning an object to multiple clusters while the hard/non-overlapping clustering techniques like k-means and k-means++ can assign an object to only one class. However, FCM still has some drawbacks in terms of accuracy and performance: (a) FCM results are sensitive to initial cluster centers and noise, (b) FCM objective function use magnitude-based Euclidean distance to calculate the similarity between the data points which is not capable to capture the overall similarity as it does not consider orientation of data points. Taking the advantage of these findings, a modified FCM with OICCaEOF techniques has been proposed that will improve the accuracy and decrease the iterations for clustering process, thereby increasing the performance of clustering. As clustering results are input for classification, this helps in increasing the performance and accuracy of multilabel classification.

The proposed technique has been shown (in Python) to maximize the overlapping clustering accuracy, thereby maximizing the classification accuracy which is critical for medical diagnosis. Some solutions have been proposed to address this issue, but, so far, nothing has been able to provide more accurate classification accuracy in an acceptable range. The proposed technique demonstrates capabilities to decrease the number of iterations required by clustering technique and capture maximum similarity between the data points and cluster centers which improves the accuracy of the multilabel classification by 5%~7% and processing speed by 0.5~1 s. Future research will be on finding an appropriate way to automatically find the number of clusters.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Abeer Alsadoon https://orcid.org/0000-0002-2309-3540

References

1. Cerri R, Silva R and Carvalho R. Comparing methods for multilabel classification of proteins using machine learning techniques. In: Guimarães KS, Panchenko A and Przytycka TM (eds) *Advances in*

bioinformatics and computational biology (BSB 2009) (Lecture Notes in Computer Science, vol. 5676). Berlin; Heidelberg: Springer, 2009, pp. 109–120.
2. Jingfeng C. "Medicine in China." In: Selin H (ed.) *Encyclopaedia of the history of science technology, and medicine in non-western cultures*. Dordrecht: Springer, 2008, pp. 1529–1534.
3. Peng L and Liu Y. Feature selection and overlapping clustering-based multilabel classification model. *Math Probl Eng* 2018; 2018: 2814897, https://search-proquest-com.ezproxy.csu.edu.au/docview/1988141419?accountid=10344
4. Yanli S and Jizhu N. Improved FCM algorithm based on initial center optimization method. *J Intell Fuzzy Syst* 2017; 32(5): 3487–3494.
5. Pacheco F, Cerrada M, Sánchez R-V, et al. Attribute clustering using rough set theory for feature selection in fault severity classification of rotating machinery. *Expert Syst Appl* 2017; 71: 69–86.
6. Saha S, Spandana R, Ekbal A, et al. Simultaneous feature selection and symmetry based clustering using multiobjective framework. *Appl Soft Comput* 2015; 29: 479–486.
7. Jiang D, Tang C and Zhang A. Cluster analysis for gene expression data: a survey. *IEEE T Knowl Data En* 2004; 16(11): 1370–1386.
8. Karimi S and Farrokhnia M. Leukemia and small round blue-cell tumor cancer detection using microarray gene expression data set: combining data dimension reduction and variable selection technique. *Chemometr Intell Lab* 2014; 139: 6–14.
9. Wold S, Esbensen K and Geladi P. Principal component analysis. *Chemometr Intell Lab* 1987; 2: 37–52.
10. Das AK, Goswami S, Chakrabarti A, et al. A new hybrid feature selection approach using feature association map for supervised and unsupervised classification. *Expert Syst Appl* 2017; 88: 81–94.
11. Hall M. *Correlation-based feature selection for machine learning*. PhD Thesis, The University of Waikato, Hamilton, New Zealand, 2000.
12. Meyer PE, Schretter C and Bontempi G. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE J Sel Top Signa* 2008; 2(3): 261–274.
13. Hu W and Pan Q. Data clustering and analyzing techniques using hierarchical clustering method. *Multimed Tools Appl* 2015; 74(19): 8495–8504.
14. Kaushik S, Srivastava T, Dar P, et al. An introduction to clustering & different methods of clustering 2016, https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/ (accessed 17 April 2018).
15. Javadi S, Hashemy SM, Mohammadi K, et al. Classification of aquifer vulnerability using K-means cluster analysis. *J Hydrol* 2017; 549: 27–37.
16. Javadi S, Kavehkar N, Mousavizadeh MH, et al. Modification of DRASTIC model to map groundwater vulnerability to pollution using nitrate measurements in agricultural areas. *J Agric Sci Technol* 2011; 13: 239–249.
17. Shakya V and Makwana RR. Feature selection based intrusion detection system using the combination of DBSCAN, K-Mean++ and SMO algorithms. In: *2017 international conference on trends in electronics and informatics (ICEI)*, Tirunelveli, India, 11–12 May 2017.
18. Maitry N and Vaghela D. Survey on different density based algorithms on spatial dataset. *Int J Adv Res Comput Sci Manage Stud* 2014; 2(2): 362–366.
19. Hamdi A, Monmarche N, Slimane M, et al. Fuzzy rules for ant based clustering algorithm. *Adv Fuzzy Syst* 2016; 2016: 8198915.
20. Deneubourg J, Goss JS, Pasteels S, et al. Self-organization mechanisms in ant societies (II): learning in foraging and division of labor. In: Pasteels JM and Deneubourg J (eds) *From individual to collective behavior in social insects* (*Experientia* Supplementum). Basel: Birkhauser, 1987, pp. 177–196.
21. Park NH and Lee WS. Statistical grid-based clustering over data streams. *Sigmod Rec* 2004; 33(1): 32–37.
22. Sheydaei N, Saraee M and Shahgholian A. A novel feature selection method for text classification using association rules and clustering. *J Inf Sci* 2000; 41(1): 3–15.
23. Shi J and Malik J. Normalized cuts and image segmentation. *IEEE T Pattern Anal* 2000; 22(8): 888–905.
24. Tang ZH and Liao Q. A new class based associative classification algorithm. *Int J Appl Math* 2007; 36: 2.

25. Barak A and Gelbard R. Classification by clustering using an extended saliency measure. *Expert Syst* 2016; 33(1): 46–59.
26. Aitkenhead MJ. A co-evolving decision tree classification method. *Expert Syst Appl* 2008; 34: 18–25.
27. Chen X, Lu C, Tan Q, et al. Semi-supervised classification based on clustering adjusted similarity. *Int J Comput Appl* 2017; 39(4): 210–219.
28. Wang B, Zhang L, Wu C, et al. Spectral clustering based on similarity and dissimilarity criterion. *Pattern Anal Appl* 2017; 20(2): 495–506.
29. Zhun-Ga L, Quan P, Jean D, et al. Hybrid classification system for uncertain data. *IEEE T Syst Man Cy: S* 2017; 47: 2783–2790.
30. Liu ZG, Pan Q and Dezert J. A new belief-based K-nearest neighbor classification method. *Pattern Recogn* 2013; 46(3): 834–844.
31. Kannan SR, Devi R, Ramathilagam S, et al. Some robust objectives of FCM for data analyzing. *Appl Math Model* 2011; 35(5): 2571–2583.
32. Mulan: a Java library for multilabel learning, 2018, http://mulan.sourceforge.net/index.html

Appendix 1. Accuracy and processing time based on the variables.

S.N.	Author details	k-means	K-NN	K-means++	FCM	Feature association map	Mutation and crossover	Neighbor information sharing	Entropy and mutual information	Eigen vectors	Rough set theory	BACA	Bit-Apriori	SMO	Hierarchical clustering	Euclidean distance	Pearson correlation coefficient	PCA	SOM	Bounded rationality theory	Kohonen self-organizing map algorithm	SAD criterion
1	Pacheco et al.[5]																					
2	Sheydaei et al.[22]																					
3	Saha et al.[6]																					
4	Hu and Pan[13]																					
5	Javadi et al.[15]																					
6	Barak and Gelbard[25]																					
7																						
8	Wang et al.[28]																					
9	Das et al.[10]																					
10	Peng and Liu[3]																					
11	Shakya and Makwana[17]																					
12																						
13	Yanli and Jizhu[4]																					
14	Zhun-Ga et al.[29]																					
15	Hamdi et al.[19]																					
16	Chen et al.[27]																					

K-NN: K-Nearest Neighbor; FCM: fuzzy C-means; PCA: principal component analysis; SOM: self-organizing map.

NOTE: The Green highlighter of reference 10 indicates this state of art solution has the best features from all the literature in the table. The Orange highlighter of reference 13, indicates some features of this solution can improve the solution in reference 10 by solving it's limitations. The Yellow highlighter indicates the features that have each solution in this table.

Mathematical justification.

State of art	Basic and theories	Proposed solution
Initial cluster centers in FCM Initial cluster centers are randomly assigned as $V=\{v_1, v_2, v_3, \ldots\ldots\ldots v_i\}$ In each iteration, the cluster center is updated as $$v_i = \frac{\sum_{j=1}^{n} (w_{ij})^m x_j}{\sum_{j=1}^{n} (w_{ij})^m}$$ Where, V_i = cluster center of ith cluster X_j = object/data point in the cluster W_{ij} = membership degree n = number of objects in the cluster j = any positive integer variable m = any real number greater than 1	Initial cluster center is optimized by the following equation (2) $$v_i = \min_{x_i \in cluster} \sum_{j=1}^{n} \left\| x_i - x_j \right\|$$ Where, v_i = cluster center of ith cluster $x_i, x_j \in$ cluster i n = the number of data in the cluster i i and j are two positive integer variables $\| \cdot \|$ = the Euclidean norm between x_i and x_j	Initial cluster center is given by, $$Ov_i = \min_{x_i \in cluster} \sum_{j=1}^{n} \left\| x_k - x_j \right\|$$ Where, Ov_i = optimized cluster center of ith cluster $x_i, x_j \in$ cluster i n = the number of data in the cluster i i and j are two positive integer variables $\| \cdot \|$ = the Euclidean norm between x_i and x_j In each iteration, the cluster center is updated as $$v_i = \frac{\sum_{j=1}^{n} \left(w_{ij} \right)^m x_j}{\sum_{j=1}^{n} \left(w_{ij} \right)^m}$$ Where, V_i = cluster center of ith cluster X_j = object/data point in the cluster W_{ij} = membership degree n = number of objects in the cluster j = any positive integer variable m = any real number greater than 1

(Continued)

Appendix 1. (Continued)

State of art	Basic and theories	Proposed solution
Objective function in FCM In FCM, the objective function is given by $J_m = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m} \left\|x_i - c_j\right\|^2$ Where, J_m = objective function i and j are two positive integer variables n = the total number of data points c = the total number of clusters u_{ij} = the membership value of the jth object and ith cluster m = any real number greater than 1 xi = the ith of d-dimensional measured data cj = the d-dimension center of the cluster. $\|\cdot\|$ = the Euclidean norm between x_i and x_j	Cosine similarity metric is given by equation (4) $S_m = \sum_{i=1}^{n}\sum_{j=1}^{c} \frac{x_i c_j}{\left\|x_i\right\| \left\|c_j\right\|}$ Where, S_m = cosine similarity of the jth object and ith cluster i and j are two positive integer variables n = the total number of data points c = the total number of clusters m = any real number greater than 1 x_i = the ith of d-dimensional measured data cj = the d-dimension center of the cluster.	Thus, the enhanced objective function is; $EJ_m = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m} \frac{x_i c_j}{\left\|x_i\right\| \left\|c_j\right\|}$ Where, EJ_m = enhanced objective function i and j are two positive integer variables n = the total number of data points c = the total number of clusters u_{ij} = the membership value of the jth object and ith cluster m = any real number greater than 1 x_i = the ith of d-dimensional measured data cj = the d-dimension center of the cluster.