ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(44), DOI: 10.17485/ijst/2016/v9i44/101286, November 2016* ISSN (Online) : 0974-5645

A Novel Weighted Class based Clustering for Medical Diagnostic Interface

Sunila Godara*, Rishipal Singh and Sanjeev Kumar

Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology,
Hisar - 125001, Haryana, India;
sunilagodara@gmail.com, sandeep1642@gmail.com, sanju.khambra@gmail.com

Abstract

Background/Objectives: Medical Decision Support System (MDSS) is a diagnostic interface which provides computer assisted information retrieval as well as may support excellence decision making, to stay away from human error. Even if human decision-making is frequently most advantageous, but it is poor when there are vast amounts of data to be classified. Also capability and accuracy of decisions will decrease when humans are set into pressure and massive work. Forever there is a need and scope for a better MDSS. **Methods/Statistical Analysis:** Cluster analysis is a method of grouping of objects keen on different groups, has proved to be a valuable tool for identifying co-expressed genes, biologically related groupings of genes and patterns. K-means, Hierarchical and Fuzzy c-means are various clustering techniques have been employed to work as core part of MDSS. **Findings:** Proposed Weighted Class Based Clustering (WCBC) method is dependent on classifying properties of medical data itself. Weights are calculated on the basis of class value consequently increases separability by placing more number of instances of same class in same cluster. In this paper, the clustering algorithms K-means, Hierarchical, Fuzzy and Weighted Class based K-Means are examined for medical domains. Our finding is that on medical domains the Proposed Weighted Class Based Clustering outperforms others. **Application/Improvements:** The application of Proposed Weighted Class Based Clustering on medical datasets gave an insight into predictive ability of Machine Learning in medical diagnosis and there is a wide liberty that proposed approach can be used in RBF Neural Network for center calculation and data base Kernel Learning which is open area of research these days.

Keywords: Clustering, Fuzzy, Hierarchical, K-Means, MDSS, Weighted Class Based Clustering

1. Introduction

Disease prediction and diagnosis is a multifaceted task which necessitates a great deal of practice and knowledge. Predication should be completed to diminish risk. Diagnosis is generally based on signs, symptoms and physical assessment of a patient. Approximately all the doctors predict by learning and experience. The diagnosis of disease is a complicated and tiresome task in medical field. It is a multi-layered matter which may escort to false presumptions and unpredictable effects. Regardless of many years of investigation and millions of dollars of overheads on medical diagnostic systems no single system is widely used till date. But research is growing in this area and technology will put back 80% of what doctors do today[1,2], it will make a new milestone by recalling complex medical information than a hotshot Harvard MD. Technology will compensate cognitive limitations and human deficiencies. There are two main types of Medical Decision Support System:

Knowledge Based MDSS: Knowledge Based MDSS consists of three parts: Knowledge base, inference engine and process to correspond. The knowledge base holds the rules and associations of processed data in the form of IF-THEN rules. The inference engine combines the rules taken from knowledge base through the patient's data[1,3].

Non Knowledge Based MDSS: MDSS's that does not have a knowledge base but useses machine learning, which permits computers to learn from past experiences and to discover patterns in medical data[1,2]. In medical domains,

* *Author for correspondence*

the data is continuously and rapidly changing with time[1]. In addition to the biological factors, this change can also be attributed to evolving lifestyle, increased rate of migration of people from one region to another resulting in greater variety among the collected medical samples, weather, pollution etc. Therefore, despite the advancement of technology and development of numerous Medical Decision Support Systems, there is always need and scope of a better solution.

Clustering process is heart of Medical Decision Support Systems[4]. It is a method of grouping the instances into different parts. Author in[4,5] presented an outline of pattern clustering methods. K-means, Hierarchical, Fuzzy c-means, Fuzzy SOM, density based are various clustering techniques that have been successfully employed in medical classification problems[6] K-means is simple, fast, unsupervised and iterative method proved to be very effective as it outperforms Hierarchical, Fuzzy c-means, Fuzzy SOM, density based methods to produce good clustering results for MDSS[7]. K-means clustering is used to provide input in RBF neural Networks[8]. There are various improvements in K-means algorithm and these methods reduce the impact of isolated points, seed selection, local minima, initial center selection enhances the efficiency of clustering. Author[9] Presented Enhanced Moving K-means algorithm and author[10] proposed improved version of the K-means algorithm based on distance and cluster evenness.

The proposed research work of this paper presents Weighted Class Based Clustering (WCBC) inspired from K-means algorithm. WCBC enhances seprability of data objects on the basis of proposed weighted distance measure in which weights are calculated from class labels and class labels further contains classifying properties of data itself. Consequently maximum objects within clusters are of same type hence will belong to same class. As a result of this error within cluster is minimized along with increased classification accuracy.

This paper comprises five parts: The second part details the related work. The third part demonstrates various clustering algorithms. Fourth part presents proposed WCBC clustering algorithm, the very last part of this paper presents the experimental results and conclusions on various UCI medical data sets.

2. Related Work

Author in[5] proposed energy consumption and hot spot problem of nodes in field of wireless sensor networks have been successfully tackled by the adoption of distributed hybrid energy efficient clustering algorithm. Unequal sized clusters involved that are placed on the distance of cluster head and from the base station. Clustering controls the flow of data and allows aggregation in the clusters. Sensor nodes are combined into clusters to progress network performance. Author in[6] proposed developed a new adaptation known as, the modified global K-means algorithm and is not applicable to big data sets. Algorithms which are found suitable to big data sets can locate local minima of the problem and these local minima can differ from global solutions remarkably when number of clusters increases. Number of clusters, are not known to go forward. So an incremental approach is used in this paper to find a confined solution which is close to universal solution.

Author in[8] proposed an extended version of the CFA algorithm known as ECFA which uses the actual output of the network to make the approximation. ECFA transfers clusters to the area of the input space where the estimation error is larger, consequently trying to homogenously hand out the whole distortion in each and every cluster, producing an enhanced distribution of clusters for the given data. This paper demonstrates that ECFA outperforms the CFA algorithm in context of final estimation and execution time. Author[9] presented modified version of the Moving K–means algorithm to enhance the performance. Author in[11] proposed this paper proposed a new data driven kernel method reliant on classifying properties of the data for SVM. The new kernel contemplates on the resemblance of joined data in classes. Proposed kernel has better classification performance than polynomial and Gaussian kernels. Author in[12] explained that kernel methods are algorithms in which inner product is replaced with an appropriate positive definite function, implicitly do a nonlinear mapping of the input space into a high-dimensional space.

Author in[13] analyzed that K-means clustering algorithm can perform well for compact super-spherical distributions, but fails in case of unidentified shapes and paper proposed a novel density-sensitive distance metric, which can explain characteristic of the distribution and can recognize complex and non-convex clusters. The results are validated on real-world and artificial data sets. Author in[14] proposed Gaussian low-pass frequency domain filtering, Histogram equalization, Image enhancement, Morphological processing, Nonlinear spatial filter, Wavelet filter image enhancement algorithms were applied to 10 ultrasound (US) liver images. Morphological filtering which used concept of class based filtering outperformed other techniques with 76% accuracy. Author in[15] presented medical image segmentation using the mixing of two different multi-kernels in Fuzzy C-Means algorithm. Multi-kernels outperform the single kernels. Author in[16] proposed that Fuzzy association rule mining is better than traditional classifiers but rules produced in it increases exponentially. This work presented an enhanced gain ratio based fuzzy weighted approach for distinct diseases classification on benchmark datasets with increased accuracy. In paper[17] flow identification method based on k-means is proposed to identify network flows based on traffic statistic. This approach adopted improved k-means cluster algorithm (SA-k-means) to classify traffic, and examine the impact factor of cluster. Also, experiment results show SA-k-means method is effective than k-means. Author in[18] presented an approach for damage detection of heart muscle from echocardiography. Statistical pattern recognition and clustering is done to identify the heart muscle damage. Author in[19] presented a comprehensive review on Lung auscultation computer-based respiratory sound analysis techniques, provides precious information concerning the patient's respiratory function. Signal processing methods, classification methods, clustering and statistical methods are employed for the analysis of lung sounds are studied and concluded that computer-based respiratory sound analysis which performs as an immense method to diagnose abnormalities and disorders in the lung. Author in[20] proposed Content Based Medical Image Retrieval (CBMIR) based on Fuzzy clustering is proposed to efficiently retrieve the most relevant medial images. Author in[21] used Classification Tree, Naive Bayes, Random Forest and Support Vector Machine algorithms for training the Prediction System. The research has been conducted using Orange tool and the scores have been evaluated and concluded that Classification Trees are efficient in Prediction. Research work suggested that Hospitals and Health Research Institutes can be uploaded in Cloud and the data analysis can be done comprehensively to get an precise Predictions which can be used crossways many hospitals and research institutes the whole time across the world. Author in[22] proposed learning technique of sub-spaces and evaluates a series of different methods of updated distributions. Further it is concluded that in case when number of properties is high, the proposed hybrid classification based on genetic algorithm can be used as the most excellent method for stable results in error prone environments.

3. Clustering Techniques

This section will discuss K-means, Hierarchical and Fuzzy clustering techniques.

3.1 K-Means Algorithm

Mac Queen in 1967, proposed K-means. It is an iterative partitioning clustering algorithm in which instances are classified into k different clusters to converge at local minima. Euclidean distance is used to calculate distance between each instance and the cluster centers. The algorithm works in two separate steps. K centers are selected in first step. Each data object is placed to the nearest center in second step.

This iterative process continues until the decisive function attains the minimum[4]. Let us consider t the target instance is x, xi allocates the average of cluster Ci , decisive function is calculated as given below[13]:

$$E = \sum_{i=1}^{k} \sum_{x \epsilon ci} \|x - xi\|^{2}$$

E is the sum of the squared error of all instances in database.

The process of k-means algorithm is[4]:

Input: Given k is number of clusters and D= {d1, d2,…dn}containing n data instances is given database.

Output: A set of k clusters.

Steps:

Step 1. Choose k data instances as initial cluster centers from dataset.

Step 2. Find out the distance between each data instance di (1 <i<=n) and all k cluster centers cj(1<=j<=k) .Place data instance di to the nearest cluster.
Step 3. New cluster center is calculated for each cluster.
Step 4. Steps 2and 3 are repeated until there is no change in clusters center.

Computational complexity for locating the optimal solution for n instances and d dimensions is $O(n^{dk+1}\log n)$.Here n is total number of instances to be clustered[4,6].

3.2 Hierarchical Clustering

Hierarchical clustering was described by S. C. Johnson in 1967 continues consecutively by either merging smaller clusters or by dividing larger clusters. This algorithm results dendrogram which is a tree of clusters tells how the clusters are related. By cutting the dendrogram at a required level, a clustering of the data objects into disjoint groups can be done[4].

The process of hierarchical clustering is[6]:

Input: Given k is number of clusters and D = {d1, d2,…dn} containing n data instances is given database.

Output: A set of k clusters.

Steps:
Step 1. Start by putting each instance to a cluster, It means n instances n clusters.
Step 2. Form a single cluster by merging closest group of clusters.
Step 3. Distances between the new cluster and each of the old clusters are calculated.
Step 4. Steps 2 and 3 are repeated until k clusters formed.

The merging criteria for hierarchical clustering is single link, average link and complete link, average and maximum distances between the members of two clusters, correspondingly[4].

3.3 Fuzzy clustering

The FCM algorithm attempts to partition a finite group of instances D = {d1, d2,…dn} containing n data instances in c fuzzy clusters depending upon given criterion. For n instances, the algorithm gives a list of C cluster centers C= {c1, c2,…cn} and a partition matrix , $W = w_{i,j} \in [0,1], \quad i = 1, \dots, n, \quad and\ j = 1 \dots, c$ where each element w_{ij} gives the degree with which an element x_i belongs to cluster C_j. FCM intends to minimize an objective function[5]:

$$\arg c^{min} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{f} \left\| x_i - c_j \right\|^2$$

Objective function contains the membership values w_{ij} and the fuzzifier $f \in R$, with $f \geq 1$ which were absent in case of K-mens, The f determines the degree of cluster fuzziness[11,22].

The process of FCM algorithm clustering is[4-]:

Step 1. Arbitrarily select number of clusters. Randomly choose coefficients for each instance to be a part of the clusters.
Step 2. Repeat until the coefficients' change between two iterations is less than required threshold.
Step 3. Centroid for each cluster and its coefficients f are calculated.
Step 4. Steps 2 and 3 are repeated until k clusters formed.

Intra-cluster variance is minimized and the results depend on the initial picking of weights.

4. Proposed WCBC Clustering Technique

Firstly, it calculates range of values of all attributes within each class and gives maximum and minimum value of each attribute for each class. We have calculated new useful ranges for all attributes by taking maximum values from set of minimum values and taking minimum values from set of maximum. Weights are calculated as shown in Step 5 below. Here importance is given to attributes that distinguishes the class, this makes members of a cluster more similar and more different from non members. The process of WCBC algorithm is:

Input: Number of clusters k, A = {A1, A2, A3...A d} set of d attributes , dataset D = {D1, D2,…Dn} having n data instances and C ={ C1, C2, ...,Cm} be a set of m classes within dataset. V is find function over set of domain.

Output: Set of k clusters.

Steps of proposed algorithm are as follow:
The Euclidean distance between one instance x = (x1 ,x2,…xd) and another instance y = (y1 ,y2 ,…yd) is d(xi, yi). WCBC arbitrarily select k data instances from dataset

D to act as initial cluster centers. A is set of attributes and C is set of classes.

Repeat

Step 1. Calculate range function

$$f(\mathrm{range}) = \bigvee_{i=1}^{C}\bigvee_{j=1}^{A} \mathrm{select}(A_{j\to min}, A_{j\to max})$$

Step 2. Calculate

$$\mathrm{find}(\max, \min) = \bigvee_{j=1}^{A}\bigvee_{i=1}^{C} C_{i\to(min,max)}$$

Step 3. Calculate

$$\mathrm{find}(\max) = \bigvee_{i=1}^{C} C_{i\to max}$$

Step 4. Calculate

$$\mathrm{find}(\min) = \bigvee_{i=1}^{C} C_{i\to min}$$

5 Calculate weight function

$$W_j = \bigvee_{j=1}^{A} = \frac{\sum_i^C \left(C_{i+1\,A_{j\to min}} - C_{i\,A_{j\to min}} + C_{i+1\,A_{j\to max}} - C_{i\,A_{j\to max}}\right)}{C}$$

Step 6. Calculate distance for each data object and reassign it.

$$d(x, y) = \sqrt{\sum_{j=1}^{n} W_j (x_j - y_j)^2}$$

Step 7. For each cluster, cluster center is recalculated until no change in the center of clusters found.

Component of class based vector Wj is the degree corresponding to each feature. Larger value of Wj means more significant the jth feature . When W = (1,1,1….) all feature are of equal importance and space is a hypersphere with radius r. In the original space {dx,y <= r} represents that the axes would be extended or shrunk in accordance with wj and space is hyper-ellipse. Lower value of Wj shows high flattening extent. Computational complexity for finding the optimal solution using the Proposed WDBC clustering having n instances, d dimensions and c number of classes is of order O (n^{dkc+1} log n).

5. Experimental Setup and Results

5.1 Performance Measures

A recognized confusion matrix was achieved to compute accuracy. Confusion matrix depicts the classification results. Table 1 depicts confusion matrix. Correctly classified instances, incorrectly classified instances, Accuracy True Positive Rate, False Positive Rate, Precision, Recall and ROC are various measures used to measure performance[22].

Table 1. Confusion matrix

	Classified as Healthy	Classified as not healthy
Actual Healthy	TP	FN
Actual not healthy	FP	TN

5.2 Performance Evaluation

Various Clustering techniques are evaluated on Diabetes, Lung Cancer, Hepatitis, Liver Disorder, Breast Cancer Wisconsin, Mammographic and Cardiovascular Cleveland Heart disease datasets using WEKA tool. All datasets are downloaded from UCI machine learning repositories. 10 V fold Cross Validation is used to validate results. 10 V-fold cross validation consists in arbitrarily partitioning the existing data into 10 subparts and then training 10 classifiers using all data but one subpart which is always taken different for all 10 classifiers is used for testing the performance of the classifiers. The approximation of the error of the classifier built from the entire data is the average error over the subparts. Table 2 shows Confusion Matrix for various techniques.

5.2.1 Diabetics Dataset

This data set enclosed 9 attributes and 768 instances. Eight attributes are conditional and one attribute is class having values 0 or 1. Distribution of the attributes pregnant, pedigree and age decreases when values of these attribute increases. Attributes plasma, diastolic, triceps and mass are bell-shaped. The assortments with the maximum cardinality are positioned in the center and decreases in the direction of end of distribution. All attributes are multi-valued. Table 2 shows TP, TN, FP and FN for Diabetics dataset and these are used to calculate values as shown in Table 3.

Table 2. Confusion matrix for k-means clustering, hierarchical , fuzzy ,weighted class based clustering on diabetes, lung cancer, hepatitis, liver disorder, breast cancer wisconsin, mammographic and cardiovascular cleveland heart disease datasets

Dataset	Simple K-means		Hierarchical		Fuzzy C Means		WCBC	
Diabetics	302	198	380	120	279	221	417	83
	137	131	138	130	115	153	228	40
Lung Cancer	5	3	1	9	7	2	7	2
	8	14	7	15	12	11	9	14
Hepatitis	29	3	52	32	7	25	3	29
	37	86	23	48	25	98	13	110
Liver-Disorders	65	80	28	117	97	48	79	63
	105	95	41	157	111	89	90	113
Breast Cancer	384	74	440	31	454	4	450	13
	31	210	210	28	49	192	40	197
Mammographic	358	87	16	400	349	96	367	78
	127	389	45	500	222	294	119	397
Cleveland Heart	113	51	162	01	110	54	121	41
	31	108	136	03	71	68	27	114

Table 3. ROC, precision, recall and RMS error, of various techniques for diabetics dataset

	Diabetics Dataset							
	Total number of Instances	Correctly Classified Instances	Incorrectly Classified Instances	ROC	Accuracy	precision	recall	RMS Error
K-means	768	433	335	0.546	56.385	.534	.564	0.6666
Hierarchical	768	430	338	0.539	55.52	.527	.559	0.7010
Fuzzy	768	432	336	0.628	56.25	.612	.563	0.6614
WCBC	768	457	311	0.692	59.45	.684	.595	0.6664

Figure 1 show that WCBC performs better among all in case of precision and recall. ROC of proposed approach is .692 which is again high among all and RMS Error is also low. Figure 2 depicts accuracy is 3% high from K-means due to class based weight calculation method.

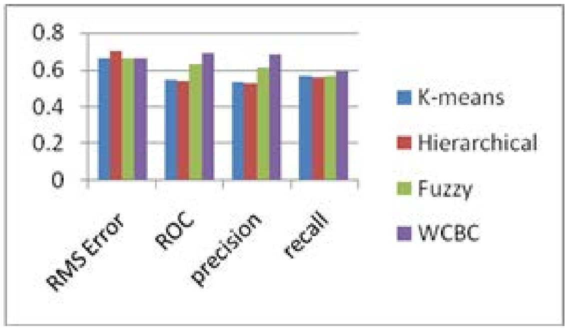

Figure 1. ROC, precision, recall and RMS error of various clustering techniques for diabetics dataset.

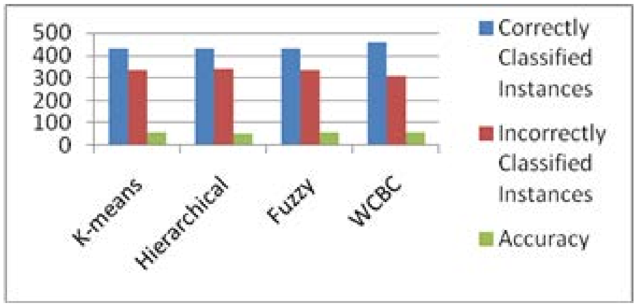

Figure 2. Correctly classified instances, incorrectly classified instances and accuracy of various techniques for diabetics dataset learning.

5.2.2 Lung Cancer Dataset

It contains 57 attributes and 32 instances. In this dataset class attribute is binary. Values of all the attributes are normally distributed and maximum attributes have three values. Any attribute is not having continuous value and all conditional attributes are multi-valued. Table 2 shows TP, TN, FP and FN for Lung Cancer dataset. These are used to calculate various parameters as shown in Table 4 for Lung Cancer dataset. Figure 3 and Figure 4 show that RMS Error is low among all and WCBC has 65.45% accuracy which is 6% high from K-means and 9% high as compared to Hierarchical and Fuzzy clustering. By keeping error function to its minimum value. ROC, Precision and Recall values are .692, .712 and .763 respectively which outperforms other clustering techniques.

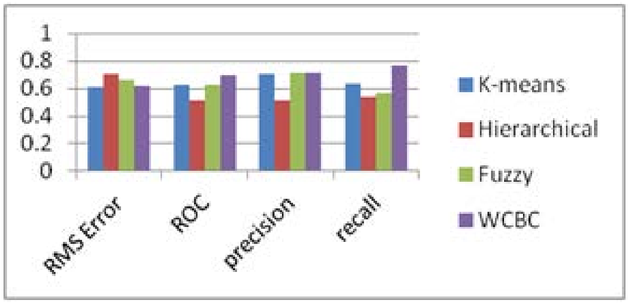

Figure 3. ROC, precision, recall and RMS error of various techniques for lung cancer dataset.

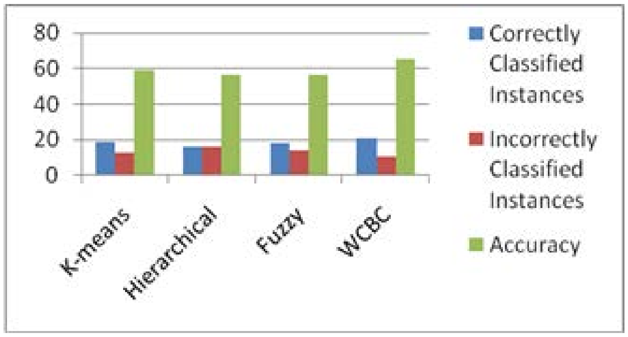

Figure 4. Correctly classified instances, incorrectly classified instances and accuracy of various techniques for lung cancer dataset dataset.

5.2.3 Hepatitis Dataset

It contains 20 attributes and 155 instances. The hepatitis database consists of 17 attributes. Four attributes are multi-valued, the rest are binary. The class attribute takes binary value 0 or 1. The distributions of attributes bilirubin and sgot decrease as their values increases. The distribution is bell-shape for attributes age and albumin. TP, TN, FP and FN values for Breast Cancer dataset taken from Table 2 are used to calculate correctly classified instances, incorrectly classified instances, precision, recall and accuracy of various clustering techniques as shown in Table 5.

Figure 5 shows that ROC of WCBC is .802 which is high among all and close to 1 is due to the fact that in WCBC class which contains classifying properties of data .WCBC has attained precision value .884 and recall value .895 which are high among all clustering techniques, implies a high degree of agreement between actual and predicted class. RMS Error is also low among all techniques.

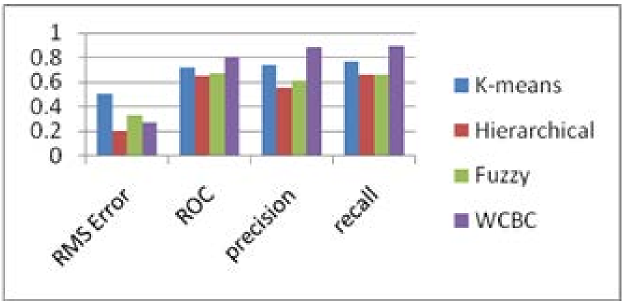

Figure 5. ROC, precision, recall and RMS Error of various techniques for Hepatitis dataset.

Figure 6 shows accuracy of WCBC is 73.2% which is 3.1% high from K-mean and 5% high from Hierarchical. This is due to class based classification approach used in

Table 4. ROC, precision, recall and RMS error, of various techniques for lung cancer dataset

		Lung Cancer Dataset						
	Total number of Instances	Correctly Classified Instances	Incorrectly Classified Instances	ROC	Accuracy	precision	recall	RMS Error
K-means	32	19	13	0.629	59.375	.706	.633	0.6066
Hierarchical	32	16	16	0.51	56.52	.517	.539	0.7010
Fuzzy	32	18	14	0.628	56.25	.712	.563	0.6614
WCBC	32	21	11	0.692	65.45	.712	.763	0.6214

Table 5. ROC, precision, recall and RMS error of various techniques for hepatitis dataset

Hepatitis Dataset								
	Total number of Instances	Correctly Classified Instances	Incorrectly Classified Instances	ROC	Accuracy	precision	recall	RMS Error
K-means	155	110	45	0.72	70.1	.734	.764	0.508
Hierarchical	155	100	55	0.65	68.2	.5527	.659	0.206
Fuzzy	155	105	50	0.668	67.74	.612	.663	0.326
WCBC	155	115	40	0.802	73.2	.884	.895	0.271

WCBC in which more number of instances of same class are placed in same cluster Proposed WCBC approach is suitable for hepatitis dataset.

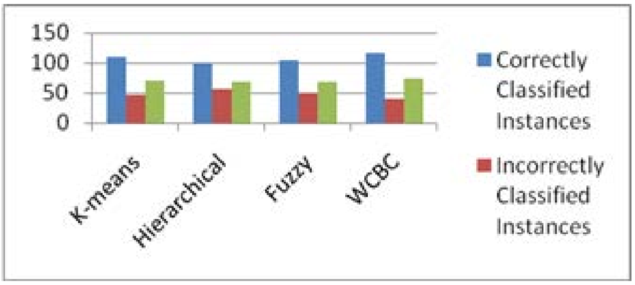

Figure 6. Correctly classified instances, incorrectly classified instances and accuracy of various techniques for hepatitis dataset.

5.2.4 Liver-Disorders Dataset

It contains 7 attributes and 345 instances. Class attribute is binary. First four attributes are bell shaped. Distribution of the attributes number 5 and 6 decreases with the increase of their values. Table 2 shows TP, TN, FP and FN for Breast Cancer dataset. These are used to calculate values as shown in Table 6. Figure 7 shows ROC, precision and recall values are high among all which implies that WCBC perform well among all clustering techniques. RMS Error of WCBC is minimum among all because proposed WCBC makes class based clusters for Liver-Disorders dataset by minimizing error function. Figure 8 shows that accuracy of proposed approach is 56.20% which is 9.83% high as compared to K-means. High value of correctly classified instances tells misclassification rate is low and WCBC has good classification capability.

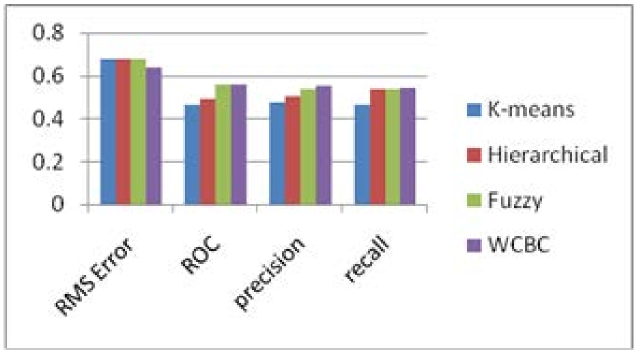

Figure 7. ROC, precision, recall and RMS error of various techniques for liver-disorders dataset.

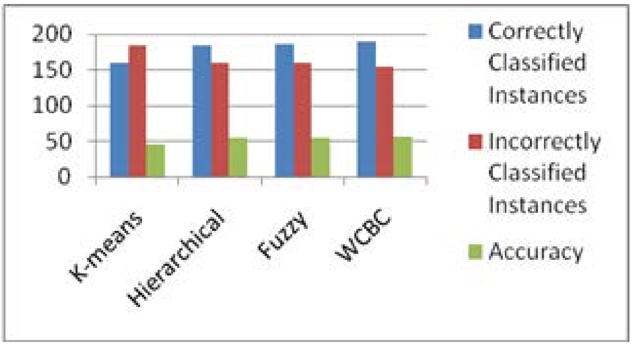

Figure 8. Correctly classified instances, incorrectly classified instances and accuracy of various techniques for liver-disorders dataset.

Table 6. ROC, precision, recall and RMS error, of various techniques for liver-disorders dataset

Liver-Disorders Dataset								
	Total number of Instances	Correctly Classified Instances	Incorrectly Classified Instances	ROC	Accuracy	precision	recall	RMS Error
K-means	345	160	185	0.462	46.37	.475	.464	0.6766
Hierarchical	345	185	160	0.491	53.62	.502	.539	0.6787
Fuzzy	345	186	159	0.557	53.91	.537	.539	0.6789
WCBC	345	190	155	0.562	56.20	.554	.542	0.6367

5.2.5 Breast Cancer Wisconsin Dataset

It contains 10 attributes and 286 instances. The class attribute is binary. Almost all attributes, the number of instances in which the attributes take the lowest values is the greatest. All conditional attributes are multi-valued. Table 2 shows TP, TN, FP and FN for Breast Cancer dataset. These are used to calculate values as shown in Table 7. Figure 9 shows that high value of precision and recall for WCBC which shows that agreement between actual and predicted class is high. Low value of RMS error of WCBC among all shows instances within one cluster are of same class and error within cluster is reduced. ROC value of proposed WCBC is also nearly equal to 1 which indicates good classification results. Figure 10 shows accuracy and correctly classified Instances of WCBC is high among all for Breast Cancer dataset.

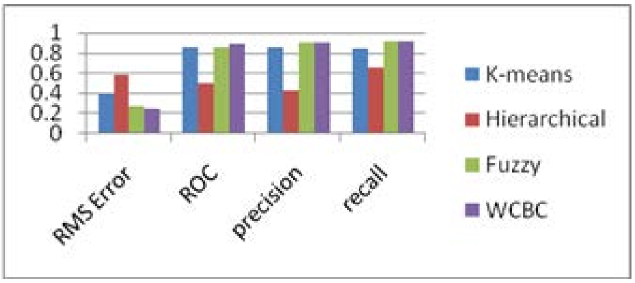

Figure 9. ROC, precision, recall and RMS error of various techniques for breast cancer wisconsin dataset.

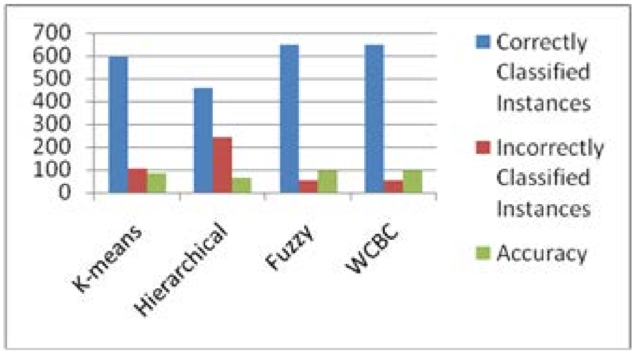

Figure 10. Correctly classified instances, incorrectly classified instances and accuracy of various techniques for breast cancer dataset.

5.2.6 Mammographic Dataset

It contains 6 attributes and 761 instances. Class attribute is binary. Attribute a has maximum instances near small values, attribute c and d have four values, e has two values and b is of bell shaped. Table 2 shows TP, TN, FP and FN. These are used to calculate various values for Mammographic Dataset as shown in Table 8. Figure 11 shows that ROC value is .895 which is high among all indicates good predictive capability of WCBC. Precision is .890 and is recall .891, which again indicates high level of agreement between actual and predicted class. Low implies errors within cluster are minimized because more number of instances within similar clusters belongs to same class.

Table 7. ROC, precision, recall and RMS error, of various techniques for breast cancer dataset

Breast Cancer Wisconsin Dataset								
	Total number of Instances	Correctly Classified Instances	Incorrectly Classified Instances	ROC	Accuracy	precision	recall	RMS Error
K-means	699	594	105	0.855	84.98	.852	.850	0.3876
Hierarchical	699	458	241	0.500	65.52	.429	.655	0.5872
Fuzzy	699	646	053	0.854	92.42	.909	.914	0.2754
WCBC	699	647	052	0.895	92.51	.910	.917	0.2554

Table 8. ROC, precision, recall and RMS error of various techniques for mammographic dataset

Mammographic Dataset								
	Total number of Instances	Correctly Classified Instances	Incorrectly Classified Instances	ROC	Accuracy	precision	recall	RMS Error
K-means	961	747	214	0.779	77.73	.781	.777	0.4719
Hierarchical	961	516	445	0.500	53.69	.288	.537	0.6805
Fuzzy	961	643	318	0.677	66.91	.688	.669	0.5752
WCBC	961	764	197	0.895	80.01	.890	.891	0.3219

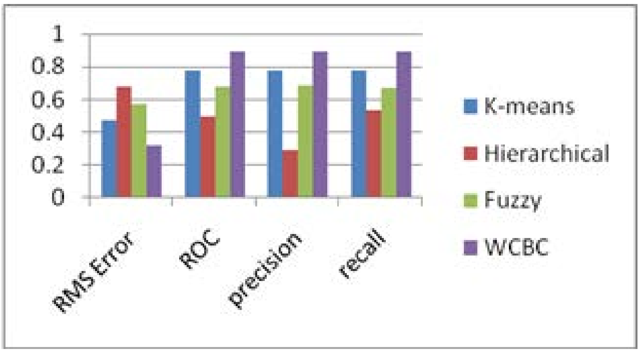

Figure 11. ROC, precision, recall and RMS error of various techniques for mammographic dataset.

Figure 12 shows correctly classified instances of WCBC is 235. This is high among all. Accuracy of WCBC is 80.01% which is 2.28% high from K-means, 26% high from hierarchical and 13% high from Fuzzy clustering.

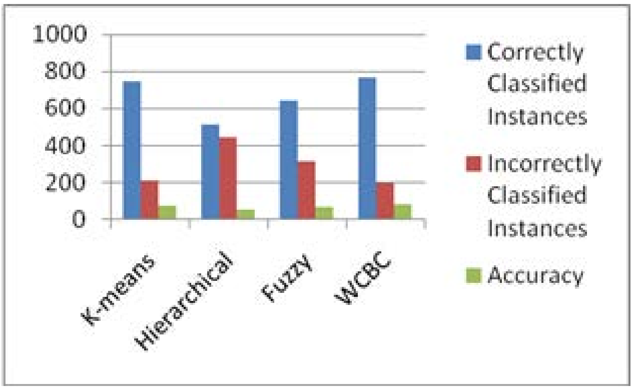

Figure 12. Correctly classified instances, incorrectly classified instances and accuracy of various techniques for mammographic dataset.

5.2.7 Cleveland Heart Dataset

The heart diseases database consists of 13 conditional attributes. The class attribute is binary. Attributes a, d, e and h are of bell-shaped; b, h and I are binominal and g, k and m have three values. Table 2 shows TP, TN, FP and FN for Cleveland Heart Dataset, used to calculate various values as shown in Table 9. ROC value for WCBC is .895 as shown in Figure 13 which is close to 1 due to class based separation. High value of precision and recall tells that agreement between actual and predicted class is high. RMS error of WCBC is low among all shows error function is minimized to attain good classification. Figure 14 shows correctly classify instances for WCBC is high among all. Accuracy is 76.01% which is 3% high from K-means, nearly 21% high from hierarchical and 11% high from Fuzzy clustering.

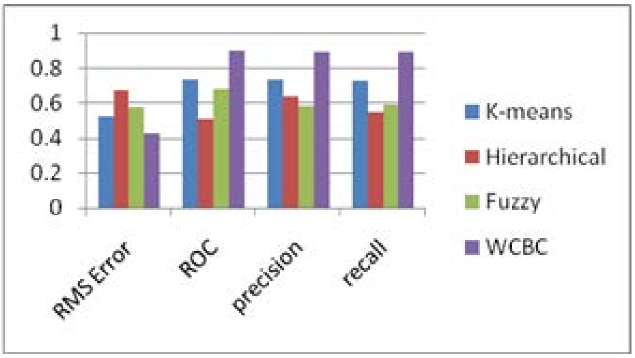

Figure 13. ROC, precision, recall and RMS error, of various techniques cleveland heart dataset.

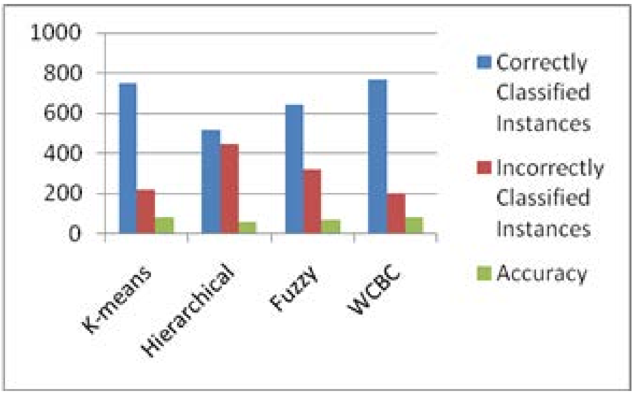

Figure 14. Correctly classified instances, incorrectly classified instances and accuracy of various techniques for mammographic.

Table 9. ROC, precision, recall and RMS error of various techniques for cleveland heart dataset

		Cleveland Heart Dataset						
	Total number of Instances	Correctly Classified Instances	Incorrectly Classified Instances	ROC	Accuracy	precision	recall	RMS Error
K-means	303	221	082	0.731	72.93	.736	.729	0.5202
Hierarchical	303	165	138	0.506	54.45	.639	.546	0.6735
Fuzzy	303	178	125	0.677	58.75	.585	.587	0.5752
WCBC	303	235	068	0.895	76.01	.889	.891	0.4279

6. Conclusion and Future Scope

In this research work, we have proposed new Weighted Class Based clustering (WCBC). WCBC has proved to be successful on various medical datasets having different distributions as compared to K-means, Hierarchical and Fuzzy clustering technique. Proposed clustering method has used classifying properties of medical data itself to calculate distances consequently clusters have more class evenness. Accuracy, Precission, Recall and ROC all are increased and RMS value is decreased by WCBC in all cases. The results showed that WCBC clustering technique has incredible capacity for medical domains as it is not highly dependent on data distributions and can act as core part of RBF neural network. Further we will try to find new class based Kernel method for RBF neural Network which will facilitate the physician to take competent and trustworthy decisions.

7. References

1. Berner ES, editor. Clinical decision support systems. New York, NY: Springer; 2007.
2. Finlay PN. Introducing decision support systems. Cambridge, MA: Blackwell Publishers; 1994.
3. Miller R. Medical diagnostic decision support systems–past, present, and future. Journal of the American Medical Informatics Association. 1994 Jan-Feb; 1(1):8–27.
4. Research. 2001; 47(1-2).
5. Pujaria AK, Rajesha K, Reddy DS. Clustering techniques in data mining- A survey. IETE Journal of Godara S, Singh R. Evaluation of predictive machine learning techniques as expert systems in medical diagnosis. Indian Journal of Science and Technology. 2016; 910.
6. Bagirov AM, Mardaneh K. Modified global k-means algorithm for clustering in gene expression data sets. Workshop on Intelligent Systems for Bioinformatics (WISB2006); Hobart, Australia. 2006.
7. Kaur SP, Sharma M. Radially optimized zone-divided energy-aware Wireless Sensor Networks (WSN) protocol using BA (Bat Algorithm). IETE Journal of Research. 2015; 61(2).
8. Awad M, Pomares H, Rojas I. Enhanced clustering technique in RBF neural network for function approximation. Mathematical and Computer Modelling. 2012 Feb; 55(3-4):286–302.
9. Siddiqui FU, Isa NAM. Enhanced Moving K-Means (EMKM) algorithm for image segmentation. IEEE Transactions on Consumer Electronics. 2011 May; 57(2):833–41.
10. Su M, Chou C. A modified version of the K-Means algorithm with a distance based on cluster symmetry. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2001 Jun; 23(6):674–80.
11. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficientk-mean clustering algorithm: Analysisan implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002 Jul; 24(7):881–91.
12. Li DC, Liu CW. A class possibility based kernel to increase classification accuracy for small data sets using support vector machines. Expert Systems with Applications. 2010 Apr; 37(4):3104–10.
13. Wang L, Bo L, Jiao L. A modified k-means clustering with a density-sensitive distance metric. Springer-Verlag Berlin Heidelberg. RSKT- LNAI. 2006; 4062:544–51.
14. Shrimali V, Anand RS, Kumar V. Comparing the performance of ultrasonic liver image enhancement techniques: a preference study. IETE Journal of Research. 2010; 56(1).
15. Venu N, Anuradha B. Multil-kernels integration for FCM algorithm for medical image segmentation using histogram analasis. Indian Journal of Science and Technology. 2015 Dec; 8(34).
16. Nithya NS, DuraiswamyK. Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface. Indian Academy of Sciences. 2014 Feb; 39(1):39–52.
17. Dong S, Zhous D, Ding W, Gong J. Flow cluster algorithm based on improved k-means method. IETE Journal of Research. 2013 Jul-Aug; 59(4).
18. Balajia GN, Subashinib TS, Chidambaramc N. Detection of heart muscle damage from automated analysis of echocardiogram video. IETE Journal of Research. 2015; 61(3).
19. Palaniappana R, Sundaraja K, Ahameda NU, Arjunana A, Sundaraj S. Computer-based respiratory sound analysis: A systematic review. IETE Technical Review. 2013; 30(3).
20. Malliga L, Bommanna RK. A novel content based medical image retrieval technique with aid of modified fuzzy c-means. Journal of Medical Imaging and Health Informatics. 2016 Jun; 6(3):700–9.
21. Puyalnithi T, Viswanatham VM. Preliminary cardiac disease risk prediction based on medical and behavioural data set using supervised machine learning techniques. Indian Journal of Science and Technology. 2016; 9(31).
22. Zolfagharifar, Ahad S, Karamizadeh F. Developing a hybrid intelligent classifier by using evolutionary learning (genetic algorithm and decision tree). Indian Journal of Science and Technology. 2016; 9(20).