Received: September 21, 2016

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Optimal Decision Tree Based Unsupervised Learning Method for Data Clustering

Nagarjuna Reddy Seelam[1*] Sai Satyanarayana Reddy Seelam[2] Babu Reddy Mukkala[3]

[1]*Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India*
[2]*Vardhaman College of Engineering, Hyderabad, Telangana, India*
[3]*Krishna University, MachiliPatnam, Andhra Pradesh, India*
* Corresponding author's Email: nagarjunareddy0884@gmail.com

Abstract: Clustering is an investigative data analysis task. It aims to find the intrinsic structure of data by organizing data objects into similarity groups or clusters. Our investigation using a pattern based clustering on numerical data set; here, we are using a Parkinson and spam dataset. These techniques are strongly related to the statistical field of cluster analysis, where over the years a large number of clustering methods has been proposed. Here, we have proposed an improved k-means clustering algorithm is used to extract patterns from a collection of an unsupervised decision tree. In our proposed research, we introduce a binary cuckoo search based decision tree. In this tree based learning technique, extracting patterns from a given dataset. Here, we have clustered the data with the aid of improved k-means clustering algorithm. The performance can be evaluated in terms of sensitivity, specificity, and accuracy.

Keywords: K-means clustering, Binary cuckoo search, Sensitivity, Specificity, Accuracy, Pattern.

1. Introduction

Data mining is one of the fast growing research fields which are used in wide areas of applications [1]. Due to the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning [2]. Text mining is the discovery of interesting knowledge in text documents [3, 4]. A Web mining system can be viewed as the use of data mining techniques to automatically retrieve, extract, generalize, and analyze Web information [5]. Most clustering done in practice is based largely on heuristic but intuitively reasonable procedures [6]. Mobile telecommunication companies have used data mining techniques to identify customers that are likely churn [7].

Clustering as applied to data mining applications encounters three additional complications: (a) large databases, (b) objects with many attributes, and (c) attributes of different types [8]. While clustering, in general, is a rather dignified problem, mainly in the last decade new approaches have been proposed to cope with new challenges provided by modern capabilities of automatic data generation and acquisition in more and more applications producing a vast amount of high-dimensional data [9].

Web usage mining provides insight about user behaviors that helps to optimize the website for increased customer loyalty and e-business effectiveness [10]. Distance-Based clustering techniques, such as k-means, and parametric mixture models, such as Gaussian mixture models, are two main approaches used in customer segmentation in the data mining literature [11]. Clustering can be informally defined as grouping data items together [12]. Recent efforts in data mining have focused on methods for efficient and effective cluster analysis in large databases [13]. The extraction of knowledge hidden in records of past observations is a common problem practically in every area of science [14]. The main intention of our proposed work is to find the intrinsic structure of data by organizing data objects into similarity groups or clusters. Here, we clustered the data with the aid of improved k-means clustering algorithm. In our research, we introduce an optimal decision tree technique for data clustering based on Binary Cuckoo Search Algorithm. The decision tree based learning technique will extract the patterns in the given data set.

2. Literature Survey

Literature presents several techniques for data clustering. A. Hatamlou [15] has presented a binary search algorithm for data clustering that not only finds high-quality clusters but also converges to the same solution in different runs. In the proposed algorithm, a set of initial centroids were chosen from different parts of the test dataset and then optimal locations for the centroids were found by thoroughly exploring around of the initial centroids.

The author, M. B. Dowlatshahi and H. Nezamabadi-pour [16] have developed the structure of GSA for solving the data clustering problem, the problem of grouping data into clusters such that the data in each cluster share a high degree of similarity while being very dissimilar to data from other clusters. The proposed algorithm, which was called Grouping GSA (GGSA), differs from the standard GSA in two important aspects.

Based on kernelized fuzzy clustering analysis, L. Liao et al. [17], have presented a fast image segmentation algorithm using a speeding-up scheme called reduced set representation. The proposed clustering algorithm has lower computational complexity and could be regarded as the generalized version of the traditional KFCM-I and KFCM-II algorithms. Moreover, an image intensity correction is employed during image segmentation process.

G. Chicco et al. [18], have provided an overview of the clustering techniques used to establish suitable customer grouping, included in a general scheme for analyzing electrical load pattern data. The characteristics of the various stages of the customer grouping procedure are illustrated and discussed, providing links to relevant literature references.

L. Galluccio et al. [19], have introduced a novel distance measure for clustering high dimensional data based on the hitting time of two Minimal Spanning Trees (MST) grown sequentially from a pair of points by Prim’s algorithm. When the proposed measure is used in conjunction with spectral clustering, we have obtained a powerful clustering algorithm that is able to separate neighboring non-convex shaped clusters

J. Pei et al. [20], have studied the problem of maximal pattern based clustering. Redundant clusters are avoided completely by mining only the maximal pattern-based clusters. MaPle, an efficient and scalable mining algorithm is developed. It conducts a depth-first, divide-and-conquer search and prunes unnecessary branches smartly.

A. Zimek et al. [21], have discussed how frequent pattern mining algorithms have been extended and generalized towards the discovery of local clusters in high-dimensional data. In particular, we have discussed several example algorithms for subspace clustering or projected clustering as well as point out recent research questions and open topics in this area relevant to researchers in either clustering or pattern mining.

3. Problem Identification

- ❖ Adapting the grouping representation, [16], we proposed a Grouping Gravitational Search Algorithm (GGSA) for data clustering in this paper.
- ❖ The main property [16], of the grouping representation which encourages us to use it, is that it has very low redundancy also it have some grouping problems.
- ❖ In the clustering [17], Clustering is a vital mechanism in data analysis to define or organize a group of patterns or objects into clusters.
- ❖ The objects in the same cluster share common properties and those in different clusters have distinct dissimilarity. Therefore, the ultimate goal of data clustering is to reach unsupervised classification of complex data where there is little or no prior knowledge of those data.
- ❖ The most widely used algorithm to solve this clustering problem is Fuzzy c-means (FCM). The goal of the FCM is to minimize the generalized least-squares objective function, in which degree of membership plays an important role to optimize the data partitioning problem.
- ❖ However, the disadvantage of this [15], an algorithm is well known: a correct number of clusters are required beforehand and the algorithm is quite sensitive to centroid initialization. Most research papers have proposed solutions for such problems by running an algorithm repeatedly with different fixed values of C (cluster) and with different initializations. However, this may not be feasible with large data sets and large C.

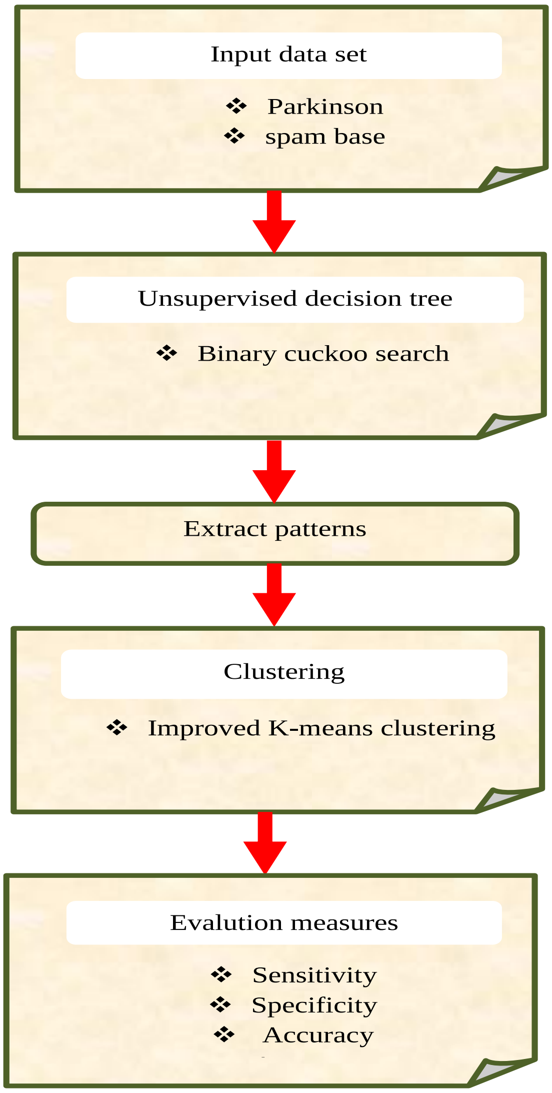

Figure.1 Proposed unsupervised learning method for data clustering

4. Proposed Methodology

In this paper, we propose a Hierarchical pattern based clustering on the numerical dataset. We investigate using a pattern based clustering approach to clustering to return a set of patterns. The algorithm will extract a small subset of patterns without applying an a priori discretization on numerical features. This algorithm is used to extract patterns from a collection of unsupervised decision trees created through a hierarchical procedure. In various researches in data mining literature has been introduced in a number of the algorithm for clustering. In our research, we introduce an optimal decision tree technique for data clustering based on Binary Cuckoo Search Algorithm. The decision tree based learning technique will extract the patterns in the given data set. The input data can be clustered based on Improved K-Means algorithm. In this research, we investigate using a pattern based clustering approach to clustering the small set of patterns. The main issues of existing studies is explained in below,

4.1. Binary Cuckoo Search-Decision Tree

Here, we utilize a binary cuckoo search algorithm is used to extract patterns from a collection of unsupervised decision trees created through a hierarchical procedure. In various researches in data mining literature has introduced a number of algorithms for clustering. In our research, we introduce an optimal decision tree technique for data clustering based on Binary Cuckoo Search Algorithm.

Cuckoo search algorithm is a metaheuristic algorithm which was presented by the reproduction activities of the cuckoos and reduces to carry out. They contain many nests in a cuckoo search. Every egg point to a resolution and an egg of cuckoo specify a fresh explanation. The new and improved explanation is replacing the majority useful explanation in the nest. The following representation system is selected by Cuckoo Search algorithm: Every egg in a nest symbolizes an explanation, and a Cuckoo egg symbolizes a novel explanation. The intent is to utilize the novel and probably improved egg to restore a not-so-good egg of Cuckoo in the nests. Though this is the fundamental situation, which means one cuckoo for each nest, but the amount of the method can be amplified by integrating the possessions that every nest can have more than one egg which corresponds to a set of explanations. The procedure of clustering is indicated below,

- The Only one egg at a time is laid by cuckoo. Cuckoo dumps its egg in a randomly chosen nest.
- The number of available host nests is fixed, and nests with the high quality of eggs will carry over to the next generations.
- In the case of a host bird discovered the cuckoo egg; it can throw the egg away or abandon the nest, and build a completely new nest.

Initialization Phase

The population (m_i, where i=1, 2…n) of host nest is commenced randomly.

Generating New Cuckoo Phase

The levy flights have used a cuckoo to select at random and it generates new explanations. After that, the created cuckoo is evaluated by the aim utility for finding the worth of the explanations.

Fitness Evaluation Phase

Assess the fitness function based on the equation and after that choose the best one.

$$fitness = \max imum\ popularity \qquad (1)$$

Updation Phase

Modify the primary explanation by levy flights in which cosine transform is engaged. The superiority of the new explanation is evaluated and a nest is selected between arbitrarily. If the superiority of new resolution in the selected nest is improved than the previous resolution, it will be alternated by the new explanation (Cuckoo). Or else, the prior explanation is put to the side as the finest explanation. The levy flights utilized for usual cuckoo search algorithm is,

$$m_i^{\;*} = m_i^{(t+1)} = m_i^{(t)} + \alpha \oplus Levy(n) \qquad (2)$$

Where t is step size, and α>0 is the step size scaling feature limit. Here, the entry wise product $\oplus$ is comparable to those utilized in PSO, $x_i^{(t+1)}$ and represents $(t+1)_{th}$ egg (feature) at nest (solution), *i=1,2,....m,* and *t=1,2,...d.* The Levy flights utilize an arbitrary level extent which is drained from a Levy allocation. So, the CS algorithm is extra competent in discovering the investigate break as its step extent is greatly longer in the long run.

In conventional COA, the explanations are reorganized in the investigate break towards continuous-respected locations. Disparate, in the BCOA for characteristic collection, the investigate break is sculpted as a dimensional Boolean lattice, in which the explanations are reorganized diagonally the angle of a hypercube. Additionally, as the difficulty is to choose or not a known characteristic, an explanation binary vector is engaged, where 1 communicates whether a characteristic will be chosen to create the novel dataset and 0 otherwise. In order to construct this binary vector, we have engaged the equation 4, which can supply only binary values in the Boolean lattice controlling the novel explanation to only binary values:

$$S\left(x_i^{(t+1)}\right) = \frac{1}{1+e^{-x_i^{(t)}}} \qquad (3)$$

$$\begin{cases} \text{if } S < rand \text{ then } & x_i^{(t+1)} = 0 \\ \text{if } S > rand \text{ then } & x_i^{(t+1)} = 1 \end{cases}$$

Here, we are including a decision tree algorithm instead of cuckoo search updation phase the updated output values are given into the decision tree.

Decision Tree

Decision tree form is quick reliable, effortless to preserve and correct in the preparation course area. In decision tree knowledge, ID3 (Iterative Dichotomiser 3) is an algorithm proposed by Ross Quinlan employed to produce a decision tree from the dataset. ID3 is classically used in the machine learning and natural language dispensation fields. The decision tree method contains build a tree to form the categorization procedure. Once a tree is constructing, it is functional to every tuple in the database and consequences in categorization for that tuple. The subsequent problems are resolved by most decision tree algorithms.

- Choosing splitting attributes
- Ordering of splitting attributes
- Number of splits to take
- Balance of tree structure and pruning
- Stopping criteria

The ID3 algorithm is a categorization algorithm depend on Information Entropy, its fundamental design is that all instances are drawn to dissimilar classed depending on dissimilar values of the state quality set; its center is to establish the finest categorization quality from state quality sets. The algorithm decides information increase as quality collection criteria; as a rule, the quality that has the uppermost information increase is chosen as the dividing quality of existing node, in order to create information entropy that the separated subsets require a minimum. According to the dissimilar values of the quality, branches can be recognized, and the procedure is recursively described on every branch to generate other nodes and branches until all the examples in a branch fit into the similar group. To choose the dividing qualities, the idea of Entropy and Information increase are utilized.

Entropy

Given probabilities $p_1,p_2,....P_s$, *where* Σp_i=1, Entropy is defined as,

$$H(p_1, p_2, ... p_s), = \sum -(p_i \log p_i) \qquad (4)$$

Entropy discovers the quantity of array in a known database state. A value of $H = 0$ recognize a completely categorized set. The sup error of the entropy means the better in the possible to develop the categorization method.

Information Gain

ID3 decide the divide quality with the maximum increase in information, where the increase is defined as dissimilarity between how much information is desirable after the divide. This is considered by formative the dissimilarity between the entropies of the unique dataset and the subjective amount of the entropies from each of the subdivided datasets. The formula used for this reason is:

$$G(D,S) = H(D) - \sum P(D_i)H(D_i) \quad (5)$$

4.2. K-Means Clustering Algorithm

One of the most widely used clustering algorithms is K-Means clustering. This minimizes the mean squared Euclidean distance from each data point to its nearest centre. Here we have a good control of the number of clusters produced.

The K-Means Clustering technique has emerged as one of the easiest unsubstantiated learning techniques which are well-equipped with the skills of finding effective solutions to the famous clustering challenges of late, the significance of the clustering approaches is gaining ground as they are extensively employed in a number of applications. The underlying process employs an easy and effortless method to categorize specified files into a specific number of clusters. The incredible proficiency of the original means technique owes a lot to the initial centroids, which have a telling impact on the number of iterations needed for executing the original k-means technique. However, a computational complication of the original k-means algorithm is found to be highly excessive, particularly in respect of gigantic massive files. The current investigation is invested in launching an improved technique for locating the top ranking cluster record.

The number of clusters *K* is deemed to be permanent in the k-means clustering technique. Let *K* prototypes

($\omega_1 \ldots\ldots, \omega_k$) be activated into one of the *n* input Data ($i_1 \ldots\ldots, i_n$). Hence,

$$\omega_j = i_1,\, j \in \{1,\ldots\ldots, k\},\, \iota \in \{1\ldots\ldots.n\} \quad (6)$$

The suitable selection of *k* invariably depends on the issue and domain and habitually a user attempts a number of values of *k*. It is presumed that there are *n* data, each of dimensions *d*.

Step1**:** Randomly pick k points as centroids of k clusters.

Step2**:**

- For each point assign the point to the nearest cluster.
- Recomputed the cluster centroids.
- Repeat Step2 (until there is no change in clusters between consecutive iterations).

With this idea of what k-Means do now, we are going to discuss certain facts with respect to cluster behavior.

- The idea of clustering is to group data items having high similarity and to separate from dissimilar data items.
- The quality of a cluster is defined as high intra-cluster similarity and low inter-cluster similarity.

Having clustered our data, we now need some mechanism to choose datasets from each cluster. In the next section, we used an algorithm called Pickup cluster that does the selection work. Here, the features are extracted from the binary cuckoo search based decision tree from that we are clustering these data with the aid of k-means clustering algorithm.

5. Result and Discussion

Our proposed modified K-Means for the effective clustering of data is implemented using the MATLAB platform on the Parkinson and spam base data from the dataset. The data set description is given below in detail.

5.1 Dataset Description

Parkinson's Dataset

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

Spam Based Dataset

The "spam" concept is diverse: advertisements for products/web sites, make money fast schemes, chain letters, pornography. Our collection of spam e-mails came from our postmaster and individuals who had filed spam. Our collection of non-spam e-mails came from filed work and personal e-mails, and hence the word 'George' and the area code '650' are indicators of non-spam.

5.2. Performance Evaluation

By utilizing the performance measures namely False Positive Rate, False Negative Rate, Sensitivity, Specificity and Accuracy, the performance of the system is estimated. The basic count values such as True Positive (*TP*), True Negative (*TN*), False Positive (*FP*) and False Negative (*FN*) are used by

Received: September 21, 2016

these measures.

False Positive Rate (FPR)

The percentage of cases where an image was classified to normal images, but in fact it did not.

$$FPR = \frac{FP}{FP+TN} \tag{7}$$

False Negative Rate (FNR)

The percentage of cases where an image was classified to abnormal images, but in fact it did.

$$FNR = \frac{FN}{FN+TP} \tag{8}$$

Sensitivity

The proportion of actual positives which are correctly identified is the measure of the sensitivity. It relates to the ability of the test to identify positive results.

$$Sensitivity = \frac{No.of\ TP}{No.\ of\ TP\ +\ No.\ of\ FN} \times 100 \tag{9}$$

Specificity

The proportion of negatives which are correctly identified is the measure of the specificity. It relates to the ability of the test to identify negative results.

$$Specificity = \frac{No.\ of\ TN}{No.\ of\ TN\ +\ No.\ of\ FP} \times 100 \tag{10}$$

Accuracy

We can compute the measure of accuracy from the measures of sensitivity and specificity as specified below.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \tag{11}$$

5.2.1. Results of Segmentation Evaluation

The efficiency of this clustering of all the data is examined by the metrics Sensitivity, Specificity and Accuracy as specified in the following table 1.

Table 1. Explains the Parkinson dataset evaluation results for cluster 2

	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Cluster 2	74	0	1	0	0.986667	0.98	0.95
	75	0	2	0	0.974026	0.99	0.95

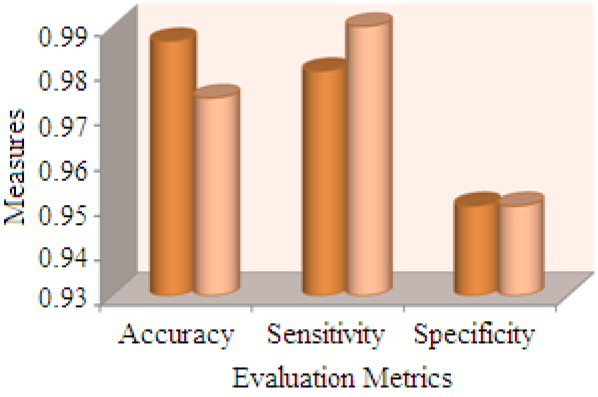

Figure.2 Cluster 1 Evaluation Measures of Parkinson dataset for cluster 2

The corresponding graph of table 1 is designed in Fig. 2 with Parkinson dataset for clustering2 measures.

We can attain the clustering based efficiency from the above table I and its corresponding graph in fig. 2. The specificity for the cluster 2 is 95%. The sensitivity for the cluster 2 is 98% sensitivity metrics, respectively. These values are also high for our proposed K-means clustering which guides to make the high accuracy of clustering. Therefore, we can obtain very good accuracy values of Parkinson dataset 98%. Generally, our proposed work gives 0.98% of accuracy for the Parkinson dataset for 2 cluster2. Table 2 shows the cluster3 evaluation measures, they are given below.

Table 2. Explains the Parkinson dataset evaluation results for cluster 3

	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Cluster 3	51	2	15	10	0.996942	0.98555	0.976666
	58	3	1	11	0.983051	0.9755	0.98

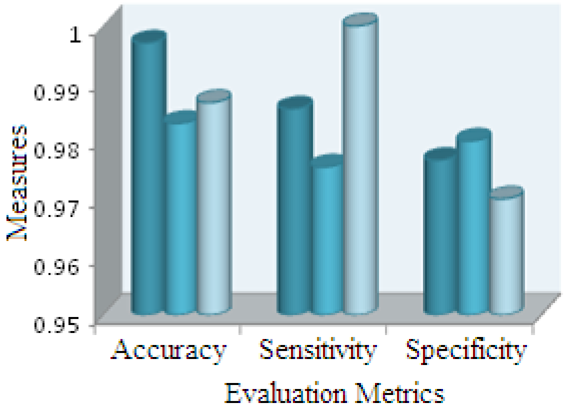

Figure.3 Evaluation Measures of Parkinson dataset for cluster 3

DOI: 10.22266/ijies2017.0430.14

The corresponding graph of table 2 is designed in Fig. 3 with Parkinson dataset for clustering 3 measures.

We can attain the clustering based efficiency from the above table II and its corresponding graph in fig. 3. The specificity for the cluster 3 is 97%. The sensitivity for the cluster 3 is 98% sensitivity metrics, respectively. These values are also high for our proposed K-means clustering which guides to make the high accuracy of clustering.

Therefore, we can obtain very good accuracy values of Parkinson dataset 98%. Generally, our proposed work gives 0.98% of accuracy for the Parkinson dataset for cluster 3. Table 3 shows the cluster 3 evaluation measures, they are given below.

We can attain the clustering based efficiency from the above table III and its corresponding graph in fig 4. The specificity for the cluster 4 is 98%. The sensitivity for the cluster 4 is 98% sensitivity metrics, respectively. These values are also high for our proposed K-means clustering which guides to make high accuracy of clustering. Therefore, we can obtain very good accuracy values of Parkinson dataset 98%. Generally, our proposed work gives 0.98% of accuracy for the Parkinson dataset for cluster 4. Table 3 shows the cluster 4 evaluation measures, they are given below. The corresponding graph of table 3 is designed in Fig. 4 with Parkinson dataset for clustering 4 measures.

Table 3. Explains the Parkinson dataset evaluation results for cluster 4

	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Cluster 4	38	8	7	17	0.975864	0.975864	0.972066
	51	10	2	5	0.962264	0.987556	0.982299

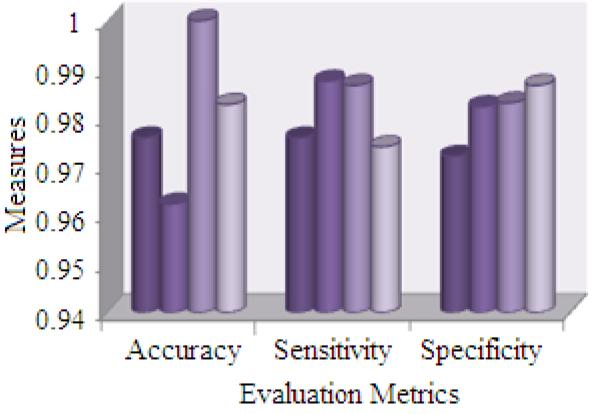

Figure.4 Evaluation Measures of Parkinson dataset for cluster 4

Table 4. Explains the Spam base dataset evaluation results for cluster 2

	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Cluster 2	326	2153	1	20	0.996942	0.98	0.97555
	2405	350	11	57	0.976848	0.999	0.9855

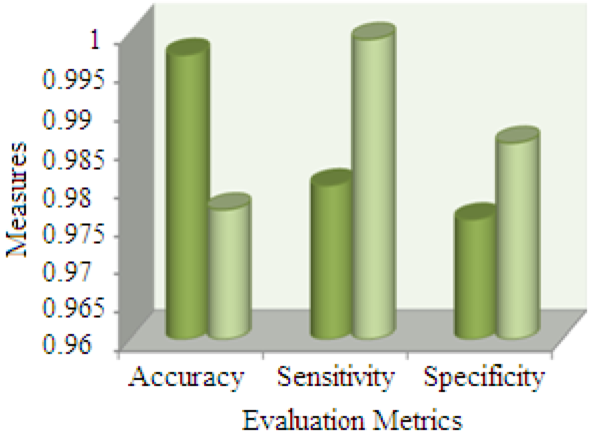

Figure.5 Evaluation Measures of Spam base dataset for cluster 3

The corresponding graph of table 4 is designed in Fig. 5 with spam base dataset for clustering 2 measures.

We can attain the clustering based efficiency from the above table VI and its corresponding graph in fig. 5. The specificity for the cluster 2 is 98%. The sensitivity for the cluster 2 is 98% sensitivity metrics, respectively. These values are also high for our proposed K-means clustering which guides to make the high accuracy of clustering. Therefore, we can obtain very good accuracy values of Spam base dataset 98%. Generally, our proposed work gives 0.98% of accuracy for the Spam base dataset for 2 cluster2. The table V shows the cluster 2 evaluation measures, they are given below.

The corresponding graph of table 5 is designed in Fig. 6 with Spam base dataset for clustering 3 measures.

Table 5. Explains the Spam base dataset evaluation results for cluster 3

	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Cluster 3	708	0	2	21	0.971193	0.97	1
	1801	38	8	12	0.972066	0.999	0.96
	255	6	5	28	0.982299	0.999	0.98

We can attain the clustering based efficiency from the above table V and its corresponding graph in fig 6. The specificity for the cluster 3 is 98%. The sensitivity for the cluster 3 is 98% sensitivity metrics, respectively. These values are also high for our proposed K-means clustering which guides to make the high accuracy of clustering. Therefore, we can obtain very good accuracy values of Spam base dataset 97%. Generally, our proposed work gives 0.97% of accuracy for the Spam base dataset for clusters 3.

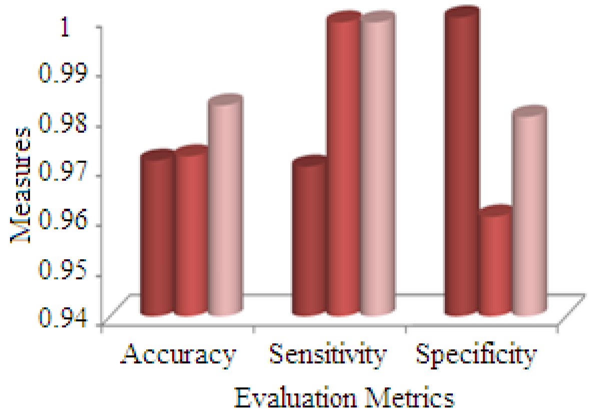

Figure.6 Evaluation Measures of Spam base dataset for cluster 3

Table 6. Explains the Spam base dataset evaluation results for cluster 4

	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Cluster 4	796	181	30	23	0.971917	0.982299	0.996942
	837	120	297	12	0.980094	0.97555	0.982299
	262	300	1573	10	0.992424	0.98555	0.971193
	468	98	98	11	0.989429	0.975862	0.972066

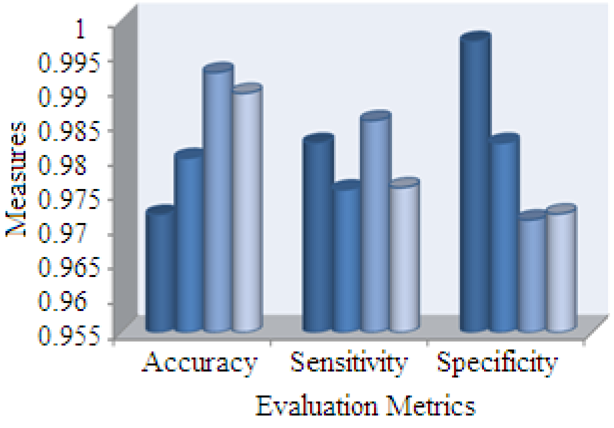

Figure.7 Evaluation Measures of Spam base dataset for cluster 4

The corresponding graph of table 6 is designed in Fig. 7 with Spam base dataset for clustering 4 measures.

We can attain the clustering based efficiency from the above table VI and its corresponding graph in fig 7. The specificity for the cluster 4 is 98%. The sensitivity for the cluster 4 is 97% sensitivity metrics, respectively. These values are also high for our proposed K-means clustering which guides to make the high accuracy of clustering. Therefore, we can obtain very good accuracy values of Spam base dataset 98%.

5.2.2. Comparison Analysis for Our Proposed Work and Existing Work

Here, we take FCM as an existing technique for comparison because FCM has to be used in many articles and it given high outcomes so we would take FCM as our existing technique compare our result prove our proposed study would outperform and given a better result for that we have compared a clustering accuracy for both methods and finally we prove that our proposed technique will give a better result. For the clustering of data of our proposed work makes use of modified K-means clustering. We can establish that our proposed work helps to attain very good accuracy for the clustering of data. And also we can establish this k-means clustering accuracy outcome by comparing other FCM clustering method. We have utilized modified FCM for comparison in this study. The comparison outcomes are presented in the following table 4.

Table 7. Comparison of proposed and existing Accuracy results

Clusters	Proposed K-Means Clustering		Existing FCM	
	Accuracy for Parkinson dataset	Accuracy for Spam Dataset	Accuracy for Parkinson Dataset	Accuracy for Spam Dataset
2	98%	98%	85%	90%
3	98%	97%	90%	87%
4	98%	98%	91%	89%

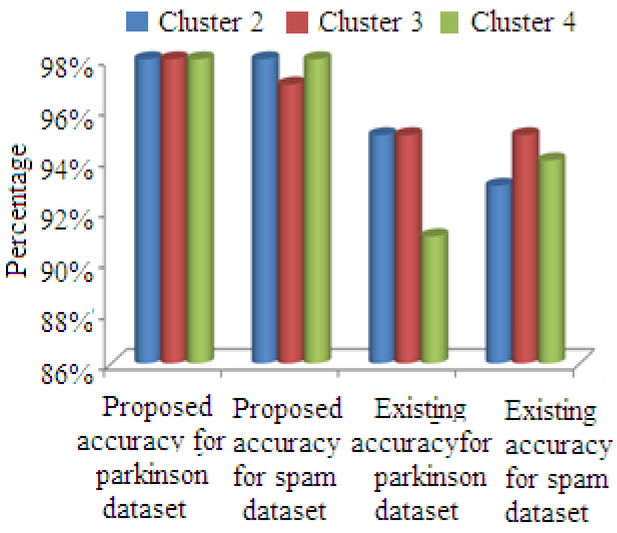

Figure.8 Comparison result for Clustering accuracy for proposed and existing methodology

Above specified fig. 8 explains the comparison outcomes of the clustering for the data Parkinson and spam dataset.

The improved K-Means outcomes of clustering of data are presented by our proposed work. We will compare to our k-means clustering technique to fuzzy c-means technique in this comparison our proposed technique will give very high accuracy values for clustering of data. the accuracy for the existing FCM is 88% accuracy for Parkinson dataset which is low when we compare to this result to our proposed k-means clustering technique it gives 98% accuracy for Parkinson dataset. However, in the spam dataset, the existing technique gives 88% accuracy which is low when we compare to this result to our proposed k- means clustering technique it will give 98% accuracy. From these outcomes, it is known that our proposed modified k-means clustering give improved accuracy outcomes. Therefore our work shows that it is worth for the clustering.

6. Conclusion

In this paper, we present a pattern based clustering technique to return a set of patterns. Initially, we will create a decision tree with the aid of binary cuckoo search algorithm in this tree will create in a hierarchical procedure. These algorithms are used to extract patterns from a decision tree. The decision tree based learning technique will extract the patterns in the given data set. The data can be clustered based on Improved K-Means algorithm. The performance measures sensitivity, specificity and accuracy were evaluated by our proposed method. The efficiency of the clustering of data is very high by presenting very good accuracy outcomes and also the clustered data. From the outcomes, we have shown that the modified K-Means utilized in our proposed work outperforms the other Fuzzy C-Means by the facilitated very good accuracy of 98% for both Parkinson and spam base dataset. Therefore by utilizing this technique, our proposed modified K-Means based clustering technique. In the future, we will use an enhanced clustering method for data clustering to improve a clustering accuracy.

References

[1] A. Pal, P. Shraddha and J. Maurya, “Classification and Analysis of High Dimensional Datasets using Clustering and Decision tree”, *International journal of Computer Science and Information Technologies*, Vol. 5, No. 2, pp. 2329-2333, 2014.

[2] N. Zhong, Y. Li and S. T. Wu, “Effective Pattern Discovery for Text Mining”, *In the Proceeding of IEEE Transaction on Knowledge and Data Engineering*, Vol. 24, No. 1, pp. 30-44, 2012.

[3] P. Berkhin, “A Survey of Clustering Data Mining Techniques”, *In the Proceeding of Springer on grouping Multidimensional Data*, pp. 25-71, 2006.

[4] J. Srivastava, R. Cooley, M. Deshpande and P. N. Tan, “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data”, *In Proceedings of ACM Digital Library on SIGKDD Exploration* , Vol. 1, No. 2, pp. 12-23, 2000.

[5] Y. Li and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs”, *In Proceedings of IEEE Transaction on Knowledge and Data Engineering*, Vol. 18, No. 4, pp. 554-568, 2006.

[6] C. Fraley and A. E. Raftery, “Model-Based Clustering, Discriminant Analysis, and Density Estimation”, *Journal of the American Statistical Association*, Vol. 97, No. 458, pp. 611-631, 2002.

[7] I. Bose and X. Chen, “Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn”, *In Proceedings of the International Multi Conference and Computer Science*, Vol. 19, No. 2, pp. 133-151, 2009.

[8] K. Mythili and K. Yasodha, “A Pattern Taxonomy Model with New Pattern Discovery Model for Text Mining”, International Journal of Science and Applied Information Technology, Vol. 1, No. 3, pp. 88-92, Aug 2012.

[9] H. P. Kriegel, P. Kröger and A. Zimek, “Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering”, *Journal Of Acm Transaction on Knowledge Discovery From Data (Tkdd)*, Vol. 3, No. 1, pp. 1-8, 2009.

[10] Q. Zhao, S. S. Bhowmick and L. Gruenwald, “Cleopatra: Evolutionary Pattern-based Clustering of Web Usage Data”, *In Proceedings of Advances ikn Knowledge Discovery and Data mining*, Vol. 3918, pp. 323-333, 2006.

[11] Y. Yang and B. Padmanabhan, “GHIC: A Hierarchical Pattern-Based Clustering Algorithm for Grouping Web Transactions”, *In Proceedikng of IEEE Transaction On Knoeledge and Data Engineering*, Vol. 17, No. 9, pp. 1300-1304, 2005.

[12] M. M. Ozdal and C. Aykanat, “Hypergraph Models and Algorithms for Data-Pattern-Based Clustering”, *Springer Journal of Data Mining and Knowledge Discovery*, Vol. 9, No. 1, pp. 29-57, 2004.

[13] H. Wang, W. Wang, J. Yang and P. S. Yu, “Clustering by Pattern Similarity in Large Data Sets”, *In Proceedings of ACM SIMMOD International Conference on Management of Data*, pp. 394-405, 2015.

[14] G. Alexe, S. Alexe and P. L. Hammer, “Pattern-Based Clustering and Attribute Analysis”, *Journal of Soft Computing*, Vol. 10, No. 5, pp. 442-452, 2006.

[15] A. Hatamlou, “In search of optimal centroids on data clustering using a binary search algorithm”, *Pattern Recognition Letters*, Vol. 33, No. 13, pp. 1756-1760, 2012.

[16] M. B. Dowlatshahi and H. Nezamabadi-pour, “GGSA: A Grouping Gravitational Search Algorithm for data clustering”, *Engineering Applications of Artificial Intelligence*, Vol. 36, pp. 114-121, 2014.

[17] L. Liao, X. Shen and Y. Zhang, “Image Segmentation Based on Fast Kernelized Fuzzy Clustering Analysis”, *In Proceedings of IEEE Eighth International Conference on Fuzzy System and Knowledge Discovery (FSKD)*, Vol. 1, pp. 438-442, July 2011.

[18] G. Chicco, “Overview and performance assessment of the clustering methods for electrical load pattern grouping”, *ELSEVIER Journal of 8th World Energy System*, Vol. 42, No. 1, pp. 68-80, 2012.

[19] L. Galluccio, O. Michel, P. Comon, M. Kliger and A. O. Hero, “Clustering with a new distance measure based on a dual-rooted tree”, *ELSEVIER Journal of Information Sciences*, Vol. 251, No. 1, pp. 96-113, 2013.

[20] J. Pei, X. Zhang, M. Cho, H. Wang and P. S. Yu, "MaPle: A Fast Algorithm for Maximal Pattern-based Clustering", *In Proceedings of IEEE International Conference on Data Mining*, pp. 259-266, Nov 2003.

[21] A. Zimek, I. Assent and J. Vreeken, “Frequent Pattern Mining Algorithms for Data Clustering”, *International Springer Journal of Frequent Pattern Mining*, pp. 403-423, 2014.