Turkish Journal of Computer and Mathematics Education *Vol.12 No.2 (2021), 1396-1400*

Research Article

An Efficient Class-Based Data Clustering Through K-Means And Knn Approach

Kailash Patidar [a], Dhanraj Verma [b],

[a] *Department of Computer Science and Engineering,Dr. A.P.J. Abdul Kalam University, Indore – 452016*
[b] *Department of Computer Science and Engineering,Dr. A.P.J. Abdul Kalam University, Indore – 452016,*

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: In this paper an efficient class-based data clustering through k-means and KNN approach were applied. It has been applied for proper data grouping and efficient classification. In this approach three dataset have been considered for the experimentation. Data preprocessing has been performed for the removal of unmatched or empty entry. Then weight assignment and normalization has been performed. K-means and k-nearest neighbor (KNN) have been applied on the dataset for the data grouping and classification purpose. Different splitting variations have been considered with higher variance. Data selection is completely random. The results obtained shows the strength of our approach though classification and class-based clustering.

Keywords: K-means, KNN, Clustering, Classification

1Introduction

In the present scenario in every field there is huge amount of data gathered, processed and extracted for different purposes [1-3]. To acquire the data in the meaningful way is very important [4-6]. The main areas are health care, education, business enterprises and so on. If we think of data arrangement and data management in the way that meaningful extraction is possible then the data mining (DM) and machine learning algorithms may find to be useful for different purposes [7, 8].

DM algorithm provides the way to extract the meaningful insights from the huge dataset. There are several DM algorithms which are useful in the direction of patter extractions for example association rule mining, clustering, classification, etc. [7−10]. The primarily operation needed in the arrangement of the data is data clustering. In which data can be arranged in similar groups of the same properties. K-means and fuzzy c-means algorithms are widely used clustering algorithms [11, 12]. In the current scenario there is different complex solution where there is the need of proper thresholding and refinement [13]. In this case clustering alone may fail and evolutionary algorithms may be helpful with soft computing techniques. Some of the famous evolutionary algorithms are ant colony optimization (ACO), particle swarm optimization (PSO), teaching learning-based optimization (TLBO), Cuckoo Search, etc. [14,15].Different machine learning algorithms are found to be useful in the extraction and data categorization process [16, 17]. Some of the machine learning algorithms are support vector machine (SVM), logistic regression (LR), naïve Bayes (NB), K-Nearest Neighbor (KNN), etc. [18-20].

In 2020, Chug and Baweja [21] discussed the approach of how to retain the customers. They took the data of 150 customers and made the three clusters. They found the age an important parameter for the clustering. They discussed two techniques of clustering. In the first phase, they applied the k-means algorithm and for the second approach, they applied the agglomerative clustering. Their approach is helpful in segmentation for the existing customers, which may be helpful in grouping of psychographic data.

In 2020, Brown and Shi [22] discussed the different steps used to implement the clustering algorithm i.e. Fast Density-Grid. They used the Apache Spark with this so that they can parallelize it. They performed it in three stages like Grid Space Density, Determination by Densest Neighbor and Generation of cluster. There results may be helpful in parallel implementation and efficient even if the data exceeds some limit.

In 2020, Chebanenko et al. [23] proposed the Fuzzy system, which can be used to estimate the patient's compliance for primary as well as further consultancy system. They performed the cluster analysis by using Fuzzy c-average algorithm on the patients' data. According to this rate, they formed the three groups. There proposed method may be helpful for clinicians to discover cardiac patients in advance. The efficiency of cardiologists may be improved.

In 2020, Gong et al. [24] discussed about the visualization by which the validity of data clustering of electricity may improve. Their approach is found to be impactful in the improvement of the clustering of electricity databy their proposed a visual cluster analysis framework. They have suggested different characteristics of electricity as the future recommendation.

In 2020, Kang et al. [25] discussed the importance of cluster analysis in identifying the faults of automobile. They proposed a fault analysis model. It considered the various factors like deterioration failure, environmental factor and various human factor in their model. They applied a clustering algorithm and form a pedigree map. After pre-processing the data, they convert that into a form of ratio matrix. Further, they applied the k-value

clustering. Their results showed that cluster analysis provides and analysis of driving habits. They showed that it may be good for auxiliary role and can be used for fault detection in later cases.
In 2020, Kesheng et al. [26] analysed the behaviour of 3,245 students at B grade universities. They used the network data from campus usage. They used the data mining for cleaning the data. They pre-process the data by using various methods like Kernel method, linear interpolation and spline method. They used the R language with the use of Rstudio software. Their result grouped them into four groups and found around 350 students have the more internet usage. This data is helpful for student affair management, which provides support to counsellor to improve the professional level.
In 2020, Roy et al. [27] proposed a classification approach, which is helpful in detecting the disease and segmentation. The acquired the MRI images as input and pre-processed it. The used the multimodal Naïve Bayes for the classification. They prepared the confusion matrix for the performance. Further, they applied the segmentation. There proposed method is Modified-C Clustering method. This study emphasized on different classification techniques so that they can perform pattern detection. They suggested to apply this method in Hospital so that they can automate the system for image classification. Further, they suggested to embed it with AI applications for optimizing the medical facility in society.
In 2020, Huijuan and Zhenjiang [28] discussed the significance of clustering in data mining. They used the k-means clustering in the mining process of data available in huge amount. They performed a survey of 1,369 students over various learning centres. The have used the SPSS tool for displaying and analysing the results in online teaching. Their results showed that 96.7% students are satisfied and 2.7% dissatisfied.
In 2020, Shi et al. [29] proposed a wind power data-clustering algorithm for solving the problem of cluster centres and their uncertainty of cluster numbers. This method is based on density peak modified fuzzy c-means of soft clustering. They used this method for the prediction of wind power. This improved version of algorithm shows the more stability of clustering quality, effectiveness and convergence.
In 2020, Sudhagar and Renjith [30] proposed an approach of Enhanced Design Weighted Mean Shift Ensemble Clustering for dealing with high dimensional records. This approach gives the higher accuracy rates of clustering results. They showed the comparison of various clustering algorithms available and then proposed the ensemble approach, which is a combination of one or more than approaches of clustering. This approach is capable in during clustering process of handling the incomplete basic partitions available in the data.
The main objective of this paper is to apply data clustering efficiently and allow heterogeneous environment for accepting wider dataset from the data repository. This paper also explores the k-centroid classification based on the variations of data clustering applied.

2.Materials and Methods

The dataset has been considered from the UCI repository [19]. The following dataset has been used for the experimentation.

1. Breast Cancer Wisconsin Dataset (D1): The number of instances in this dataset are 699. Total number of attributes are 10. Class labels are two.
2. Statlog Dataset (D2): The number of instances in this dataset are 270. Total number of attributes are 13. Class labels are three.
3. Hepatitis Dataset (D3): The number of instances in this dataset are 155. Total number of attributes are 19. Class labels are two.

In this paper an efficient class-based data clustering through k-means and KNN approach was presented for the proper data categorization. It includes grouping and classification. Our approach is divided into the following parts:

1. Preprocessing
2. Feature selection
3. Splitting ratios
4. Validations

In data preprocessing cleaning of data has been performed as there is the possibility of missing values, null and other related things which may affect the performance in the further procedure. In feature selection procedure only, relevant domains have been considered for extraction which have significant variance for the next procedure. Then different variable splitting ratios have been considered for the training and testing of the data. Then finally the model has been evaluated based on precision, recall and accuracy. K-means and KNN have been used for the purpose of clustering and classification purpose. Random centroid selection has been performed for the unbiased selection. The proposed working flowchart depicts the complete working procedure (Figure 1). It explores the complete working scenario with the clustering and classification procedures.
The algorithms of these approaches are as follows:

K-means algorithm

Step 1: Final input set from the selected dataset. Preprocessed input has been processed.

Step 2: Random centroid initialization has been performed for unbiased selection.
Step 3: The weight has been assigned randomly.
Step 4: Distance have been calculated based on Euclidean distance algorithm. It is as follows:

X(c)= $\sum_{j=1}^{k}\sum_{i=1}^{n}||d_i^{(j)} - c_2$

d_i $c_j$shows the group distance
k shows the cluster numbers
n denotes the total iterations
Step 5: The above process is repeated till the epoch or the repletion of similar results

c_i = ( ) $\sum_{j=1}^{n_i}$
Step 4: Grouping has been performed based on similarity.
Step 6: Final grouping based on the data point has been performed.

KNN algorithm

Step1: Final input set from the selected dataset. Preprocessed input has been processed.
Step 2: K selection has been performed for the neighbors
Step 3: Then distance calculation has been performed from the selected K neighbors. In our case Euclidean distance has been considered.
Step 4: Consider the nearest neighbor for the further processing.
Step 5: Data point has been considered and calculated from the above procedure.
Step 6: Data point assignment and updating have been performed.
Step 7: Final classified data has been obtained.

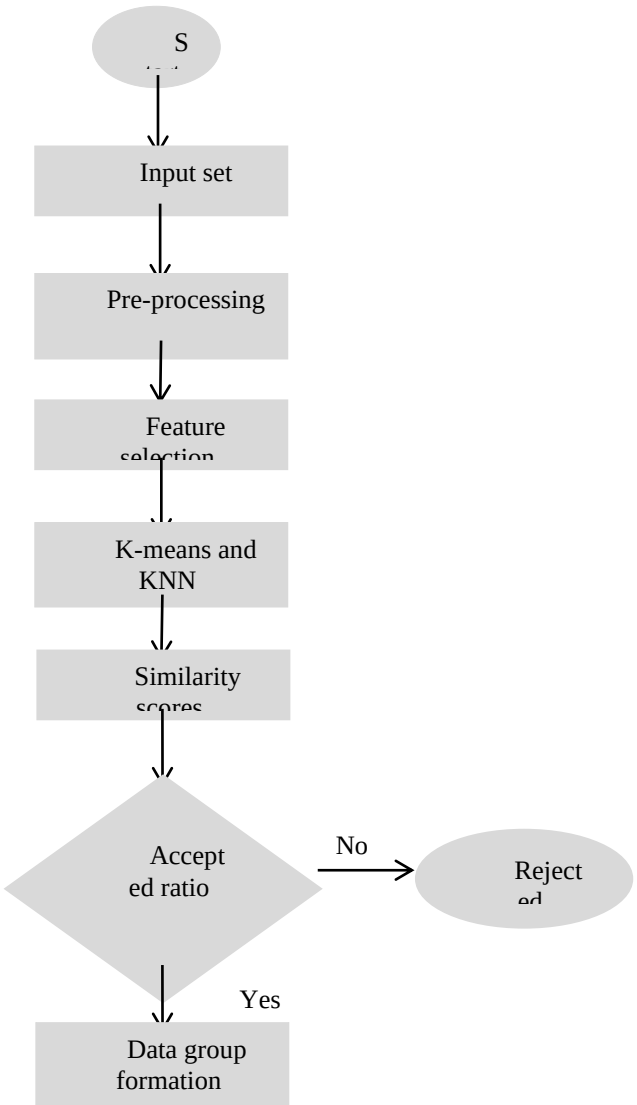

Figure 1: Flowchart for the working procedure

3.Results

In this section results were evaluated based on the k-means clustering and KNN classifier. The measures considered for the performance evaluation are as follows:

Accuracy = (True Positive +True Negative)/ (True Positive +False Positive +False Negative +True Negative)

Precision = True Positive/ (True Positive +False Positive)

Recall = True positive / (True Positive +False Negative)

For the unbiased system random selection from the dataset were considered. Splitting ratio considered was 30% to 20%.Figure 2 shows the precision based on k-means and KNN. Figure 3 shows the Recall based on k-means and KNN. Figure 4 shows the accuracy based on k-means and KNN.

It is clear from the results that the validation ratio variations have very negligible impact if the variations are in the range of 30% to 20%. KNN was found to be prominent in all cases. The prominence is due to the classification based on efficient grouping.

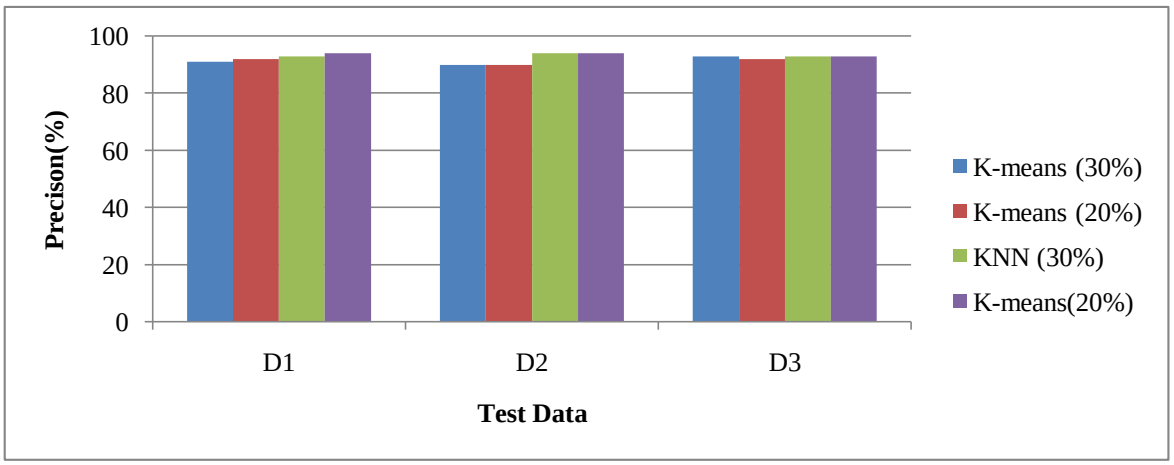

Figure 2: Precision basedon k-means and KNN

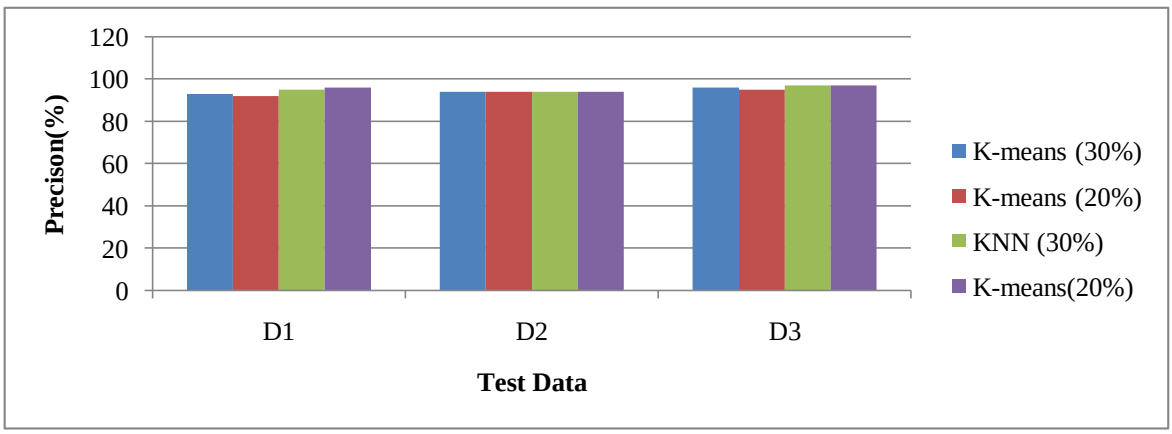

Figure 3: Recall based on k-means and KNN

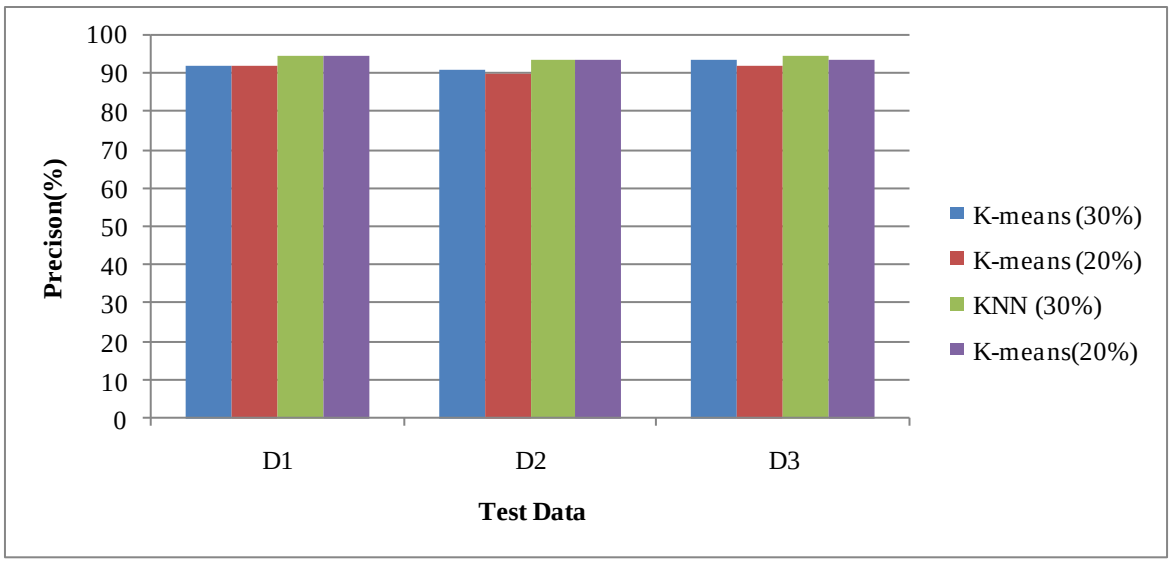

Figure 4: Accuracy based on k-means and KNN

4.Conclusion

In this paper a combination of clustering and classification algorithm were used for the purpose of class-based grouping. This approach is found to be helpful in the data classification along with the splitting variations. For the result evaluation accuracy, precision and recall values were considered. It has been found that the KNN algorithm was better in terms of the performance measures. In case of splitting minor variations have been observed which can be neglected. The results also indicate the need of different other classifier with optimization techniques for the betterment and future framework design.

References

1. Dubey AK, Kumar A, Agrawal R. An efficient ACO-PSO-based framework for data classification and preprocessing in big data. Evolutionary Intelligence. 2020 Sep 9:1-4.
2. Dubey AK, Gupta U, Jain S. Computational Measure of Cancer Using Data Mining and Optimization. InInternational Conference on Sustainable Communication Networks and Application 2019 Jul 30 (pp. 626-632). Springer, Cham.
3. Agarwal R, Srikant R. Fast algorithms for mining association rules. InProc. of the 20th VLDB Conference 1994 Sep 12 (pp. 487-499).
4. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. ACM sigmod record. 2000 May 16;29(2):1-2.
5. Jamil A, Salam A, Amin F. Performance evaluation of top-k sequential mining methods on synthetic and real datasets. International Journal of Advanced Computer Research. 2017 Sep 1;7(32):176.
6. Kumari I, Sharma V. A review for the efficient clustering based on distance and the calculation of centroid. International Journal of Advanced Technology and Engineering Exploration. 2020 Feb 1;7(63):48-52.