Predicting Class Label Using Clustering-Classification Technique: A Comparative Study

Aseel faysal Alshaibanee
Department of Computer Science
Faculty of CS and Mathematics
University of Kufa, Iraq
aseelfaysalalshaibanee@gmail.com
https://orcid.org/0000-0002-3190-0314

Kadhim B. S. AlJanabi
Department of Computer Science
Faculty of CS and Mathematics
University of Kufa, Iraq
kadhim.aljanabi@uokufa.edu.iq
https://orcid.org/0000-0001-9529-642X

DOI: http://dx.doi.org/10.31642/JoKMC/2018/100101

Received Apr. 11, 2022. Accepted for publication Jul. 17, 2022

Abstract:

Among different techniques, algorithms and applications of Data Mining, predicting the class label of unlabeled objects(undefined class label) is a crucial term in the field. The most common approaches in this area is the use of classification technique (DT, Bayes, SVM, KNN and others) that represent what is known as supervised learning. However, in many cases no target class labels and the boundaries are available to perform the prediction, so the new approach Clustering-classification technique is used.

The work in this paper presents a survey of the most common researches conducted in this field and discuss their experiments, the algorithms they used, the types of data they utilized, the data sizes used, and the results they discovered.

According to the results, applying the clustering techniques before classification improved classification accuracy and reduced experiment execution time. The Cluster Classifier was proven to be a suitable approach to summarize data by some of the researchers. It achieves a summarization rate of over 50%, which represents a considerable reduction in the size of the test datasets .

The findings of the researches indicated that, in addition to feature selection and feature extraction, data preprocessing (handled missing data and effective outlier detection techniques) enhanced the classifier performance and accuracy while reducing the classification error.

keyword: Cluster-Classifier, Clustering, Classification, Data Mining

1. Introduction

In today's world, large amounts of data are generated constantly. Analyzing such data is a critical need. Data mining can provide tools for meeting this need. It's a process of finding novel and potentially useful information from large amounts of data. It is a process that extracts potentially useful information from raw data, commonly known as knowledge discovery in databases (KDD). An Outline of the Steps of the KDD Process is shown in figure (1).[1],[2],[3].

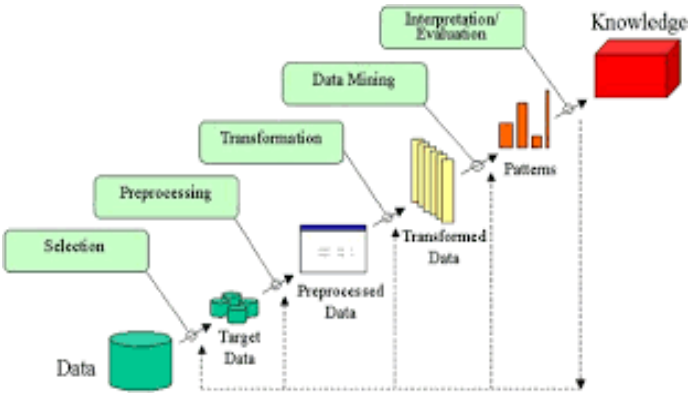


Figure (1) KDD Stages

1-Identify the goal and gain a better grasp of the application domain.

2-Build a target dataset that you want to analyze.

3-Cleaning and preprocessing of data

4-Reduction and projection of data

5-Choosing the right data-mining task

6-Data-mining algorithm(s) selection

7-Data mining

8-Visualization and interpretation

9-Consolidating discovered knowledge

The criteria that can be used to classify a data mining system are :

Visualization of data, databases, statistics, and machine learning Additional Subjects.

DM is a multidisciplinary technique as shown in figure (2). [4]

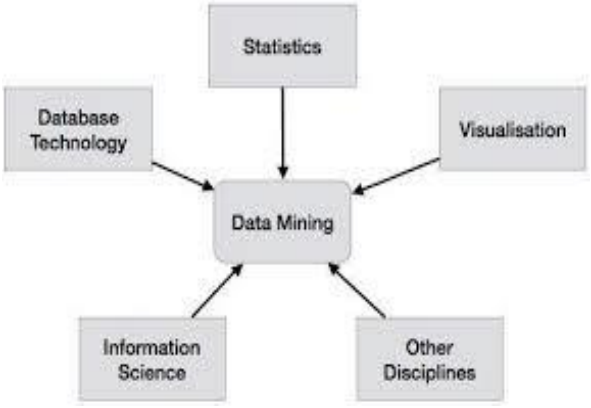
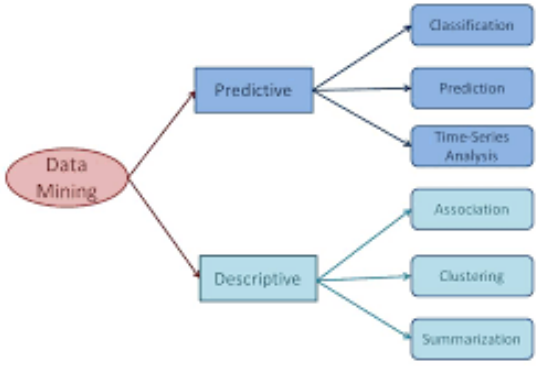


Figure (2) Data Mining multidisciplinary.

Also, data mining system can be classified based on the type of (a) databases mined, (b) knowledge mined, (c) methodologies used, and (d) applications adopted, among other things.

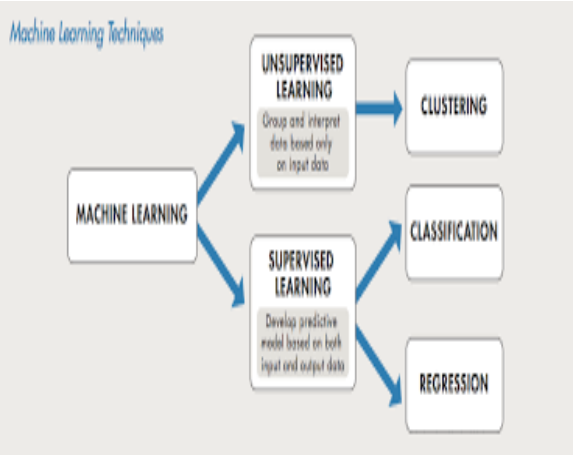
Data mining tasks include classification, prediction, time-series analysis, association, clustering, summarization, etc. Predictive and descriptive data mining are the two types of data mining tasks. A data mining system can perform one or more of the functions described above as part of data mining.[5]



Figure(3) Data Mining Task Categorization

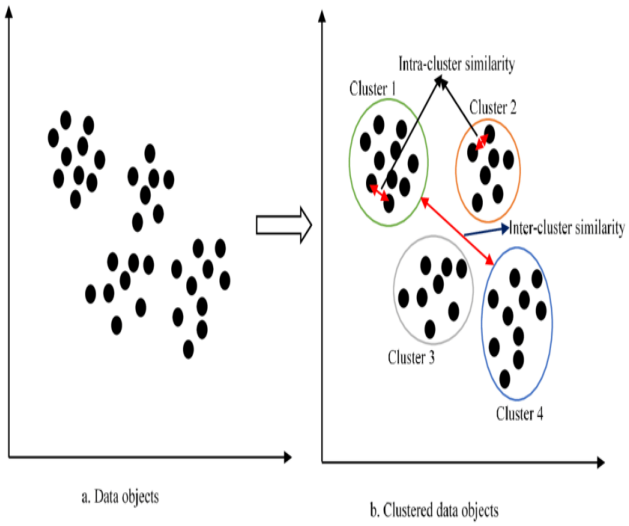
Predictive data mining tasks create a model from a data set that can predict unknown or upcoming results in a different collection of data. Descriptive data mining is used to find patterns in data and produce new, helpful info from it.

Data mining is a multidisciplinary field that focuses on discovering data set properties. Finding properties of data sets can be done in a variety of ways. One of them is Machine Learning. Machine learning is a subdivision field of data science concerned with developing algorithms to learn from and predict data. These algorithms can be employed in data mining. Supervised and Unsupervised Learning are two types of machine learning methodologies. Machine learning uses supervised learning, which includes training a model to anticipate future outcomes using known input and output data, and unsupervised learning, which involves uncovering hidden structures in unlabeled input data.[6][7]



Figure(4) Machine Learning techniques

1.1. Clustering is the process of grouping together similar data points using unsupervised learning algorithms. In clustering, many data points are assigned to a smaller number of groups; accordingly, data points in the same group share similar properties, while those in other groups do not.[8][9]



Figure(5) Clustering Process

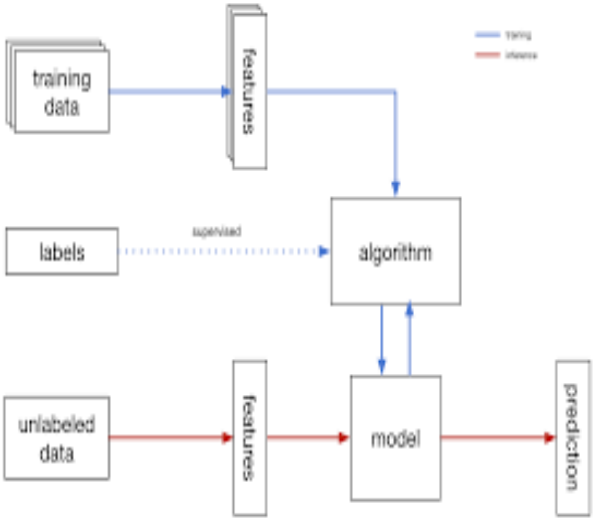
1.2. Classifying each item in a sample of data into one of a specified set of classes or groups is known as classification; classification in data analysis is the process of building a classifier that can predict classification labels (the class labels). In data mining, Classification is the process of assigning objects in a collection to certain types or categories. The goal is to predict the target class appropriately for each occurrence. as an example It might be used to classify loan applicants as high, medium, or low risk

The classification of data consists of two steps.

I. a classifier is constructed to describe a predetermined set of data classes or concepts.

II. A model is used to classify the data.

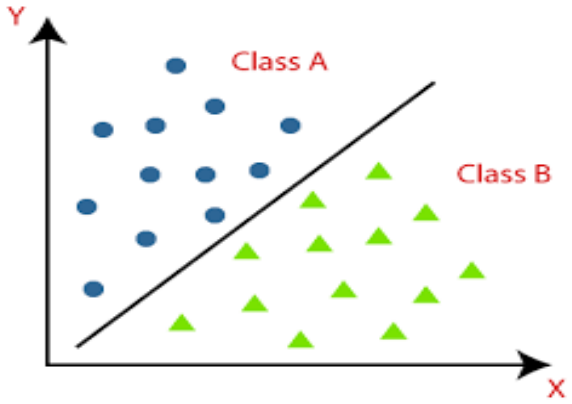
Classification is the most widely used technique to predict the class labels of the unknown objects where predefined attribute values are available. However, many cases require assigning labels to unidentified objects without these attribute values. Classification is a technique in supervised Machine learning where the algorithm builds the predictive model from the training data after that tries to predict unknown labels. The fitted model would try to predict the most likely labels for a new set of samples in the testing set as the figure(6) [10].



Figure(6) Building the Predictive Model

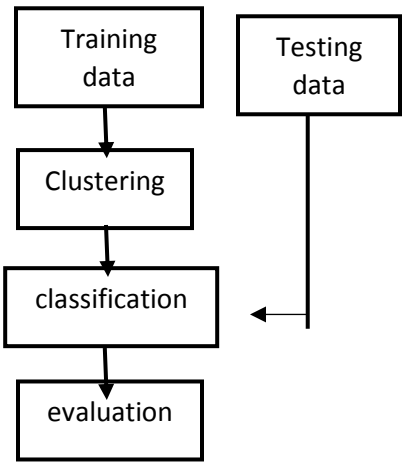
The process of categorizing a set of data into groups is known as classification. Both organized and unstructured data can be used. The initial phase in the technique is to predict the class of data points presented. The classes are described using terms like target, label, and categories.[11]

Predictive modeling is the process of calculating the mapping function from discrete input variables to discrete output variables. The basic objective is to determine which category or class the new information falls into.[12]



Figure(7) Classification Process

1.3. Clustering_Classifier is a model to classify objects after performing the clustering process and construct the class label to solve such problems. A clustering algorithm aims to group data with similar characteristics into groups and build the classes labels from the centroids of the clusters to do the classification next. Also, recent studies have demonstrated that combining classification with clustering models can improve classification results.



Figure(8):Clustering Classifier Process

In this paper, we will present some research that covers that approach.

2. Literature Survey

I. Sakib Khan and Shakim Ahamed, et al, 2020 [13], they proposed an approach for Classifying Big Data Based on Similarities. The idea of the research aims to categorize huge data by creating a classifier based on the samples' similarities rather than their class labels. The proposed research approach divides large amounts of data into groups or clusters and then creates a classification system based on the clusters. They used several decision tree learning algorithms to design and build classifiers: ID3, C4.5, and CART. The decision tree (DT) is a widely used machine learning algorithm for data mining tasks. The DT technique can be used in supervised learning to solve many real-life classification and regression problems.

The authors offer a classification strategy for generating classifiers that ignore the class labels of data samples, using the Classification by Clustering (CbC) method. At firstThe huge data is broken down into smaller sub-datasets. After that, each sub-dataset is subjected to a clustering technique. A similarity-based clustering algorithm is used in the proposed method, as it is a reliable way for grouping instances groups with similar characteristics Cluster numbers are generated automatically using similarity-based clustering. and forms different groups of clusters. Following that, sub-datasets are used to construct numerous decision trees. Lastly, all the DT are reviewed and combined into a single decision tree that is highly similar to the tree that would have been formed if the enormous data had fit in memory.

the authors used 10 real datasets from the UCI machine learning repository.

For experimental setup and design, they employed Python's scikit-learn machine learning toolkit, as well as Spyder 3.2.6 for Python coding.

For all datasets except the Character Font dataset, the proposed technique increases classification results compared to traditional DT algorithms (ID3, C4.5, and CART). The classification accuracy of the Dota2 Games dataset improves from 52% to 98% , while the Adult dataset improves from 81% to 99% .

II. Reuben Evans, et al, 2011[14], they proposed an approach for Clustering for Classification. The research goal was to show how clustering classifiers may summarize massive datasets by employing cluster centroids to create a more compact representation.. The researchers use the center of a cluster as a meta-data point to characterize the data points that make up that cluster.meta data points. Some clusters, such as KMeans, create easily accessible cluster centroids. The Cluster Classifier uses these centroids as meta-data points directly. On the other hand, if the cluster produces unweighted or no cluster centers at all, and so is the situation with model-based approaches such as Expectation-Maximization, the Cluster Classifier can construct them from the cluster predictions. To produce the meta-data element for that cluster, the Clustering Classifier utilizes the mean values of each characteristic across all data points given to that cluster.

They employed five different clustering algorithms, each of which could produce a specific number of clusters (First K, K-

Means, FF, Bisecting KM, and Expectation-Maximization). Researchers employed 19 datasets (9 nominal datasets like KR-vs-KP, hypothyroid, Agrawal, Waveform, Waveform40, and ten numeric datasets like 2DPlanes, fried, mv, ailerons, and house8L).

The research uses two different methods using nominal datasets: Nave Bayes and Logistic Regression. Naive Bayes was selected also as a simple classifier due to its speed and efficiency. The complex classifier was chosen since logistic regression is often parameter-free and fast.

Numeric datasets are used in the studies of Linear Regression and Model Trees, Because linear regression is a rapid and well-known regression approach that employs just simple data relationships, it was chosen as the simple classifier. MT was also chosen for its speed and ability to find more complicated links.

The results show that the Cluster Classifier is a good way to summarize data. Clustering frequently achieves over 50% summarization, This is a considerable reduction in the study's massive datasets. However, the results demonstrate that when using the Cluster Classifier, it is critical to choose a cluster properly. In the experiments, three clusters have proven to be useful inside this framework. On some datasets, the Farthest First, Bisecting KM, and First K all perform well. None of these clustering algorithms, however, is always preferred to the others. Based on the dataset's format to be grouped, the appropriate clustering method differs.

III. Rasoul Kiani, et al, 2015[15], they proposed an approach for Analysis and Prediction of Crimes by Clustering and Classification. The main goal of this research is to categorize clustered crimes based on their frequency of occurrence across various years. They analyzed an actual crime dataset gathered by the police in England and Wales between 1990 and 2011 using a theory based on DM techniques like clustering and classification.

The preprocessing phase included (reading the crime dataset using the Read Excel operator, filtering the dataset according to requirements, replacing missing values with the Replace Missing Value operator, detecting outliers using the Outlier Detection(Distance) operator, and optimizing the Outlier Distance operator parameters using the Genetic Algorithm (GA), and Store a new dataset).

The researcher used the K-means algorithm in clustering and the Decision tree operator in the classification Phase.

After optimizing the variables of the Outlier Detection operator, classification accuracy increased, classification error decreased, and the fitness function derived was optimal when the number of clusters was identical.

The GA was employed in this framework to improve outlier detection during the preprocessing phase, and the fitness function was determined using accuracy and classification error parameters. To optimize the clustering process, the features were weighted and low-value features were deleted using an acceptable threshold. The suggested approach was used to compare the quality and efficacy of optimized and non-optimized parameters.

The following are the main goals of the new framework for crime clustering and classification:

1. Training and testing data collecting
2. Dealing with high-dimensional data issues by removing low-value features using a weighting approach3. Using GA, optimize the Outlier operator parameters.

IV. Asha Gowda Karegowda , et al 2012[16], they proposed an approach for Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients. The construction of a hybrid model for classifying the Pima Indian diabetic database (PIDD) is presented in this research. There are three steps to the model. K-means clustering is used in the initial stage to discover and delete instances that have been wrongly classified. In the second step, essential attributes are extracted using a combination of genetic techniques and correlation-based feature selection (CFS), with GA rendering a global search of characteristics and CFS affecting fitness evaluation. Lastly, in the third step, employing K-nearest neighbor (KNN), a fine-tuned classification is conducted by feeding the KNN the suitably grouped instance from the first step as well as the feature subset obtained in the second step.

KNNs are lazy learners or instance-based. It postpones the modeling of the training data until the test samples must be classified. It can both classify and predict. The training samples are described by n-dimensional numeric properties, and they are stored in n-dimensional space. The KNN classifier looks for the k closest adjacent occurrences when given a test sample (with an unknown class label). Distance is frequently calculated using the Euclidean distance.

In the first stage of the proposed model, 392 diabetic patient samples obtained are clustered using the Weka tool using simple K-means clustering (with k = 2). The samples that were incorrectly classified are removed from the final 299 samples. In the second stage, cascaded GA CFS identifies significant features. Feature selection in supervised learning tries to enhance classification accuracy. Finally, using the 70-30 ratio partitioning strategy, the correctly classified samples from the first stage and the relevant features found in the second stage were fed into the Weka KNN classifier in the third stage (training-test).

Experiments were conducted for various k values ranging from 1 to 15.

With k = 5, the diabetic data set utilizing the suggested technique without feature selection achieves 95.56 percent. The accuracy of the presented method is determined to be 96.67 percent with k= 5, sensitivity and specificity are 100 and 88, respectively, with feature selection using GA CFS.

V. Yaswanth Kumar Alapatie, et al, 2016[17], they proposed an approach for Combining Clustering with Classification: A Technique to Improve Classification Accuracy. The proposed method in the research demonstrates that the classifier works well with clustered data, i.e., cluster the data before applying any classification algorithm to the dataset and then use the classification algorithm. As a result, the classifier's accuracy is enhanced. Apply the Feature Selection Algorithm first for high-dimensional datasets. It is critical to choose a clustering algorithm carefully for each dataset.

The suggested framework contains several phases.

1) Selection of Features

The process of identifying a subset of the most valuable features that deliver the same results as the whole collection of features is known as feature selection.

2) Creating clusters

Apply a clustering technique to the reduced dataset after it has been reduced in dimensionality. Add the cluster id to the dataset after clustering. Kmeans and hierarchical clustering are the clustering

techniques employed in the suggested framework.

3) Classification

Work with clustered data to apply a classification technique. The classification methods employed in the suggested system are naive Bayes Classifier and Neural Network Classifier.

Lung Cancer, Coil2000, Mfeat-Fourier, and Arrhythmia are some of the benchmark datasets used to test the strength of the suggested technique.

The studies suggest that clustering techniques are better than classification algorithms.

According to the findings, using Feature Subset Selection Algorithms enhances a classifier's accuracy.

VI. Norsyela Muhammad Noor, et al 2018[18], they proposed an approach for Improving Classification Accuracy Using Clustering Technique. The research employed a large- to evaluate the efficiency of clustering algorithms on a large-scale real-world data collection in improving classification models. Before using the clustering technique, the study employs standard text classification procedures such as extracting features, extraction of features, and preprocessing

The data for this study was acquired from Tesco stores using prototype web scrapers produced by the Department of Statistics Malaysia under the STATSBDA project titled Price Intelligence (PI) (DOSM). This study's data corpus is the baby products data corpus. It includes information on baby food, infant amenities, diapers and wipes, and milk powder, among other things. There are 11419 products in all of these nodes, and there are 401 features in the product descriptions.

Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF) are some of the classification techniques used in the study. In clustering, hierarchical clustering and the K-means algorithm are utilized.

The findings show the accuracy of a classifier by utilizing hierarchical and K-means clustering techniques before recognizing items in the dataset. When feature selection is applied to the clustered dataset, The number of features has been lowered from 275 to 257 and 260, respectively, for hierarchical and K-means clustering. Compared to the other three classification models, the accuracy of the KNN model is still the best model for classifying baby products. The K Nearest neighbor model is a straightforward and useful model for a variety of applications, including textual data mining. Similarly, on the Reuters corpus of newswire items, the K Nearest neighbor model is one of the most important classification methods. a benchmark corpus for text categorization. In clustering, hierarchical clustering and the K-means algorithm are utilized.

VII. Raksha K.Mundhe, et al 2014[29]. They proposed an approach to the automatic Labelling and Clustering of documents for Forensic Analysis. First, they use the database of crime files. The database contains various text files, images, audio and video files. Then, they have to apply the DBSCAN clustering to the database. Once the file goes to the specified location, they get the linked data to the forensic analysis, such as Robbery files, murder files, etc.; for this purpose, K-mean is employed to meet the generated files. After these steps, the automatic labelling for the documents was done.

Table (1) Summary of the mentioned Related Work

	Author(s)	Data used	Algorithms and techniques	Results	Evaluation criteria	advantages	recommendations
1	Sakib Khan and Shakim Ahamed et al , 2020	UCI machine learning repository	ID3, C4.5, and CART	The experimental results demonstrated that the suggested strategy of classifying instances based on their similarities improved classification results.	Accuracy	In comparison to classic decision tree algorithms, the suggested technique improves classification results.	It's used when the size of data is big and the number of features is between 40 to 650 and the type of data are categorical, integer, and real
2	Reuben Evans, et al, 2011	UCI machine learning repository	First K, K-Means, Farthest First, Bisecting K-means, and Expectation-Maximization, Nave Bayes and Logistic Regression	The results show that the method used achieves over 50% summarization	Reduction Achieved from applying the method	The presented technique demonstrates that the Cluster Classifier is a useful data summarizing tool.	It's used when the number of features is between 9 to 65 and the type of data is nominal and numeric and used when the summarization of data is needed.
3	Rasoul Kiani, et al, 2015	real world dataset	GA, K-means , Decision tree	The accuracy of classification increased from 85.74% to 91.64%	Accuracy	the classification accuracy increased and classification error decreased	It's used when the preprocessing phase and outliers detection is an important issue
4	Asha Gowda Karegowda , et al 2012	The PIMA diabetic database	K-means, (GA) , (CFS), KNN	Increase the accuracy from 95.56% to 96.76%	Accuracy	Improve the classification accuracy	It's used when the preprocessing phase and feature selection is an important issue and the dataset used is The PIMA diabetic database.

5	Yaswanth Kumar Alapatie, et al, 2016	Lung Cancer, Coil2000, Mfeat-Fourier, Arrhythmia	Kmeans, hierarchical clustering, naive Bayes, Neural Network	The accuracy improved after applying the clustering and feature selection before classification	Accuracy	improve the accuracy of a classification algorithm	It's used when the number of features is between 50 to 300 and the size of data is small.
6	Norsyela Muhammad Noor, et al 2018	Tesco online stores	Naive Bayes, Support Vector Machine(SVM), K-Nearest Neighbor(KNN), Random Forest (RF), hierarchical clustering, K-means	clustering strategies aid in enhancing the accuracy of classification algorithms. K-means clustering, on the other hand, when compared to Hierarchical clustering, K-means clustering appears to give better computation time. features reduced from 401 to 257 and 260	Accuracy, Execution Time	Clustering algorithms assist in the improvement of classification algorithms' accuracy.	It's used when the data size is small and from one type of 'categorical ' data type.
7	Raksha K.Mundhe, et al 2014	Crime files database	DBSCAN, K-means clustering algorithms	Using the document clustering approach, the suggested system yields the optimal result, In the old system, searching for files	Searching Time	The automated labeling technique enables quick and accurate analysis, reduces human work	it is used when the information retrieval of document files is needed-

3. CONCLUSION

From the survey conducted of different clustering-classification approach to predict the class label, it show that using The classification results are improved when the cases are similar. Applying the clustering before the ID3, C4.5, and CART on benchmark datasets improves the classification results and the performance of the algorithms.

The Cluster Classifier is a successful approach to data summarization, according to the results of Section IV. Clustering produces over 50% summarization in many situations, which is a considerable decrease in the enormous datasets utilized in the tests.

According to the authors in section V, Feature Selection Algorithms can improve the accuracy of a classifier and decrease the classification time by using them before the k-means clustering method, and it's similar to the results of the authors in section VI they found that by choosing the important features with K-means clustering, the time for classifying the datasets may be reduced, and K-means clustering appears to give faster computation time than Hierarchical clustering.

Outliers' impact on data mining preprocessing, as well as the number of features, were determined to be important by several authors. The GA was utilized to improve outlier detection and classification accuracy during the preprocessing phase.

All of the researchers in this publication discovered that clustering before classification enhances classification accuracy and the ability to find class labels.

4. RECOMMENDATIONS

1- Dealing with a massive amount of data in different data mining techniques requires data reduction, feature selection and extraction to improve the algorithms performance.

2- When automatic labelling of unsupervised data is required, the proposed approaches are suggested

5.REFRENCES

[1] J. Han, M. Kamber, And Jian Pei, “Data Mining: Concepts And Techniques,” The Morgan Kaufmann Series In Data Management Systems, 2012.

[2] Shrishrimal, P & Deshmukh, Ratnadeep & Waghmare, Dr. Vishal, “Multimedia Data Mining: A Review,” International Conference On Recent Trends And Challenges In Science And Technology, 2014.

[3] Ryan S.J.D. Baker, “Data Mining For Education,” International Encyclopedia Of Education, 2010.

[4] Rabia Saleem, Sania Shaukat, “Denormalization To Enhance Effciency In Data Mining,” International Journal Of Scientific & Engineering Research, Vol. 7, Issue 9, September 2016.

[5] Mirela Danubianu, Et Al, “Model Of A Data Mining System For Personalized Therapy Of Speech Disorders,” Journal Of Applied Computer Science And Mathematics, 2018.

[6] Manish Kumar Aery And Chet Ram, “A Review On Machine Learning: Trends And Future Prospects,” An International Journal Of Engineering Sciences, Vol. 25, November 2017.

[7] Vladimir Nasteski, “An Overview Of The Supervised Machine Learning Methods,” Horizons, Vol. 4, 2017, Pp. 51-62.

[8] Sankar Rajagopal, ”Customer Data Clustering Using Data Mining Technique,” International Journal Of Database Management Systems (Ijdms), Vol.3, No.4, November 2011.

[9] Absalom Ezugwu, Et Al ,”Automatic Clustering Algorithms: A Systematic

Review And Bibliometric Analysis Of Relevant Literature," Neural Computing And Applications, 2021.

[10] Syed Muhammad Raza Abidi, Et Al, "Popularity Prediction Of Movies: From Statistical Modeling To Machine Learning Techniques," Springer Science & Business Media, January 2020.

[11] G. Kesavaraj And S. Sukumaran, "A Study On Classification Techniques In Data Mining," In 2013 Fourth International Conference On Computing, Communications And Networking Technologies (ICCCNT), Tiruchengode, India, 2013, Pp. 1-7. Doi:10.1109/ICCCNT.2013.6726842

[12] Md Mustafa Md-Muziman-Syah, Et Al, "Machine Learning Cases In Clinical And Biomedical Domains," International Medical Journal Malaysia, July 2018.

[13] S. S. Khan, S. Ahamed, M. Jannat, S. Shatabda, And D. Md. Farid, "Classification By Clustering (Cbc): An Approach Of Classifying Big Data Based On Similarities," Springer Nature Singapore, 2020, Pp. 593-605. Https://Doi.Org/10.1007/978-981-13-7564-4_50 .

[14] Reuben Evans, Bernhard Pfahringer, And Geoffrey Holmes, "Clustering For Classification," IEEE 7th International Conference On IT In Asia (CITA), 2011.

[15] Rasoul Kiani, Siamak Mahdavi, And Amin Keshavarzi, "Analysis And Prediction Of Crimes By Clustering And Classification," (IJARAI) International Journal Of Advanced Research In Artificial Intelligence, Vol. 4, No.8, 2015.

[16] Asha Gowda Karegowda , M.A. Jayaram, And A.S. Manjunath, "Cascading K-Means Clustering And K-Nearest Neighbor Classifier For Categorization Of Diabetic Patients," International Journal Of Engineering And Advanced Technology (IJEAT), Vol. 1, Issue. 3, February 2012.

[17] Yaswanth Kumar Alapati And Korrapati Sindhu, "Combining Clustering With Classification: A Technique To Improve Classification Accuracy," International Journal Of Computer Science Engineering (IJCSE), Vol. 5, No.06, Nov 2016, Pp. 336-338.

[18] Norsyela Muhammad Noor Mathivanan, Nor Azura Md.Ghani, And Roziah Mohd Janor, "Improving Classification Accuracy Using Clustering Technique," Bulletin Of Electrical Engineering And Informatics, Vol. 7, No. 3, September 2018, Pp. 465-470 .

[19] Mundhe, Ms Raksha K., and Ankush Maind. "Automatic labelling and document clustering for forensic analysis." Int J Recent Innovation Trends Comput Commun 2.9 (2014): 2934-41.