International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 07, July 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Algorithms in Clustering and Classification Data Mining

Mohd Vasim Ahamad[1], Misbah Urrahman Siddiqui[2],

Tariq Ahmed[3], Asia Mashkoor[4]

[1,2,3,4]*Aligarh Muslim University (India)*

ABSTRACT

With the information technology advancements, huge amount of data is available around each corner. To extract meaningful information and hidden patterns from available data, data mining techniques can be applied. There are many techniques in data mining to find hidden patterns such as clustering, classification, association rule mining and regression analysis, etc. Among all of them, clustering and classification techniques are more popular than others. The main objective of this work is to provide researchers with the insight into various clustering and classification algorithms.

Keywords: ***Data Mining, Classification, Clustering, K Means, DBSCAN Algorithm, k-NN, Artificial Neural Network***

I. INTRODUCTION

Nowadays, because of the significant evolutions in the information technology and computers, there is huge availability of data in every sector. However, without converting this huge amount of data into some useful information, it'll not be beneficial to us. To analyse and figure out hidden patterns and useful information, data mining techniques are used. It can be defined as the method of extracting hidden patterns and meaningful information from the huge amount of data available [1] [2]. Data mining is one of the core processes in Knowledge Discovery in Databases (KDD).

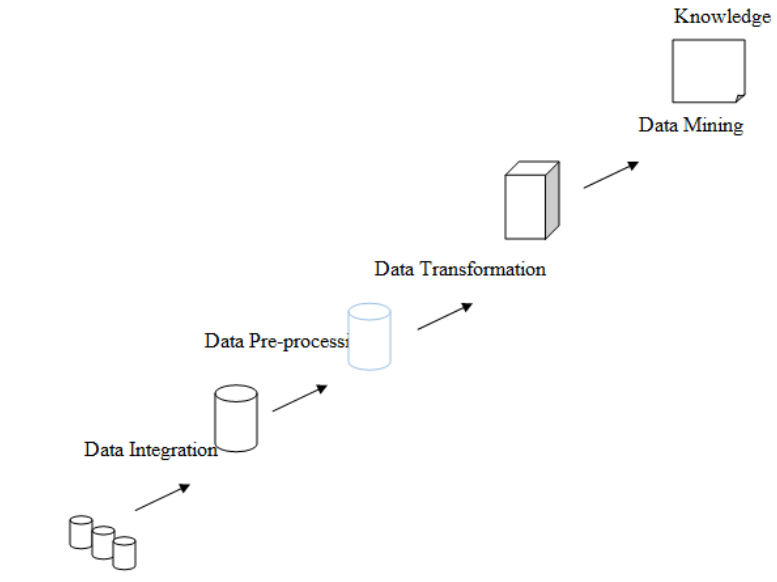

Fig. 1: The KDD Process

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Fig. 1 describes the process of Knowledge Discovery in Databases (KDD). In this process, data are gathered from a numbers of sources and integrated into unified schema. Next step in KDD is to pre-process the data to eliminate inconsistencies, ambiguities, errors and incompleteness of data. After that, data is transformed into a structure which is suitable for data mining tasks. Data mining techniques are applied to transformed data. The knowledge and hidden patterns are figured out after applying different data mining algorithms. Techniques used in data mining are association rules discovery, clustering, classification, sequential pattern mining, etc. Clustering and classification are mostly used techniques in data mining. After getting hidden patterns and useful information, it can be used in the applications such as market research, target marketing, fraud detection, customer profiling, product customizing, science exploration etc.

In this paper, we have provided a critical review of different types of clustering and classification algorithms. The remaining sections of this paper is organized as follows: Different types of clustering algorithm are explained in section II. The classification algorithms are presented in section III. In the last section, we have provided the conclusion.

I. INTRODUCTION TO CLUSTERING

A cluster is basically a collection of data items clubbed together into the same group which are similar, dissimilar data items are scattered into different groups. Clustering is the process of grouping a set of data items into classes of similar data items. Clustering is an example of unsupervised learning, which do not require predefined class labels to identify the cluster of a data item. In clustering, data items are grouped together based on their similarity, unlike the classification where data items are grouped based on the class labels. Based on the application area, clustering techniques can be categorized as Partitioning Based Methods, Soft computing Methods, Density Based Methods, Model Based Approaches, etc. [1][2]

A. Partitioning Based Methods

A partitioning method constructs k partitions of n data items, where each partition represents a cluster such that k<n. Partitioning based methods groups these n data items into k clusters. In partitioning method, each group must contain at least one data item and each data item must be associated with one of the k clusters. Examples of partitioning methods are K-Means Clustering Algorithm, K-Medoids Clustering Algorithm, etc. [3]

1) K-Means Clustering

K – Means Clustering is an example of partitioning based algorithms and second most popular data mining algorithm. It tries to partition n data items into k partitions, such that, data item associated to its closest partition [1] [3]. These partitions can be considered as cluster prototypes. To form efficient clusters, intra-cluster similarity should be as high as possible whereas the inter cluster similarity should be as low as possible [3]. Cluster similarity is measured by calculating the mean value of data items in the cluster, which can also be considered as the cluster's centre. Following are the basic steps of K-Means clustering algorithm.

- Choose the required number of clusters (say k) beforehand.
- Randomly selects k cluster centers.
- The next step is to choose each data item of the input data set and compare its distance to all of the k cluster centers. The data item is associated with the cluster whose centre is nearest to the data item.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

- The cluster centres are re-calculated after each iteration by taking the mean of all data items in each cluster. The new mean values is used as new cluster centers
- This process iterates until the criterion function converges or there are no significant changes in the structure of clusters. -Means clustering algorithm is very simple and efficient for low dimensional data. The dark side of this algorithm is selection of number of required clusters beforehand and it cannot find clusters of arbitrary shape. The K-Means clustering algorithm is applicable when mean is defined.

2) K-Medoids Clustering

K-Means clustering algorithm is only applicable when data items are such that their mean is defined. To deal with the data whose mean is not defined, K-Medoids clustering algorithm is used. In K-Medoids clustering algorithm, a cluster is represented by one of its data item, and hence, it can be used with any type of data. A medoid can be defined as the most centrally located data item of a cluster. The K-Medoid clustering algorithm starts by choosing k medoids randomly. Then it tries to associate other data items in clusters whose medoid is closer to them. After that, it swaps these representative medoids with non-medoids. This process continues until the quality of the result is improved [4]. It is more efficient algorithm than K-Means while dealing with outliers and works well for relatively small datasets. K-Medoids clustering has its two variations namely PAM (Partitioning around Partitioning) and CLARA (Clustering LARge Applications) [5].

B. Hierarchical Methods

In Hierarchical methods of clustering, clusters are formed by hierarchical decomposition of the input data set using some criterion. It works by grouping data items into a tree of clusters. The hierarchical method can be categorized into two classes namely agglomerative and divisive. [3]

1) Agglomerative Approach

Agglomerative approach is a bottom up approach which starts with assigning one data item, from the dataset, in each cluster. Then it tries to merge the clusters which are closer to each other. This process repeats until we achieve top most level of hierarchy (one cluster) or termination criteria encountered. AGNES (AGglomerative NESting) follows agglomerative hierarchical clustering approach [3].

2) Divisive Approach

Divisive hierarchical method follow top down approach. It is a reverse of agglomerative approach and starts with assigning each and every data item of the input dataset into a single clusters. After that, the cluster is broken into smaller clusters based on dissimilarity measures, until we have every data object into separate cluster or termination criteria encountered. DIANA (DIvisive ANAlysis) follows the divisive hierarchical approach. [3]

C. Density Based Methods

In partitioning based clustering methods, clusters are formed by data items based on distance between them. These methods can only find the clusters of spherical shape [1] [3]. It is very difficult to find arbitrary shaped clusters using partitioning based clustering methods. To deal with this scenario, Density Based Methods are used. They discover the clusters of arbitrary shapes efficiently. In density based methods, clusters are formed by the dense region of objects surrounded by sparse regions. A well-known example of density based method is DBSCAN algorithm.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

1) *DBSCAN (Density-Based Spatial Clustering of Applications with Noise)*

DBSCAN algorithm follows density based clustering approach. The general idea behind the DBSCAN algorithm is to grow the cluster until the number data items (Density) are greater than some "Threshold" [1] [3]. For each data item in a given cluster, within a given radius there must be at least a minimum number of data points. It discovers the clusters of highly density connected data items. DBSCAN starts by calculating the number of data items within the radius with respect to a data item. There are some terminologies that needs to be discussed before figuring out the exact algorithm steps.

D. Soft Computing Methods

In traditional clustering approaches like partitioning approach, each data object belongs to only one cluster. It is also called as "Hard Clustering". Hard clustering discovers the disjoint clusters [3] [6]. There cannot be a common data point in any two clusters. Unlike the traditional partitioning methods, in soft computing method, every data point belongs to every cluster by some membership value. Hence, every cluster is a fuzzy set of all the data points. Fuzzy C-Means clustering algorithm is most popular method under the soft computing clustering approaches [6].

1) Fuzzy C-Means Clustering

The Fuzzy C-Means algorithm starts by assigning random membership values to each of the data object with respect to each cluster, based on the difference between the cluster center and the data item [6]. Membership value will be more if the data point is closer to a cluster. The sum of membership values of each data object should be equal to one [6]. At each iteration, the membership values and cluster centers are updated. It terminates when there are no significant changes in the structure of resulting clusters or minimum objective function is achieved.

II. CLASSIFICATION

Classification is a supervised machine learning algorithm which predicts the class of categorical data. It classifies data based on the training dataset and the class labels in a classifying attribute and uses it to classify new data [3]. Classification algorithms works in two steps: model construction and model usage.

1) Model Construction

In training dataset, each record belongs to a predefined class label, which is determined by the class label attribute in the dataset. Then a classifier model is constructed by using different techniques such as classification rules, decision trees, or some mathematical formulae.

2) Model Usage

Classification make use of two types of datasets namely training dataset and testing dataset. Training dataset is used in classifier model construction, whereas testing dataset is used for predicting the class label of previously unseen records. The accuracy of the classifier model is calculated by comparing the known label of test samples with classified results of classifier model. Accuracy rate can be defined as the percentage of test dataset samples that are correctly classified by the classifier model [3].

Typical applications of classification are medical diagnosis, fraud detection, target marketing, performance prediction, manufacturing, etc [3]. There are some of many classification algorithms.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

A. Decision Tree based Classification

A decision tree is a flow chart like structure that includes a root node, branches, internal nodes, and leaf nodes. The topmost node in the tree is the root node. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. Decision tree are commonly used for gaining information for the purpose of decision making in classification. Decision tree starts with a root node. From this node, users split each node recursively according to decision tree induction algorithm. A decision tree represents a procedure for classifying categorical data based on their attributes. It is also efficient for processing large amount of data, so is often used in data mining application.

Decision tree induction algorithm starts by constructing a tree with top down recursive divide and conquer approach. Initially, all training examples are at the root node of the tree. Then these example are divided based on selected attributes. These test attributes are selected on the basis of heuristics measures such as information gain, gini index, etc. Decision tree induction algorithm stops when all samples for a given node belongs to same class or there are no attributes remaining for partitioning or there are no samples anymore.

B. ID3

ID3 is a decision tree based classification algorithm. In ID3, a decision tree is constructed with each internal node representing the selected attribute and leaf nodes representing the class label of the final cluster. The basic steps of ID3 algorithm are as follows [3] [7].

- Calculate the entropy or information gain of each attribute of training example
- maximum
- Create an internal node containing that attribute
- Apply above steps recursively on remaining attributes in the attribute set
- Terminate when all samples belongs to same class or there are no attributes remaining for partitioning or there are no samples anymore
- Split the example into subsets using the attribute for which the entropy is minimum or information gain is

C. C4.5

C4.5, also called as statistical classifier, is a successor of ID3 and mostly used data mining algorithm [8] [9]. This algorithm starts with choosing the attribute of the sample dataset, at each node of the tree, which splits the samples into sublists most effectively. The attributes are splitted whose normalized information gain is maximum [9]. The C4.5 algorithm then iterates on the smaller sublists until termination criteria is achieved. C4.5 terminates when

- All the training samples belong to the same class
- None of the attribute features provide any information gain
- Instance of previously unseen class is encountered [9]

D. Naïve Bayes Classifier

The conditional probabilities are the foundation of the Naive Bayes Classification algorithm. The Naive Bayes Classifier technique is particularly used when the input dataset is of high dimensionality. The Bayesian Classifier is capable of calculating the most possible output based on the input. It is also possible to add new raw

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

data at runtime and have a better probabilistic classifier. It uses Bayes' Theorem for calculating the conditional probability using following formula

$$P(x|y) = P(y|x) * P(x) / P(y)$$

Where x represents the prior event and y represents the dependent event. Bayes' Theorem calculates the probability of an event occurring given the probability of already occurred event [8]. Naive Bayes classifier is simple, easy to develop and highly efficient for large data sets.

E. Support Vector Machine

Support vector machines are based on supervised machine learning models that analyse data used for classification and regression analysis [10]. Support vector machine algorithm splits the n-D data into two regions in a hyper plane. The algorithm tries to find an optimal hyper plane which classifies new examples. The support vector machine represents the examples in dataset as data points in space. It maps the examples of the different categories are divided by a clear distance as wide as possible. After mapping the examples of dataset, previously unseen examples belong to a class based on which side of the space they fall.

- Support vector machines are most useful in text and hypertext classification.
- Image classification can also be done by using support vector machines.
- Support vector machines can also recognize the hand written characters.
- It has widespread applications in the areas of biological and other sciences.

F. K Nearest Neighbour

K Nearest Neighbour (k-NN) algorithm is the simplest algorithm under supervised machine learning algorithms. It follows the lazy learning principle in which classifier model generation is delayed until test example is presented. In k-NN classifier, the class of previously unknown data point is calculated on the basis of its nearest neighbours with known class labels. In k-NN, the class label of unknown data item is calculated on the basis of k nearest to that data item. In k-NN classification, an object is classified by a majority vote of its neighbours, and assigned to the most common class among its k nearest neighbours [11]. The values of k is generally a small integer. If the value of the k is set to 1, then this algorithm is treated as nearest neighbour classification.

G. Artificial Neural Networks

The term neural network is generally used to describe a network of biological neurons that are functionally connected to the central nervous system of living things. The basic element of the neural network is called a neuron. A neuron (or nerve cell) is a special biological cell that processes information. Artificial neural network (ANN) is a machine learning approach that models human brain and consists of a number of artificial neurons. An ANN is composed of artificial neurons that are connected together to form a network. Neuron in ANNs tend to have fewer connections than biological neurons. Each neuron in ANN receives a number of inputs.

An activation function is applied to these inputs to obtain the output value of the neuron. The neurons are trained by using the knowledge about the domain in the form of training examples. An Artificial Neural Network is specified by the following components:

1) ***Neuron Model:*** The information processing unit of the ANN

2) ***An Architecture:*** A set of neurons and links connecting neurons. Each link has a weight.

3) ***A Learning Algorithm:*** Used for training the ANN by modifying the weights in order to model a particular learning task correctly on the training examples.

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 07, July 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

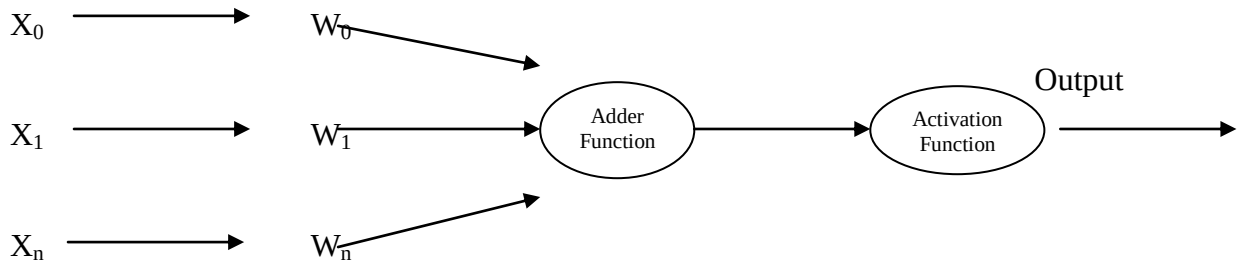

Fig. 2: The neuron model

The neuron model, as given in fig. 1, consists of a set of weighted links coming from different neurons. Adder function calculates the weighted sum of n number of inputs. If the result of adder function exceeds a threshold, then activation function comes into action and classifies the input dataset.

Perceptron is considered as simplest of all neural network architectures. It is a single layer feed forward neural network used for binary classification. The perceptron model classifier starts by training a perceptron for a classification task. To train a perceptron for classification, find suitable weights in such a way that the training examples can be correctly classified. The perceptron can only model linearly separable classes. The activation function used in case of a perceptron is a STEP function which returns the value 1 if weighted sum of its input is greater than or equal to a defined threshold, else it returns the value -1. Given training examples having class labels C1 and C2, the objective of perceptron model is to train the perceptron in such a way that

- If the output of the perceptron is +1 then the input is assigned to class C1
- If the output is -1 then the input is assigned to C2 The basic steps of learning algorithm for perceptron model are as follows:
- Start with assigning random weights to inputs ranges between -0.5 to 0.5
- Training data is given to perceptron one by one and its output is calculated.
- If output is incorrect, the learning algorithm adjust weights using the following formula.
- $w_i = w_i + (a * x_i * e)$, where 'e' is error produced and 'a' $(-1 < a < 1)$ is the learning rate
- If output is correct, then set 'a' as 0
- If output is too low, then set 'a' as some +ve value
- If output is too high, then set 'a' as some –ve value
- Once the weights are adjusted, the next training data samples are presented to the perceptron in the same way.
- This process continues in iterations until all the weights are correct and all errors become zero.

III. CONCLUSIONS

To extract meaningful information and hidden patterns from huge amount of available data, data mining techniques can be applied. There are many techniques in data mining to find hidden patterns such as clustering, classification, association rule mining and regression analysis, etc. Among all of them, clustering and classification techniques are more popular than others. To achieve overall objective of this paper, we have discussed various classification and clustering algorithms. In clustering, we have given an overview of partitioning based algorithms such as K-Means and K-Medoids, hierarchical based algorithms such as AGNES and DIANA, DBSCAN algorithm under density based clustering and Fuzzy C-Means in soft computing

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

based algorithms. In classification, we have discussed about most popular classification algorithms namely C4.5, ID3, k-NN and decision tree induction. We have also discussed Naïve Bayes classifier, support vector machine and artificial neural network based classifiers.

REFERENCES

[1] Nadeem Akhtar, Mohd Vasim Ahamad, Shahbaz Khan "MapReduce Model of Improved K-Means Clustering Algorithm Using Hadoop MapReduce", In 2nd IEEE International Conference on Computational intelligence and communication technology (ICICT-2016), ISSN/ISBN No. 978-1-5090-0210-8/16, DOI 10.1109/CICT.2016.46

[2] Nadeem Akhtar, Mohd Vasim Ahamad, Shahbaz Khan "Clustering on Big Data Using Hadoop MapReduce", in 7th IEEE International Conference on Computational Intelligence and Communication Networks (CICN 2015), ISSN/ISBN No. 978-1-5090-0076-0/15, DOI 10.1109/CICN.2015.161

[3] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000

[4] Rashi Chauhan "Clustering Techniques: A Comprehensive Study of Various Clustering Techniques", International Journal of Advanced Research in Computer Science, Volume 5, No. 5, May-June 2014

[5] Swapnil R. Isod, Amit M. Sahu "Clustering Techniques", International Journal of Advanced Research in Computer Science, Volume 4, No. 6, May 2013 (Special Issue)

[6] Nadeem Akhtar, Mohd Vasim Ahamad, "A Modified Fuzzy C Means Clustering Using Neutrosophic Logic", IEEE fifth International Conference on Communication Systems and Network Systems 2015, ISSN/ISBN No. 978-1-4799-1797-6/15, DOI 10.1109/CSNT.2015.164

[7] Nikam S. S. "A Comparative Study of Classification Techniques in Data Mining Algorithms". Orient.J. Comp. Sci. and Technol;8(1)

[8] Umd.edu - Top 10 Algorithms in Data Mining

[9] C4.5 Algorithm, https://en.wikipedia.org/wiki/C4.5_algorithm

[10] Support Vector Machine, https://en.wikipedia.org/wiki/Support_vector_machine

[11] K Nearest Neighbour Algorithm, https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm