Chapter – 3

Clustering and Classification

The study proposes the integration of clustering and classification to solve several significant machine-learning problems, which occurs due to the intrinsic nature of real-world datasets. This chapter gives a detailed description of clustering and classification techniques and several tools available to solve clustering and classification tasks.

3.1 Classification

Classification is a predictive data mining technique used to predict the class to which an instance belongs to predefined classes. Classification is a supervised learning technique. It tries to assign previously unseen records to a class as accurately as possible. Classification is a two-stage method where at the first stage we built/trained models from historical training data sets with labeled class attributes. In the second stage, it tries to maximize the classification accuracy rate which is the ratio of the number of correct predictions to the total number of predictions in the test dataset (Kesavaraj, 2013). Fundamentally, it is a mapping from target function to attribute set of already labeled class (Ningtan, 2008).

Classification is necessary to segregate existing unseen data into predefined classes. It is done based on attributes and features present in the data. Users can get a better understanding and description of the data for each class of the database. Classification provides a model that can describe future data (Duda et al., 2001).

A classification algorithm used for summarizing the performance- the technique used is the confusion matrix (Deng, 2016). Figure 3.1 is a confusion matrix for binary-classification problems.

	Positive Prediction	Negative Prediction
Positive Class	**True Positive** (TP)	**False Negative** (FN)
Negative Class	**False Positive** (FP)	**True Negative** (TN)

Figure 3.1 Confusion matrixes for binary classification

The concept of True Positive, True Negative, False Positive, and False Negative as an example:

- **True Positive (TP)** – A positive and is correctly classified as positive
- **True Negative (TN)** – A negative and is correctly classified as negative
- **False Positive (FP)** – A negative but is wrongly classified as positive
- **False Negative (FN)** – A positive but is wrongly classified as negative

3.2 Classifiers Performance measuring Parameters

A confusion matrix is used to evaluate the performance of the classifier. Table 3.1 provides the formulas to calculate these parameters to evaluate the performance of the classifier.

Table 3.1 Performance Measuring Parameters

S. No	Parameters	Formulas
1	Accuracy	$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN}$
2	Precision	$\text{Precision}=\frac{TP}{TP+FP}$
3	Recall	$\text{Recall}=\frac{TP}{TP+FN}$
4	F-measure	$F=\frac{2*\text{Precision}-\text{Recall}}{\text{Precision}+\text{Recall}}$
5	True positive rate	$TP_{rate}=\frac{TP}{Total\_P}$
6	False Positive rate	FP_rate=FP/(Total_N)
7	AUC	$\frac{1+TP_{rate}-FP_{rate}}{2}$
8	G-Mean	$\sqrt{\frac{TP}{TP+FN}*\frac{TN}{TN+FP}}$

3.3 Classification techniques

“A data mining algorithm is defined as a procedure that takes an input of data and gives output in the form of models or patterns”. The Top 10 Algorithms for classification are C4.5, K-Means, KNN, Apriori, EM, PageRank, AdaBoost, SVM, naive Bayes, CART (Wu et al. ,2007).

The classification technique is of five categories, based on different mathematical concepts. These categories are distance-based, neural network-based, rule-based decision tree-based, statistical-based and each category consists of several algorithms, but the most popular from each category that is used extensively are C4.5, Naıve Bayes, K-Nearest Neighbors, and Backpropagation Neural Network (Dimitoglou et al. ,2012 & Huang et al. ,2003).

Figure 3.2 is a tree diagram representing some of the significant classification techniques with the categories they belong to based on the type of approach they adopt while performing the classification task.

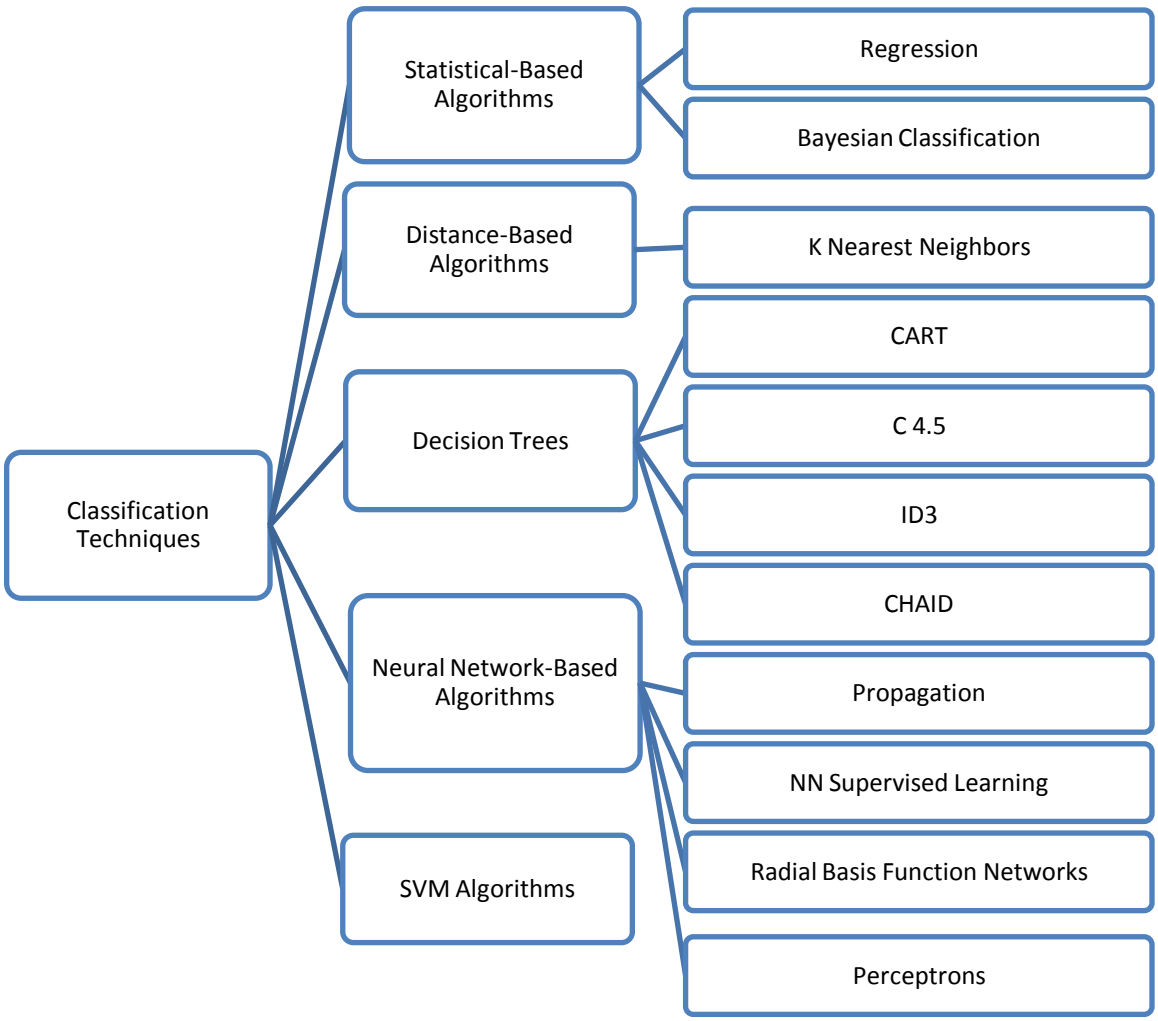


Figure 3.2 Classification Techniques

3.3.1 Regression

Regression is a statistical tool used to investigate the relationships among variables. Estimation of an output value based on input values is what Regression problems deal with. It is generally used to predict the values of the future based on past values by fitting a set of points on a curve. A linear relationship exists between the input data and the output data in linear regression. Regression can be used to solve classification problems. Analysis of Regression is used to the two continuous (scale) variables. It is better suited for studying functional dependencies between factors.

i) Linear Regression

The most basic regression is linear regression. It allows us to derive the relationship between two continuous variables.

ii) Logistic Regression

Logistic regression estimates the probability of an event to occur based on the data provided. It is used for a binary variable where two values, 0 and 1, represent the outcomes.

3.3.2 Bayesian Classification

It uses probability to predict a class based on a set of features. It classifies each value as an independent value. It is a directed acyclic graph where nodes represent random variables, Edges represent dependencies. Classifiers Based on Bayesian networks have strength, model interpretability, and fine-tuning with data while handling complex classification problems (Bielza & Larranaga,1986).

Naïve Bayes is a Bayes theorem based probabilistic classifier which helps in developing machine learning tool for improving classification accuracy.

Bayes' theorem gives a relationship for marginal and conditional probabilities of events A and B:

$$P(A|B) = \frac{(P(A)|P(B)) * P(A)}{P(B)}$$

Where

• P(A) is the prior or marginal probability of A. prior, as it does not take into account any information about B.

• P(A|B) is the conditional probability of A, given B. It is also known as the posterior probability because it is derived from or depends upon the specified value of B.

• P (B|A) is the conditional probability of B given A.

• P (B) is the prior or marginal probability of B and is a normalizing constant.

In this probabilistic classification theorem assumption is that the available features in the dataset are not dependent on the classification label. It means that on a given class each feature has an independent impact .it is also known as conditional independence.

Naïve Bayes classifiers can handle categorical as well as continuous values. Despite its simplicity, the classifier outperforms over sophisticated classification methods.

3.3.3 Decision Tree

It is a supervised predictive model that uses the divide and conquer method to extract knowledge and present them in a tree structure format. Based on the decision rules and received input, the prediction of discrete or continuous output is done by answering the questions. Decision trees are directed trees having root nodes, a link between the input to output nodes, subsets of internal nodes, and leaf nodes. The splitting of data performed repeatedly based on some decision criteria till the leaf node which is the class label is achieved. Decision Trees classify the tree from the root to the leaf node, while the leaf node provides a Class label. A node in the tree is a test case of some attribute, and each edge from that node

would be one of the possible answers to the test case. For every sub-tree, the process is recursive and repeated. In a given dataset, algorithms of Decision tree inducers construct a decision tree automatically. The generalization errors are minimized to find the optimal decision tree. Reducing the number of nodes & the average depths would be the other target functions.

The well-known splitting criteria are Binary, Impurity, and Normalized impurity. Popular decision tree algorithms are ID3, C4.5, CART, and CHAID.

i) ID3 (Iterative Dichotomiser 3)

The decision tree algorithm was developed by Quinlan. Iterative Dichotomiser 3 or ID3 is a simple decision tree learning algorithm (Quinlan,1986). The entropy-based information gain approach is used in the determination of a suitable class for each node by the decision tree method. The test attribute of the current node is the attribute with the highest information gain. When all instances belong to a single value the growth stops of target feature or when best information gain is not more than zero. It does not use any pruning procedures on numeric attributes or missing values.

ii) C4.5

Decision Tree C4.5 algorithm is a well-known algorithm for data classification developed by Ross Quinlan. It is an extension of Quinlan's earlier ID3 algorithm. The decision trees of C4.5 can be used for classification and is known as a statistical classifier. Missing values are easily handled by the C4.5 algorithm. It uses gain ratio as splitting criteria. When the number of instances to be split is less than a particular threshold the splitting stops. The difference between ID3 & C4.5 algorithm is that the C4.5 algorithm uses multi-way splits, While ID3 uses binary splits.

iii) CART

CART an acronym for Classification and Regression Trees (Breiman et al., 1984). It constructs binary trees, where every internal node has two outward edges. The splits selection is done by using the two criteria Gini index and information gain and the tree obtained is pruned by cost complexity Pruning. In the tree induction, misclassification costs are considered by CART and help to provide probability distribution as a priority. CART can generate regression trees. Regression trees where the leaves predict a real number but not a class. In regression, CART splits that minimize the prediction squared error. Each leaf prediction is based on the weighted mean for node.

iv) CHAID

Chi-square **A**utomatic **I**nteraction **D**etector(CHAID) is a well known statistical tree construction techniques for generating decision trees. The type of target attribute decides the type of statistical test. For a continuous target attribute - an F test is used. For nominal attribute - Pearson chi-squared test is used. For the ordinal attribute - the likelihood–ratio test is used. For each selected pair, CHAID checks if the p-value obtained is greater than a certain merge threshold. For a positive answer, the values are merged and search for a potential additional pair for merging. The process is continued until no significant pairs are available. The best input attribute for splitting the current node is selected so that each child node has a group of homogeneous values for the selected attribute. Note that no split is performed if the adjusted p-value of the best input attribute is not less than a certain split threshold. This procedure ceases when one of the following conditions is satisfied (Wilkinson,1992).

1. Maximum tree depth is reached.
2. The minimum number of cases in a node for being a parent is reached, so splitting further is not possible.
3. The minimum number of cases for being a child node is reached.

CHAID handles missing values by treating them all as a single valid category. CHAID does not perform pruning.

3.3.4 The KNN (K-Nearest Neighbor)

K-Nearest Neighbor (KNN) algorithm can be used is also known as the lazy learning approach and is a Known method in which no prior classification model is built (Han et al. ,2011). It keeps whole training data for classification as test data. K-Nearest derives using distance metrics like Euclidean, Manhattan, or Minkowski. KNN is considered under the top 10 most important data mining algorithms (Wu,2008). It is easily understandable, simple in programming, and convenient in implementation. It can easily handle missing values.

An important factor of KNN is that the error rate is bounded above by twice the Bayes error rate Cover and Hart (1967). An efficient algorithm performing well in the applications having no proper information about data in rare instances. It is an algorithm introduced by the Nearest Neighbor algorithm that is designed to find the nearest point of the observed object. The main purpose of the KNN algorithm is to find K-nearest points (Larose,2005).

KNN is well suited for multi-modal classes and also applicable where an object has many class labels. Nearest neighbor (KNN) is a simple, very popular, efficient, and effective algorithm for pattern recognition. It is a straight forward classifier, where samples are classified based on their nearest neighbor.

3.3.5 Support Vector Machines

This algorithm is a supervised learning model that is associated with learning algorithms to analyze data with recognizable patterns relying on statistical learning theory. An extremely nonlinear mapping of the input vectors present in the high-dimensional feature space gives optimal separating hyperplanes, forms an SVM binary classifier (Tong & Koller, 2001) It is widely used in pattern recognition, text classification, marketing, and medical diagnosis.

It is a binary classifier aimed helps in finding large margin hyperplane that separating two class values according to the feature space (Hearst,1998). The margin between positive and negative observations of the training dataset obtained by N observations is maximized when the data is linearly separable in the optimal hyperplane. Following the optimization problem the learning task of in the general case, the task of learning an SVM from a dataset is formalized.

3.3.6 NN supervised learning

To find important relationships or patterns from the information neural networks are used to analyze complex information that is imprecise, incomplete. These are intricate patterns, which are not easily detected by humans & computer-based analysis. It is an information processing system, comprising of graphical representations of the processing system & various other algorithms that access the graph. Similar to the human brain, the NN consists of multiple connected processing elements. Its approach requires decision trees, with a graphical structure, to built a model and be applied to the data. The NN is represented as a graph including the source (input), sink (output), and internal (hidden) nodes (Singh & Chouhan,2009).

It is a computational model comprising of three parts:

1. A graph that defines the data structure of the neural network.
2. An algorithm indicates how learning takes place.

3. Techniques recalling that determine how information is obtained from the network.

i) Propagation:

A tuple with input values X= (X_J, . . . , X_h) with one input value at each node in the input layer. The functions of summation and activation are applied at each node, the output value is created for each output arc from that node. These values are in turn sent to the subsequent nodes. Until a tuple of output values, Y = (y_1, . . . Y_m), is produced from the nodes in the output layer this process continues(Haykin,2009).

ii) Perceptrons :

The simplest form of NN is termed as the perceptron. It is a single neuron with many inputs and one output. the step activation function is the original use of a perceptron proposed, another type of function which is very common is a sigmoidal function. A simple perceptron can be classified into two classes. In a unipolar activation function, an output of 1 will be used to classify into one class and an output of 0 will be used to pass it in the other class.

iii) Radial Basis Function Networks:

A Radial basis function (RBF) is a class of functions whose value increases or decreases with reference to distance from a central point. An RBF network is typically a neural network (NN) with three layers and has a Gaussian shape. The input layer is used to input the data. When a linear activation function is used at the output layer a Gaussian activation function is used at the hidden layer. The purpose of having hidden nodes is to learn to respond only to a subset of the input, viz. the centered Gaussian function. This is accomplished through supervised

learning. The nodes can be sensitive to a subset of the input values when RBF functions are used as the activation functions on the hidden layer.

3.3.7 Random Forests

Random forests are an ensemble learning method, it combines multiple algorithms and improves the performance of the classifiers. Each independent classifier is weak, but when combined with other classifiers, can improve the results. The algorithm takes a decision tree as an input, entered at the top. It then traverses down the tree; the data is being segmented into smaller sets, depending on specific variables.

Table 3.2: Comparative analyses of Classification Techniques

Parameters	Decision Trees	Neural Networks	Naïve Bayes	KNN	SVM	Rule-learners
Accuracy in general	**	***	*	**	****	**
Speed of learning with respect to the number of attributes and the number of Instances	***	*	****	****	*	**
Speed of Classification	****	****	****	*	****	****
Tolerance to missing values	***	*	****	*	**	**
Tolerance to Irrelevant values	***	*	**	**	****	**
Tolerance to redundant	**	**	*	**	***	**

attributes						
Tolerance to highly interdependent attributes(Parity problem)	**	***	*	*	***	**
Dealing with Discrete /Binary/Continuous attribute	****	*** (not discrete)	***(not continuous)	***(not directly discrete)	**(not discrete)	***(not directly continuous)
Tolerance to Noise	**	**	***	*	**	*
Dealing with the danger of Overfitting	**	*	***	***	**	**
Attempts for incremental learning	**	***	****	****	**	*
Explanation ability/transparency of knowledge/classification	****	*	****	**	*	****
Model parameter handling	***	*	****	***	*	***

* : Poor performance

** :Average performance

*** :Good Performance

**** :Excellent performance

Source: S. B. Kotsiantis (2007)

3.4 Clustering

The most efficient technique that is applied to raw data for the extraction of useful information. Similar and dissimilar types of data are clustered for analysis of useful information from the dataset. When data is sorted into groups of similar objects it is known as Clustering. Each one of such groups is known as a cluster. Every cluster has similar objects but differs from the objects of the other clusters (Hoppner et al., 1988). Clustering is not predefined but an unsupervised learning method & class labels are not present.

Clustering similar to classification in which objects of data are grouped without consulting a known class label. Data groupings are not pre-defined in clustering; they are generated by the similarities within the data objects based on the characteristics present in the actual data. Partition or division of datasets into clusters or many groups is based on their similarities that are done in such a way that the objects with maximum similarities belong to one group or cluster and are highly dissimilar with the other group or cluster. That means a good clustering algorithm has a maximum intra-cluster similarity and minimum inter-cluster similarity.

Clustering is studied in the field of machine learning as an unsupervised learning process, as it is “learning from observation” and not “learning from examples.” The pattern proximity matrix is being measured as a distance function defined on pairs of patterns (Jain &Dubes, 1988; Duda et al., 2001). Clustering is applied in applications in which no or little prior knowledge of the groups or classes in a database is available. The clustering usefulness is often associated with individual interpretation and on the selection of suitable similarity measure (Jain et al.,1999).

3.4.1 Clustering Techniques

The division of data into meaningful groups is done by Clustering algorithms. The patterns in the same group are similar to those that are present in dissimilar groups. Clustering is unsupervised learning (Jain, et al. ,1999).For information retrieval, from the search engine clusters of web pages divided into different groups, such as videos, news, reviews, and audios. Clustering divide points into different groups.

(i) **Hierarchical Clustering method**

It combines data objects into subgroups; these subgroups conjoin to larger and higher-level groups and ultimately form a hierarchy tree. This method has two kinds of approaches, divisive (top-down) & agglomerative (bottom-up) approaches. The agglomerative clustering starts at one-point clusters and recursively merges two or more of the clusters. The divisive clustering has a top-down strategy; it starts with one cluster comprising of all data points and recursively splits that cluster into proper sub-clusters (Navarro et al., 1997).

(ii) **Partitioning Algorithms**

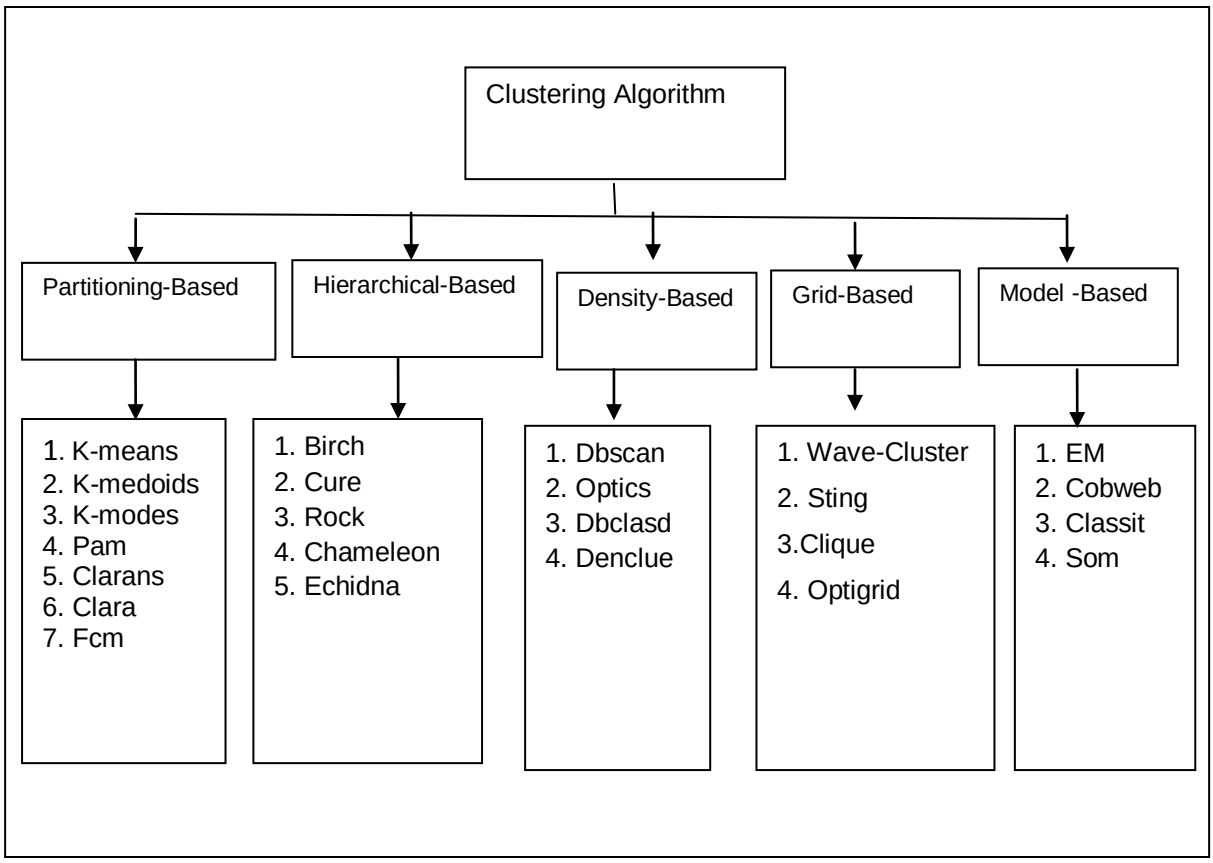
It helps in discovering clusters either by distinguishing areas heavily populated with data or by iteratively relocating points in the subsets (Swarndeep & Pandeya,2016).

(iii) **Density-based partitioning methods**

Discovers low-dimensional data, that is densely connected, and it is known as spatial data. (Ester et al. ,1996).

(iv) **Grid-based Partitioning**

Hierarchical agglomeration in one phase is used by Grid-based partitioning algorithms for processing and performing space segmentation, then aggregate appropriation of segments (Liao et al., 2004). Faces scalability problems for computing time and memory requirements in Scalable clustering research. Uses high dimensionality data clustering methods meant for designating & handling of data with hundreds of attributes.



Fahad (2014)

3.5 K-means algorithm

The widely used clustering algorithm is the K-means algorithm, it is a straightforward algorithm. In a set of data sets the objects are divided into

“clusters” with similar objects as compared to objects of other clusters. Clustering algorithms keep similar data in one cluster while dissimilar data is segregated into other clusters. In supervised tasks such as classification or regression, there is an associated class label but with clustering, the objects do not come with an associated target.

Hence, clustering is referred to as unsupervised learning. As there is no need for labeled data, for many applications un-supervised algorithms are suitable where labeled data is difficult to obtain. To explore and characterize the dataset Unsupervised tasks such as clustering are also often used prior to running a supervised learning task. some notion of similarity must be defined based on the attributes of the objects, as clustering makes does not use of class labels.

A simple iterative clustering algorithm is the K-means algorithm that partitions a dataset into K clusters. The algorithm works by iterating at its core, involves two steps:

(1) Based on the distance between each point clustering all points in the dataset and its closest cluster representative
(2) Re-estimation of the cluster representatives. k-means algorithm has some limitations that include the sensitivity of k-means for the initialization and determining the value of K, K-means remains the most widely used partition clustering algorithm in practice even it has some drawbacks.

It is a simple algorithm, reasonable, understandable, and scalable, and is easily modified to work with different scenarios namely, semi-supervised learning or streaming data. The basic algorithm has continual improvements and generalizations that ensured its continued relevance by gradually increasing its effectiveness (Elkan, 2004).

3.6 Distance metrics

Young M., et al**.** (2004), described in their Distance Metrics Overview, various distance metrics were used to determine the distance between two data points. The distance can be calculated by following distance metric between xi and yj is given by

$$\textbf{Sum of Squared Error}=\sum_{i=1}^{k}\sum_{j=1}^{ik}\mathbf{dist}(x_i,y_j)$$

i. Euclidean distance:

To compute the distance between data points Euclidean distance is the most regularly used metric. The square root of the sum between two points is Euclidean distance. N-dimensional data distance is given by the formula;

$$\textbf{Euclidean distance}=\sqrt{\sum_{k=1}^{n}(x_{ik},y_{ik})^2})$$

ii. Manhattan distance:

A well-known function for measuring distance is Manhattan distance. It is calculated by summing the absolute value of the difference between data points. Manhattan distance is less costly to calculate in comparison to Euclidean distance. Manhattan distance is given by formula;

$$\textbf{Manhattan distance}=\sum_{k=1}^{n}|x_{ik}-x_{jk}|$$

iii. Minkowski distance

Minkowski function is a geometric distance between two points with a scaling factor, r. The important use is to find the similarity between objects. When r=2 then

it becomes the Euclidean distance. When r=1 then it become the Manhattan distance. The distance is given by the Formula

$$\textbf{Minkowski} = \Sigma(|x_{ij} - x_{jk}|)\ 1/q$$

iv. Canberra distance

The sum of absolute values of the differences between ranks divided by their sum is Canberra distance. A weighted version of the Manhattan distance function, where d is the distance, x and y are two sets in the dataset, n is the total number of cases in the dataset.

$$d_{ij} = \sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}$$

v. Mahalanobis distance

The Mahalanobis distance considers covariance among the variables for distance calculations. The problems of Euclidean distance (scale and correlation) can be resolved. The Mahalanobis distance can be applied to model problems such as radial basis function neural networks and detection of outliers (Shah & Greenville, 1989).

$$\text{Mahalanobis distance} = [(a_i, b_i)^t S^{-}(a_i, b_i)^t]$$

3.7 Summary

This chapter gives a detailed description of the clustering and classification techniques. It lists out the parameters used to measure the performance of any classifiers with their formulas. This chapter also provides a comparative analysis of popular classification techniques. Several categories of the clustering technique and the significant distance measures are also discussed in this chapter.