CLUSTERING AND CLASSIFICATION EVALUATION USING RAPID MINER

[1]Varsha C. Pande, [2]Dr. Abha S. Khandelwal
[1]Research Scholar, [2]Former HOD,
[1]Electronics and Computer Science,
[1]RTMNU, Nagpur (MH), India.

Abstract: Clustering and classification of data is a difficult problem that is related to various fields and applications. Challenge is greater, as input space dimensions become larger and feature scales are different from each other. The term "classification" is frequently used as an algorithm for all data mining tasks. Data Mining (DM) is best to use the term to refer to the category of supervised learning algorithms used to search interesting data patterns. While classification algorithms have become very popular and ubiquitous in DM research, it is just but one of the many types of algorithms available to solve a specific type of DM task. This paper deals with the clustering and classification evaluation of various algorithms using Rapid Miner.

IndexTerms **- Data Mining, clustering, classification, supervised learning, scalability.**

I. INTRODUCTION

The methods used in data mining for analysing the data sets and split them on the basis of some particular classification rules or the association between objects are **Clustering and Classification**. Classification classifies the data with the help of provided training data. On the other hand, clustering uses different similarity measures to group the data. These two types of learning methods, characterize objects into groups by one or more features. These processes seem to be similar, but there is a difference between them in background of data mining. The major difference between clustering and classification is, clustering is used in unsupervised learning where similar instances are grouped, based on their features or properties and on the other hand, classification is used in supervised learning technique where predefined labels are assigned to instances by properties.

The technique of organising a group of data into classes and clusters where the objects reside inside a cluster will have high similarity and the objects of two clusters would be dissimilar to each other is known as **Clustering**. Here the two clusters can be considered as disjoint. The main target of clustering is to divide the whole data into multiple clusters. Unlike classification process, here the class labels of objects are not known before, and clustering pertains to unsupervised learning.

In clustering, the similarity between two objects is measured by the **similarity function** where the distance between those two object is measured. Shorter the distance higher the similarity, conversely longer the distance higher the dissimilarity.
The example of clustering, there are two clusters named as mammal and reptile. A mammal cluster includes human, leopards, elephant, etc. On the other hand, reptile cluster includes snakes, lizard, Komodo dragon etc.

The process of learning a model that elucidate different predetermined classes of data is called as **Classification**. It is a two-step process, comprised of a **learning** step and a **classification** step. In learning step, a classification model is constructed and classification step, the constructed model is used to **prefigure** the class labels for given data. For **example**, in a banking application, the customer who applies for a loan may be classified as a safe and risky according to his/her age and salary. This type of activity is also called supervised learning. The constructed model can be used to classify new data. The learning step can be accomplished by using already defined training set of data. Each record in the training data is associated with an attribute referred to as a class label that signifies which class the record belongs to. The produced model could be in the form of a decision tree or in a set of rules.

A **decision tree** is a graphical depiction of the interpretation of each class or classification rules. **Regression** is the special application of classification rules. Regression is useful when the value of a variable is predicted based on the tuple rather than mapping a tuple of data from a relation to a definite class.

II. CLUSTERING ALGORITHMS

The Algorithms mainly used in cluster analysis are k-means, k-Medoids, density based, hierarchical and several other methods

- **k-means:**

One of the most popular heuristics for solving the k-means [1] problem is based on a simple iterative scheme for finding a locally optimal solution. This algorithm is often called the k-means algorithm. There are a number of variants to this algorithm, so to clarify which version we are using, we will refer to it as the naïve k-means algorithm as it is much simpler compared to the other algorithms described here.

- **k-Medoids:**

The K-Medoids algorithm is a partitional clustering algorithm which is slightly modified from the K-means algorithm. They both attempt to minimize the squared-error but the K-medoids algorithm is more robust to noise than K-means algorithm. In K-means algorithm, they choose means as the centroids but in the K-medoids, data points are chosen to be the medoids. A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal. The difference between k-means and k-medoids is analogous to the difference between mean and median: where mean indicates the average value of all

data items collected, while median indicates the value around that which all data items are evenly distributed around it. The basic idea of this algorithm is to first compute the K representative objects which are called as medoids. After finding the set of medoids, each object of the data set is assigned to the nearest medoid. That is, object i is put into cluster v_i, when medoid mv_i is nearer than any other medoid m_w.

III. CLASSIFICATION ALGORITHMS

Some common classification **algorithms** are decision tree, K-Nearest Neighbor, naïve Bayes, support vector machine, random Forest, neural networks, logistic regression, etc.

- **Decision tree**

A decision tree [1] is a classification scheme which generates a tree and a set of rules, representing the model of different classes from a given data set [2]. The set of records available for developing classification methods is generally divided into two disjoint subsets as follows:

(i) A training set - used for deriving the classifier

(ii) A test set - used to measure the accuracy of the classifier

The accuracy of the classifier is determined by the percentage of the test examples that are correctly classified. The attributes of the records are divided into two types as follows:

- Numerical attributes- attributes whose domain is numerical.
- Categorical attributes- attributes whose domain is not numerical.

There is one distinguished attribute called the class label. The goal of the classification is to build a concise model that can be used to predict the class of the records whose class label is not known. A decision tree is a tree where the internal node - is a test on an attribute, the tree branch - is an outcome of the test, and the leaf node - is a class label or class distribution. There are two phases of decision tree generation:

Tree construction

- At start, all the training examples are at the root,
- partition examples based on selected attributes,
- Test attributes are selected based on a heuristic or a statistical measure.

Tree pruning

- Identify and remove branches that reflect noise or outliers.
- One rule is generated for each path in the tree from the root to a leaf.
- Each attribute-value pair along a path forms a conjunction.
- The leaf node holds the class prediction.
- Rules are generally simpler to understand than trees.

- **K-Nearest Neighbor:**

K nearest neighbor (KNN)[3] is a simple algorithm, which stores all cases and classify new cases based on similarity measure.KNN algorithm also called as 1) case based reasoning 2) k nearest neighbor 3) example based reasoning 4) instance based learning 5) memory based reasoning 6) lazy learning [4].KNN algorithms have been used since 1970 in many applications like statistical estimation and pattern recognition etc. KNN is a nonparametric classification method which is broadly classified into two types1) structure less NN techniques 2) structure based NN techniques. In structure less NN techniques whole data is classified into training and test sample data. From training point to sample point distance is evaluated, and the point with lowest distance is called nearest neighbor. Structure based NN techniques are based on structures of data like orthogonal structure tree (OST), ball tree, axis tree, nearest future line and central line [5]

- **Naïve Bayes Classifiers**:

Naive Bayes classifiers [6] are linear classifiers that are known for being simple yet very efficient. The probabilistic model of naive Bayes classifiers is based on Bayes' theorem, and the adjective *naïve* comes from the assumption that the features in a dataset are mutually independent. In practice, the independence assumption is often violated, but naive Bayes classifiers still tend to perform very well under this unrealistic assumption [7]. Especially for small sample sizes, naive Bayes classifiers can outperform the more powerful alternatives [8].

Being relatively robust, easy to implement, fast, and accurate, naive Bayes classifiers are used in many different fields. Some examples include the diagnosis of diseases and making decisions about treatment processes [9], the classification of RNA sequences in taxonomic studies [10], and spam filtering in e-mail clients [11].

- **SVM Classifiers:**

SVM Classifiers attempt to partition the data space with the use of linear or non-linear delineations between the different classes. The key in such classifiers is to determine the optimal boundaries between the different classes and use them for the purposes of classification.

- **Pattern (Rule)-based Classifiers:**

In rule-based classifiers we determine the word patterns which are most likely to be related to the different classes. We construct a set of rules, in which the left hand side corresponds to a word pattern, and the right-hand side corresponds to a class label. These rules are used for the purposes of classification.

- **Neural Network Classifiers:**

Neural networks are used in a wide variety of domains for the purposes of classification. In the context of text data, the main difference for neural network classifiers is to adapt these classifiers with the use of word features. We note that neural network

classifiers are related to SVM classifiers; indeed, they both are in the category of discriminative classifiers, which are in contrast with the *generative classifiers*.

- **Bayesian (Generative) Classifiers:**

In Bayesian classifiers (also called generative classifiers), we attempt to build a probabilistic classifier based on modeling the underlying word features in different classes. The idea is then to classify text based on the posterior probability of the documents belonging to the different classes on the basis of the word presence in the documents.

Vector-based Methods:

There are two types of vector-based methods. The centroid algorithm and support vector machines. One of the simplest categorization methods is the centroid algorithm. During the learning stage only the average feature vector for each category is calculated and set as centroid-vector for the category. A new document is easily categorized by finding the centroid-vector closest to its feature vector. The method is also inappropriate if the number of categories is very large. Support vector machines (SVM) need in addition to positive training documents also a certain number of negative training documents which are untypical for the category considered.

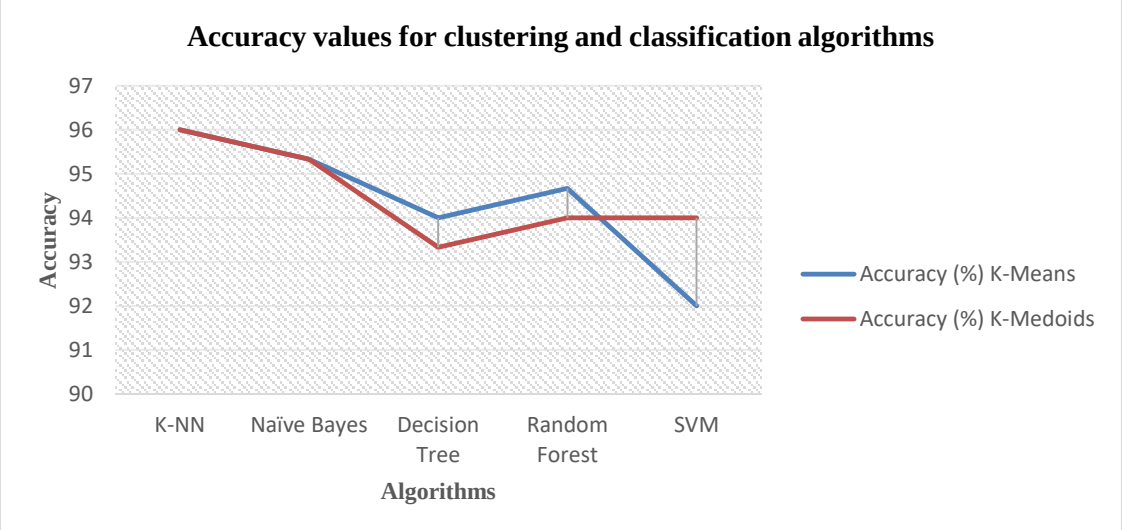
IV. RESULTS AND DISCUSSION

No target attribute (label) can be defined and the data should be automatically grouped. This procedure is called "Clustering". Rapid Miner supports a wide range of clustering schemes which can be used in just the same way like any other learning scheme. This includes the combination with all preprocessing operators. In this process, the well-known Iris data set is loaded (the label is loaded, too, but it is only used for visualization and comparison and not for building the clusters itself). The clustering algorithms i.e. **K-Means** and **K-Medoids** are applied to an **IRIS data set**. Afterwards, a dimensionality reduction is performed in order to better support the visualization of the data set in two dimensions, then the five Classification algorithms i.e. **K-NN, Naïve Bayes, Decision tree, Random Forest and Support Vector Machine** algorithms are applied on the output of clustering algorithms. The process is executed and the results of all five algorithms are given in following table.

Table 4.1: Accuracy and Kappa values for different algorithms

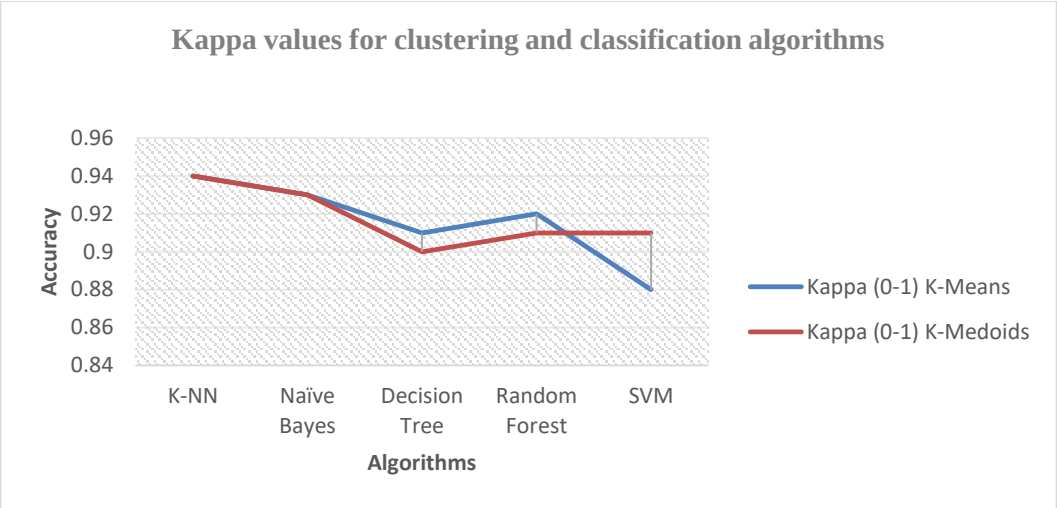
Algorithm	Accuracy (%)		Kappa (0-1)	
	K-Means	K-Medoids	K-Means	K-Medoids
K-NN	96.00	96.00	0.940	0.940
Naïve Bayes	95.33	95.33	0.930	0.930
Decision Tree	94.00	93.33	0.910	0.900
Random Forest	94.67	94.00	0.920	0.910
SVM	92.00	94.00	0.880	0.910

The following graph 1 shows the accuracy values for classification algorithms when applied on K-means and K-Medoids clustering algorithms.



Graph 1: **Accuracy values for clustering and classification algorithms**

The following graph 2 shows the Kappa values for classification algorithms when applied on K-means and K-Medoids clustering algorithms.



Graph 2: **Kappa values for clustering and classification algorithms.**

V. CONCLUSION

In Data Mining, Clustering and classification techniques organizing a collection of documents into groups based on similarity. Classification is to accurately predict the target class for each case in the data. In this paper the Accuracy and Kappa values of various algorithms are measured and analyzed. The K-means and K-Medoids clustering algorithms are applied on the given data set then on the resultant clusters the classification algorithms K-NN, Naïve Bayes, Decision Tree, Random Forest and Support Vector Machine (SVM) are applied. The results shows that the accuracy and Kappa values for K-NN algorithm is better for clustering algorithms.

REFERENCES

[1] Valsala, S. George, J. Parvathy, P. 2011. A Study of Clustering and Classification Algorithms Used in Data mining. IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.10.
[2] Holsheimer, M., Kersten, M., Mannila, H., Toivonen, H. A Perspective on Databases and Data Mining. Proceedings KDD '95.
[3] Jabbar, M.A. Deekshatulu, B.L Chandra, P. 2013. Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) pp. 85 – 94.
[4] Dr saed sayad,"University of toronto http://chem-eng.utoronto.ca/data mining.
[5] Bhatia, v. 2010. Survey on nearest neighbor techniques. IJCSIS,Vol 80,no 2.
[6] Raschka, S. 2014. Naïve Bayes and Text Classification.
[7] Rish, I. 2001. An empirical study of the naive bayes classifier. IJCAI workshop on empirical methods in artificial intelligence, pp. 41–46.
[8] Domingos, P. and Pazzani, M. 1997. On the optimality of the simple bayesian classifier under zero-one loss. Machine learning, vol. 29, no. 2-3, pp. 103–130.
[9] Kazmierska, J. Malicki, J. 2008. Application of the naıve bayesian classifier to optimize treatment deci- sions. Radiotherapy and Oncology, vol. 86, no. 2, pp. 211–216.
[10] Wang, Q. Garrity, G. Tiedje, J. Cole, J. 2007. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. Applied and environmental microbiology, vol. 73, no. 16, pp. 5261–5267.
[11] Sahami, M. Dumais, S. Heckerman, D. Horvitz, E. 1998. A bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop, vol. 62, pp. 98–105.