A Survey On Data Mining Algorithm

Rohit Jacob Mathew[1] Sasi Rekha Sankar[1] Preethi Varsha. V[2]

[1]*Dept. of Software Engg.,* [2]*Dept. of Electronics & Instrumentation Engg.*
SRM University India

***Abstract*— This paper puts forward the 8 most used data mining algorithms used in the research field which are: C4.5, k-Means, SVM, EM, PageRank, Apriori, kNN and CART. With each algorithm, a basic explanation is given with a real time example, and each algorithms pros and cons are weighed individually. These algorithms are seen in some of the most important topics in data mining research and development such as classification, clustering, statistical learning, association analysis, and link mining.**

I. Introduction

Data is produced in such large amounts that today the need to analyze and understand this data is of the essence. The grouping of data is achieved by clustering algorithms and can then further be analyzed by mathematicians as well as by big data analysis methods. This clustering of data has seen a wide scale use in social network analysis, market research, medical imaging etc. This grouping of data is seen in many different graphical forms such as the one shown below.

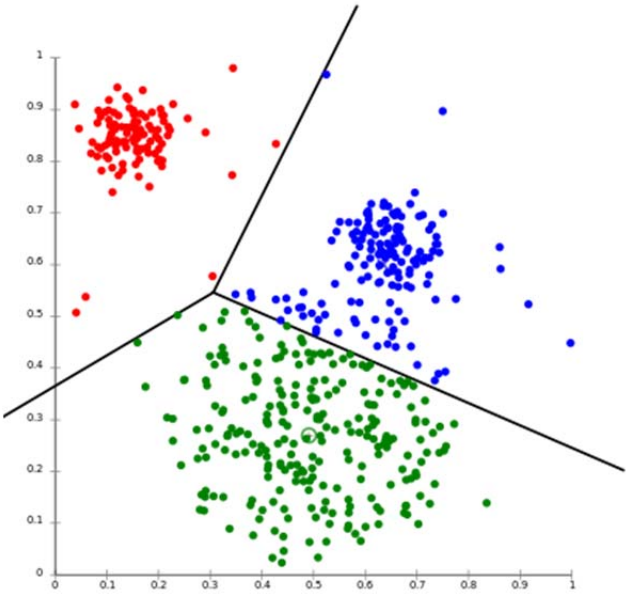


Figure 1 : K-MEANS

Here we have taken an instance to better understand and identify the most influential algorithms that have been widely used in data mining. Most of these were identified during the ICDM ’06 in Hong Kong. We deal with a wide variety of algorithms such as clustering, classification, link mining, association analysis and statistical learning. We have analyzed these algorithms in depth and have put forward a simple explanation of these concepts with real world examples to help in the better understanding of the chosen algorithms and have also weighed in each one’s pros and cons individually to help with implementation of the algorithms.

II. Survey

A. C4.5 :

A classified set of data representing things is given to C4.5. With this data, c4.5 constructs a classifier in the form of a decision tree. Usually data mining uses classifier as a tool to classify a bunch of data representing things and predicts which class the data may be grouped to. Example: To predict whether the patient will get cancer or not. Hence performing C4.5 algorithm for the dataset which contains a bunch of patients details like age, blood pressure, and pulse rate, family history, VO$_2$max etc. These are called as attributes. Now the patient’s data has to be grouped under two classes, Class 1 & class 2: Whether the patient will get cancer or not. From these attributes C4.5 can predict whether the patient will get cancer. A decision tree is built from patients attributes and corresponding classes. This decision tree can further predict the class for new patients based on their attributes. The work of a decision tree is to create something similar to a flowchart to classify the data. The example can be further understood by explaining one path of the flowchart. The patient has tumors, history of cancer in the family, expressing a gene highly correlated with cancer patients; tumor size is greater than 5cm.The data gets classified depending on certain value of the attribute. C4.5 is supervised learning as the data set is labeled with classes. From the above stated example, C4.5 is not self-learning. It can only predict if the patient will get cancer or not only from the decision tree classification. C4.5 differs from the other decision tree systems as it uses information gain when generating the decision tree. Although other systems can integrate pruning, C4.5 uses a single-pass pruning process to move around over-fitting. Pruning results in many improvements. It can work with both discrete data and continuous data. Incomplete data or missing data can be dealt in its own way. The main reason to go for C4.5 would be its bestselling point of decision trees in their ease of analysis and explanation. It also gives a fast response and is human readable. Easily interpreted models can be built and implementation is easy. Categorical values and continues values can be used and C4.5 deals with noise. The limitations of this approach are that when the variable has close values or if there is a small variation in the data, different decision trees are formed. Not suitable while working with small training set. It is used in the decision tree classification for open source java implementation at opentox.

B. *K-MEANS :*

It is a famous clustering analysis technique for handling and exploring dataset. It first creates a k group from the set of objects such that the members of the group are similar. Cluster analysis is a family of algorithms intended to form groups such that group members are more similar versus non-group members. In clustering analysis, clusters and groups are synonyms. Example for k-means: For a dataset which consists of patient information: In cluster analysis, the data set is called as observations. The patient's information includes age, pulse, blood pressure, cholesterol, etc. This is a vector representation of the data. Vector representation is in the form of a multi-dimensional plot. The list is interpreted as coordinates. Where cholesterol can be one dimension and age can be another dimension. From the set of vectors, K means does the clustering. The user only has to mention the number of clusters that are needed. K-means clustering operation has a different types of variations to optimize for certain types of data. At a high level, k-means picks the different points and represents each of them k clusters. These are points are called as centroids. Every patient will be closest to any one of the centroids. They won't be closest to the same one. Hence they will form a cluster around their nearest centroids. There are totally 'k' clusters. All the patients will be a part of a cluster. Now k-means finds the center of each cluster based on its cluster member using patient vectors. This center is now the new centroid for the cluster. Due to change in the centroid, patients may now be closer to a different centroid. In other words, they may change their cluster membership. Steps 2-6 are repeated such that a point occurs where the centroids no longer shift position and the cluster membership stabilize. This is called convergence. K-means algorithm can either be supervised or unsupervised. But mostly we would classify k-means as unsupervised. We call it unsupervised because k-means algorithm learns about the clusters on its own without any information about the cluster from the user. The user only has to mention the number the clusters that are required. The key point of using k means is its simplicity. It is faster and more efficient than other algorithms, especially for large datasets. k-means can be used in Apache Mahout, MATLAB, SAS R, SciPy, Weka, Julia,. The advantages of k-means are:

If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls. K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular. The limitations are: Difficult to predict K-Value, with global cluster. Another limitation encountered is that different initial partitions can result in different final clusters. It does not work well with clusters (in the original data) of Different size and Different density.

C. *Support vector machine (SVM) :*

It classifies the data into two classes from the hyperplane. SVM performs a similar task like C4.5 at higher level. But SVM doesn't use decision tree. A hyperplane is a function similar to the equation of a line, y=mx+b. It's a simple classification of a task with just two features. SVM figures out the ideal hyperplane which separates the data into two classes. Example for SVM algorithm: There are a bunch of red and blue balls on the table. The balls aren't too mixed together and you could take a stick and separate the balls without moving the stick. So by this way when a new ball is added to the table, by knowing which side of the stick the ball is on, the color of the new ball can be predicted. Similarly, the balls represent the data points and the red & blue balls represent the two classes. The hyperplane is the stick. SVM figures out the function of the hyperplane all by itself. The problem is when the balls are mixed and a straight stick won't work. Here is the solution. Throw the balls in the air and use a paper to divide the balls in the air. Lifting up the table is equivalent to mapping your data in higher dimensions. In such cases, we go from two dimensions to three dimensions. The second dimension is the table surface and the third dimension is the balls in the air. SVM does this by using kernel which operates in higher dimension. The large sheet of paper is a function for a plane rather than a line. Therefore, SVM maps the things into higher dimensions and finds a hyperplane to separate the classes. Margins are generally associated with SVM. It is the distance between the hyperplane and the two closest data points from the respective class.

Advantage: Produce very accurate classifiers, Less over fitting and robust to noise. The limitations are: SVM is a binary classifier. To do a multi-class classification, pair-wise classifications can be used (one class against all others, for all classes). Computationally expensive, thus runs slow.

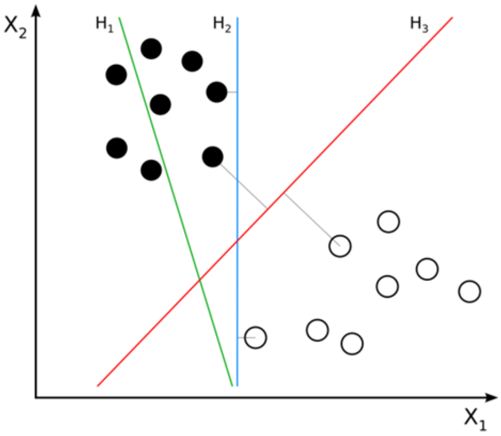
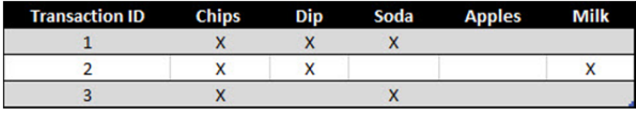


Figure 2 : SVM

D. *Apriori :*

It is applied to a dataset containing a large number of transactions. The algorithm learns associate rules. In data mining, associate rules are techniques for learning relations and correlations among variables in database. Example for Apriori algorithm: Consider the dataset of a supermarket transaction to be a giant spreadsheet. Each row of the spreadsheet is a customer transaction and every column represents a different grocery item. By using Apriori algorithm, we can analyze the items that are purchased together. We can also find the items that are frequently purchased than the other items. Together, these items are called item sets.

Table 1: Apriori Example:



Transaction ID	Chips	Dip	Soda	Apples	Milk
1	X	X	X		
2	X	X			X
3	X		X		

[1]

The main aim of this is to make the shoppers buy more. For example: You can see that chips & dip and chips & soda seem to frequently occur together. These are called 2-itemsets. With a large enough dataset, it is much harder to “see” the relationships especially when you’re dealing with 3-itemsets or more. Working of Apriori: three things need to be defined before starting with the algorithm. They are: The size of the item set (If you want to see the patterns for 2-itemset, 3-itemset, etc.?), the number of transactions containing the item set divided by the total number of transactions. A frequent item set is one which meets the support and Confidence or conditional probability. Apriori algorithm has 3 steps of approach: Join, Prune and Repeat. Apriori is an unsupervised learning approach. It discovers or mines for interesting patters and relationships. Apriori can also be supervised to do classification on labeled data. Apriori can be used for ARtool, Weka, and Orange. The advantages are: it uses large item set property, easily Parallelized, easy to implement. The limitations are: Assumes Transaction database is memory resident and requires many database scans.

E. Expectation-maximization (EM) :

It is a commonly used as clustering algorithm for knowledge discovery in data mining. It is similar to k-means. While figuring out the parameters of a statistical model with unobserved variables, the EM algorithm optimizes the likelihood of seeing the observed data. Statistical model is describing how observed data is generated. For example, the grades for an exam are normally distributed using a bell curve. Assume that this is the model. Distribution is generally the probabilities for all measurable outcome. The normal distribution of the grades for an exam represents all the probabilities of a grade. Or it is the determination of how many exam takers are expected to get that grade. A normal distribution curve has two parameters: The mean and the variance. In certain cases, the mean and the variance may not be known. But still we can calculate the normal distribution using the sample case. For example, we have a set of grades and are told the grades follow a bell curve. However, we’re not given the grades but only a sample. Using these parameters, the hypothetical probability of the outcomes is called likelihood. Keeping in mind that it’s the hypothetical probability of the existing grades and not the probability of a future grade. Probability is estimating the possible outcomes that should be observed. Observed data is the data that you recorded. Unobserved data is data that is missing. There are many reasons that the data could be missing (not recorded, ignored, etc.). By optimizing the likelihood, EM generates a beautiful model that assigns class labels to the different data points. EM algorithm helps in the clustering of data. It begins by taking a guess at the model parameters. Then it follows a three step process: E-step: The probabilities for assignment of each data point to a cluster are calculated. This is done based on the model parameters. M-step: updates the model parameters based on the cluster assignment from the previous step (E-step). The process is repeated until the model parameters and cluster assignments stabilize. The EM algorithm is unsupervised as it is not provided with labeled class information. The reason to go for EM algorithm is that it is simple and straight-forward to implement, it can iteratively make guesses about missing data and can be optimized for model parameters. The limitation are: EM always doesn’t find the optimal parameters and gets stuck in local optima rather than global optima. EM is quick in the early few iterations, but slower in the latter iterations.

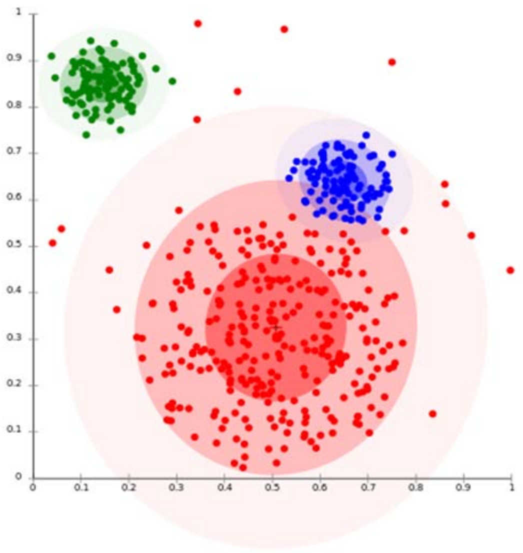


Figure 3 : EM

F. Page rank :

It is a link algorithm used to determine the relative significance of certain object linked within a network of objects. Link analysis is similar to network analysis which is looking to search the association among links. The most common example for page rank is Google’s search engine. Though, the search engine doesn’t solely rely on page rank. It’s one of the measures goggle uses to determine a web page importance. Advantages are: Robust against spam, global measure and query independent. The limitations are: Favors older pages

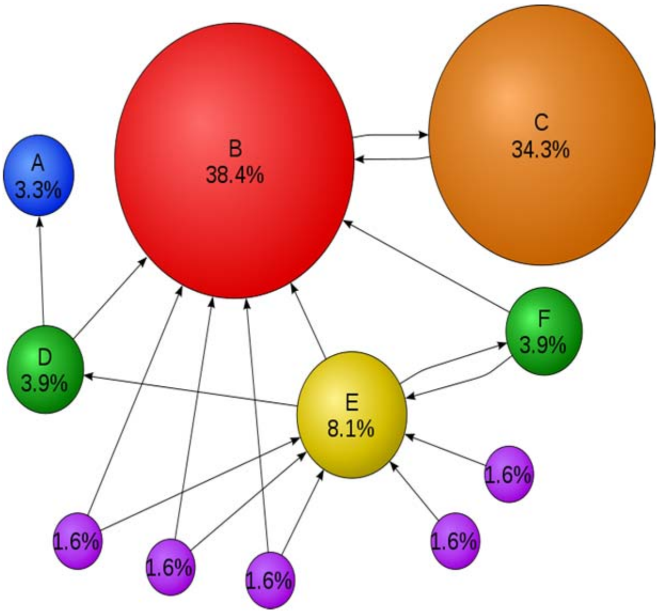


Figure 4 : Page Rank

G. kNN :

It stands for k-Nearest Neighbors. It is a classification algorithm which differs from the other classifiers previously described because it's a slow learner. A slow learner only stores the training data during training process. It classifies only when a new unlabeled data is given as input. On the other hand, a fast learner builds a classification model during training. When new unlabeled data is given as input, this type of learner feeds the data into the classification model. C4.5 and SVM are both fast learners because: SVM builds a hyperplane classification model during training. C4.5 builds a decision tree classification model during training.

kNN builds no such classification model as seen above. Instead, it just stores the initial labeled training data. When new unlabeled data comes in, kNN operates in 2 basic steps: First, it looks at the k closest labeled training data points. Second, using the neighbors' classes, kNN gets a better idea of how the new data should be classified. For figuring out the closest data, kNN uses a distance metric like Euclidean distance. The choice of metric for the distance largely depends on the data. Some even suggest learning a distance metric based on the training data. There's a lot of detail and many papers on kNN distance metrics. For data that is discrete, the idea is to transform the obtained discrete data into continuous data. 2 examples of this are: Using Hamming distance as a metric for the "closeness" of two text strings and hence transforming discrete data into binary features. kNN has an easy time when all the neighbors are of the same class. The basic feeling is if all the neighbors agree, then the new data point is likely to fall in the same class. Two common techniques for deciding the class when the neighbors don't have the same class is to take a simple majority vote from the neighbors. Whichever class has the greatest number of votes becomes the class for the new data point. Take a similar vote, except this time give a heavier weight to those neighbors that are closer. A simple way to implement this is to use reciprocal distance. For example, if the neighbor is 5 units away, then the weight of its vote is 1/5. As the neighbor gets further away, the reciprocal distance gets smaller and smaller. kNN is a supervised learning algorithm as it is provided with labeled training dataset. A number of kNN implementations exist in MATLAB k-nearest neighbor classification, scikit-learn KNeighborsClassifier and k-Nearest Neighbor Classification in R. The advantages are: Depending on the distance metric, kNN can be quite accurate. Ease understanding and implementing. The disadvantages are: Noisy data can throw off kNN classifications. kNN can be computationally expensive when trying to determine the nearest neighbors on a large dataset. kNN generally requires greater storage requirements than faster classifiers. Selecting a good distance metric is crucial to kNN's accuracy.

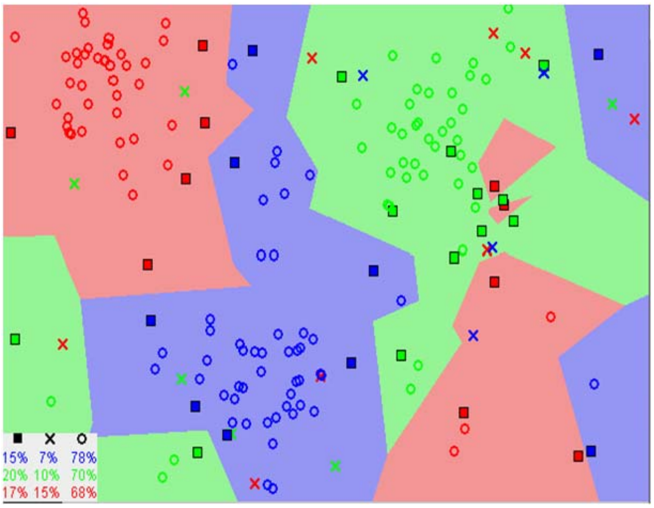


Figure 5 : kNN

H. CART:

CART stands for Classification and Regression Trees. It is a decision tree learning technique similar to C4.5 algorithm. The output is either a classification or a regression tree. In simple, CART is a classifier.

A classification tree is a type of decision tree. The output of a classification tree is a class. A regression tree predicts numeric or continuous values. Classification tree outputs classes, regression tree outputs numbers. For example, given a patient dataset, you might attempt to predict whether the patient will get cancer. The class would either be "will get cancer" or "won't get cancer." The numeric or continuous value will be the patient's length of stay or the price of a smart phone. The classification of data using a decision tree is similar to that of C4.5 algorithm. CART is a supervised learning technique as it is providing with a labeled training data set in order to construct the classification or regression tree model.

CART is used in: scikit-learn implements, R's tree package and MATLAB. The advantages are: Builds models that can be easily interpreted. Easy to implement. Can use both categorical and continuous values. Deals with noise.

III. CONCLUSION

Data mining is a broad area that deals in the analysis of large volume of data by the integration of techniques from several fields such as machine learning, pattern recognition, statics, artificial intelligence and database management system. We have observed a large number of algorithms to perform data analysis tasks. We hope this paper inspires more research in data mining so as to further explore these algorithms, including their many impact and look for new new research issues.

REFERENCES

[1] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, "Top 10 algorithms in data mining", Springer-Verlag London limited, 2007.

[2] http://rayli.net/blog/data/top-10-data-mining-algorithms-in-plain-english/

[3] http://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html

[4] http://dianewilcox.blogspot.com/

[5] http://www.slideshare.net/Tommy96/top-10-algorithms-in-data-mining

[6] http://janzz.technology/glossary-en/

[7] http://dsguide.biz/reader/tag/data?page=2