Data Mining Algorithms

Aakash Patel[1] Juhil Zalavadiya[2]
[1]Department of Computer Engineering [2]Department of Information and Technology Engineering
[1,2]A. D Patel Institute of Technology, New Vidhya Nagar, Anand, Gujarat, India

Abstract— This paper presents the 5 of the top data mining algorithms: Apriori, Decision Tree, Association Rule Mining, Linear Progression, K-mean Clustering. These algorithms are among the most influential data mining algorithms in the research community. With each algorithm we have given a brief description of the algorithm, the impact of the algorithm and why is consider one of the best algorithms. The classification, clustering, statistical learning, association analysis and link mining are also discussed which are among the most important topics of the data mining community.

Keywords: Algorithms, Data Mining, Clustering, Analysis, Progression, Association

I. INTRODUCTION

Data mining refers to extracting or mining the knowledge from large amount of data. It is an interdisciplinary field, whose core is at the intersection of machine learning, statistics, and databases. The word "Data Mining" can also be referred to as mining of knowledge. As of now tech firms like Amazon, Facebook, Google are considered as the giants of the tech world they have trillions of user data stored in their databases. This data can be used for a good cause, many useful information related to the users can be mined and can be used to generate a tremendous profit. Here data mining comes into play. Data Scientist analysis this data sets apply the various data mining algorithms and try to extract the useful information from those data sets which can be used ahead.

A. Definition:

Data Mining is a process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order or to discover the meaningful patterns and rules.

Data Mining is also called by some other names like Knowledge Discovery in Database, Knowledge Extraction, Data or Pattern Analysis, Data Archeology, Data Dredging, Information Harvesting, BusinessHarvesting etc.

Their many applications of data mining but it is most importantly used for market analysis and management, risk analysis and finding out the best possible alternatives and fraud analysis and management.

B. Related Works:

Many papers have been published by different authors as an attempt to discuss the top, most frequently used and most efficient algorithms when it comes to testing of a huge data set. Byung-Hoon Park et al. [1] of University of Maryland Baltimore County published their research work on Distributed Data Mining. They discussed theiralgorithms, systems and applications. Nikita Jain et al. [2] of Arya College of Engineering and IT published a paper on discussing the various Data Mining Algorithms. A brief discussion on Neural Networks and Genetic Algorithms is discussed and the ways in which the data mining can be achieved on Neural networks and Genetic Algorithms and also the paper conducts a formal review of the area of rule extraction from ANN and GA. Agarwal Srikant.[3] had published a paper discussing the algorithms for Mining Association Rules. Nesma Settoutietal. [4] had published a research paper which made a Statistical Comparisons of the top 10 algorithms in Data Mining for Classification Task. Moloud Abdaret al. [5] published a research paper for comparing the performance of data mining algorithms in prediction of heart diseases. Abdullah H. Wahbehet al. [6] published a research paper making a comparison between the Data Mining Tools Over some Classification Methods. Rafael S. Parpinelliet al. [7] discussed the process of Data Mining with an Ant Colony Optimization Algorithm.

C. Methodologies of Data Mining

In this paper we have discussed five types of data mining algorithms based on their popularity among different fields and their specifications which can be applied on large data set.

1) Apriori Algorithm:

Apriori algorithm, a classic algorithm, is useful in mining frequent item sets and relevant association rules. Usually, you operate this algorithm on a database containing a large number of transactions. One such example is the items customers buy at a hardware store. It helps the customers buy their items with ease, and enhances the sales performance of the departmental store. Mostly this algorithm is largely used in the field of medical science where it is used in detecting adverse drug reactions (ADR) by producing a set of association rules which is used to find out the combination of medications available and patients characteristics which can lead to Adverse Drug Reactions. (ADR).

Apriori algorithm has three significant components:

- Support
- Confidence
- Leaf

Let's look deep into it by taking an example into consideration:

First of all, you need a huge database. Let us consider you have 6200 customer transactions in a hardware store. You have to find the Support, Confidence, and Lift for two items, say paint and brush. It is because people frequently bundle these two items together. Out of the 6200 transactions, 1500 contain paint whereas 1000 contain brushes. These 1000 transactions include a 500 that includes brushes as well as paints. Using this data, we shall find out the support, confidence, and lift.

D. Support:

Support is the default popularity of any item. You calculate the Support as a quotient of the division of the number of transactions containing that item by the total number of transactions. Hence, in our example,

Support (Paint) = (Transactions involving Paint) / (Total Transactions)

= 1500/6200 ≈ 24.2%

E. Confidence:

In our example, Confidence is the likelihood that customer bought both Paint and brushes. Dividing the number of transactions that include both Paint and brushes by the total number of transactions will give the Confidence figure.
Confidence = (Transactions involving both Paint and brushes) / (Total Transactions involving Paint)
= 500/1500 ≈ 33.3%
It implies that approximately 33.3% of customers who bought Paint bought Brushes as well.

F. Lift:

According to our example, Lift is the increase in the ratio of the sale of Brushes when you sell Paint. The mathematical formula of Lift is as follows.
Lift = (Confidence (Paint – Brushes)) / (Support (Paint))
= 33.3 / 24.2 ≈ 1.38

It says that the likelihood of a customer buying both Paint and brushes together is 1.38 times more than the chance of purchasing Paint alone. If the Lift value is less than 1, it entails that the customers are unlikely to buy both the items together. Greater the value, the better is the combination.

G. Apriori Algorithms Pros:

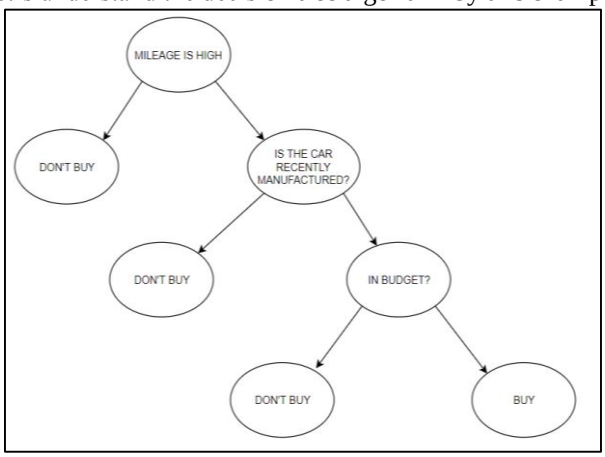
- Easy to understand and implement
- Can use on large datasets.

H. Apriori Algorithms Cons:

- It can become computationally expensive because sometimes it requires a large number of candidate rules.
- The calculation of SUPPORT is quite expensive because the calculation needs to be passed through the entire database. So it is a quite time consuming and expensive process.

II. DECISION TREE ALGORITHM

Let's understand the decision tree algorithm by this example.



This is a classic case of decision tree algorithm example.

Suppose you have decided to buy a used car, but you are stressing on the requirement that the mileage should be high. Thus, if initially, you come to know that the mileage is low, you would reject the idea of buying the car, but if the mileage is high, you will go one step further and see how old the car is. If it is recent, you will buy it, if not, you won't. Such kind of decision making is done by humans on a regular basis, and the tree shown above that represents the decision-making process is called a predictive model, which predicts a further step on the basis of existing data of previous steps. The importance of the decision tree algorithm in machine learning is often stressed upon stating this similarity with humans' decision-making methodology. With various languages being popular for data science and machine learning, decision tree algorithm in Python is an attractive combination.
The Terminology of Decisiontree:
The decision tree contains seven main features.

A. Root Node:

The root node is the largest (undivided) sample of data at the beginning, which further gets split into two or more homogeneous sets.

B. Branch:

It is the sub-section of tree.

C. Parent Nodes and Child Nodes:

Any node split into sub-nodes is a parent node and the split nodes are called child nodes.

D. Decision Nodes:

Any node that splits into sub-nodes can be called a Decision Node.

E. Leaf Nodes:

The final nodes that do not split further are called Leaf Nodes or Terminal Nodes.

F. Splitting:

The process of splitting a Node into two sub-nodes is called splitting. It occurs at all nodes except leaf nodes.

G. Pruning:

When we reduce the size of a tree (even literally) by removing any nodes/sub-nodes, it is called pruning.

H. The Algorithm

1) Place the best attribute of the dataset at the root of the tree.
2) Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3) Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

I. Pros:

It follows the same approach as humans generally follow while making decisions.

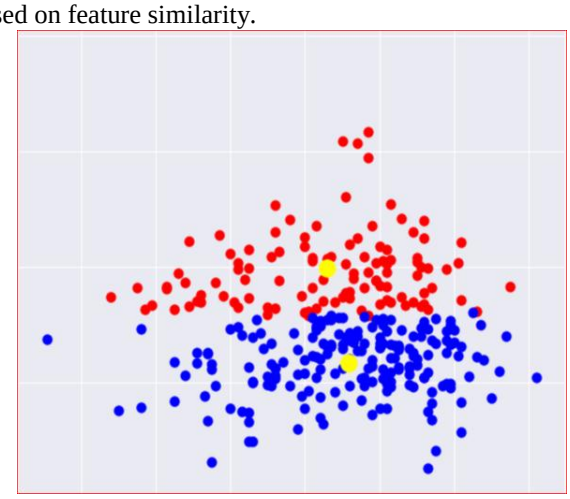
Interpretation of a complex Decision Tree model can be simplified by its visualizations. Even a naive person can understand logic.

J. Cons:

It gives low prediction accuracy for a dataset as compared to other machine learning algorithms.
Calculations can become complex when there are many class labels.

III. K-MEANS CLUSTERING

K-means clustering is a type of unsupervised learning, which is used when you have unlabelleddata. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.



A. *The Algorithm*

It has 4 basic steps:

1) Initialize Cluster Centroids (Choose those 2 books to start with)
2) Assign data points to Clusters (Place remaining the CDs one by one)
3) Update Cluster centroids (Start over with 2 different books)
4) Repeat step 2–3 until the stopping condition is met.

B. *Results:*

1) The centroids of the K clusters, which can be used to label new data
2) Labels for the training data (each data point is assigned to a single cluster)

C. *Pros:*

K-means is a fast method because it does not have many computations.

D. *Cons:*

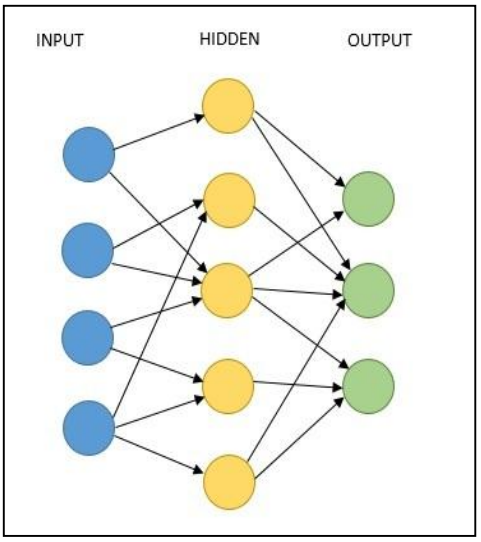
Identifying and classifying the groups can be a challenging aspect.

As it starts with a random choice of cluster centres, the results can lack consistency.

IV. ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks are the computational models that are inspired by the human brain. Many of the recent advancements have been made in the field of artificial intelligence, including voice recognition, image processing and robotics using artificial neural networks. Artificial neural networks, in general is a biologically inspired network of artificial neurons configured to perform specific tasks.

A neural network may contain the following 3 layers:



1) Input layer – It contains those units (Artificial Neurons) which receive input from the outside world on which network will learn, recognize about or otherwise process.
2) Output layer – It contains units that respond to the information about how it's learned any task.
3) Hidden layer – These units are in between input and output layers. The job of the hidden layer is to transform the input into something that output unit can use in some way.

A. *Training Algorithms for Artificial Neural Networks*

1) *Gradient Descent Algorithm*

This is the simplest training algorithm used in case of supervised training model. In case, the actual output is different from target output, the difference or error is find out. The gradient descent algorithm changes the weights of the network in such a manner to minimize this mistake**.**

2) *Back Propagation Algorithm*

It is an extension of the gradient-based delta learning rule. Here, after finding an error (the difference between desired and target), the error is propagated backward from the output layer to the input layer via the hidden layer. It is used in case of Multilayer Neural Network.

a) Pros:

– They can work fine in case of incomplete information
– Process information in a highly parallel way

b) Cons:

– Hardware dependence

V. SUPPORT VECTOR MACHINE ALGORITHM

A support vector algorithm is performed by plotting each acquired value of data as a point on an n-dimensional space or graph. Here "n" represents the total number of a feature of data that is present. The value of each data is represented as a particular coordinate on the graph.

After distribution of coordinate data, we can perform classification by finding the line or hyper-plane that distinctly divides and differentiates between the two classes of data.

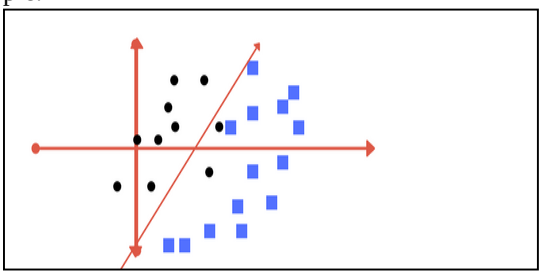
Support vector machines are a tool which best serves the purpose of separating two classes. They are a kernel-based algorithm.

- A kernel refers to a function that transforms the input data into a high dimensional space where the question or problem can be solved.
- A kernel function can be either linear or non-linear. Kernel methods are a type of class of algorithms for pattern analysis.
- The primary function of the kernel is to get data as input and transform them into the required forms of output.

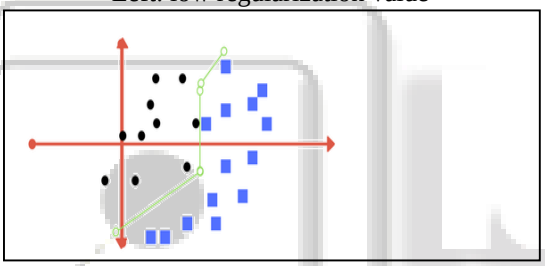
A. Tuning Parameters for Support Vector Machine Algorithm

1) Regularization

The Regularization parameter (often termed as C parameter in python's sklearn library) tells the SVM optimization how much you want to avoid misclassifying each training example.



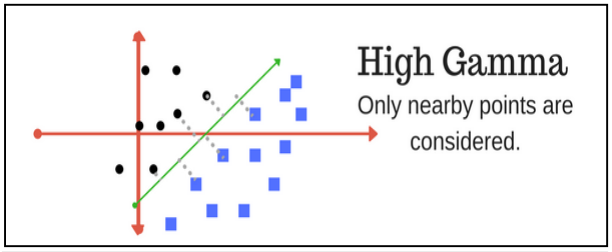
Left: low regularization value

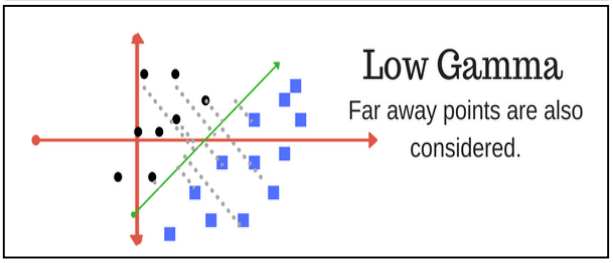


Right: high regularization value

2) Gamma

The gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. In other words, with low gamma, points far away from plausible seperation line are considered in calculation for the seperation line. Whereas high gamma means the points close to plausible line are considered in calculation.

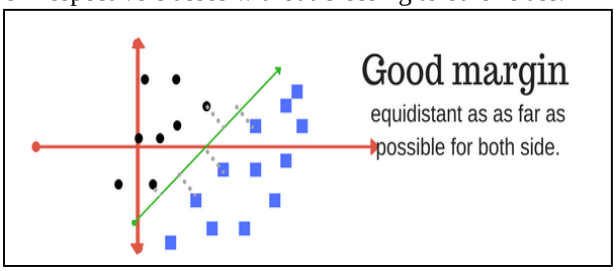


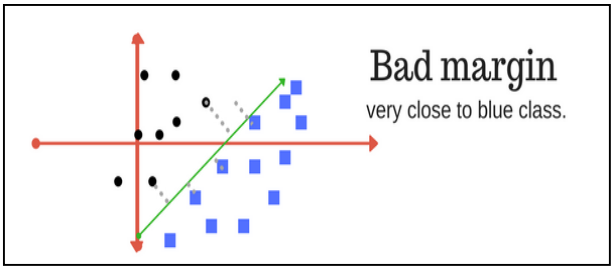


3) Margin

A margin is a separation of line to the closest class points.

A good margin is one where this separation is larger for both the classes. Images below gives to visual example of good and bad margin. A good margin allows the points to be in their respective classes without crossing to other class.





4) Pros:

SVM works relatively well when there is clear margin of separation between classes.

SVM is relatively memory efficient

5) Cons:

SVM algorithm is not suitable for large data sets.

In cases where number of features for each data point exceeds the number of training data sample, the SVM will underperform.

VI. Conclusion

So these were the best suitable algorithms and most frequently used algorithms in the computing world. The areas in which these methodologies can be applied is very astonishing. At present data mining is very new and important area of research it is very suitable for solving problems of data because of the characteristics of robustness, self-organizing, adaptive, parallel processing, distributed storage and high degree of fault tolerance. The commercial, educational and scientific applications are increasingly dependent on these methodologies.

References

[1] Byung-Hoon Park and HillolKargupta proposed a paper on Distributed Data Mining: Algorithms, Systems and applications.
[2] Nikita Jain and Vishal Srivastava proposed a paper on Data Mining Techniques.
[3] Agrawal and Srikantproposed a paperFast Algorithms for Mining Association Rules.
[4] NesmaSettouti, Mohammed El Amine Bechar and Mohammed Amine Chikhproposed a paperStatistical Comparisons of the Top 10 Algorithms in Data Mining for Classification Task.
[5] MoloudAbdar, Sharareh R. NiakanKalhori, ToleSutikno, Imam Much IbnuSubroto, GoliArjiproposed a

paperComparing Performance of Data Mining Algorithms in Prediction Heart Diseases.
[6] Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfaproposed a paperA Comparison Study between Data Mining Tools over some Classification Methods.
[7] Rafael S. Parpinelli,Heitor S. Lopes and Alex A. FreitasData Mining With an Ant Colony Optimization Algorithm.

