© 2018 IJSRSET | Volume 4 | Issue 4 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Evaluating the performance of SVM and Apriori Algorithms for Bigdata

Sudha M[1], Saravana Kumar E[2]

[1]P.G. Scholar, Dept of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, India

[2]Associate Professor, Dept of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, India

ABSTRACT

Data Mining is the process of extracting the required information from large data sets to obtain accurate value. This is the study paper about the Machine Learning algorithms. Supervised Machine Learning Algorithm is One type of Machine Learning Algorithm Which is used for classifications and regressions, but SVM algorithms are commonly used in classifications. Association Rule Mining(ARM) is a technique which is used to find the frequent patterns. The Frequent patterns are which occurs frequently in transactional data. In Frequent Pattern Mining, Apriori algorithm is used to find the frequent items in the transaction Data. Now a days Market Basket Analysis Association rules are used in many applications.

Keywords: ARM, SVM, Apriori, Dip-svm

I. INTRODUCTION

Association Mining is a technique that finds its usage in the market basket analysis. This technique, as can be said in general terms, is used in order to bring together items of the same type [2]. Association analysis can be used to improve decision making in a wide variety of applications such as: market basket analysis, medical diagnosis, biomedical literature, protein sequences, survey data, logistic failure, deception detection in web [2]. Apriori algorithm is, the most classical and important algorithm for mining frequent item sets, proposed by Agrawal and R. Srikant in 1994. Apriori is used to find all frequent itemset in a given database DB. The key idea of Apriori algorithm is to make multiple passes over the database [3]. Frequent patterns are patterns like as item sets, subsequences or substructures that come along in a data set subsequently. Behalf of the transactional database, we can suppose the behavior of the products purchased by the customers. For example a set of items Mobile and Sim card that appear frequently as well as together in a transaction set is a frequent item set. Subsequences means if a customer buys a Mobile he must also buy a Sim card and then head phone etc. From the overall structure of the database these transactions are occurs sequentially is called sequential patterns. The Substructure is concerned to different structural forms such as sub graphs, sub trees which may be manipulate along with item sets or sequences [7].

Some List of Machine Learning Algorithms

1. Artificial Neural Networks
2. Random Forests
3. Decision Trees
4. Naïve Bayes Classifier Algorithm
5. K Means Clustering Algorithm
6. Support Vector Machine Algorithm
7. Apriori Algorithm
8. Linear Regression
9. Logistic Regression

10. Nearest Neighbor's [5].

II. MACHINE LEARNING ALGORITHMS ARE CLASSIFIED AS

1. **Supervised Machine Learning Algorithms:**
   a. Make predictions on given set of samples.
   b. Searches for patterns within the value labels assigned to data points.
2. **Unsupervised Machine Learning Algorithms**
   a. No labels associated with data points.
   b. Organize the data into a group of clusters to describe its structure for analysis.

Support Vector Machine (SVM) Learning Algorithm

SVM is a supervised machine learning algorithm for classification or regression problems, SVM can classify any new data. It works by classifying the data into different classes by finding a line which is called as Hyperplane. It separates the training data set into classes.

SVM's are classified into

Linear SVM's: - Training data are separated by a hyperplane to classify.

Non-Linear SVM's: - It is not possible to separate the training data using a hyperplane.

SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes, as shown below [9].

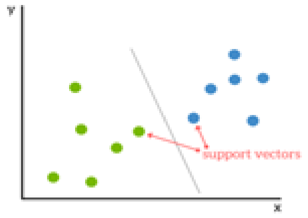

Figure 1. Support Vector Machine [9]

Support Vectors

Support vectors are the data points nearest to the hyperplane, the points of a data set that, if removed, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of a data set.

Hyperplane

As a simple example, for a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data.

Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it[9].

Advantages of SVM:

- ✓ It offers best classification performance (accuracy) on the training data.
- ✓ More efficiency for correct classification of the future data.
- ✓ It does not make any strong assumptions on data.
- ✓ It does not over-fit the data.

Dip-SVM

Data classification is the process of organizing data into categories for its most effective and efficient use. The goal of classification is to accurately predict the target class for each case in the data. Dip-SVM – algorithm is used to obtain the minimal loss in classification accuracy and low communication overhead, two phases [6]:

1. Distribution preserving partitioning (DPP)
2. Distributed learning phase(DL)

III. ASSOCIATION RULE MINING

Association rule learning is a method for determining relations among variables in large databases. The formal definition of the problem of association rules given by Rakesh Agrawal, the President and Founder of the Data Insights Laboratories. The traditional association rule mining techniques provides predefined support and confidence values. By using this algorithm, we can create association rules depending upon the dataset available in the database.

Association rules mining is to find the associations and relations among item sets of large data. Association rules mining is an important technique used in data mining research, and association rules is the most typical style of data mining.

There are many algorithms for finding frequent patterns. Association rule mining first consider by Agrawal has now become one of the main pillars of data mining and knowledge discovery tasks. An association rule is an expression $X \rightarrow Y$, where X is a set of items and Y is usually a single item. It means in the set of transactions, if all the items in X exist in a transaction, then Y is also in the transaction with a high probability or in other word It is a method of finding relationships of the form x→y item sets that occur together in a database where X and Y are disjoint item sets [2]. Support and confidence are two key measures for association rule mining. The association rule point that the transactions that contain X, tend to contain Y support. Let I = {I1, I2 …Im} be a set of items. Let D, the task relevant data, be a set of database transactions where each transaction T is a set of items such that T C [4]. Association rules mining tend to produce a large number of rules. The goal is to find the rules that are useful to users. There are two ways of measuring usefulness, being objectively and subjectively. Objective measures involve statistical analysis of the data, such as support and confidence (Agrawal et al., 1993) [8].

$$support(x \rightarrow y) = \frac{Number\ of\ transaction\ in\ which\ x\ appears}{Total\ number\ of\ Transactions}$$

$$confidence(x \rightarrow y) = \frac{support(x \cup y)}{support(x)}$$

Apriori algorithm:

```
$Ck$: Candidate item set of size k
$Lk$: frequent item set of size k
$L1$= {frequent items};
for ($k$= 1; $Lk$! =∅; $k$++) do begin
$Ck+1$= candidates generated from $Lk$;
for each transaction $t$in database do Increment the count of all candidates in $Ck+1$that are contained in $t$
$Lk+1$= candidates in $Ck+1$with min_support
End
return∪$kLk$;[4].
```

Measures used in Apriori such as support, confidence, lift and conviction. Other related measures are

- All-confidence
- Collective strength
- Leverage

In association rule mining two threshold values are required. As given below,

Minimum support: The rule X ⇒ Y holds with support 's' if s% of transactions in D contains $X \cup Y$ Rules that have a 's' greater than a user specified support is said to have minimum support[8].

Minimum confidence: The rule X ⇒ Y holds with confidence 'c' if c% of the transactions in D that contain X also contain Y . Rules that have a 'c' greater than a user-specified confidence is said to have minimum confidence [8].

APRIORI ALGORITHM

The algorithm [7] is designed to find associations in sets of data in a database. Apriori is a definitive algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation).Apriori uses breadth-first search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length *k* from item sets of length *k* − 1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent *k*-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates. Candidate generation generates large numbers of subsets (the algorithm attempts to load up the candidate set with as many as possible before each scan). Bottom-up subset exploration (essentially a breadth-first traversal of the subset lattice) finds any maximal subset S only after all 2 | *S* | − 1 of its proper subsets.

ITEMSET

Item set is collection of items in a database which is denoted by D= {x1, x2,……..…,xn},Here 'n' is the number of items.

CANDIDATE ITEMSET

Candidate item sets are items which are only to be considered for the processing. Candidate item set are all the possible combination of item set. It is usually denoted by 'Ci' where 'i' indicates the i-item set.

TRANSACTION

Transaction is a database entry which contains collection of items. Transaction is denoted by and T⊆D. A transaction contains set of items T= {x1, x2, ……………., xn}.

MINIMUM SUPPORT

Minimum support is basically condition which should be satisfied by the given items so that further processing of that item can be completed. Minimum support can be considered as a condition which helps in removal of the in-frequent items in any database. Usually the Minimum support is given in terms of percentage.

FREQUENT ITEMSET

Frequent item set is commonly large item set i.e. the item sets which satisfies the minimum support threshold value are known as frequent item sets. It is usually denoted by 'Li' where 'i' indicates the i-item set.

CONFIDENCE

Confidence indicates the certainty of the rule. This argument lets us to count how often a transaction's item set couple with the left side of the implication with the right side. The item set which does not satisfies the above condition can be discarded. Consider two items X and Y. To calculate confidence of X->Y the following formula is used, Conf(X->Y) = (number of transactions containing both X& Y) (Transactions containing only)[7].

- **Itemset:** A collection of one or more items.
  - ✓ Example
  - ✓ {Milk, bread, jam}. K-itemset that contains k-items [3].
- **Frequent Itemset:** An itemset whose support is greater than or equal to a min_sup threshold. In association rule mining task from a set of transactions T, the goal of association rule mining is to find all rules having [3].Support >= min_sup; threshold and Confidence>= min_conf threshold[3].

Simple Example:

Frequent itemset: I_1= {A, B, C}, I_2= {B, C}

Here I_1, I_2 are transactions, (B, C) repeatedly occurred in two transactions so is called frequent itemset.

Table 1

Transaction	Items
T_1	A, B, C
T_2	A, B, D
T_3	B, C
T_4	A, C
T_5	B, C, D

Total number of Transaction =5

For Example user-Specified Support and confidence are 50% Support(AB) = 2/5=40% ; Support(B,C) = 3/5=60% (ie) occurrence of AB and BC in total transactions/total number of transactions .
Confidence {A=>B}= 2/3=66. // this meets user specified confidence >=50.

Advantages of Apriori algorithm

- easy-to-implement and easy-to-understand algorithm
- It can be used on large item sets

Disadvantage of Apriori algorithm

- Need to find a large number of candidate rules which can be computationally expensive.

IV. EXPERIMENTAL STEPS

Dataset: In this research the accidental datasets are used to predict the accuracy and lifetime.

Process involved

- Preprocessing
- Partitioning
- Clustering
- Classification (Dip-SVM- Apriori)

Preprocessing

- Transforming raw data into an understandable format(structured format). Eliminating redundant data to save memory and Eliminating NULL values in dataset.

Partitioning

- Splitting Dataset in to manageable partitions, Training SVM on each partition. Apply some logic and in this module and split our file randomly. Partition should be done based on the size of file. User defined partitioning.

Clustering

- The process of organizing objects into groups whose members are similar in some way. A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. In proposed system we are using K-MEANS algorithm for clustering.

Classification (Dip-svm- Apriori)

- Data classification is the process of organizing data into categories for its most effective and efficient use. The goal of classification is to accurately predict the target class for each case in the data. Dip-SVM –algorithm is used to obtain the minimal loss in classification accuracy and low communication overhead. Two phases:

✓ Distribution preserving partitioning (DPP)
✓ Distributed learning phase(DL)

In our dataset the state-based classification is going to perform by using Dip-SVM algorithm. Apriori Algorithm finds the frequent patterns in large dataset. A two process are join and prune.

V. CONCLUSION

Datamining is the one the technique to mine the datasets, for research we can use any large datasets to do process. Here accidental dataset are used.To use Datamining as a tool to find the accuracy and frequent pattern. Apriori and Dip-svm both are supervised machine Learning algorithm to find lifetime, accuracy in large dataset. In this paper two supervised machine learning algorithms-SVM(Dip-svm), Apriori algorithm are discussed. Language used in this result is java, possible to try in other languages too like python etc.

VI. REFERENCES

[1]. Varsha Mashoria, Anju Singh, "Literature Survey on Various Frequent Pattern Mining Algorithm ", IOSR Journal of Engineering (IOSRJEN) , (Jan. 2013).

[2]. Amit Kumar Gupta Dr. Ruchi Rani Garg Virendra Kumar Sharma, "Association Rule Mining Techniques between Set of Items", ,

International Journal of Intelligent Computing and Informatics, Vol. 1, Issue 1, January 2014.

[3]. Charanjeet Kaur, "Association Rule Mining using Aprior Algorithm: A Survey", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 6, June 2013

[4]. Varsha Mashoria, Anju Singh Department of computer science, Barkatullah University Bhopal (M.P) India, "Literature Survey on Various Frequent Pattern Mining Algorithm" IOSR Journal of Engineering (IOSRJEN) e-ISSN: 2250-3021, p-ISSN: 2278-8719 Vol. 3, Issue 1 (Jan. 2013), ||V1|| PP 58-64.

[5]. www.analyticsvidhya.com/blog/2017/09/common-machine-learning- algorithms/.

[6]. Dinesh Singh, Student Member, IEEE, Debaditya Roy, Student Member, IEEE,and C. Krishna Mohan, Member, IEEE, "DiP-SVM : Distribution Preserving Support Vector Machine for Big Data" IEEE Transactions On Big Data,Vol. 3, No. 1, January-March 2017.

[7]. Peeyush Kumar, Shukla, "Review Paper On Mining Association Rule And Frequent Patterns Using Apriori Algorithm ".

[8]. Kenneth Lai, Narciso Cerpa," Support vs Confidence in Association RuleAlgorithms.

[9]. https://www.kdnuggets.com/2016/07/support-vector-machines- simple-explanation.html