ISSN No: - 2456 – 2165

A Survey on Data Mining Association Rules By Effective Algorithms

Sivarani Jalmanayani,
Dept of Computer Science,
Sri Padmavathi Womens University, Tirupati, India.
sivaranijalamanayani@gmail.com

T.Subramanyam,
Asst professor, Dept of Computer Science,
Sri Padmavathi Womens University, Tirupati,India
subbu637@gmail.com

Abstract: Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers and develop more effective marketing strategies as well as increase sales and decrease costs. Classification is a Data Mining technique, to build a concise model that can be used to predict the class of records whose class label is not known. Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases.This paper, we provide the basic concepts about classification, association rule mining, mining algorithms and Survey association rule mining techniques.

Keywords: Data mining, Single Dimensional, Multi-Dimensional, Hybrid Dimensional, Apriori algorithm, Éclat algorithm, FP growth algorithm.

I. INTRODUCTION

Data Mining [1] is the process of discovering knowledge, such as patterns, association, changes, anomalies, and significant structures, from large amount of data stored in database, data warehouses, or other information repositories. Data mining has been popular for knowledge discovery in database. The information or knowledge extracted so can be used for any of the following applications, operations, techniques and algorithms.

Applications	Market Analysis and Management Corporate Analysis and Risk Management Fraud Detection
Operations	Classification and Prediction Clustering Association Analysis Fore Casting
Techniques	Neural Networks Decision Trees K-nearest neighbour algorithms Naive Bayesian Cluster Analysis

Data mining can also be used in the areas of production control [2], customer retention [3], science exploration [4], sports [5], astrology [6], and Internet Web Surf-Aid [7].

A. Data Mining Applications

a). Market Analysis and Management

- *Customer Profiling* –Data mining helps determine what kind of people buy what kind of products.
- *Identifying Customer Requirements* –Data mining helps in identifying the best products for different customers. It uses prediction to find the new customers.
- *Cross Market Analysis* –Data mining performs association[8]/correlations[9] between product sales.
- *Target Marketing* –Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc.
- *Determining Customer Purchasing Pattern* –Data mining helps in determining customer purchasing pattern.
- *Providing Summary Information* –Data mining provides us various multi-dimensional[10] summary reports.

B. Corporate Analysis and Risk Management

- *Finance Planning and Asset Evaluation* –It involves cash flow analysis and prediction[11], contingent claim analysis to evaluate assets.
- *Resource Planning* –It involves summarizing and comparing the resources and spending.
- *Competition* –It involves monitoring competitors and market directions

C. Fraud Detection

Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyses the patterns that deviate from expected norms.

ISSN No: - 2456 – 2165

II. DATA MINING OPERATIONS

A. Classification

Classification is a most familiar and a most popular data mining technique. Classification[12][13] application includes image, pattern recognition, loan approval and detecting faults in industrial applications. All approaches to performing classification assume some knowledge of the data. Training set is used to develop specific parameters required by the technique. The goal of classification is to build a concise model that can be used to predict the class of records whose class label is not known.Classification rule mining is to find a small set of rules in the database to form an accurate classifier. There is one and only one pre-determined target: class.The data classification is a two-step process.

- The first step is learning: analyze data using a classification algorithm and build a model represented in the form of classification rules, decision trees, or mathematical formulae.
- The second step is classification: test the accuracy of the classification rules, if the accuracy is acceptable, the rules or the model is used for the classification of new data.

B. Clustering

- Clustering[14] is the process of making a group of abstract objects into classes of similar objects. Two types of clustering's:
- Intra class similarity - Objects are similar to objects in same cluster**.**
- Inter class dissimilarity - Objects are dissimilar to objects in other clusters.

C. Association

Association[15]refers to uncovering relationship among data. Used in retail sales community to identify the items (products) that are frequently purchased together.

D. Fore Casting

Estimate future values based on patterns with large sets of data. Data mining is the process of extracting patterns or correlations[16] among dozens of fields in large relational data bases[17]. With the amount of data doubling every three years, it is becoming increasingly important for transforming data into in-formation, which in turn, can be used to increase revenues, cut costs, or both. Data mining uses simple and multi-variety linear [18] and non-linear regression [19] models as well as hypothesis testing [20].

III. DATA MINING TECHNIQUES

A. Neural Networks

An artificial neural network (ANN) [21], is known as "neural network", it is also known as mathematical model or computational model based on biologic neural network. It maps a set of input data onto a set of appropriate output data. It consists of three layers input layer, hidden layer, and output layer. There is connection between each layer and weights are assigned to each connection. The primary function of neurons of input layer is to divided into xi into neurons in hidden layer. Neuron of hidden layer adds input signal xi with weight why of respective connections from input layer. The output yj is function of yj=f(wjixi), where f is a simple threshold function.

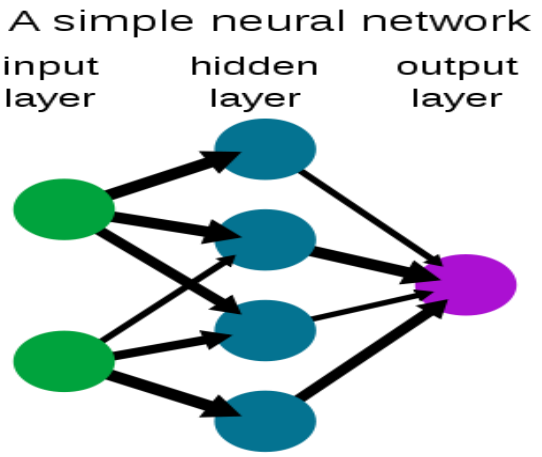


Fig. 1 A Simple Neural Network

B. Decision Trees

A decision tree [22] is a structure that includes a root node, branches, and leaf nodes. Each internal node represents a test on an attribute, each branch represents the outcome of a test, and each leaf node holds a class label. The top node in the tree is the root node.

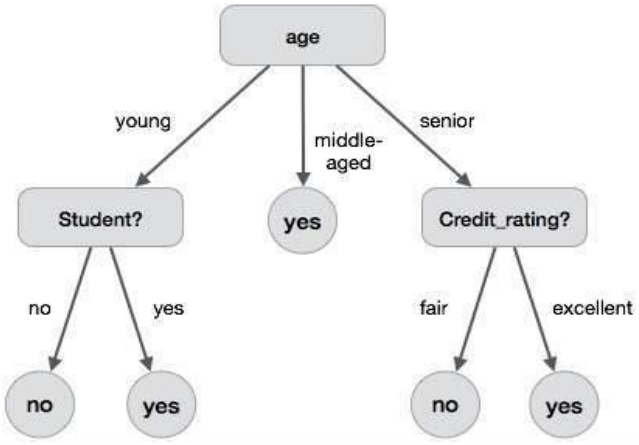


Fig. 2 K-Nearest Neighbour Algorithms

ISSN No: - 2456 – 2165

A powerful classification algorithm used in pattern recognition. K nearest neighbour [23] stores all available instances and classifies new instances based on a sameness measure (e.g. distance function).One of the top data mining algorithms used today. A non-parametric lazy learning algorithm (An instance based learning method).

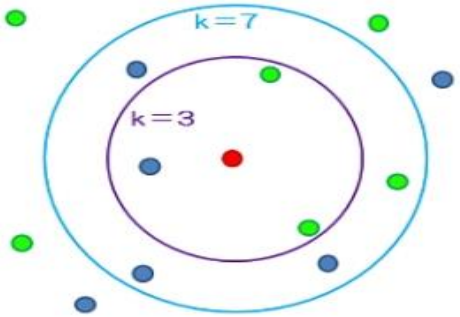


Fig.3 *Naive Bayes Algorithm*

The Naïve Bayesian classifier [24] is based on Bayes' theorem [25] with independence assumptions between predicators. A Naïve Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. In spite of its simplicity, the Naïve Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

C. Cluster analysis

In cluster analysis [26], we first partition the set of data into groups based on data similarity and then assign the labels to the groups. The main pros of clustering over classification are that, it is flexible to changes and helps single out useful features that distinguish dissimilar groups.

a). Applications of Clustering Analysis Are

- *Marketing*: Finding groups of customers with similar behaviour given a large database of customer data containing their properties and past buying records.
- *Biology:* Classifying of plants and animals given their features.
- *Libraries:* Book ordering.
- *Insurance:* Recognize groups of motor insurance policy holders with a high average claim cost; identifying frauds.
- *City-planning:* Recognize groups of houses according to their house type, value and geographical location.
- *Earthquake studies:* Clustering observed earthquake to identify dangerous zones.
- *WWW:* Record classification; clustering weblog data to discover groups of sameness access patterns.

IV. ASSOCIATION RULES

Association rules [27] are useful for analyses and predict customer behaviour. Association rules for if/then statements that help uncover the relationship between unrelated data in a relational data base or other information repository.
Ex:buys {bread} → buys {butter}
Buys {potatoes, onions}→buys{tomatoes}

A. Frequent Patterns and Association Rules

- Item set k={x1… xk}
- Find all the rules X →Y with minimum support and confidence
- Support: S, probability that a transaction contains both X and Y.
- Confidence: C, Conditional Probability that a transaction having X also contains Y.

Ex: Consider, in a Super market
Total Transactions: 100
Bread: 20
So,
20/100=20% which is support In 20 transactions, butter: 9 transactions
So, 9/20=45% which is confidence

B. Types of Association Rules

a). Single Dimensional Association Rule

In Single-Dimensional [28] rules have one dimension or predicate. That is items in a rule referred to only one dimension or predicate (e.g. , items are all in "product" dimension)

Ex: Buys(X, "milk")→ Buys(X, "bread")
Dimension: Buying

b). Multi-Dimensional Association Rule

In Multi-Dimensional rules [10] with two or more predicates or dimensions (i.e., items in ≤ 2 dimensions or predicates) and the predicates or dimensions should not be repeated.

Ex: Occupation (I.T), age (age>22) → buys (laptop)
Dimensions: occupation, age, buys

c). Hybrid Dimensional Association Rule

In Hybrid Dimensional rule [29] with repetitive predicates or dimensions.
Ex: time(5'O clock), buys(tea) → buys (biscuit)

d). Association Rules Are Used In Various Fields

- Web Usage Mining
- Banking
- Bio informatics

ISSN No: - 2456 – 2165

- Market based analysis
- Credit/Debit card analysis
- Product Clustering
- Catalogue Design

e). V Algorithms are used in Association Rule

Association algorithms are

- Apriori algorithm
- Éclat Algorithm
- FP- Growth Algorithm

C. Apriori Algorithm

Apriori algorithm is a classical algorithm in data mining. it is a "bottom up" approach. It is used for mining frequent item sets [30] and relevant association rules. It is devised to operate on a database containing a lot of transactions, for instance, items brought by customers in a store.

It is very important for effective Market Basket Analysis and it helps the customers in purchasing their items with more ease which increases the sales of the markets. It has also been used in the field of healthcare for the detection of adverse drug reactions. It produces association rules that indicate what all combinations of medications and patient characteristics lead to ADRs.

a). Key Points

- Frequent Item sets: The sets of item which has minimum support(denoted by Li for ith-Item set).
- Apriori Property: Any subset of frequent item set must be frequent.
- Join Operation: To find Lk , a set of candidate k-item sets is generated by joining Lk-1 with itself.

b). Steps to Perform Apriori Algorithm

1. Scan the transaction data base to get the support of S each 1-itemset, compare S with min-sup, and get a support of 1-itemsets, L1.
2. Use Lk-1 joins Lk-1 to generate a set of candidates' k-item sets. And use Apriori property to prune the unfrequented k-item sets from this set.
3. Scan the transaction database to get the support S of each candidate k-item set in the find set, compare S with min-sup, and get a set of frequent k-item sets Lk.
4. Compare the candidate set is equal to null or not. Candidate set is not null repeat step 2 otherwise go to next step5.
5. For each frequent item set 1, generate all nonempty subsets of 1.

Example: Find the frequent items sets in a data base 9 transactions. With a minimum support 50% and confidence 50%

Transaction Id	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Step1: Scan D for count of each candidate. The candidate list is {A,B,C,D,E,F}and find the support.C1=

Items	Sup
A	3
B	2
C	2
D	1
E	1
F	1

Step2: Compare candidate support count with minimum support count {50%}.L1=

Items	Sup
A	3
B	2
C	2

Step3: Generate candidate C2 from L1.C2

Items
{A,B}
{A,C}
{B,C}

Step4: Scan D for count of each candidate in C2 find the support.C2=

Items	Sup
{A,B}	1
{A,C}	2
{B,C}	1

ISSN No: - 2456 – 2165

Step5: Compare candidate(C2) support count with the minimum support count.L2=

Items	Sup
{A,C}	2

Step6: So the data contain frequent item. I(A,C) Therefore the association rule that can be generated from L, with the support and confidence [30].

Association Rule	Support	Confidence	Confidence %
A→C	2	2/3=0.66	66%
C→A	2	2/2=1	100%

Minimum confidence threshold is 50%(given).then both the rules are output as the confidence is above 50%.So final rules are: Rule1: A→C Rule2: C→A

c). Methods to Improve Apriori's Efficiency

- *Hash-based itemset counting:* A k-itemset whose corresponding hashing bucket count is below the threshold cannot be frequent.
- *Transaction reduction:* A transaction that does not contain any frequent k-itemset is useless in subsequent scans.
- *Partitioning:* Any itemset that is potentially frequent in DB must be frequent in at least one of the partitions of DB.
- *Sampling:* mining on a subset of given data, lower support threshold + a method to determine the completeness.
- *Dynamic itemset counting:* add new candidate itemsets only when all of their subsets are estimated to be frequent.

d). Advantages

- Uses large item set property.
- Easy parallelized.
- Easy to implement

e). Disadvantages

- Assume transaction database is memory resident
- Requires many database scans.

D. Éclat Algorithm

Equivalence Class Clustering and bottom up Lattice Traversal –ECLAT. Method for frequent item set generation Searches in a DFS manner. Represent the data in vertical format.

- *Steps*
  1. Get tidlist for each item (DB scan).
  2. Tidlist of {a} is exactly the list of transactions containing{a}.
  3. Intersect tidlist of {a} with the tidlist of all other items, resulting in tidlist of {a,b},{a,c},{a,d}…={a}-conditional database (if {a}removed).
  4. Repeat from 1 on {a}-conditional database.
  5. Repeat for all other items.

Example: Frequent itemset1 And min sup-2

Item Set	Tid Set
Bread	1,4,5,7,8,9
Butter	1,2,3,4,5,6,9
Milk	3,5,6,7,8,9
Coke	2,4
Jam	1,3

Frequent item set2

Item Set	Tid Set
{bread, butter}	1,4,8,9
{bread, milk}	5,7,8,9
{bread, coke}	4
{bread, jam}	1,8
{butter, milk}	3,6,8,9
{butter, coke}	2,4
{butter, jam}	1,8
{milk, jam}	8

Frequent Item Set 3

Item Set	Tid Set
{bread, butter, milk}	8,9
{bread, butter, jam}	1,8

- This process repeats, with k incremented by 1 each time, until no frequent items or no candidate item sets can be found.

a). Advantages

- Depth first search reduces memory requirements.

- Usually (considerably) faster than Apriori.
- No need to scan the database to find the support of (k+1) item sets, for k>=1.

b). Disadvantages

- The TID set can be quite long, hence expensive to manipulate

E. FP-Growth Algorithm

FP-Growth is an Apriori algorithm to find patterns in the data using a special data structure, a frequent pattern tree, and let it grow to mine patterns.

- *Steps:*

a). Build a compact data structure from FP-tree.

- Scan the transaction DB for the first time, find frequent items.
- Order them into a list L in frequency descending order.
- For each transaction, order it frequent items according to the order in L.
- Scan DB the second time, construct FP-tree byputting each frequency ordered transaction onto it.
- Extract frequent item sets directly from the FP-tree.

Example: Find all frequent item sets or frequent patterns in the following database using FP-growth algorithm. Take minimum support as 30%.

Tid	Items
1	E,A,D,B
2	D,A,C,E,B
3	C,A,B,E
4	B,A,D
5	D
6	D,B
7	A,D,E
8	B,C

Step 1 - Calculate Minimum support

Minimum support count(30/100 * 8) = 2.4
Minimum support count is ceiling(30/100 * 8) = 3

Step 2 - Find frequency of occurrence

Item	Frequency
A	5
B	6
C	3
D	6
E	4

Step 3 - Prioritize the Items

Item	Frequency	
A	5	**3**
B	6	**1**
C	3	**5**
D	6	**2**
E	4	**4**

The frequent item list for the table will be B:6, D:6, A: 5, E:4, C: 3.

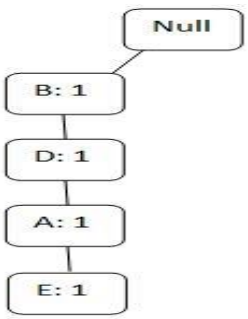
Step4-Order the Items According to Priority

Tid	Items	Ordered Items
1	E,A,D,B	B,A,D,E
2	D,A,C,E,B	B,D,A,E,C
3	C,A,B,E	B,A,E,C
4	B,A,D	B,D,A
5	D	D
6	D,B	B,D
7	A,D,E	D,A,E
8	B,C	B,C

Step5: Draw the FP Tree

Row1

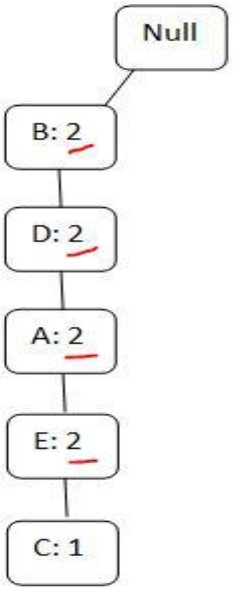
All FP trees have 'null' node as the root node. So draw the root node first and attach the items of the row 1 one by one respectively.



ISSN No: - 2456 – 2165

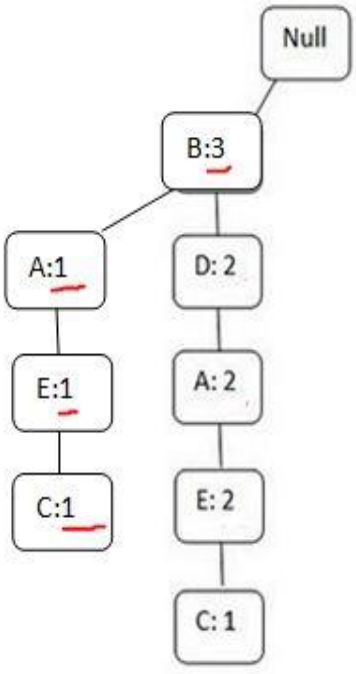
Row2

Then update the above tree by entering the items of row 2.



Row3

In row 3 you have to visit B,A,E and C respectively. So you may think you can follow the same branch again by replacing the values of B,A,E and C.

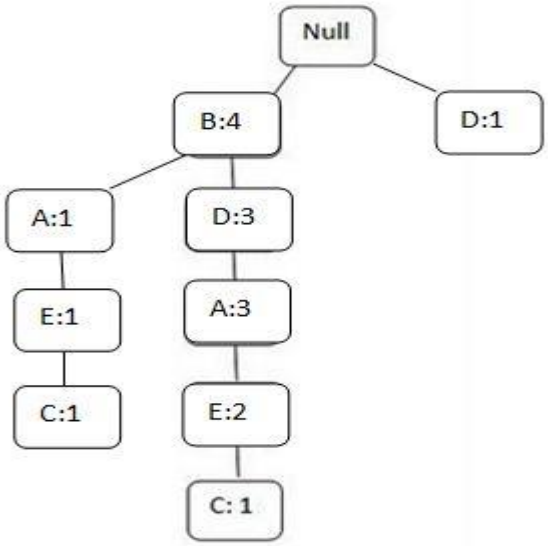


Row4

Then row 4 contain B,D,A. Now we can just rename the frequency of occurrences in the existing branch. As B:4,D,A:3.

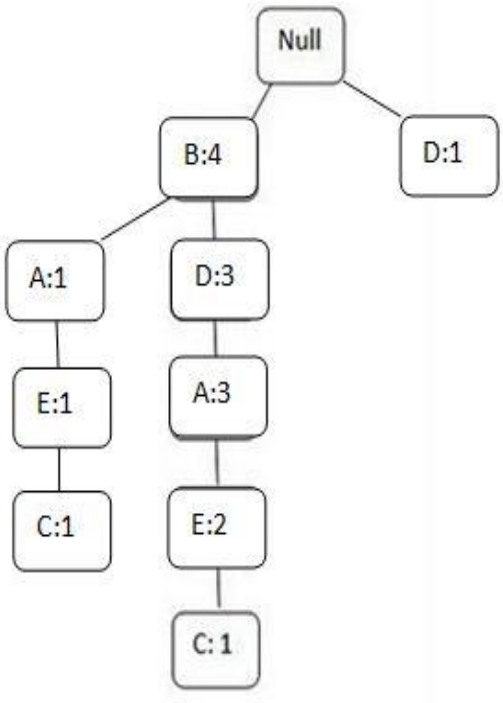
Row5

Only item D



Row6

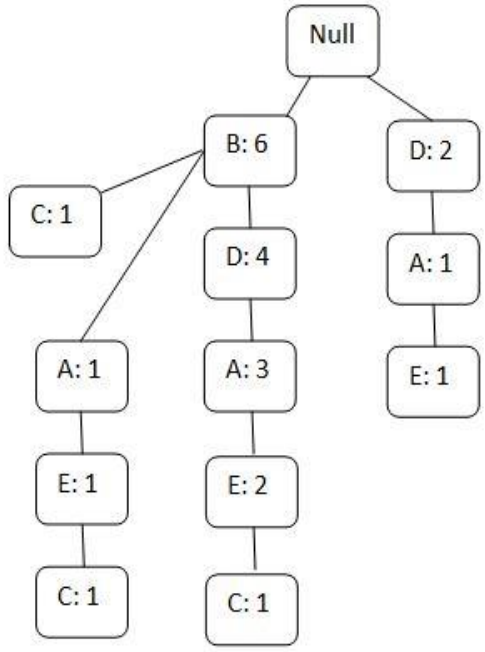
B and D appears in row 6. So just change the B:4 B:5 and D:3 to D:4.



Row7

Attach two new nodes A and E to the D node which hanging on the null node. Then mark D,A,E as D:2,A:1 and E:

ISSN No: - 2456 – 2165

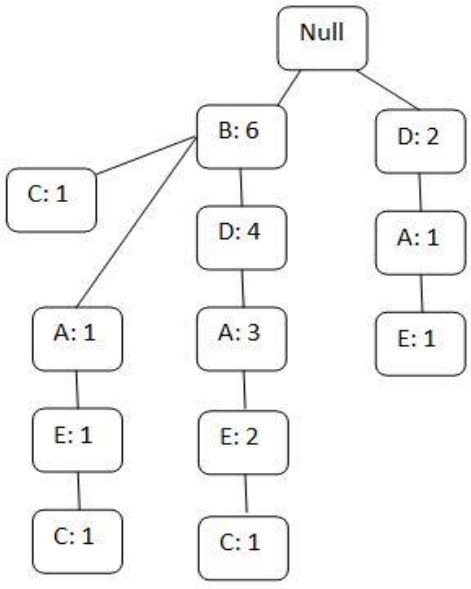


Row8

Attach new node C to B. Change the traverse times.(B:6,C:1)

V. VALIDATION

Item	Frequency
A	5
B	6
C	3
D	6
E	4



A. *Advantages*

- Only 2 passes over data-set.
- "Compresses" data-set.
- No candidate generation.
- Much faster than Apriori.

B. *Disadvantages*

- FP-Tree may not fit in memory.
- FP-Tree is expensive to build.

VI. CONCLUSION

In this paper, we presented a survey of the data mining applications, operations and techniques. We can also conclude the types of association rules and solutions for the association algorithms are Apriori algorithm, Éclat Algorithm, FP-Growth algorithm.Éclat algorithm is the fastest algorithm among three association rule mining algorithms based on execution time versus support and confidence. Éclat algorithm is the execution time decreases with increasing confidence and support.
The above algorithm can be used in other domains to bring out interestingness among the data present in the repository. Association rule produced by these three algorithms can be combined to form efficient algorithms for better results for any real life application algorithms can also combined to for an efficient algorithm.

REFERENCES

[1]. https://www.slideshare.net/akannshat/data-mining-15329899
[2]. https://www.academia.edu/5988472/Data_mining_in_production_planning_and_scheduling_A_review
[3]. https://www.ijircce.com/upload/2014/january/21_Data%20mining.pdf
[4]. https://www.ercim.eu/publication/ws-proceedings/12th-EDRG/EDRG12_Re.pdfhttps://computation.llnl.gov/projects/sapphire/overview.html
[5]. https://vlebb.leeds.ac.uk/bbcswebdav/orgs/SCH_Computing/MSCProj/reports/1011/de_marchi.pdf
[6]. https://www.researchgate.net/publication/281995140_Astrological_Prediction_for_Profession_Doctor_using_Classification_Techniques_of_Artificial_Intelligence
[7]. https://www.ischool.utexas.edu/~i385e/readings/Srivatta.pdf
[8]. https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/market_basket.htm#DMCON009
[9]. https://www.slideshare.net/dataminingcontent/mining-associations-and-correlations
[10]. https://link.springer.com/chapter/10.1007/3-540-48252-0_15
[11]. http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-8.html

[12]. http://d.researchbib.com/f/cnnJcwp21uYzAioF9jqJWfnJAuqTyioaZiMzIvpaIupaxlZQR0Y1LlFGVkAP5jMTL.pdf
[13]. http://www.ijcse.com/docs/IJCSE11-02-02-53.pdf
[14]. http://www.cc.gatech.edu/~isbell/reading/papers/berkhin02survey.pdf
[15]. https://www.users.cs.umn.edu/~kumar/dmbook/dmslides/chap6_basic_association_analysis.pdf.
[16]. https://www.cs.purdue.edu/homes/neville/papers/jensen-neville-interf2001.pdf
[17]. http://people.cs.pitt.edu/~chang/156/04reldb.html
[18]. http://www.public.iastate.edu/~maitra/stat501/lectures/MultivariateRegression.pdf
[19]. http://www.stat.colostate.edu/regression_book/chapter9.pdf
[20]. http://www.sci.utah.edu/~arpaiva/classes/UT_ece3530/hypothesis_testing.pdf
[21]. https://www.cse.unr.edu/~bebis/MathMethods/NNs/lecture.pdf
[22]. http://www.saedsayad.com/decision_tree.htm.
[23]. https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/
[24]. http://file.scirp.org/pdf/JSEA_2013042913162682.pdf
[25]. https://www.slideshare.net/salahecom/08-classbasic
[26]. http://wwwusers.cs.umn.edu/~han/dmclass/cluster_survey_10_02_00.pdf.
[27]. https://www.slideshare.net/zafarjcp/data-mining-association-rules-basics.
[28]. http://www.interlinepublishing.com/user-content-detail-view.php?cid=5672.
[29]. http://www.publishingindia.com/GetBrochure.aspx?query=UERGQnJvY2h1cmVzfC8xMjkwLnBkZnwvMTI5MC5wZGY=.
[30]. http://infolab.stanford.edu/~ullman/mmds/ch6.Pdf