2017 International Conference on Intelligent Computing and Control Systems (ICICCS 2017)

Madurai, India
15-16 June 2017

Pages 1-685

IEEE Catalog Number: CFP17K74-POD
ISBN: 978-1-5386-3901-6

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number: CFP17K74-POD
ISBN (Print-On-Demand): 978-1-5386-3901-6
ISBN (Online): 978-1-5386-2745-7

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

International Conference on Intelligent Computing and Control Systems (ICICCS 2017)

TABLE OF CONTENTS

Association Rule Mining with Regression for Optimization Approach

Moksha Shridhar
Department of Computer Science & Engineering
Madhav Institute of Technology & Science
Gwalior, India
Moksha.itm@gmail.com

Mahesh Parmar
Department of Computer Science & Engineering
Madhav Institute of Technology & Science
Gwalior, India
maheshparmarcse@gmail.com

Abstract— Association rule mining is an amazingly basic and critical piece of data mining. It will be utilize to the entrancing plans from exchange databases. Apriori count will be a champion among those intents and purposes built up computations from guaranted association rules, yet all it require the bottleneck previous adequacy. In this paper, we recommended a prefixed-itemset based data structure to create filtered itemset, with the help of structure we made sense of how to improve the viability of the conventional Apriori computation. Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. Association rule is one of the popular techniques used for mining data for pattern discovery is the KDD. Rule mining is an important component of data mining. To find regularities/patterns in data, the most effective class is association rule mining. Mining has been used in many application domains. In this paper, an efficient mining based algorithm for rule generation is presented. By using Apriori algorithm, we improve the precision and recall system.

Keywords— Apriori, Improved Apriori, Frequent itemset, Support, Candidate itemset, , Time consuming, Regression.

I INTRODUCTION

The role of data mining is easy and has been describe as "extracting knowledge from large amount of data".
Association rule mining is a ruling data mining technique. Association rule mining is a transform for finding acquaintance or relations between data items or attributes in large datasets. It permits mainstream designs also associations, correlations, or connections Around examples with determined negligible human effort, bringing paramount data of the surface to utilization. Association rule mining required demonstrated to be a great technique to weed out suitable data from huge datasets.
Different calculations or models were produced a number from claiming which have been connected on different requisition domains that incorporate telecommunication networks, market analysis, hazard official suite, account control and many others. The attainment of applying those concentrated rule for cracking real world issues is very frequently classified by the selection of rules.

However, the character of the concentrated guidelines need not drawn sufficient consideration. Measuring those rank of Association rules decides may be additionally troublesome. Furthermore current systems will make unsuitable, particularly when multi-level (rules whose things / to hail starting with first taxonomy level, yet the lay of decides compass more than you quit offering on that single taxonomy level) and cross level (rules whose things / topics turned starting with more than one scientific classification level) guidelines would involved. Regression analysis is a measurable strategy for deciding the relationship between the reliant factors and at least one autonomous variable. The subordinate factors is the one whose qualities you need to predict, whereas the free factors are the factors that you construct your forecast with respect to.
The regression utilizing known information positions like straight or calculated expect the future information arrangement will fall into the data structure. If then tries to foresee the incentive by applying some numerical calculation on the informational collections.

II ASSOCIATION RULE MINING

Association rule mining is an intriguing data mining method i.e. endure with figure out engaging patterns or association among the data put away in the database. Support and Confidence are two measures for the extract patterns. These are client enrich parameters and conflict from client to client. Association rule mining generally utilized within showcase data analysis or retail information analysis.
In market basket analysis, we discriminate non-identical elevate style of customer What's more analyze should be organized around the things of individuals would secured toward purchaser. Things that need habitually obtain helpfully eventually candidate set might be recognized. Association analysis is worn to help retailers to arrange diverse sorts for marketing, thing placement furthermore stock administration methodologies.
The point when we do Association rule mining in social database oversaw economy frameworks we change the database under (tid, item) format, the place tid remains for

transaction id furthermore items stands for different items purchased by the clients. There will a chance to be various sections for a specific transaction ID, since you quit offering on that one transaction id demonstrates buy for one specific client furthermore a client could buy Similarly as a number things Similarly as he have any desire. An association rules can look like this:

X(buys, Computer) X(buys, Window OS CD) [support=1% confidence=50%]where:

$$\text{Support} = \frac{\text{The number of transactions that contain computer and window OS CD}}{\text{The total number of transactions}}$$

$$\text{Confidence} = \frac{\text{The number of transactions that contain windows OS CD}}{\text{The number of transactions that contain computer}}$$

The above rule will retention with the help of support and confidence are equivalent to alternately more excellent over the client specified minimum support and confidence. The investigation of claiming association rules may be moving by all the more applications for example, such that telecommunication, banking, human services What's more manufacturing, and so forth throughout this way, observing and stock arrangement, etc.

III Related Work

In 1993 Agrawal, Imielinski, Swami [4] put ahead one step for man, which conduct a giant leap for computer science applications suggest an algorithm AIS forebear of the algorithms should begin those frequent itemsets & confident association rule. It holds two phases. The primary stage constitutes the provoke of the frequent itemsets need support in acutate the initially phase and in the next stage confident and frequent association rules are produced. In 1995 SETM (SET-oriented Mining of association rules) might have been convience by the passion to utilize SQL should figure large itemsets. It utilized best easy database primeval, viz.sorting and consolidate-scan join. It might have been easy, fast furthermore tough over the variety of framework use.

It demonstrated that exactly parts of facts mining could be a chance to be carried out towards utilizing general query languages, for example such as SQL, an opposed to creating specific black-box algorithms. Those set-oriented characteristic for claiming SETM eased the blooming of extensions Apriori. On 1994-95 those ignoring algorithms were improved by Agrawal et al by operate the monotonicity property of the support of itemsets and the confidence of association rules. In 1995 Park et al. [4] arranged an optimization, known as Direct Hashing and Pruning (DHP) pointed towards controlling those number from claiming candidate itemsets.

They provided for DHP algorithm to proficient large itemset generation. The proposed algorithm need two principle traits: one is proficient generation for large itemsets and other is agent reduction on transaction database span. DHP will be exceptionally skilful for the campanion of candidate set for large 2-itemsets, over requests of magnitude, lesser than that by past methods; it may be with the goal to operate the hash techniques thus resolving the operational bottleneck.

In 1996 Agrawal et al, recommended that the finest features of the Apriori & Apriori Tid algorithms could a chance to be joined together under a hybrid algorithm, known as Apriori Hybrid. Scale up tests indicated that Apriori Hybrid scrabbles linearly for the amount of transactions. Previously, the execution time fall a little as the number of Items in the database upon surge.

As those normal transaction measure upsurge (however manage those database size constant), the execution time upsurges singular gradually.

IV Apriori Algorithm

Apriori algorithm is frequently utilize algorithm for finding frequent item sets. This algorithm also determine association rules in the transactional database. It begins by identifying the single frequent item and then proceeds to combine the items to form larger item-sets as extensive item-sets exist in the database. Thus, it is called as Bottom Up approach. The frequent sets structured would expose those association rules starting with a large database. The principle of mining methodology is to recover from a dataset and then convert it into a form that is understandable and can be reused further. The guideline of apriori algorithm is the subsets of regular item-sets are frequent itemsets and the supersets of rare item-sets are infrequent item-sets. Apriori algorithm utilize level wise search item-sets for item k are used to extent item-sets of size k+1. Finding out those frequent item-sets fundamentally includes two steps:

- Join Operation: In sequence to frequent set in pass k signified by Lk, candidate set, signified by Ck, is formed by adhere Lk-1 with itself.

- Prune Operation: The figure dependent upon each subset of Ck is computed in sequence to find the frequent set since all the representative of Ck may not be frequent. Thus, all the members with count less than support value are removed. Rest of the members form the frequent set. Also if some subset of Ck of size k-1 is not present in Lk-1 the it's not a frequent candidate. Thus, it is removed from Ck.

Clampdown of Apriori Algorithm

Apriori algorithm experiences some shortcoming despite being clear and straightforward. The primary constraint is expensive squandering of time to hold countless laydown with much continuous itemsets, low least support or extensive itemsets.

For instance, though there are 104 from unremitting 1-itemsets, it must transform more than 107 hopeful under 2-length which along these lines will a chance to be attempted furthermore aggregate [2]. Besides, will perceive visit outline in size 100 e.g, v1, v2… v100, it needs to process 2100 hopeful itemsets [1] that yield on expensive and What's more misusing of duration of the time about candidate period. Through these lines, it will examine for certain sets from hopeful itemsets, moreover it will analyze database conventionally through and once more on discover candidate itemsets. Apriori will be low and more wastefulness when memory cutoff will be compelled for reaching the number of exchanges. In this paper, we recommended an algorithm to decrease the time spent for chasing the database exchanges from frequent itemsets.

V REGRESSION TECHNIQUE

In this methodology, the following definitions are needed:

Definition 1: Assume T={T1, T2, … , Tm},(m=1) is situated for transactions, Ti= {I1, I2, … , In},(n=1) may be those situated about items, Also k-itemset = {i1, i2, … , ik},(k=1) is likewise the situated for k items, and k-itemset?.

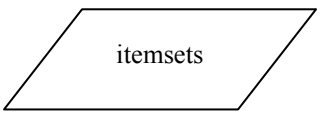
Definition 2: Suppose s (itemset), may be the support count of itemset or the recurrence of event from claiming an itemset in transactions.

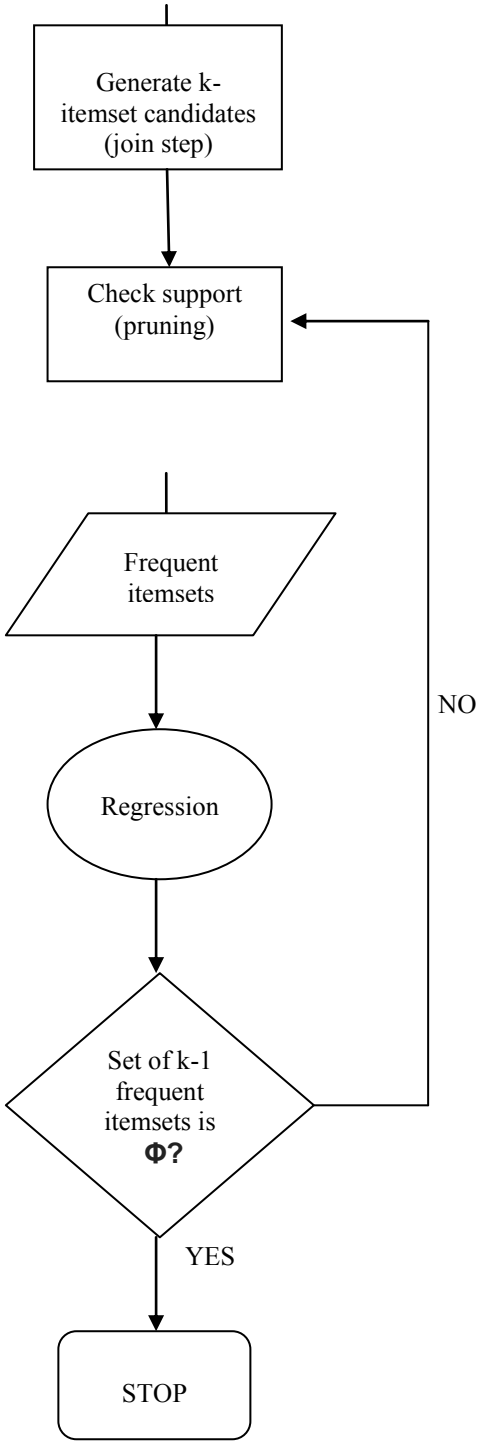
Definition 3: assume Ck is the candidate itemset about size k, and Lk is the frequent itemset for span k.

By performing these changes in Apriori algorithm, we achieve the regression which is described in below figure.

C_k denotes the set of candidate k-itemsets and F_k denotes the set of frequent k-itemsets

Step 1: Scan the database D, generating candidate 1- setC1;
Step 2: According to the min_sup, frequent item 1- set Lk is generated from the candidate 1- setCk;
Step-3: According to the min_sup, frequent item (k+1) – set Lk+1 is generated from the candidate (k+1) - set CK+1;
Step-4: Get frequent item sets Lk;
Step-5: Generate frequent itemset LK from candidate items CK.;
Step-6: Prune the results to find the frequent item sets using Linear Regression Lk from candidate (k+1) – set Ck+1 are generated.
Step-7: Goto Step-4 till outliers are removed.
Step-8: Generate strong association rules from frequent item sets. According to the min_sup, frequent item (k+1) - set Lk+1 is generated from the candidate (k+1) - set CK+1;
Step-9: A Rule which satisfy the min. support and min. confidence threshold.





Flowchart of Apriori algorithm with Regression

PSEUDO CODE

i= 1

$F_k = \{i|I \in I \wedge \sigma(\{i\}) \geq N \times \min sup\}$

{Find all frequent 1-itemsets}
Repeat

K = k+1

C_k = apriori-gen (F_{k-1})

{Generate candidate itemsets}
For each transaction t Є T do

C_t = subset (C_k, t)K

{identify all candidates that belong to t}
for each candidate itemset

C Є C_t do
σ (c) = σ (c) + 1

{increment support count}
end for

end for

F_k = {c|c Є C_k ∧ σ ({c}) ≥N x min sup}

{extract the frequent k – itemsets By Pruning Using Linear Regression }
Read Fk
sumx = Ø

sumxsq = Ø

sumy = Ø

sumxy = Ø

For each transaction do
Read Fk, Uk
Sumxq = sumxq +2

Sumy = sumy + U_k * sumxsq

Sumxy = sumxy + F*U end for

Denom = n * sumxsq – sumx *sumx

F_k Ø = (sumy * sumxsq – sumx * sumxy) / denom

U_k = (n * sumxy – sumx * sumy) / denom

write U_k, F_k Ø

until F_k = Ø

Result = U_kF_k

VI RESULTS

Those main test collates the time devour of claiming first Apriori, and our progressed calculation towards applying the five groups of transactions in the implementation. Those come about is indicated in Figures:

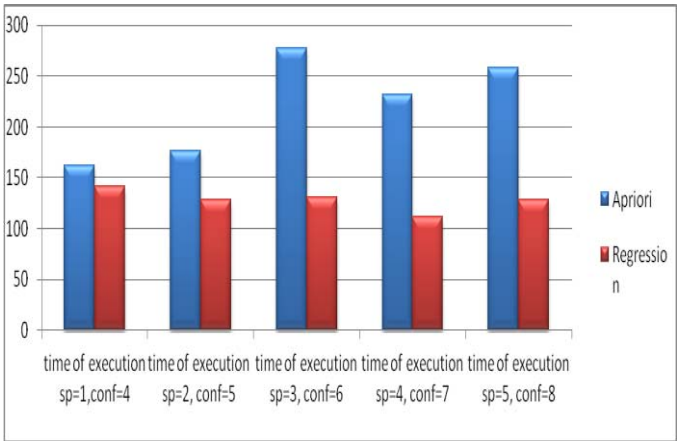


Fig. 1 Time execution with different support and confidence.

The above graph shows the difference of time in between Apriroi algorithm and proposed algorithm i.e. Apriori with Regression on different support and confidence.

Time of Execution (In msec)	
Apriori	278
Regression	141

Table 1 Comparision of Apriori and Regression on highest value of time.

Fig.2 Time optimization of Apriori with Regression

The above graph shows the difference of time in between the Apriori algorithm and implemented algorithm i.e. Apriroi with regression algorithm based on Table 1.

VII IMPLEMENTATION

SIMULATION OF REGRESSION IN DOT NET

Fig.3 Regression implemented for optimization

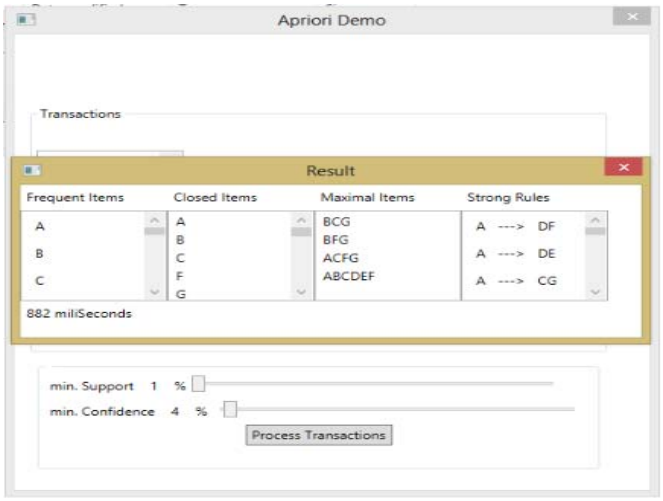


Fig.4 Snap Shot of Apriori with Regression in frame work.

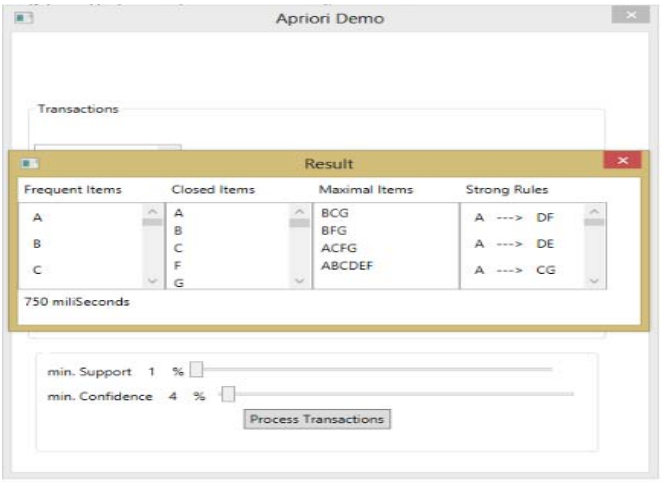


Fig. 5 Snap Shot of Apriori with regression in frame work.

VIII CONCLUSION

Data mining is also introduce as comprehension finding within the databases (KDD). It is a very crucial research area in today's time. It is one of the vital technique in facts mining is frequent pattern discovery. Finding co-occurrence a relationship between items is that the focus of this method.
The active analysis topic for KDD is association rules mining and many algorithms are developed on this. This algorithm is employed for locating associations within the item-sets. Efficiency has been an issue of concern for assorted years in mining association rules. Apriori is establish on the approach of finding helpful patterns from varied datasets. Although it's a conventional approach, it still has several shortcomings.

It also from the inadequacy of redundant examine of the database whereas glance for frequent item-sets as there's frequent generation of candidate item-sets that aren't needed. Conjointly there are sub item-sets generated which are redundant and algorithm involves repetitive looking out within the database. After implementing the developed approach get the conclusion that the modified Apriori algorithm is proposed an effective algorithm to diminish the consumtion of time.

The work is carried out on partitions of a dataset rather than applying on full dataset which results in reduction of time taken by the Apriori Algorithm. Instead of repeated scan of the original database, it is scanned only once to form large 1 item-set from which further computations are carried out. This reduces the time involved in scanning the dataset which in turn reduces the overall time to a greater extent. The minimum support value is also calculated at each pass which removes the unnecessary formed sets. Although the algorithm is simple, it carries out more effective pruning.

In this paper we have combined two association rule algorithms i.e. Apriori algorithm and regression to reduce the time. We have analyzed the frequent itemsets generation and number of cycle performed over the Apriori algorithm and Filter Associated in the context of association analysis.

IX FUTURE SCOPES

In this paper, we depicted the Apriori computation especially, and pointed out a couple of restrictions of the conventional Apriori estimation among the two phases of the count, to be particular the association. Besides, the paper cutting steps, and proposed the technique for prefixed-itemset-based data structure and the upgrades in light of it.

With those support from asserting prefixed-itemset-based data structure, we made sense of how to complete those interfacing step and the pruning data of the Apriori calculation substantially speedier, other than we can store the applicant itemsets with more minor lspace. Toward keep going, we focus on the capability of all inclusive Apriori algorithm. Furthermore update Apriori algorithm with respect to bolster check and the total number of itemset, and the test goes something like investigating the perspectives shown those credibility of the prefixed-itemset based computation.

X REFERENCES

[1] Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ and Steinberg D, Top 10 algorithms in data mining, Knowledge and Information Systems, vol. 14, no. 1, pp. 1–37, Dec. 2007.

[2] Wang Feng, Li Yong-hua, An Improved Apriori Algorithm Based on the Matrix, fbie, pp.152- 155, 2008 International Seminar on Future BioMedical Information Engineering, 2008.

[3] Agrawal R, Imieli ski T, and Swami A, "Mining association rules between sets of items in large databases," in Acm Sig Mod Record, vol. 22, pp. 207–216, 1993.

[4] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In Proceeding of the 20th International Conference on VLDB, pp. 478-499, 1994.

[5] M. S. Chen, J. Han, P.S. Yu, "Data mining: an overview from a database perspective", IEEE Transactions on Knowledge and Data Engineering, vol. 8, no.6, pp. 866– 883, 1996.

[6] Irena Tudor, Universitatea Petrol-Gaze din ploeiesti, (2008)"Association Rule Mining as a Data Mining Technique", Bd.Bucuresti 39, ploeiesti,Catedra de Informatica, Vol-LX,No.1.

[7] Goswami D.N., Chaturvedi Anshu and Raghuvanshi C.S. (2010), "An Algorithm for Frequent Pattern Mining Based On Apriori", International Journal on Computer Science and Engineering, Vol. 02, No. 04, 942-947 [22]Jyothi Pillai (2011) , "User centric approach to itemset utility mining in Market Basket Analysis" , International Journal on Computer Science and Engineering, Vol. 3 No. 1.

[8] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," Knowledge and Information Systems, vol. 14, no. 1, pp. 1–37, Dec. 2007.

[9] VLDB Journal2007, pp: 507-521, 2007. [12]. Li N., Zeng L., He Q. & Shi Z. Parallel Implementation of Apriori Algorithm Based on MapReduce. In Proc. of the 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD '12), Kyoto, IEEE: 236 – 241, 2012. [13].

[10] Lin M., Lee P. & Hsueh S. Apriori-based Frequent Itemset Mining Algorithms on MapReduce. In Proc. of the 16th International Conference on Ubiquitous Information Management and Communication (ICUIMC '12), New York, NY, USA, ACM: Article No. 76, 2012.

[11] Klosgen W and Zytkow J M (eds.), Handbook of data mining and knowledge discovery, OUP, Oxford, 2002.

[12] Provost, F., & Fawcett, T., Robust Classification for Imprecise Environments. Machine Learning, Vol. 42, No.3, pp.203-231, 2001.

[13] Scheffer T. Finding Association Rules That Trade Support Optimally against Confidence. Unpublished manuscript.

[14] Larose D T, Discovering knowledge in data: an introduction to data mining, John Wiley, New York, 2005. [4].

[15] Kantardzic M, Data mining: concepts, models, methods, and algorithms, John Wiley, New Jersey, 2003.

[16] Sunita B. Aher and Lobo L.M.R.J., "Combination of Clustering, Classification and Association Rule Based Approach for Course Recommender System in E-learning" In International Journal of Computer Applications(0975- 8887), Volume 39- No. 7, February 2012.

[17] Tunedit Machine Learning Repository, Available http://tunedit.org/repo/UCI/spect_test.arff

[18] M.Shridhar, M.Parmar: Survey Paper on Association Rule Mining and its Approaches. International Journal of Computer Science and Engineering (IJCSE), March 2017, Volume-5, Issue-3, pp.-129-135, E-1SSN:2347-2693.

[19] Kritika Yadav, Mahesh Parmar: Review Paper on Data Mining and its Techniques of Mahatma Gandhi National Rural Employment Guarantee Act. International Journal of Computer Science and Engineering (IJCSE), April 2017, Volume-5, Issue-4, pp. 68-73, E-ISDN: 2347-2693.

[20] Prateeksha Tomar, Amit Kumar Manjhvar: Survey Report on Various Decision Tree Classification Algorithm Weka Tool. International Journal of Computer Science and Engineering (IJCSE), March 2017.