Combination of Clustering, Classification and Association Rule based Approach for Course Recommender System in E-learning

Sunita B. Aher
M.E. (CSE) -II
Walchand Institute of Technology,
Solapur University India

Lobo L.M.R.J.
Associate Professor, Head, Department of IT
Walchand Institute of Technology,
Solapur University India

ABSTRACT
Data mining also known as Knowledge Discovery in Database is the process of discovering new pattern from large data set. E-learning is the electronically learning & teaching process. Course Recommender System allows us to study the behavior of student regarding the courses. In Course Recommender System in E-learning, we collect the data regarding the student enrollments for a specific set of data i.e. the courses which the students like to learn. After collection of data, we apply three data mining techniques namely clustering, classification & association rule to find the best combination of courses. Here we compare the result of this combined approach with result obtained using only association rule & present how this combined approach is better than only the association rule algorithm.

KEY WORDS
Weka, Moodle, Simple K-means Algorithm, ADTree Classification Algorithm, Apriori Association Rule Algorithm

1. INTRODUCTION
The course recommendation system in e-learning is a system that suggests the best combination of courses in which the students are interested [10].

In this Course Recommendation System, we have considered the 13 course category. Under each category there will courses. So there are about 82 courses. Student first logs in the learning management system e.g. Moodle & enrolled for those courses in which they are interested. The activity chart for student is shown in figure 2. This data is stored in the moodle database which we use to find out the best combination. After collecting the data from student which is stored in Moodle database, the next stage is to gather & prepare the data. In this step, first we select the data from database which is relevant. To test the result using Weka i.e. the best combination courses, first we need to preprocess the data & find out the result. The step-build the model, we directly select the relevant data from Moodle database. After collecting the data from Moodle database, we clustered the data using clustering algorithm e.g. Simple K-means algorithm. After clustering data, we classify that data using ADTree algorithm. We apply the Apriori Association Rule algorithm on classified data to find the best combination of courses. To find the result using only Apriori association rule algorithm, we need to preprocess the data from Moodle database but if we consider the combined approach then there is no need to preprocess the data. The preprocessing technique is explained in paper [8].

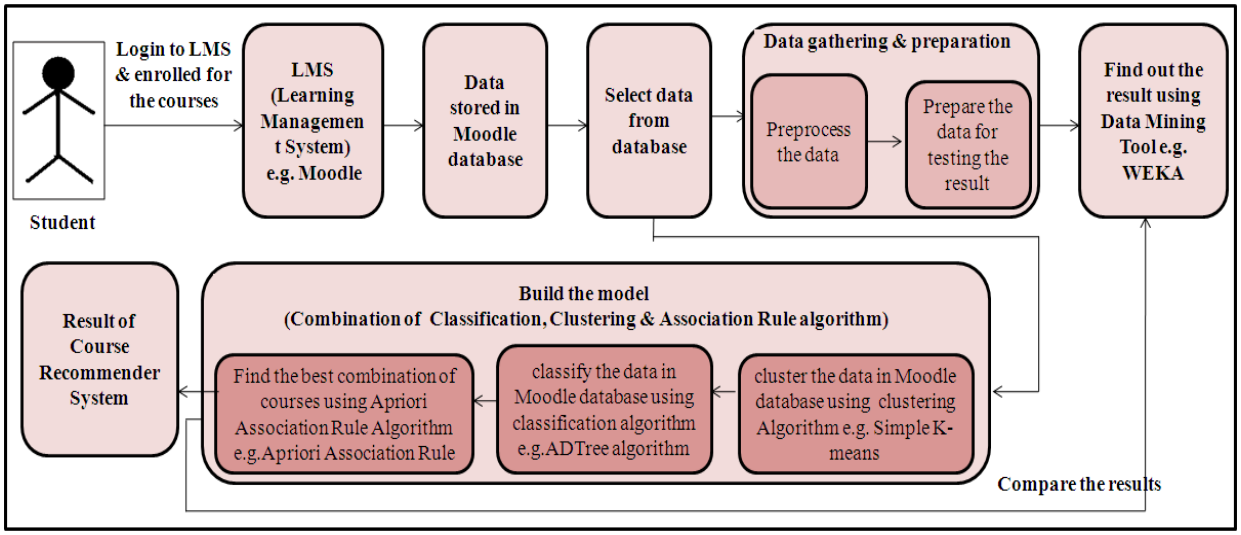

Figure 1: Architecture for recommendation of courses in E-learning System

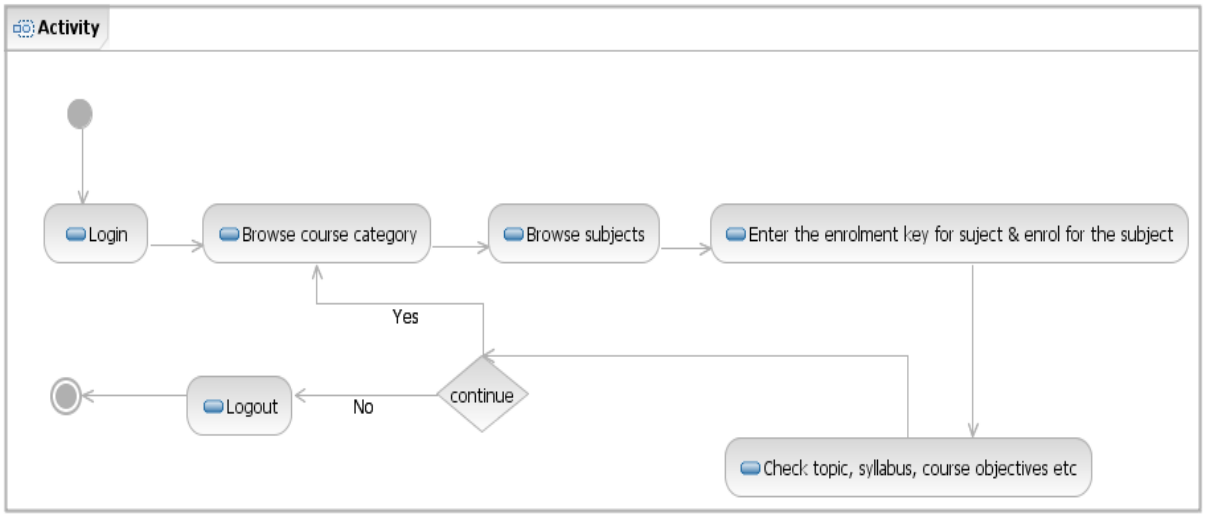

Figure 2: Activity chart for student in Course Recommendation System

2. LITERATURE REVIEW

The research [1] aim is to extract the significant prevention factors for particular types of cancer. They used three association rule mining algorithms, Apriori, Predictive Apriori and Tertius algorithms in order to discover most of the significant prevention factors against these specific types of cancer.

In paper [2], they focused on the cases where the items of a large domain correlate with each other in a way that small worlds are formed, that is, the domain is clustered into groups with a large number of intra-group and a small number of inter-group correlations. This property appears in several real-world cases, e.g., in bioinformatics, e-commerce applications, and bibliographic analysis, and can help to significantly prune the search space so as to perform efficient association-rule mining. They developed an algorithm that partitions the domain of items according to their correlations and they describe a mining algorithm that carefully combines partitions to improve the efficiency. Their experiments show the superiority of the proposed method against existing algorithms, and that it overcomes the problems (e.g., increase in CPU cost and possible I/O thrashing) caused by existing algorithms due to the combination of a large domain and a large number of records.

In paper [3], they propose a method for grouping and summarizing large sets of association rules according to the items contained in each rule. They used hierarchical clustering to partition the initial rule set into thematically coherent subsets. This enables the summarization of the rule set by adequately choosing a representative rule for each subset, and helps in the interactive exploration of the rule model by the user. Rule clusters can also be used to infer novel interest measures for the rules. Such measures are based on the lexicon of the rules and are complementary to measures based on statistical properties, such as confidence, lift and conviction.

In paper [4], they proposed a system that integrates Web page clustering into log file association mining and uses the cluster labels as Web page content indicators. It is demonstrated that novel and interesting association rules can be mined from the combined data source. The rules can be used further in various applications, including Web user profiling and Web site construction. We experiment with several approaches to content clustering, relying on keyword and character n-gram based clustering with different distance measures and parameter settings. Evaluation shows that character n-gram based clustering performs better than word-based clustering in terms of an internal quality measure (about 3 times better). On the other hand, word-based cluster profiles are easier to manually summarize. Furthermore, it is demonstrated that high-quality rules are extracted from the combined dataset.

The goal of research [5] is to experimentally evaluate association rule mining approaches in the context of XML databases. Algorithms are implemented using Java. For experimental evaluation different XML datasets are used. Apriori and FP Tree algorithm have been implemented and their performance is evaluated extensively.

In paper [6], they proposed a approach called Associative classification which is a classification of a new tuple using association rules. It is a combination of association rule mining and classification. The accuracy can be achieved by producing all types of negative class association rules.

3. DATA MINING ALGORITHMS

Here we consider brief idea about each data mining algorithm.

3.1 Simple K-means Clustering Algorithm

Simple K-means algorithm is a type of unsupervised algorithm. In this algorithm items are moved among the set of cluster until required set is reached. This algorithm is used to classify the data set, provided the number of cluster is given in prior. This algorithm is iterative in nature.

The flowchart for Simple K-means algorithm is shown in figure 2.

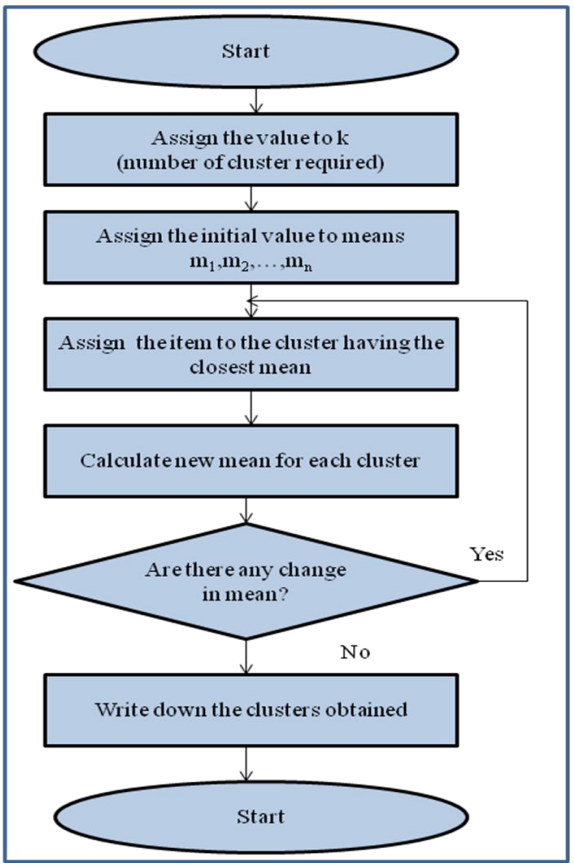

Figure 3: Flowchart for Simple K-means Algorithm

3.2 ADTree Classification Algorithms

An alternating decision tree (ADTree) is a machine learning method for classification which generalizes decision trees.

An alternating decision tree consists of two nodes:

- ➔ Decision nodes: specify a predicate condition.
- ➔ Prediction nodes: contain a single number. ADTree always have prediction nodes as both root and leaves.

An instance is classified by an ADTree by following all paths for which all decision nodes are true and summing any prediction nodes that are traversed.

The fundamental element of the ADTree algorithm is the rule as it comes under the tree category. A single rule consists of

- Condition: is a predicate of the form "attribute <comparison> value.
- Precondition: is simply a logical conjunction of conditions and Two scores.

A precondition Evaluation of a rule involves a pair of nested if statements [11]:

```
if(precondition)
    if(condition)
        return score_one
    else
        return score_two
    end if
else
    return 0
end if
```

3.3 Apriori Association Rule Algorithms

Apriori is designed to operate on databases containing transactions. The algorithm attempts to find subsets which are common to at least a minimum number C (the cutoff, or confidence threshold) of the itemsets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time. This step is known as *candidate generation*, and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a hash tree structure to count candidate item sets efficiently. The Apriori Association Rule algorithm is given in as [9]:

Input : Database of Transactions D={$t_1, t_2, \ldots, t_n$}
Set if Items I= {$I_1, I_2, \ldots I_k$}
Frequent (Large) Itemset L
Support,
Confidence.

Output : Association Rule satisfying Support & Confidence

Method :

1. C_1 = Itemsets of size one in I;
2. Determine all large itemsets of size 1, L_1;
3. i = 1;
4. Repeat
5. i = i + 1;
6. C_i = Apriori-Gen(L_{i-1});
7. Apriori-Gen(L_{i-1})
    1. Generate candidates of size i+1 from large itemsets of size i.
    2. Join large itemsets of size i if they agree on i-1.
    3. Prune candidates who have subsets that are not large.
8. Count C_i to determine L_i;
9. until no more large itemsets found;

4. RESULT & IMPLEMENTATION

Here we are considering the sample data extracted from Moodle database as shown in Table 1. Here we consider 45 student & 15 courses. We consider the courses like C-programming (C), Visual Basic (VB), Active Server Pages (ASP), Computer Network (CN), Network Engineering (NE), Microprocessor (MP), Computer Organization (CO), Database Engineering (DBE), Advanced Database System (ADS), Operating System (OS), Distributed System (DS), Finite Automata System (FSA), Data Structure (DS-I), Software Engineering (SE), and Software Testing & Quality assurance (STQA). In this table yes represent that the student is interested in that course & no represent that student do not like that course . In preprocessing step, we delete those rows & columns from sample table shown in table 1, having very less student count & less course count. After preprocessing of data we got 8 courses & 38 rows i.e. 38 students. These 8 courses are C-programming (C), Visual Basic (VB), Active Server Pages (ASP), Computer Network (CN), Network Engineering (NE), Operating System (OS), Distributed System (DS), Data Structure (DS-I) [12]. The graph for sample data before preprocessing & after preprocessing is shown in figure 4 & 5 respectively. In these graph the X-axis represents the courses & Y-axis represents the course count.

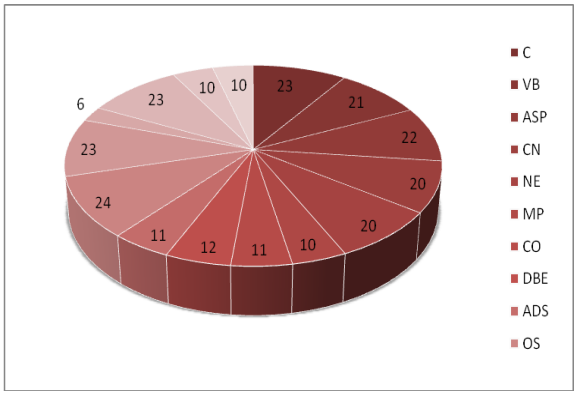

Figure 4: Graph for table 1 Before preprocessing

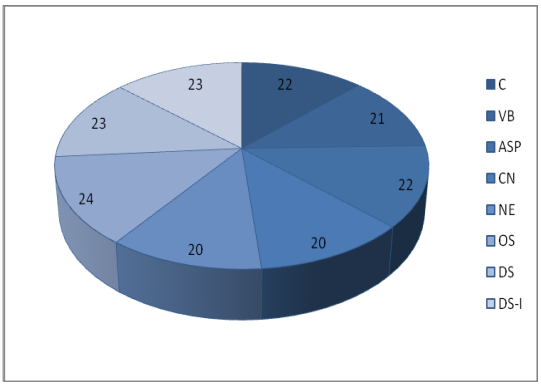

Figure 5: Graph for table 1 after preprocessing

Table 1: Sample table from Moodle Database [10]

Courses → Roll_No ↓	C	VB	ASP	CN	NE	MP	CO	DBE	ADS	OS	DS	FSA	DS-I	SE	STQA
1	yes	yes	yes	yes	yes	no	no	no	no	no	no	no	yes	no	no
2	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no
3	yes	yes	yes	yes	yes	no	no	no	no	yes	yes	yes	yes	yes	yes
4	no	no	no	yes	yes	no	yes	no	no	no	no	no	no	no	no
5	yes	yes	yes	yes	yes	no	no	yes	no	yes	yes	no	yes	no	no
6	yes	yes	yes	no	no	no	no	no	no	yes	no	no	yes	no	no
7	no	no	no	yes	yes	yes	yes	no	no	no	no	no	no	yes	no
8	no	no	no	no	no	no	no	yes	yes	yes	yes	no	yes	no	no
9	no	no	no	yes	yes	yes	yes	no	no	no	no	yes	no	no	no
10	yes	no	no	no	no	no	no	no	no	no	no	no	no	no	no
11	yes	yes	yes	no	no	no	no	no	no	yes	yes	no	yes	no	no
12	yes	yes	yes	yes	yes	no	no	no	no	no	no	no	no	no	no
13	no	no	no	no	no	no	no	yes	yes	yes	yes	no	yes	yes	yes
14	yes	yes	yes	yes	yes	no	no	no	no	yes	yes	no	no	no	no
15	yes	yes	yes	no	no	no	no	no	no	no	no	no	yes	no	no
16	no	no	no	yes	yes	no	no	yes	yes	yes	yes	no	yes	no	no
17	yes	yes	yes	no	no	no	no	no	no	yes	yes	no	yes	yes	yes
18	yes	yes	yes	no	no	no	no	no	no	no	no	no	no	no	no
19	no	no	no	yes	yes	yes	yes	yes	yes	no	no	no	no	no	no
20	yes	no	no	no	no	no	no	no	no	yes	yes	no	yes	yes	yes
21	yes	no	yes	no	no	yes	yes	no	no	yes	yes	yes	no	no	no
22	no	no	no	no	no	no	no	yes	yes	yes	yes	no	yes	no	no
23	yes	yes	yes	yes	yes	yes	yes	no	no	yes	yes	no	yes	no	no
24	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
25	no	yes	yes	no	no	yes	yes	yes	yes	yes	yes	no	no	no	no
26	yes	yes	yes	no	no	no	no	no	no	yes	yes	no	yes	no	no
27	yes	yes	yes	yes	yes	no	no	no	no	no	no	no	no	no	no
28	no	no	no	yes	yes	no	no	no	no	yes	yes	no	yes	no	no
29	no	no	no	no	no	yes	yes	yes	yes	no	no	no	no	no	no
30	yes	yes	yes	yes	yes	no	no	no	no	no	no	no	no	yes	yes
31	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no
32	yes	yes	yes	no	no	no	no	yes	yes	yes	yes	no	yes	no	no
33	no	no	no	yes	yes	no	no	no	no	yes	yes	no	yes	no	no
34	yes	yes	yes	no	no	no	no	no	no	no	no	no	no	no	no
35	no	no	no	no	no	no	no	no	no	yes	yes	no	no	no	no
36	no	no	no	yes	yes	no	no	no	no	no	no	no	yes	no	no
37	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	no	no	no	no	no
38	no	no	no	no	no	no	no	no	no	yes	yes	yes	yes	yes	yes
39	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes

40	no	no	no	no	no	no	no	no	no	no	no	no	no	yes	yes
41	yes	yes	yes	no	no	no	no	no	no	yes	yes	no	yes	no	no
42	no	no	no	yes	yes	no	no	no	no	no	no	no	no	no	no
43	no	no	no	no	no	no	no	no	no	yes	yes	no	yes	no	no
44	no	no	no	no	no	no	no	no	no	no	no	no	no	no	yes
45	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no

Table 2: After application of clustering algorithm to table1

Courses → Roll_No | v	C	VB	ASP	CN	NE	MP	CO	DBE	ADS	OS	DS	FSA	DS-I	SE	STQA
1	yes	yes	yes	yes	yes	no	no	no	no	no	no	no	yes	no	no
3	yes	yes	yes	yes	yes	no	no	no	no	yes	yes	yes	yes	yes	yes
5	yes	yes	yes	yes	yes	no	no	yes	no	yes	yes	no	yes	no	no
6	yes	yes	yes	no	no	no	no	no	no	yes	no	no	yes	no	no
11	yes	yes	yes	no	no	no	no	no	no	yes	yes	no	yes	no	no
12	yes	yes	yes	yes	yes	no	no	no	no	no	no	no	no	no	no
14	yes	yes	yes	yes	yes	no	no	no	no	yes	yes	no	no	no	no
16	no	no	no	yes	yes	no	no	yes	yes	yes	yes	no	yes	no	no
17	yes	yes	yes	no	no	no	no	no	no	yes	yes	no	yes	yes	yes
23	yes	yes	yes	yes	yes	yes	yes	no	no	yes	yes	no	yes	no	no
24	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
26	yes	yes	yes	no	no	no	no	no	no	yes	yes	no	yes	no	no
27	yes	yes	yes	yes	yes	no	no	no	no	no	no	no	no	no	no
28	no	no	no	yes	yes	no	no	no	no	yes	yes	no	yes	no	no
30	yes	yes	yes	yes	yes	no	no	no	no	no	no	no	no	yes	yes
32	yes	yes	yes	no	no	no	no	yes	yes	yes	yes	no	yes	no	no
33	no	no	no	yes	yes	no	no	no	no	yes	yes	no	yes	no	no
37	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	no	no	no	no	no
39	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
41	yes	yes	yes	no	no	no	no	no	no	yes	yes	no	yes	no	no

Table 3: After application of classification algorithm to table 2

Courses → Roll_No | v	C	VB	ASP	CN	NE	MP	CO	DBE	ADS	OS	DS	FSA	DS-I	SE	STQA
3	yes	yes	yes	yes	yes	no	no	no	no	yes	yes	yes	yes	yes	yes
5	yes	yes	yes	yes	yes	no	no	yes	no	yes	yes	no	yes	no	no
23	yes	yes	yes	yes	yes	yes	yes	no	no	yes	yes	no	yes	no	no
24	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
39	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes

The result of applying Apriori association rule before & after preprocessing of data is shown in first & second row of table 4. Before preprocessing of data, we got the association rule containing "no" only. As we are recommending the course, we preprocess the data. The result after preprocessing of data is shown in second row of table 4. Now the association rule contains only "yes". The meaning of the association rule "DS=yes -> OS=yes" is that we can recommend to new student who has recently enrolled for DS course, the operating system as a course to be opted.

If we consider combination of, clustering, classification & association rule then there is no need to preprocess the data. First we apply the Simple K-means clustering algorithm to data selected from Moodle database. We consider two cluster out of which the cluster 1 gives the correct result & cluster 0, the incorrect result. After clustering of data, we got the table 2. We apply the ADTree classification algorithm on correct cluster i.e. cluster 1 & we got the table which is shown in table 3. Last step is to apply the association rule to this classified data. The result of application of Apriori association rule to this clustered & classified data is shown in fourth row of table 4.

Table 4: Result after application of machine learning algorithms

Course considered	Parameter Considered	Results
Result of Apriori Association Rule before preprocessing & application of combination of Clustering & Association Rule		
C, VB, ASP, CN, NE, MP, CO, DBE, ADS, OS, DS, FSA, DS-I, SE, STQA	Minimum support: 0.7 Minimum metric <confidence>: 0.9	Best rules found: 1. CO=no → MP=no 2.DBE=no → ADS=no 3.CO=no FSA=no → MP=no 4.MP=no → CO=no 5.STQA=no → SE=no 6.SE=no → STQA=no 7.ADS=no → DBE=no 8.MP=no FSA=no →CO=no 9.FSA=no STQA=no → SE=no 10. FSA=no SE=no → STQA=no
Result of Apriori Association Rule after preprocessing & before application of combination of Clustering & Association Rule		
C, VB, ASP, CN, NE, OS, DS, DS-I	Minimum support: 0.5 Minimum metric <confidence>: 0.9	Best rules found: 1. DS=yes →OS=yes 2. VB=yes →ASP=yes 3. NE=yes → CN=yes 4. CN=yes → NE=yes 5. C=yes VB=yes → ASP=yes 6. DS=yes DS-I=yes → OS=yes 7. OS=yes →DS=yes 8. ASP=yes →C=yes 9. C=yes →ASP=yes 10.ASP=yes →VB=yes
	Minimum support: 0.6 Minimum metric <confidence>: 0.9	Best rules found: 1. DS=yes →OS=yes 2.OS=yes →DS=yes
Result After Application of Clustering algorithm-Simple K-means		
C, VB, ASP, CN, NE, MP, CO, DBE, ADS, OS, DS, FSA, DS-I, SE, STQA	Number of Cluster:2 Seed: 10	Cluster 0 Mean/Mode: no no no no no no no no no no no no no no no Std Devs: N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A Cluster 1 Mean/Mode: yes yes yes yes yes no no no no yes yes no yes no no Std Devs: N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A Clustered Instances 0 25 (56%) 1 20 (44%) 2
After Application of Clustering algorithm-Simple K-means , Classification Algorithm- ADTree & Association Rule-Apriori Association Rule		
C, VB, ASP, CN, NE, MP, CO, DBE, ADS, OS, DS, FSA, DS-I, SE, STQA	Minimum support: 0.95 Minimum metric <confidence>: 0.9	Best rules found: 1. VB=yes → C=yes 2. C=yes → VB=yes 3. ASP=yes → C=yes 4. C=yes → ASP=yes 5. CN=yes → C=yes 6. C=yes → CN=yes 7. NE=yes → C=yes 8. C=yes → NE=yes 9. OS=yes → C=yes 10. C=yes → OS=yes

When we apply the Simple K-means Clustering algorithm to sample table 1 we got two cluster out of which cluster 1 is the correct cluster. The graph for courses after application of Simple K-means Clustering algorithm on on cluster 1 & cluster 0 is shown in figure 6 & 7 respectively. After application of classification algorithm on clustered data we obtain the graph which is shown in figure 8.

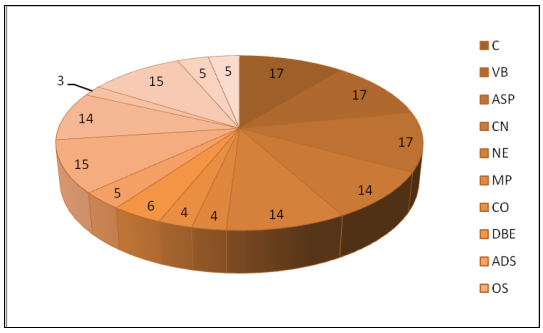

Figure 6:Graph after application of Simple K-means clustering algorithm on Table 1 (correct result using cluster 1)

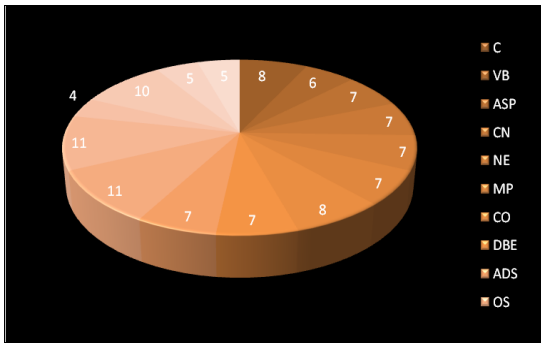

Figure 7:Graph after application of Simple K-means clustering algorithm on Table 1

(incorrect result using cluster 0)

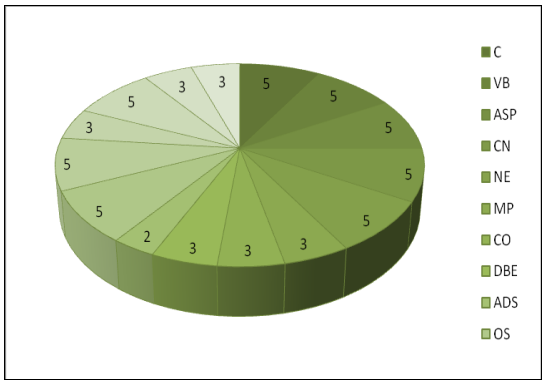

Figure 8:Graph after application of ADtree classification algorithm on clustered data on Table 2

The overall graph is shown in figure 7.The line ST, SKM1 Correct, SKM0 & SKM1ADT in graph represents courses considered in sample table, after application of Simple K-means clustering algorithm (Cluster 1-correct cluster), after application of Simple K-means clustering algorithm (cluster 0-incorrect cluster) & after application of ADTree classification algorithm on clustered data respectively.

5. CONCLUSION AND FUTURE WORK

In this paper, we consider three data mining technique i.e. Simple K-means clustering algorithm, ADTree classification algorithm & Apriori Association rule algorithm to recommend the course to the student. We compare the result of this combined approach with the result of Apriori association rule algorithm. We found that this combined approach is better than only Apriori association rule algorithm as there is no need to preprocess the data. This combined approach increases the strength of the association rule. Future work includes the atomization of this combined approach & to test result on huge amount of data.

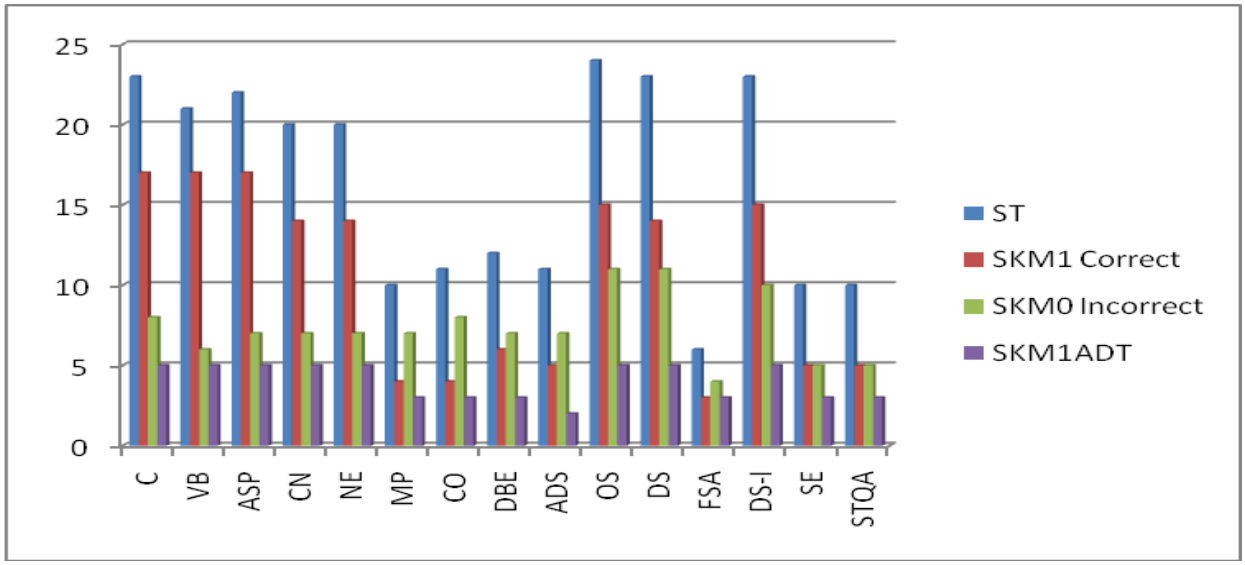

Figure 9: Graph for courses after application of various data mining algorithms

6. REFERENCES

[1] Nahar J, Tickle KS, Ali AB, Chen YP: "Significant cancer prevention factor extraction: an association rule discovery approach".in Journal of Medical Systems Volume 35, Number 3, 353-367, DOI: 10.1007/s10916-009-9372-8

[2] Alexandros Nanopoulos, Apostolos N. Papadopoulos, Yannis Manolopoulos:" Mining association rules in very large clustered domains" Information Systems 32 (2007) 649–669

[3] Al¶³pio Jorge:"Hierarchical Clustering for thematic browsing and summarization of large sets of Association Rules" Supported by the POSI/SRI/39630/2001/Class Project

[4] Jiayun Guo, Vlado Keˇselj, and Qigang Gao:"Integrating Web Content Clustering into Web Log Association Rule Mining?" supported by NSERC

[5] Gurpreet Kaur, Naveen Aggarwal:" Association Rule Mining in XML databases: Performance Evaluation and Analysis" IJCST Vol. 1, Issue 2, December 2010 ISSN :2229-4333(Print) |ISSN : 0976 - 8491 (Online)

[6] B. Ramasubbareddy, A. Govardhan & A. Ramamohanreddy: "Classification Based on Positive and Negative Association Rules" International Journal of Data Engineering, (IJDE), Volume (2): Issue (2) : 2011 84

[7] Sunita B Aher and Lobo L.M.R.J.. Data Mining in Educational System using WEKA. *IJCA Proceedings on International Conference on Emerging Technology Trends (ICETT)* (3):20-25, 2011. Published by Foundation of Computer Science, New York, USA (ISBN: 978-93-80864-71-13)

[8] Sunita B Aher and Lobo L.M.R.J.: "Preprocessing Technique for Association Rule Based Course Recommendation System in E-learning" selected in ICECT-12, proceeding published by IEEE

[9] "Data Mining Introductory and Advanced Topics" by Margaret H. Dunham

[10] Sunita B Aher and Lobo L.M.R.J. Article: A Framework for Recommendation of courses in E-learning System. *International Journal of Computer Applications* 35(4):21-28, December 2011. Published by Foundation of Computer Science, New York, USA ISSN 0975 – 8887

[11] Alternating decision tree, available at: http://en.wikipedia.org/wiki/alternating_decision_tree Accessed on 13-02-2012

[12] Sunita B Aher and Lobo L.M.R.J.:" Mining Association Rule in Classified Data for Course Recommender System in E-Learning" Selected in *International Journal of Computer Applications ((IJCA))* Published by Foundation of Computer Science, New York, USA ISSN 0975 – 8887