International Journal of Soft Computing 9 (2): 103-108, 2014
ISSN: 1816-9503

OLS-Association Rule for Optimal Learning Sequence Using K-means in Educational Data Mining

[1]Muruganathan Velayutham and [2]B.L. ShivaKumar
[1]Research and Development Centre, Bharathiar University, Coimbatore, India
[2]Department of Computer Applications, Sri Ramakrishna Engineering College, Coimbatore, India

Abstract: Education data mining is one of the new emerging research areas in intra data mining domain. The main objective of applying data mining to educational data is to analyse educational data contents, models to summarize/analyse the learner's discussions, etc. Education data mining concentrates on the computing process models which focus on education context. Researchers proposed a new approach in deriving new association rules for optimal learning sequence of students and tutors using K-means Clustering algorithm; here data's are visualized and processed. The methodology increases the performance with the fast support calculation and other significant techniques are introduced to improve the efficiency of the association rule based mining process using K-means. The new approach is compared with Apriori algorithm and the comparison results presented here shows the algorithm is optimal than the traditional Apriori algorithm.

Key words: Educational data mining, K-means, learning, sequence, optimal

INTRODUCTION

Educational data mining fully focus on educational context and to resolve the context with the research issues. Now a days education relational databases are developed in large numbers, various data stores to store student data's in repositories about that how student learn effectively. Education system through online learning (e-Learning) has reached the market level at high instances. Even though the smart pervasive devices focus towards the educational content which was already pre- fetched in the device is for sale in the market (for example, HCL U1 Tablet with IIT JEE 2013, Penta 703C with IIT JEE 2013, Karbon smart Tabwith IIT JEE). By means of various research methods EDM seeks the traditional methods to make use of large data base in learning and learning methods. Various countries follow the e-Learning and field learning methods for students. Basic educational system follows various researches in learning methods which are short listed as:

Traditional teaching methods: Traditional education tries to communicate knowledge and skills based on person to person contact and also study internally on how humans learn (Chen *et al*., 2012).

e-Learning: e-Learning Method provides online learning, training, instructions to the students. Learning is revised through visual communication, collaboration, report generation, etc. Data sets are mined in web to maintain the log's and databases for student data and learning resources (Chen *et al*., 2012).

Online tutoring system: Online tutoring system is widely followed by various universities all over the global. Approach fully focus on web based where all the study materials are hosted in the web, student can avail and make use of it (Chen *et al*., 2012).

Intelligence tutoring system: Intelligence tutoring system is an alternative approach to web based learning, data mining technique is used to maintain the data, logs and databases.

The EDM process converts raw data coming from educational systems into useful information which have a great influence on educational research and training. It follows the same process of traditional data mining technique such as pre-processing data and post processing data in DM. EDM allows to discover new knowledge on particular domain and its domain constraints based on student's active presentation in the domain (Chen *et al*., 2012; Carmona *et al*., 2010; Brtka *et al*., 2012; Parack *et al*., 2012; North *et al*., 2007) (Table 1).

Corresponding Author: Muruganathan Velayutham, Research and Development Centre, Bharathiar University, Coimbatore, India

Table 1: EDM participants adopted from EDM review

Users	Objective for using EDM
Learners	To generalize the web based learning methods, e-Learning Methods, etc. To recommend necessary activities to the learners and researchers. To provide ultimate resource to all the learners, to provide hints and t develop learning path
Teachers	To get necessary statement from the learners, getting back the feedback about the instructions. To analysis the student's performance, student support and to classify the student and group them to find the frequent mistake and analysis
Researchers	To improve student performance, educational quality measures, to evaluate the course content, to evaluate student activity, to specify the particulars and data, to design the data model
Administrators	To develop the organization in the best way and to utilize the resources. To develop the organization by enhancing the research motto's and learning bench marks
Universities/Colleges	To make perfect decisions, maintaining training data sets for both students and tutors. To make effective decision making process and streamlining it

RECENT TRENDS TOWARDS EDUCATIONAL DATA MINING

The basic generalized traditional data mining techniques:

- Clustering-determines the separation or grouping of data
- Classification-classifies orders into predefined classes
- Association rule mining-used to determine the data that can be classified into the same group. This technique is also known as modelling data
- Visualization-graphical representation of the data

Data mining techniques and their applications are widely recognized as powerful tools in various domains (Brtka *et al*., 2012). Chen *et al*. (2012) proposed research reveals, the development of a predictive model that can predict student performance in a class to assist lecturers in improving student's learning process. The predictor variables that can be used in the predictive model (Chen *et al*., 2012). The predictor variables of this model are based on attributes from different educational settings such as coursework marks, psychosocial factors and Course Management System (CMS) log data (Chen *et al*., 2012; Brtka *et al*., 2012; North *et al*., 2007; Ding *et al*., 2008). Carmona *et al*. (2010) says about the application ofsub group discovery which scope is to extract rules describing relationships between the use of the different activities and modules available in the elearning platform and the final mark obtained by the students.

CLUSTERS

EDM focuses fully on educational context hence, cluster formation are also performed in same way (Ahmad and Shamsuddin, 2010). Here, grouping is made by means of categorizing learners/tutors and researchers based on their performance, skills and domain:

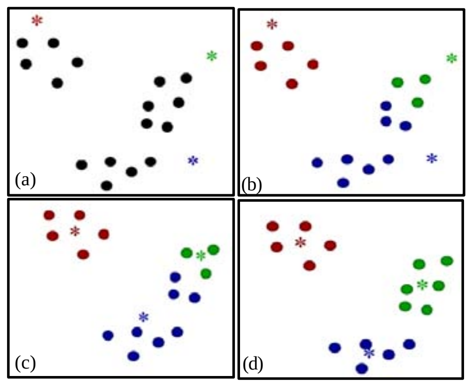

Fig. 1: K-means basics; a) Initializing mean value; b) assign to the nearest representatives; c) re-estimating mean value and d) clustered part

- Learners/Students: clustered by means of class and subjects
- Tutors: clustered by means of experience and domain
- Researchers: clustered by means of area of research
- Universities/Colleges: clustered based on level sets (active levels)
- Learner analysis: active learner, average learner, slow learner

K-means Clustering algorithm

Basic K-means: In this study, researchers summarize about the popular K-means clustering algorithm which is used in various data mining applications. Given a set of n data points on Sd and an integer K, the problem is to outline a fixed set of k points Sd called centers, so as to reduce the mean squared distance from each data to its nearest center. For example, K-means is shown in Fig. 1.

K-means operation:

Algorithm:
Step 1: Predict the number of clusters and their mean distance.
Step 2: Assign the interval Data.
Step 3: Initialize the mean by picking k samples at random.
Step 4: Perform iteration with following:
  Step 4.1: Assign each point to nearest mean.
  Step 4.2: Move the mean to the center of the cluster.
Step 5: Objective function:

$$\sum_{j=1}^{k}\sum_{i=1}^{x}\left\|x(i)^{(j)}-c(j)\right\|^{2}$$

Relationship mining

Deriving association rule using K-means: Researchers can formulate the association rule mining model as follows. Let I be the set of all items and T be the set of all transactions I = {(f, v)|f = Tutor/Faculty, v = value (record set 1, record set 2,..., record set n)}, T = {Learners set}.

Definition 1: Allowable item sets (OLS sets) are item sets of the form, Dataset 1×Dataset 2×...×Dataset n = $\prod_{i=1,\ldots,n}$ Dataset i, where Dataset i is an interval of values in tutor content i (some of which may be un restricted, i.e., [0-26; 0-9; 0-255]). A k-Tutor OLSset (K-dataset) is an OLSset in which k of the Dataset i intervals are restricted (i.e., in k of the Tutors(n) the intervals are not all of [n values]). Researchers use the notation [a, b] i for the interval [a, b] in tutors i. For example, [00, 10] indicates the interval [00, 10] (which is [0, 150] in decimal) in Tutor i. It clearly states that the interval between [00, 10] denotes the multimedia as well as text files in sequence (Ding *et al*., 2008; Derrac *et al*., 2011).

The root count of an OLSset is equivalent to the root count of its K-means: Various kind users are interested in various kinds of rules and some users may be interested in some specific kinds of rules. For an Educational DM (Chen *et al*., 2012; Carmona *et al*., 2010; Brtka *et al*., 2012) based rule prediction, there is a little interest in rules of the type. Initially the record sets where Alphabets<<26, Numbers<<10, RGB-Red<48, Blue<74, Green<134. Researchers define a rule restrictions known as rules of interest fordistinct rules such rule may be of interesting fact or may be of non-interesting fact, all depending upon the measures of support and confidence.

Researchers propose a new algorithm called OLS, to mine association rules on learning sequences from the given trained data's. The algorithm is similar to Classic Apriori algorithm. The Apriori algorithm uses a level wise approach to raise all the frequent item sets, starting with frequent item sets level 1 of item set 1. Based on the datum, if an item set is frequent, all its subset must also be in frequent, the Apriori algorithm generates candidate (k+1)-item sets from frequent k-item sets and then calculates the support for each candidate (k+1)-item set to form frequent (k+1)-item sets (based on Classic Apriori algorithm).

Similarly, in the OLS algorithm, researchers try to find all OLSsets that are frequent and of-interest. Researchers start by partitioning the data into intervals. Then, researchers find all frequent 1-OLSsets by checking the root count of the corresponding K-means (EPIC, 2003; Yu *et al*., 2002a, b). The candidate k-OLSsets are those whose (k-1)-OLSset subsets are frequent. The essential difference between the OLS algorithm and the Apriori algorithm is how the candidate OLSsets is counted. In OLS, OLSsets are counted by performing mean value operations on corresponding basic K-means while in Apriori, it is done by scanning the entire data. In addition, a set of pruning techniques can be used to further improve the efficiency (Ding *et al*., 2008; Derrac *et al*., 2011; Sachin and Vijay, 2012; Karpuk, 2006).

The OLS algorithm assumes a fixed precision in all datas/records. In the Apriori algorithm, there is a function called "apriori-gen" (based on Classic Apriori algorithm) to generate candidate k-Datasets from frequent (k-1)-Datasets. The OLS-generate function in the OLS algorithm differs from the apriori-gen function in the way pruning is done (Ding *et al*., 2008; Derrac *et al*., 2011; Agrawal and Srikant, 1994; Aumann and Lindell, 2003; Breiman, 1984). Researchers use pruning in the OLS-generate function. Since, no value can be in multiple intervals simultaneously, joining among intervals from the same datas can be avoided. For example, even if [00, 01]1 and [11, 11]1 are frequent, there is no need to join them to form a candidate OLSset ([00, 00, 11, 11, 01, 10]1× [11, 11]1). OLS-generate only joins items from different datas (Ding *et al*., 2008). Two frequent (k-1)-OLSsets will be joined into a candidate k-OLSset only if the first (k-1) items of both OLS sets are the same. The order of the last item is compared to avoid the generation of the duplicate candidate OLS set. The rootcount function is directly used to calculate OLS set counts by predicting the appropriate basic K-means instead of scanning the whole databases. For example, in the OLSsets, {B1[0, 64), B2[44, 117)}, denoted as [00, 00]1×[10, 01, 00, 11]2, the count is the root count of P1(00)--P2(01). This provides fast calculation and is useful for huge data OLSsets and eventually improves the mining performance (Ding *et al*., 2008; Derrac *et al*., 2011).

PRUNING BASICS

Root count margin based pruning: To determine if a candidate OLSset is frequent or not, researchers need to AND appropriate K-means to get the root count. In fact, researchers can tell the margins for the root count by observing at the root counts of two K-means by without performing AND operations (Ding *et al*., 2008). Suppose researchers have two K-means for 26 alphabets *10Integers*255 bit files with the first K-means having root count 32 and the level-1 count 16, 16, 0 and 0 and the

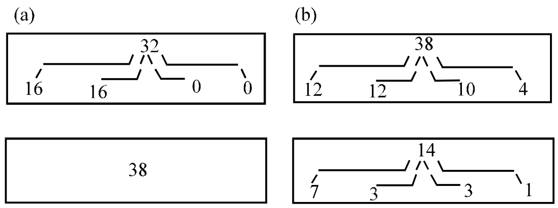

Fig. 2: Denotes the pruning root count

second K-means having rootcount 38 and the level-1 count 12, 12, 10 and 4 (Ding *et al.*, 2008; Derrac *et al.*, 2011; North *et al.*, 2007). By looking at the root level, researchers know that the root count of ANDing result will be at most 38. If researchers go one more level, researchers can say that the root count will be at most seven, calculated by min(16, 1)+(12, 5)+(0, 16)+(1, 14) where min(x, y) gives the minimum of x and y. If the support threshold is 30%, the corresponding OLS set will not be frequent since 9/75<0.3 (Ding *et al.*, 2008; Derrac *et al.*, 2011). As researchers progress to a deeper level, the range to estimate the root count narrows but the cost increases (Ding *et al.*, 2008; Derrac *et al.*, 2011). In the system, researchers provide an option for the user to specify the number of levels (from 0-3) used to estimate the root count before actually calculating the value (Ding *et al.*, 2008) (Fig. 2).

EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

Researchers compare the OLS algorithm with the Apriori algorithm (Agrawal *et al.*, 1993). In other words, pruning is not applied for the comparisons while the performance of pruning is separately given. Researchers obtained identical rules by running Apriori and OLS algorithms. Researchers performed comparison of the results for nearly 8 groups of various students. Total number student in a cluster/group is about 300 where how optimal the students are utilizing the resources effectively are stated. The algorithm provides the sequence of about 0.956>OLS correlation value with less error tolerance than the apriori algorithm. The OLS algorithm is more scalable than apriori for large data sets as shown in Fig. 3. In apriori, researchers need to scan the entire database each time a livelihood (i.e., probability value is to be estimated) calculated. This has a high cost for large databases. However, in OLS, researchers compute the count directly from the values of root count of a basic K-means and the AND program. When data set size is doubled or dually increased, only another single layer (one level) level is added to each basic K-means. The cost is comparatively very small when compared to the Apriori algorithm as shown in Fig. 4.

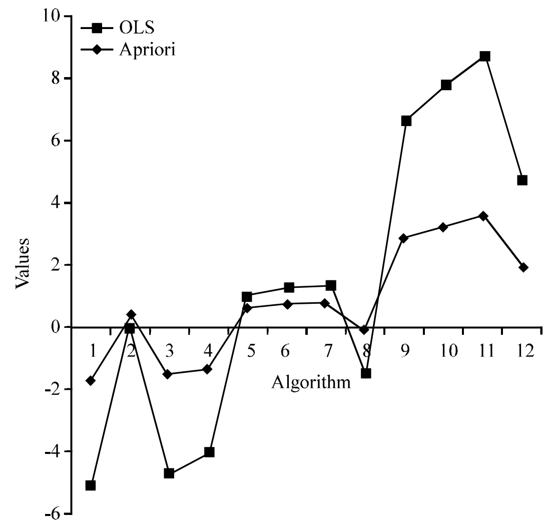

Fig. 3: Threshold sequence

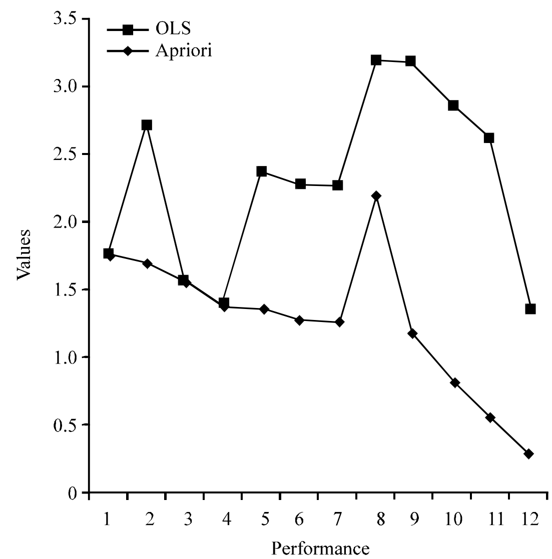

Fig. 4: Minimal support

RESULTS

The result shows the OLS performance when compared to threshold values of the datasets and minimal support sequence. Figure 3 and 4 clearly denotes the threshold sequence and the minimal support. Threshold sequences of the Apriori algorithm leads same as OLS where as there is slight deviation in OLS. Table 2 represents appropriate comparison between the two algorithms where group 4 and 5 yields higher results (Table 2). Figure 4 clearly defines the comparative results of cost estimation; cost evaluation is denoted based on minimal support value only. Minimal support of two

Table 2: Comparison between Apriori and OLS algorithm

	Apriori algorithm				OLS algorithm			
	Text based				Text based			
Groups	Off	Web	Visual	ITS	Off	Web	Visual	ITS
1	124	64	122	222	122	64	144	144
2	22	100	222	100	32	24	200	140
3	34	44	33	250	10	100	190	222
4	245	100	100	200	22	99	144	250
5	150	20	142	148	44	100	168	256
6	100	15	152	158	88	48	145	144
7	20	50	162	144	80	58	200	158
8	8	50	174	155	78	120	222	98

ITS: Intelligent Tutor System

Table 3: Data sets for minimal and threshold sequnce and OLS

Minimal support	Threshold sequence	OLS performance
1.753	-1.747	0.2380
1.709	0.297	0.3480
1.568	-1.558	0.3550
1.389	-1.373	0.4000
1.363	0.655	0.4550
1.275	0.747	0.4600
1.255	0.769	0.4660
2.182	-0.152	0.4770
1.169	2.861	0.6550
0.817	3.253	0.6880
0.563	3.555	0.7020
0.283	1.917	0.7980
0.003	-3.497	0.0248
1.003	-0.409	0.3550
0.005	-3.121	0.4020
0.008	-2.754	0.4250
1.009	0.301	0.4680
1.011	0.483	0.4700
1.012	0.526	0.5550
1.015	-1.319	0.6100
2.015	3.707	0.7770
2.035	4.471	0.8700
2.059	5.051	0.9200
1.1	2.734	1.0600

defined strategies which has less minimum support is examined which was represented by group 1-Visual and ITS (Table 2). Figure 5 denotes OLS performance which is clearly discussed in study.

Performance analysis: In this study, the performance analysis of the OLS algorithm and basic pruning techniques is given in Fig. 5. Here, OLS algorithm predicts the frequent sets of learners who are all utilizing the resources are classified and predicted. Here, rules are generated based on the learners classification and co- relation of learners/tutors/teachers and researchers. Researchers also performed trials on rootcount margin based pruning using various levels. The results show that using level-1 rootcount margin (Ding *et al*., 2008; Derrac *et al*., 2011) based pruning typically provides better performance than using level-0 or level-2 or more Table 3 and 4.

Table 4: Correlation value of Apriori and OLS data sets which are frequently used

Apriori	OLS
0.238	0.248
0.348	0.355
0.355	0.402
0.400	0.425
0.455	0.468
0.460	0.470
0.466	0.555
0.477	0.610
0.655	0.777
0.688	0.870
0.702	0.920
0.798	1.060
0.806	1.660

Total correlation value for Apriori = 0.930; correlation value<0.930; total correlation value for OLS = 0.989; correlation value<0.989

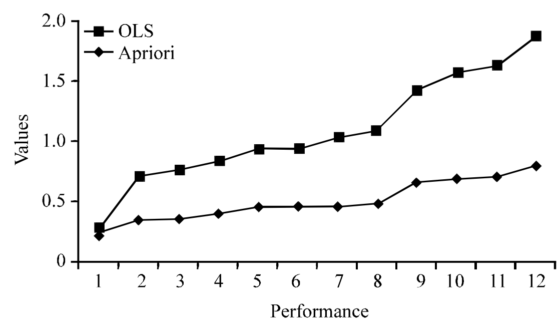

Fig. 5: OLS performance

CONCLUSION

In this study, researchers propose a new model to derive association rules for optimal learning sequence for learners using K-means. In the model, K-means structure is spatial-inherent data mining structure which is used to organize and represent datasets in the form of Clusters/groups. For association rule mining, K-means facilitate advantages such as fast computation and new pruning techniques. Similarly the OLS algorithm has a high information gain for ITS (Intelligent Tutor System) acquired with high optimality than the Apriori algorithm.

Frequent selection of datasets from the resources, proper utilization of resources is made using OLS. OLS algorithm can be applied in various data mining applications such as remote sensing, satellite communication, multimedia frame retrieval, etc. (Chaudhuri and Dayal, 1997).

REFERENCES

Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, September 12-15, 1994, San Francisco, USA., pp: 487-499.

Agrawal, R., T. Imielinski and A. Swami, 1993. Mining association rules between sets of items in large database. Proceedings of the ACM SIGMOD International Conference on Management of Data, June 1, 1993, New York, pp: 207-216.

Ahmad, N.B.H. and S.M. Shamsuddin, 2010. A comparative analysis of mining techniques for automatic detection of student's learning style. Proceedings of the 10th International Conference on Intelligent Systems Design and Applications, Novemkber 29-December 1, 2010, Cairo, pp: 877-882.

Aumann, Y. and Y. Lindell, 2003. A statistical theory for quantitative association rules. J. Intell. Inf. Syst., 20: 255-283.

Breiman, L., 1984. Classification and Regression Trees. Chapman and Hall, USA., Pages: 358.

Brtka, E., V. Brtka, V. Ognjenovic and I. Berkovic, 2012. The data visualization technique in e-Learning system. Proceedings of the IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics, September 20-22, 2012, Subotica, pp: 489-492.

Carmona, C.J., P. Gonzalez, M.J. del Jesus, C. Romero and S. Ventura, 2010. Evolutionary algorithms for subgroup discovery applied to e-Learning data. Proceedings of the Education Engineering (EDUCON), April 14-16, 2010, Madrid, pp: 983-990.

Chaudhuri, S. and U. Dayal, 1997. An overview of data warehousing and OLAP technology. ACM SIGMOD Record., 26: 65-74.

Chen, Y.Y., S.M. Taib and C.S. Che Nordin, 2012. Determinants of student performance in advanced programming course. Proceedings of the International Conference for Internet Technology and Secured Transactions, December 10-12, 2012, London, pp: 304-307.

Derrac, J., J. Luengo, J. Alcala-Fdez, A. Fernandez and S. Garcia, 2011. Using KEEL software as a educational tool: A case of study teaching data mining. Proceedings of the 7th International Conference on Next Generation Web Services Practices, October 19-21, 2011, Salamanca, pp: 464-469.

Ding, Q., Q. Ding and W. Perrizo, 2008. PARM-an efficient algorithm to mine association rules from spatial data. IEEE Trans. Syst. Man Cybernetics Part B: Cybernetics, 38: 1513-1524.

EPIC, 2003. Privacy and human rights an international survey of privacy laws and developments. Electronic Privacy Information Center.

Karpuk, D.M., 2006. From resource discovery to knowledge management. Libraries Unlimited, Westport, CT.

North, M.A., T.C. Ahern and S.B. Fee, 2007. The effect of student self-described learning styles within two models of teaching in an introductory data mining course. Proceedings of the 37th Annual Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, October 10-13, 2007, Milwaukee, WI., pp: 13-18.

Parack, S., Z. Zahid and F. Merchant, 2012. Application of data mining in educational databases for predicting academic trends and patterns. Proceedings of the IEEE International Conference on Technology Enhanced Education, January 3-5, 2012, Kerala, pp: 1-4.

Sachin, R.B. and M.S. Vijay, 2012. A survey and future vision of data mining in educational field. Proceedings of the 2nd International Conference on Advanced Computing and Communication Technologies, January 7-8, 2012, Rohtak, Haryana, pp: 96-100.

Song, C.X. and K. Ma, 2008. Applications of data mining in the education resource based on XML. Proceedings of the 8th International Conference on Advanced Computer Theory and Engineering, December 20-22, 2008, Phuket, pp: 943-946.

Yu, K., X. Xu, J. Tao, M. Ester and H.P. Kriegel, 2002a. Instance selection techniques for memory-based collaborative filtering. Proceedings of the SIGMOD International Conference Data Mining, July 23-26, 2002, New York, pp: 59-74.

Yu, H., J. Han and K.C.C. Chang, 2002b. PEBL: Positive example-based learning for web page classification using SVM. Proceedings of the ACM SIGKDD International Conference Knowledge Discovery in Databases, July 23-26, 2002, Germany, pp: 239-248.