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

Publication venues:

International Journal of Computer Applications (0975 – 8887) Volume 39– No.7
Published by Foundation of Computer Science, New York, USA in February 2012

Section I: Overview of the Paper

Problem:

Here main problem was to make efficient course recommendation system where we should get the large number of association rules that match the real world interdependencies among the courses as it was found before by increasing the min-support of apriori algorithm, we get the refined rule using Apriori association rule but that gives less number of rules as well as we have to preprocess the data to use this one algorithm only. So for avoidance of preprocessing and getting vast range of association rules we want to see effect of combination of other techniques too.

This is not exactly classical problem but somehow existing problem as already some work have done in previously published Articles: A Framework for Recommendation of courses in E-learning System and in other prior work paper named Mining Association Rule in Classified Data for Course Recommender System in E-Learning.

Also there is published work related to data mining techniques but only for recommendation system not for particularly course recommendation systems.

Here main challenge was to apply and combine 3 data mining techniques and algorithms e.g. APriori association rule algorithm, K-mean clustering and ADTree classification on data; thus designing a model where at 1st step clustering the data in moodle database using K-Mean clustering algorithm then 2nd step was to classify data in Moodle database using classification algorithm e.g. ADTree algorithm and 3rd step was finding the best combination of courses using Apriori algorithm e.g. Apriori association rule and then evaluation of results by comparison of combine effect and prior effect of only association rule mining for this system.

We need to study this whole research about course recommendation system because at then we can analyze the efficient results of combined approach on courses data set selected by students with the single effect of apriori association rule.

Application(s):

It can be used in any university, college and school as in all educational institutes its great concern to have best choice of courses. Students get to choose between hundreds of courses every time they want to take a course. In this case having many choices is a good thing, but it does make it hard for a student to wade through and read all of the information on each course. Universities usually employ guidance counselors, people who are tasked with helping students making their choice. But in practice the counselors are often overloaded with too many students and not enough time, and some students are not satisfied with the level of knowledge that the counselors have.

Methodology / Theory

Authors proposed solution of above problem for refining rules generation by using combination of 3 data mining techniques e.g. by combine effect of classification, clustering and association rule based approach where the course recommendation system in e-learning is a system that suggests the best combination of courses in which the students are interested. This solution is based on existing study of previous paper where combination of 2 algorithms (ADTree classification algorithm & Apriori association rule) is used as they get better results so in this paper now they combined another data mining technique e.g. k-mean clustering.

Algorithms

No new algorithm proposed in this paper. There are 3-already developed algorithms are proposed for this system by using as combined form.

1. Simple K-Mean Clustering

Clustering partitions the large data sets into groups according to their similarity and here it is used to classify datasets. In this algorithm items are moved among the set of cluster until required set is reached.

Key steps:

a) Assign the value to k(number of cluster required)
b) Assign the initial value to means $m_1,m_2,m_3,\ldots\ldots..,m_n$
c) Assign the item to the cluster having the closest mean
d) Calculate new mean for each cluster
e) If there is any change then go to step c else write down the clusters obtained

Complexity: $O(nkt)$, where n is the total number of objects, k is the number of clusters, and t is the number of iterations.

2. ADTree Classification Algorithm

An alternating decision tree is machine learning method for classification that generalizes decision trees and it consists of two nodes:
Decision nodes: specify predicate condition
Prediction nodes: contains single number [prediction nodes always at root and leaves]

Key steps:

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. 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:

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

Complexity:

The complexity of the algorithm is quadratic in the number of boosting iterations and this makes it unsuitable for larger knowledge discovery in database tasks.[Optimizing the induction of alternating decision tree]

3. Apriori Association Rule Algorithm

The Apriori Algorithm is an influential algorithm for mining frequent item-sets for Boolean association rules. This algorithm attempts to find subsets which are common to at least min number C (confidence threshold) of the item sets.

Input: Database of transactions D= {t_1,t_2,….t_n},Set if Items I= {I_1 ,I_2 ,….I_k}, Frequent (Large) Itemset L,Sup,confidence
Output: Association rule satisfying support and confidence

Key Steps:Join Step: C_k is generated by joining L_{k-1} with itself.
Prune Step: Any (k-1)-item set that is not frequent cannot be a subset of a frequent k-item set.
Pseudo-code: C_k – candidate itemset of size k, L_K – Frequent itemset of size k

```
$L_1$ = {Frequent itemsets}
For(k=1; $L_k$!= Ø; k++) do begin
$C_{k+1}$ = candidates generated from $L_k$
For each transaction t in database do increment the count of all candidates in $C_{k+1}$ that are contained in t
$L_{k+1}$= candidates in $C_{k+1}$ with min-support end
```

Complexity : The failure rate of apriori is studied both analytically and experimentally. The time needed by apriori is determined by the number of item sets that are output (success: item sets that occur in atleast k-baskets) and the number of item sets that are counted but not output (failures: item sets where all subsets of itemset occur in atleast k baskets but full set occurs in less than k baskets).
[http://www.cs.indiana.edu/~vgucht/AverageCaseApriori.pdf]

Experiments

In this course recommendation system author considered 13 course category, each category have courses so there are total 82 courses and sample data that was extracted from moodle database as an experiment have 45 students and 15 courses. Courses considered like C-programming(C) , Visual Basic(VB), Active server pages(ASP), Network Engineering(NE), Computer Networks(CN), Operating Sys(OS), Distributed System(DS) etc.

This experiment was conducted for evaluating the performance of combined algorithms on student's selection of different courses. Here yes represents the interest of student in course and No represent that student don't like that course. Then different tables are represented that showing results before preprocessing, after pre-processing, after clustering, classification combination and thus by visualizing diff. tables data author validated goal of this research that as if we consider combination of, clustering, classification and association rule then there is no need to preprocess the data.

Section II: Qualitative Evaluation

Quality

This paper sound a bit technically as very minute issue have been covered, only avoidance of preprocessing of data. Author only continuing his previous work without considering cost and complexities of usage of algorithms and also it must be analyzed for bulk of data on initial stages because system performance can go bitter with data volume. Here authors only pointed strength of their work and ignored the weaknesses of current work.

Clarity

Yes paper written very clearly and well-organized. Authors elaborated each algorithm and their working with experiments and readers can apply given methodology to their own data-set of institutes.

Originality

Given problem and approach is not new as this is novel combination of familiar techniques. But here it clearly differs from previous contributions and related work also here referenced honestly.

Significance

Results are significant for others to follow the given ideas and build their course recommendation system on basis of given solution. The addressed problem is not much difficult and authors follow very simple and easy way though there are some limitations that are not covered in this approach. There is not given any advance and innovative idea. And conclusions are somehow same as it were given in their previous work on same domain. So I don't think there is any unique or pragmatic representation.

Section III: Limitations and Future Work

No paper is perfect in every aspect. Please comment at least three key limitations of the paper. Please also discuss at least three kinds of future work which can be considered to address/overcome the limitations of the paper. You can also discuss what kinds of applications can be benefitted by applying the proposed techniques in this paper.

Key Limitation(s):

a) Literature survey doesn't have deep structured survey of relevant work of pre-existed course recommendation system's data mining approaches.

b) Proposed system only classifies courses as 'yes' or 'no' it doesn't rank courses on basis of user's ratings.

c) The given approach doesn't cover certain aspects of course selection as it only considering what courses students like or dislike despite of fact of other factors because course selection should not solely based on personal taste or preference rather it is heavily influenced by other factors like prerequisites, course format (number of lectures per week and other organizational features related to course, number of credits) and timetable etc.

d) Apriori algorithm is not much efficient with respect to time and system cost; there are alternatives to improve apriori efficiency and also another better algorithm is available for association rule mining.

Future work(s) to overcome limitations:

a) The first step of literature survey should be to define or follow the goals for doing literature survey. So following goals should be in consider for designing best approach of course recommendation system:
Goal 1: Give us an overview of the different types of recommender systems.
Goal 2: Find out how the different types of evaluation methods compares.
Goal 3: Give us an overview of how different types of explanations work in relation to recommender
Goal 4: Give us an overview of existing work in recommender systems for course selection.

b) There should be pre-existed record of courses-ratings by previous students so such ratings would be helpful in recommending best course to students. Here the collaborative filtering gives us recommendation based on what other seemingly similar students and friends of the user prefer. One of most popular way of implementing CF is to look at users that have a similar rating history to the user that the system is giving recommendations to, and then suggest items those users rated highly.

c) For best course selection by considering all related factor they should attempt to use social graph between users and other factors to determine the correlation between users so plan should be to weight each user correlation with each other by using these signals. So try to explore and use existing CF correlation algorithm to give better course recommendations.

d) Performance study shows that FP-growth is an order of magnitude faster than apriori as no candidate generation, no candidate test. It uses compact data structure and eliminates repeated scan. Here basic operation is counting and FP-Tree building.