International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

A Novel Way for Mining Frequent and Interesting Patterns using Genetic Algorithm

Reshu Tyagi[1], Muskaan Batra[2]

[1]Trident ET Group of Institutions, NH-58, Delhi-Meerut Road, Ghaziabad, India

[2]Franconnect Software India pvt.ltd, C-94, Sector-8, Noida, India

Abstract: ***Over the years, the process of finding interesting association rules has become a keystone in adept decision making. Among various data mining techniques, Association Rule Mining is mostly used for finding interesting associations among different products in a transactional database. However, mining association rules wind up in finding plethora of rules and hence finding the most "interesting" and "optimal" rule becomes a irksome task using generally used Apriori algorithm. However Apriori Algorithm uses Conjunctive nature of association rules, and single minimum support threshold to reveal the interesting rules. But only these factors don't seem sufficient to unearth the interesting association rules efficaciously. Hence, in this paper we have introduced a entire distinct approach for finding much optimized association rules using numerous and varied quality factors like support, confidence, comprehensibility and interestingness. The demonstration performed on copious datasets shows the much improved performance than Apriori algorithm.***

Keywords: Data Mining, Association Rule, Genetic Algorithm, Support, Comprehensibility, Apriori Algorithm

1. Introduction

Various energetic and agile research fields exist in computer science discipline. DATA Mining [1] is one of them. It is a very rapidly emerging field. In the domain of data mining and computing, Knowledge Discovery in Databases (KDD)[1] has been a very enchanting and fascinating exploration challenge. Its focus is to draw captivating and purposeful data from an massive and bulky miscellany of data kept in the transactional databases. A diagrammatic representation of knowledge discovery using data mining is shown below:

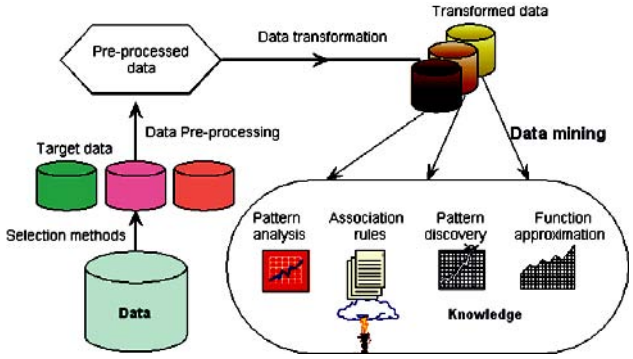


Figure 1: Various Stages of Knowledge Discovery

1.1 Association Rule Mining

Association Rule Mining[2][3] is getting immense recognition since its introduction. Association rules are aimed to discover strong rules from databases with the help of various measures of interestingness and for uncovering regularities and strong correlation among products in bulky transaction data . Association Rule Mining focuses on identifying interesting correlations, repeated occurring patterns , associations or informal structures between sets of items in the commercial databases or any other data repositories. The leading objective of ARM is to identify the set of all subgroup of items and/or attributes that repeatedly occur in various database records or business agreements, and additionally, to discover rules on how a subgroup of items impacts the existence of another subgroup. The process of mining association rules for market basket data is reviewed as a important knowledge discovery process. Different correlations between items belonging to any customer conducting business in some market -basket database can be effectively discovered using Association rule mining

Rules in ARM algorithms are generally in the form:
$X \rightarrow Y$ i.e.

IF the value of the predicting attributes is true, THEN value is predicted for goal attributes.

Both X and Y are frequent item-sets in some transactional database and $X \cap Y = \emptyset$. The rule $X \rightarrow Y$ can be elucidated as "if some item set X happens in a transaction, then some another item set B will also occur in the same transaction". For example, suppose in some database 35% of total transactions include both bread and sauce and 75% of all transactions include bread. An Association Rule Mining system will formulate the rule bread → sauce with 35% support and 75% confidence. Rule support and rule confidence are two very important virtue measures of rule interestingness.

If some rule has confidence of 75% it means that 75% of the customers who purchased bread also purchased sauce. Generally, the association rules are rated as interesting if they meet both the criteria of minimum support and minimum confidence. These two criteria are generally set by experts or by users. The rules having values of support and confidence greater than or equal to the user defined values are found by association rule discovery process.

Support: It can be defined as the probability of some item or group item in the given transactional database:
support(I) = n(I) / n

where n is the total number of transactions in the given database and n(I) is the number of transactions that include the item s I. Therefore, support (A=>B) = support(AUB).

Confidence: It is actually a conditional probability. For some association rule A=>B , confidence is defined as **confidence(A=>B) = support(AUB) / support(A).**

Frequent itemset*:* Let D be some set of items, and let TD be the transaction database and σ be the user defined minimum support. An itemset D in S (i.e., S is a subset of D) is called as frequent itemset in TD with respect to σ, **if support(D)TD >= σ.**

Finding Association Rules is not brimful of advantages but it also has some constraints, first the number of rules generated are increased as the number of items or products increases in the database , it simply means that **Association rule mining posses algorithmic complexity**. To reduce this complexity various novel algorithms have designed that can actively lower the space of search . Secondly, the difficulty in finding interesting rule from a given set of rules. So in this paper we have tried to resolve these issues. For resolving first issue we apply the evolutionary Genetic Algorithms for decreasing the number of rules generated because GA finds more fitter solution since it perform search globally and it subsist better for second issue. We can find worthful association rules from the set of rules by applying useful measures on the set of rules .
Therefore in this paper we have used proposed Genetic Algorithm on those set of rules that were generated by Apriori Algorithm for achieving over written objectives.

1.2 GENETIC ALGORITHM

Genetic Algorithm (GA)[4] is an imitation procedure of natural evolution, developed by John Holland in 1970 . It is grounded on the principle of natural selection and development. A genetic algorithm (GA) is a process heuristic that imitate the procedure of natural evolution. This process is used to obtain purposeful solutions to optimization and search related problems. Genetic algorithms are a type of evolutionary algorithms (EA), which evolve in every step to generate solutions to optimization problems by applying various techniques motivated by natural evolution like mutation , crossover , inheritance and selection. For some problem which is very little known GAs are one which has best way to solve it . Being a very widespread and commonly known algorithm , these will work dramatically in any search space provided.
GA are very well applied in various search and machine learning areas. GA evolve themselves in a monotonous manner by generating a set of new and better fitness populations from old population. Every string of population is encoded in binary. Every string is associated with a fitness function as measured by a evaluating function.
A GA explores for a best solutions to a problem by conserving candidate solutions populations and building further generations by picking the current best solutions and with the help of operators like Mutation & Crossover to generate new candidate solutions. Thus, more superior solutions then previous are “evolved” over time. Generally , the algorithm abort when either a extreme number of generations has been generated , or some expected fitness level has been gained for the population . The merits of GA becomes more clear-cut once the search area of a task is big enough . The GAs become more important when discovering association rules because rules founded by these algorithms are more common due to its global nature of search to discover the frequency of the item sets and they are less tangled as compared to other induction algorithms generally used in data mining tasks, since these algorithms performs local search. Since genetic algorithms search globally , they manage better with various attribute interactions than inductions algorithms.
GA uses monotonous manner of generating new populations from old population.

Some Terminologies

Chromosome: A chromosome or a genome is a position of variables which suggest a solution to the problem that is being tried to be solved by genetic algorithm . The chromosome is generally depicted as a simple string; though various types of data structures are also used. The Chromosome depiction is redefined for every particular problem, along with its fitness, mutate and reproduce methods.

Gene*:* A Gene is actually a chunk of chromosome. A gene contains a sub part of solution. For example if 142894 is a chromosome then 1, 4, 2, 8, 9 and 4 are its genes.

Fitness*:* Fitness is a core idea in process of evolution. It is generally denoted as ω in various genetic models.

Various genetic operators like selection, crossover and mutation are applied on an initial arbitrary population for computing generation of new strings. Hence in successive generations of the algorithm the worth of the solutions improves. The process is aborted when an satisfied or optimum solution of the problem is found. GA is more suitable for problems which needs optimization, with respect to some estimated criteria.

Overall, Genetic algorithms are a process of "multiplying" computer programs and solutions to optimization or search problems with the help of simulated evolution. The procedure is totally based on natural selection and operators like mutation and crossover are continuously applied to binary strings population representing potential solutions.

GA Work Procedure

The GA works as follows:
1) First of all, a population is created. Population is nothing but a group of selected individuals (Chromosomes) . In other words we can say that a string of genes is called chromosomes.
2) Choose chromosomes which have higher fitness value.
3) Crossover operation is performed between the selected pair of chromosomes for producing new offspring which have better higher fitness.
4) If needed, Mutation is done on the the new chromosomes.
5) Terminate the process on finding the optimum solution.

This evolutionary process is repeated and the process is terminated until a termination condition is reached.

Termination Conditions

Common terminating conditions are:

1) A minimum criteria satisfying solution has obtained.
2) Generations have reached to a fixed number.
3) Budget and/or time exceed the decided limit.
4) If further generations are not producing better results as compared to their parents chromosomes i.e. highest ranking fitness of solutions has reached.
5) Any other Combination of the above conditions.

2. Related Works

Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar, proposed a paper "Mining Frequent Itemsets Using Genetic Algorithm[7]" in October 2010 to explain the merits of using genetic algorithms to unearthen the hidden association rules over other existing methods like Pincer-Search, Border algorithm, Incremental algorithm and widely used Apriori algorithm. They also talked about the refined scenario when using GA for association rules. Their paper also throw some light on positive and negative association rules[14][15].Positive Association rules contains only those items which are listed in the list while on the other hand Negative Association rules contains negated items(not present in list). They primarily focused on Genetic Algorithm to find the quality association rules because it offers several advantages like less computing time, executing global search and its greedy nature.

Gaurav Dubey and Arvind Jaiswal proposed Identifying Best Association Rules and Their Optimization Using Genetic Algorithm[6]" in May 2013 to showcase the effectiveness of this algorithm by executing it on a toy stock management and came to a conclusion that Genetic Algorithm are easy to learn and implement, modular and supports multi objective optimization technique. One of the attractive feature of genetic Algorithm is that the result gets better with time. In this paper the authors also proposed a methodology to discover frequent itemsets. They start with placing a instance of records from some transactional database into the memory and then executing GA that constantly alters the population by performing the following steps repeatedly. The steps are:

1) Estimate fitness function
2) Selection of individuals from current population
3) Producing new individuals by recombination process
4) Replacement.

Mohit K. Gupta and Geeta Sikka developed "Association Rules Extraction using Multi-objective Feature of Genetic Algorithm [8]" in October 2013 and focused on four quality factors namely Support, Comprehensibility, Interestingness and Comprehensibility to refine association rules as obtained from Apriori.

The authors also showed experimental results of applying genetics on a number of different datasets and concluded that in most of the cases the value of all the our quality factors was much superior than that of Apriori but since it is a multi objective procedure, so we cannot prioritize one's objective over another.

Apart from this various researches have been done on genetic Algorithms to mine efficient association rules with various quality factors. Rest of the paper is organized as follows

3. Methodology Adopted

Presenting Rules and Scheme of Encoding

Presentation of rules plays a important role in Genetic Algorithms; mainly there are two methods of encoding the population of individuals into rules. One technique is Michigan approach[5] , in which every single rule is encoded into an individual and another technique is called Pittsburg approach[5] in which a set of rules is encoded into an individual.

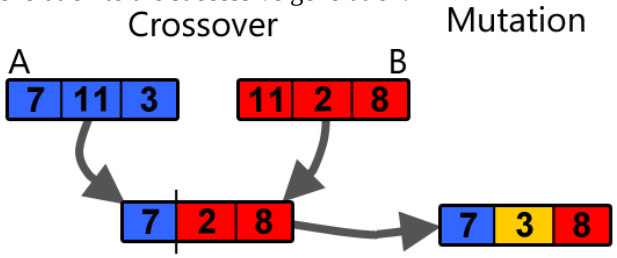
For some rule the set of attributes that forms an antecedent body and the set of attributes that forms a consequent body would be disjoint, means set of attributes situated in the antecedent body ∩ set of attributes situated in the consequent body= Ø. In mathematical form it can be represented as **:**

If X→ Y then X∩Y= Ø.

Genetic Operators

i) Selection*:* This is a very important operator where a chromosome is chosen from the current population based on its fitness function and reprint it without any changes into the new population.

ii) Crossover: With the help of this crossover operator , two new child chromosomes are produced from two parent chromosomes. This is done by swapping a part of gene of one parent with that of another parent. With this Crossover operator, the programming of a single chromosome or a group of chromosomes can be made to vary from one generation to the successive generation.



iii) Mutation : Mutation operation changes a single bit of the new solutions to add stochasticity for better solutions in the search . This is actually a possibility that a bit within a chromosome will be flipped (0 becomes 1, 1 becomes 0).

Fitness Function

We always use a mechanism so as to know that how near our design solution is with the set goal. For this to test we need a specific type of objective function, known as fitness function. Defining a good fitness function is very important. For discovering the gripping association rules, we can also use Multi-objective fitness function. In this project work, we have fostered this approach and have used four remarkable measures for finding the association rules. These

measures are support, confidence, comprehensibility and interestingness. So, in lieu of a single attribute, ARM problems can be considered as a Multi- attribute problems.

Support

The support of a item set Y, denoted as $\sigma(Y)$, can be defined as the percentage of transactions in the dataset, containing item set Y.

$$S = \sigma(A \cup B)/\sigma(N)$$

Where, $\sigma(N)$ = total number of transactions in the dataset
$\sigma(A \cup B)$ = number of transactions that contain both A and B
Support is generally used to banish non-interesting rules.

Confidence

Confidence is another measure to evaluate the precision of association rule. It computes the conditional probability.

$$C = \sigma(A \cup B)/\sigma(A)$$

Where σ(A) = number of transactions containing A.
A very effective association is considered between A and B if the confidence value is higher.

Comprehensibility

If the generated rule contains a huge number of attributes with it ,then that rule will be considered as very tough comprehend. The discovered rules should be very simple and understandable to the user, so that the user can use them effectively. So the Comprehensibility factor is required for making rules a bit easy to understand. Comprehensibility of a certain association rule can be formulated as:

$$Comp = \log(1+|B|) / \log(1+|A \cup B|)$$

Where B and |A∪B| are the number of attributes included in the consequent part of the rule and total rule respectively.
If the number of constraints in the antecedent body of the rule is less, then the rule is rated as much more simple and comprehensible.

Interestingness

Interestingness of certain rule, represented as **Interestingness X→Y**, is used to quantify how much a certain rule is "eye-opener" for the users. Since finding some concealed information is the core point of data mining, so it should reveal those rules that have relatively less occurrence in the database. Interestingness of a rule can be formulated as:

$$\text{Interestingness } X \rightarrow Y = \frac{Sup(X \cup Y)}{Sup(X)} \times \frac{Sup(X \cup Y)}{Sup(Y)} \left(1 - \frac{Sup(X \cup Y)}{\sigma(N)}\right)$$

Where σ(N) indicates total number of transactions in dataset.

As stated above, ARM is reviewed as Multi-objective problem not Single Objective one. So, its fitness function can be formulated as:

$$F = \frac{(P \times Sup) + (Q \times Con.) + (R \times Comp) + (S \times Interest.)}{(P+Q+R+S)}$$

Where P, Q, Rand S are user-defined weights.

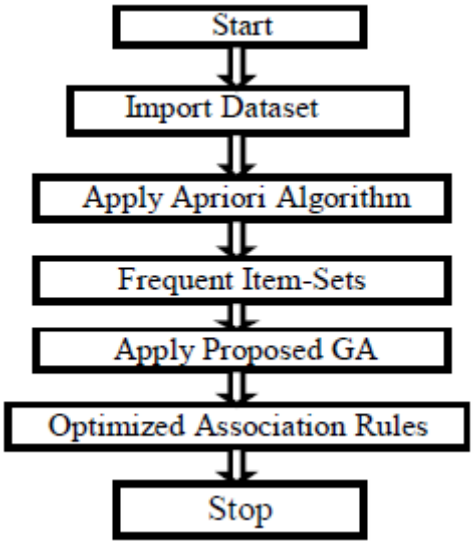
Since discovering frequent item sets is of high computational complexity so, extracting association rules can be reduced to finding frequent item sets .For this, in this project work the weight values of P= 4, Q=3, R=2 and S=1 are taken based on the relative significance of the four quality measures support, confidence, comprehensibility and interestingness. It should also be noted that the range of fitness values should be in [0…1].

Algorithmic Structure

Here, the structure of the proposed algorithm is presented. First the rules are generated using Apriori algorithm and then GA is applied on them. The process of the proposed algorithm for generating optimized association rule with the help of Genetic Algorithms is as given below:

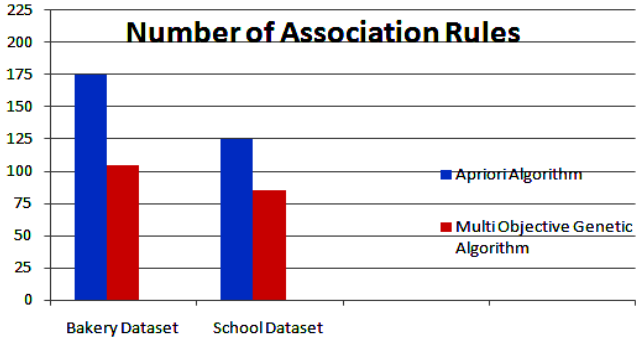
1. Begin
2. Put a dataset into the memory.
3. Apply Apriori Algorithm on it to frequent item-sets. Let F is the frequent item-set set discovered by Apriori Algorithm.
4. Set O =Φ where O is the output set containing all discovered association rules.
5. Put some terminating condition on Genetic Algorithm.
6. Represent every item set of F in some encoding scheme.
7. Then, selected members and apply Genetic Algorithm on them to generate association rules.
8. Then, calculate the fitness function of every rule A→ B.
9. If value of fitness function meet the criteria of selection then
10. Set O= O∪{A→ B}.
11. If the required number of generations is not completed, then go to step 3.
12. END.

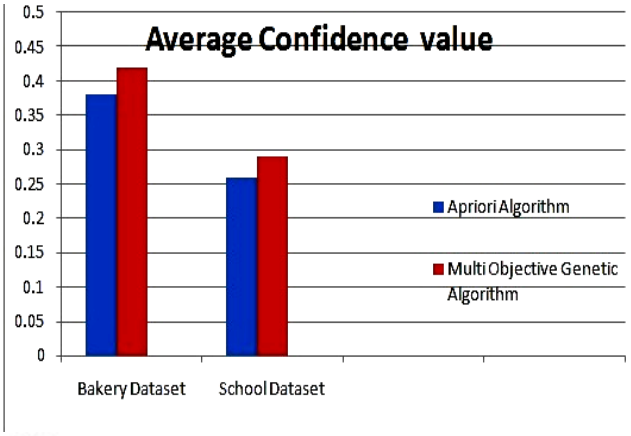
4. Block Diagram of Proposed Algorithm

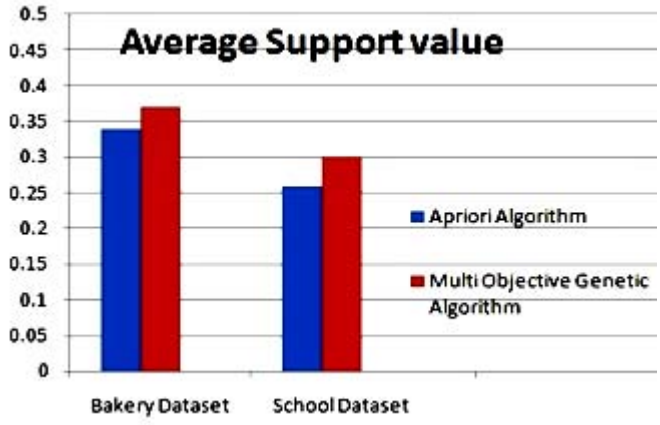


5. Results of Experiment and Analysis

The recommended multi-objective GA was implemented using MATLAB 2013b and 8GB RAM. Different datasets were collected and the running behavior of this proposed Algorithm was checked . The algorithm aborted on reaching a desired number of generations. The performance evaluation of proposed algorithm and Apriori algorithm was compared.







The above presented figures shows that the proposed algorithm's performance was much better as compared to Apriori's Algorithm Performance. But since it is a Multi-Objective approach we cannot prioritize one's objective over other.

6. Conclusion and Future Scope

In this project work we have seeked Multi-objective feature of Genetic Algorithms [9]for extracting the best association rules. When this proposed genetic algorithm was attempted on other datasets, we got rules with extreme accuracy and interestingness.

It has also been noticed that this proposed algorithm can also give good improvement in performance in extraction of association rules. Also this generates less number of rules than other algorithms. So, we can conclude that this algorithm optimize the association rule very efficaciously.

References

[1] From Data Mining to Knowledge Discovery in Databases by Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth
[2] Association Rule Mining: A Survey by Qiankun Zhao Nanyang Technological University, Singapore and Sourav S. Bhowmick Nanyang Technological University, Singapore
[3] http://www.ms.unimelb.edu.au/~odj/Teaching/dm/1%20Association%20Rules%2008.pdf
[4] http://www.boente.eti.br/fuzzy/ebook-fuzzy-mitchell.pdf
[5] Hisao Ishibuchi , Tomoharu Nakashima ,Tadahiko Murata""Comparison of the Michigan and Pittsburgh Approaches to the design of Fuzzy Classification Systems"
[6] Arvind Jaiswal, Gaurav Dubey " Identifying Best Association Rules and Their Optimization Using Genetic Algorithm" International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue-7, May 2013
[7] Soumadip Ghosh, Susanta Biswas , Debasree Sarkar, P. P. Sarkar "Mining Frequent Itemsets Using Genetic Algorithm" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2011
[8] Mohit K. Gupta and Geeta Sikka" Association Rules Extraction using Multi-objective Feature of Genetic Algorithm" Proceedings of the World Congress on Engineering and Computer Science 2013 Vol II WCECS 2013, 23-25 October, 2013, San Francisco, USA
[9] Ashish Ghosh , Bhabesh Nath "Multi-objective rule mining using genetic algorithms" Information Sciences 163 (2004) 123–133
[10] Manish Saggar, Ashish Kumar Agrawal , Abhimanyu Lad "Optimization of Association Rule Mining using Improved Genetic Algorithms" 2004 IEEE International Conference on Systems, Man and Cybernetics
[11] Peter P. Wakabi–Waiswa, Venansius Baryamureeba, Karunakaran Sarukesi "Optimized Association Rule Mining with Genetic Algorithms" 2011 Seventh International Conference on Natural Computation
[12]] Indira K, and Kanmani S "Performance Analysis of Genetic Algorithm for Mining Association Rules" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012
[13]] Anubha Sharma , Nirupma Tivari " A Survey of Association Rule Mining Using Genetic Algorithm" International Journal of Computer Applications & Information Technology Vol. I, Issue II, August 2012
[14] T Ramakrishnudu ,R B V Sbramanyam " Mining Positive and Negative Association Rules Using FII-Tree" International Journal of Advanced Computer Science and Applications, Vol. 4, No. 9, 2013
[15] Mining Positive and Negative Association Rules: An Approach for Confined Rules by Maria-Luiza Antonie Osmar R. Za¨ıane
[16] Xiaowei Yan, Chengqi Zhang , Shichao Zhang b,c, "Genetic algorithm-based strategy for identifying

association rules without specifying actual minimum support"
[17] Mrinalini Rana and P S Mann "Association Rule Mining with Multi-Fitness Function Genetic Algorithm " International Journal for Science and Emerging Technologies with Latest Trends" 8(1): 14-23 (2013
[18] Soumadip Ghosh, Susanta Biswas , Debasree Sarkar, P. P. Sarkar "Association Rule Mining Algorithms and Genetic Algorithm: A Comparative Study" 2012 Third International Conference on Emerging Applications of Information Technology (EAIT)
[19] K.Poornamala and R.Lawrance "A General Survey of Frequent Pattern Mining Using Genetic Algorithm"

Author Profile

Reshu Tyagi has received B.Tech degree in Computer Science from UPTU and M.Tech (Computer Science) from Amity University, Noida. She is currently working as Asst. Professor in Trident ET Group of Institutions, Ghaziabad. She has more than 4 yrs. of teaching experience. She has attended various seminars, workshops and conferences. Her area of research includes Data Mining, Theory of Automata and Compiler Design.

Muskaan Batra has received B.Tech degree in Computer Science from RTU and M.Tech (Computer Science) from Amity University, Noida. She is currently working as Senior Software Developer, Franconnect Software India pvt.ltd, Noida. She has more than 3 and half yrs. of development experience. She was part of various projects in company and has provided training related to programming to juniors. Her area of research includes Data Mining, Data structures and Java programming.