International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com

A DETAILED DESCRIPTION OF SEVERAL ASSOCIATION RULE MINING ALGORITHMS: A SURVEY

Vikram Rajpoot[1], Akhilesh Tiwari [2],Bharat Mishra[3]
rajputvikram022@gmail.com,atiwari.mits@gmail.com,bharat.mgcgv@gmail.com
[1]Research Scholar, MGCGV Chitrakoot SATNA, India
[2]Professor, Dept. of CS&IT, MITS Gwalior, India
[3]Associate Professor, Faculty of Science & Environment, MGCGV, Chitrakoot SATNA, India

Abstract

In this paper, a brief discussion about the various data mining fields has been discussed. Association rule mining is a vast field that has been worked upon a lot. Various works has been performed on the optimization of the ARM approach. Genetic algorithm, PSO and ACO have been performed on ARM and various papers that have been built on this have been discussed in this paper. Fuzzy and rough set implementation on the association rule mining has been done and this field has also been discussed in this paper. All the strategies discussed have a few not common issues that are mentioned in this paper.

Keywords— Data Mining, ARM, PSO, ACO, fuzzy set, tough set, GA.

Introduction

Data mining suggests Data mining from immense measure of information. Through execution huge mining, stimulating learning, regularities, or high-level knowledge can be mined from the databaseand watched or browsed from different edges. The discovered information can be applied to query processing, decision making, process control, and information management. Data mining, with its capability to successfully find significant, non- obvious knowledge from huge databases [1], is primarily defenseless against abuse. In today's world privacy assumes an imperative part by securing and protecting the sensitive data esteems from being utilized by unapproved access and subsequently it is not the same as some other information security field for example access control and data security which prevents knowledge disclosure against the illegitimate means. The main aim of privacy preservation is to prevent data or information from unauthorized access to the data. With the development of database technology and network technology, [2] a large number of useful data, which contains much individual privacy information, have been accumulated in various fields, such as patient's condition information, customer preferences, personal background information, etc. Once the information leaked, it will be harmful to individual. If they provide real information directly to diggers, it will unavoidably create private data disclosure. [3] Traditional analysis tools and techniques cannot be used because of the massive size of data.

Association rule mining

ARM is an outstanding strategy to find interesting rules and relations between different items in enormous databasesBased on the strong rules idea, RakeshAgrawal et al. [4] Introduced association rules for learning regularities between various products in the huge-scale transaction documents recorded.

Association rules are made through studying data for visit if/at that point designs and applying the criteria certainty and support to characterize the most huge connections. *Support* is a how frequently items perform in the database indicate. *Confidence* indicates the various times the statements of if/then have been found to be true [5].Numerous business enterprises accumulate big amounts of information from their everyday operations. For example, enormous measures of client buy record are gathered everyday at the grocery stores counters.

Genetic algorithm

GAs (Genetic algorithms) was composed through John Holland in the 1960s and were developed through Holland and his understudies and besides relates at the Michigan University 1960s and Seventies.The decision operator picks these chromosomes in the populace a decent method to be allowed to imitate, and on creatures; transformation arbitrarily modifications the allele estimations of different areas in the chromosome; and inversion reverse request of the bordering component of the chromosome, thusly improving the request where qualities are displayed GA [6]are optimized and also searched algorithms which is based on the principles of natural

evolution, which were first introduced through John Holland in 1970.

In genetic algorithms, term chromosome classically refers to the candidate problem solution, often encrypted as a string bit.An allele in a bit string is either 1or 0; for higher letters in order more alleles are conceivable at each locus.

GA has the accompanying advances:-

1) Introduction: GA are [7] ordinarily begin with beginning populace. In this manner a strategy is intended to give the GA a good start and speed up the developmental methodology.
2) Selection: This operator selects chromosomes in populace for replica. The more healthy chromosome, the much time it is likely to be selected to the reproduce.
3) Reproduction:- It pick out two different chromosomes according to the present selection processachieve crossover for them and find one or two children, perhaps using mutation.
4) Crossover: With a crossover probability crossover parent to form novel offspring. This administrator randomly chooses a locus and trades subsequence after and before that locus between two distinct chromosomes to create two diverse different.
5) Mutation:-After a crossover [8] this operator is performed. Mutation is the operator of genetic used to the keep genetic variety from one population generation of chromosomes to the next. This operator randomly flips several of the bits in a chromosome.
6) Replacement: Use new generated population for a further run of algorithm.

Particle swarm optimization

PSO has two sector approaches[9]. Perhaps extra obvious are its ties to artificial life (A-life) in the common, and to flocking of bird, schooling of fish, and theory of swarming in specific. It comes from the study on the fish and bird Flock movement habits. The algorithm is extensively used and quickly established for it's simply implementation and also some particles need to be tuned.

It is established from intelligence of swarm and is based on research of fish and bird flock movement behavior. While for food searching, the birds are either scattered or go together already they build up where they might have the capacity to discover the food.Whilst the birds are shopping for meals from one area to another, there is always a bird that can be food smell very well, that is, the bird is noticeable of the place where the found the food, containing the improved food resource knowledge. The particle without the value and also volume serves as all Man or woman, and simple behavioral pattern is regulated for all particles to present the complexity of the complete particle swarm.

Ant colony optimization

ACO has been widely utilized for numerous combinatorial optimization issue. ACO is the heuristic algorithm which has been proven an efficient technique and using to numerous CO problems. ACO algorithms are invigorated through subterranean ant's conduct to look for a way between their home and a standard food source. It has been experiential that ants discover such a path very rapidly through applying indirect communication via pheromones. This observed behavior is put into an algorithmic structure with considering artificial ants that build solutions for a give issue through carrying out random walks. It is a comparatively new meta-heuristic method and has been effectively used in numerous applications particularly issue in combinatorial optimization. Models of ACO algorithm ant colonies real behavior for shortest path producing between food sources and nests.Ants can speak with each other by different chemicals called as pheromones in their moment environment. The ants move as indicated by the amount of pheromones, wealthier the pheromone trail on a way is, the extra likely it would be trailed by various ants.

Rough sets

In the theory of rough set, membership isn't essential idea. Rough sets speak to different numerical technique to vulnerability and vagueness. Description of a set in the harsh set hypothesis is identified with our data, learning and discernment about components of the universe. At the end of the day, we "see" components of the universe in the setting of accessible data about them. As a result, two unique components can be incongruous in the connection of the data about them and "seen" as the same. Methodology of rough set is based on premise that dropping the precision degree in the data creates the pattern of data more observable, whereas the focal preface of the philosophy of rough set is that the data incorporate into the order capacity. The rough set results are displayed as classification or rules got from aset of cases [10]. The fundamental rough set theory advantage in data examination is that it doesn't require any extra information or any preliminary about the information [11].For ease we

initially elucidate the proposed approach instinctively, by methods for a basic instructional exercise case.

- The essential advantages of the rough set method are as per the following:
- It does not prerequisite any preparatory or additional learning about data − like likelihood in measurementsmembership grade in the theory of fuzzy set.
- It gives effective approaches, tools and algorithms for discovery hidden patterns in document.
- It permit to decrease original information, i.e. to find minimal data sets with the similar information as in the original document
- It permits to create in automatic way the decision rules sets from data.
- It offers straightforward interpretation of obtained results.
- It is suited for concurrent processing.
- Produces decision rules sets from data.

Disadvantages

1. One of the rough set theory downsides is its reliance on fulldata frameworks i.e.
2. One of the big limitations of model of the classical rough sets in real applications is the ineffectiveness in the core and reduct computation, because each intensive operations of computational are achieved in the flat files.

Fuzzy sets

The concept of empowering classical organization principles through combining them with fuzzy set idea has already been around considering the fact that several years. The main concept derives from attempts to the deal with quantitative attributes in a database, where quantitative values subdivision into crisp sets would lead to the over- or underestimating esteems close borders. Fuzzy sets can defeat that issue through allowing fractional enrollments to the different sets. Even though a lot of research has been done on the topic and algorithms have been proposed, not many programs provide the functionality yet.

Fuzzy association rule mining primary started In the type of expertise discovery in Fuzzy knowledgeable programs. A fuzzy trained approach [12] uses a group of fuzzy membership features and rules, as a substitute of Boolean common sense, to rationale about knowledge. The rules [13] in the fuzzy proficient procedure are typically of a kind similar to the next: “whether it is raining then put up your umbrella” here if phase is the antecedent section after which phase is the ensuing section. This rules kind as a set helps in the pointing towards any solution with in the set of solution. But in Boolean logic case all data attribute is measured only in yes or no terms, in the other different words negative or positive. So it never permit us to have the diverse solutions field. It has always solutions marginalizes; on the other different hand fuzzy logic keeps broad ways of solutions open for the customers [14]. Fuzzy variants of FP-Growth and Apriori algorithms also can be categorized into DFS and BFS type algorithms.

Advantages:-

1. This approach provides very fast preprocessing.
2. The approach is exceptionally powerful.

Disadvantages:-

1. Fuzzy feedback systems control represents a small valued engineering method which is used through people who contain never create an effort to learn the traditional control theory. Fuzzy logic did not succeed in offering an alternative method when it comes to systems without known model of mathematical which was its basic purpose.
2. It is difficult to fuzzy control system solidness demonstrate. When it comes to proofs which we can discover in literature, stability is often proved on the system of 'crisp' which is only a deformed fuzzy picture, while approachesfrom the system theory are utilized.
3. There is no orderly technique to fuzzy framework planning. Rather, observational s ad-hoc methods are utilized.
4. Fuzzy control approaches are appropriate only for trivial issue which do not need high accuracy..

Related work

Numerical ARM by means of multi-objective GA [15] Multi-objective GAprocedure for mining association rules for numerical information.Numerous measures are well-defined in order to determine much effective rules [16]. Lastly, the best rules is found by Pareto optimality [17]. This technique is based on the notion of irregular patterns that use irregular values defined by lower and upper intervals to represent a range or collection of values.

Numerical association rule mining approaches-

1. Discretization- means we can divide it into intervals. Eg. For 0-100, we have a value 49. So it comes under interval 45-

50. So it will have value 1 and other intervals will have 0 values.

2. Distribution of numerical value

3. Optimization.

Multi-objective rule mining problems [18]

1. **Confidence-** SUP(A union C)/ SUP(A)
2. **Comprehensibility-** log(1+ |C|)/ log(1 + |A union C|)
3. **Interestingness-** [SUP(A union C)/SUP(A)] * [SUP(A union C)/SUP(c)] * [1 – SUP(A union C)/SUP(D)]

Mutation and crossover operators

Based on the defined chromosome representation, now, mutation and crossover operators which are used in the proposed method, can be definite [19]. The place and value of chromosomes' attribute symbols (A, B, . . .)Are settled at the season of both change and crossover operations. Two label bits related with each trait are changed by bit-flip transformation.Two different numbers viewing upper and lower bounds of attribute intervals, are randomly created within the attribute range, such that the lower bound value is smaller as compared to the upper bound value. These both values can be rounded to nearest desired value (like the nearest integer). Considering the portrayal of chromosomes, for the crossover activity different discretionary kinds of 'k-point crossover' can be utilized. In the experimental outcomes section, we discuss that which type of k-point crossover leads to a perfect outcome.

Advantages- 1.Generated rules are much better than the previous rules generated.

2. We can mine numerical values also based on multi objective attributes for the first time.

3. Based on rough sets.

Mining Frequent Itemsets Using Genetic Algorithms [20]

To find frequent itemsets various algorithms like pincer, apriori have been designed till now, but they were not so efficient. So use of genetic algorithm was done. Concept of association rule mining is explained i.e. of support count, confidence, frequent patterns. GA is used for finding optimality from population, we use the concepts of crossover, mutation and selection. The fitness function is also used. Genetic keeps on going till it terminates based on a few conditions.

Performance measurement- candidate itemset $C_{k.}$

Complexity of apriori is- $C= m_k$ where $m_k= |C_k|$

On solving complexity is- $T= O(d^2n)$

Performance evaluation- 1. Likelihood of optimality

2. Average fitness value

3. Likelihood of evolution leap

$C=kr$

where k= best cut-off generation, r= no. of repeated runs

Limitations of previous work- No. of iterations are increasing due to the calculation of all itemsets thus leading to increase in time complexity. Also a factor of interestingness is considered according to user-defined value.

Advantages-

1) Greedy approach.
2) Does global search.
3) The time unpredictability is less when contrasted with different calculations.
4) Unsupervised learning

Discovering Interesting Rules from Biological Data Using Parallel Genetic Algo[21]

A new Algo has been designed using biological data and applying a PGA. Here GA plays its role , fitness function works as the threshold identifier. PGA shows that how GA works parallel. The Parallel processing concept is used in PGA.Multiple threads are created for parallel execution of fitness function and each thread then goes through the whole GA process i.e, verifying the condition, selection, crossover &mutation and generation of new population.

It summarizes the observation on master –slave GA that here additional processors are used which increases the computation time and communication time but also reduces the evaluation time of the fitness function .thus it is the trade of b/w computational & communication time[22].

Chromosomes are represented as

C1	C2	C..	C_m	C_{m+1}	C_{m+2}	C…	C_{k-1}	C_k

1. Select: It is a boolean function that checks if the value is greater than fitness(c), if yes then it one true other false.
2. Mutate: It generates the random by using rand () function.
3. Crossover: $C_i$and$C_j$are two population wher I!= j, p contains random(k+1) & q contains random(k+1) ,q contains max(p,q) and p contains(min p,q).C_3 and C_4 are generated.
4. Fitness function: {sup(C1…Ck)-sup(C1…Cm)(C_{m+1}…. Ck)}/ sup(C1….Cm) (1-SUP(C_{m+1}..Ck)
5. Population Initialization: populationTemp-> null
6. Main Algo: contains all functions defined above and thus creates a new population.

Advantages: It takes less computational cost. Parallel processing thus better to use PGA. Also gives Global optimality. Threshold value is generated on its own as it is tough to find out. It is better when the data is very large. Interesting rules are generated after evaluating the fitness value [23].

Limitations: Problem exists due to large search space. Sometimes incomplete and noisy data is produced for finding interesting rules.

ARM using Self AdaptivePSO [24]

PSO is a basic and capable populace based stochastic look algorithm for clarifying optimization issue in continuous search area. Be that as it may, the normal PSO is extra liable to stall out at a nearby ideal and along these lines prompting untimely union when solving practical issue. This paper proposes two distinctive adaptive mechanisms for altering the idleness weights to be specific self adaptive PSO1 (SAPSO1) and the particles move through the search space with a predefined speed looking for optimal solution. For D-dimensional inquiry space the speed is represented as

$v_i = (v_{i1}, v_{i2}, \dots v_{id}, \dots v_{iD})$ (1)

All moleculeskeeps up a memory which helps it in monitoring its past best position.

$x_i = (x_{i1}, x_{i2}, \dots x_{id}, \dots x_{iD})$ (2)

The personal best and global best is represented as

$p_i = (p_{i1}, p_{i2}, \dots p_{id}, \dots p_{iD})$ (3)

$p_g = (p_{i1}, p_{i2}, \dots p_{id}, \dots p_{iD})$ (4)

The particles or members of the swarm fly through a multidimensional search space solution as present figure 1.
ding velocity and position with Equations 5 and 6 as follows

$$v_i^{new} = \omega * v_i^{old} + c_1 \text{rand}(\,)(\text{pbest}-x_i) + c_2 \,\text{rand}(\,)(\text{gbest}-x_i)$$

$$x_i^{new} = x^{old} + v^{new}$$

Where v_i^{old} is the particle speed of the i^{th}particle, x_i is the present molecule, I is the molecule number, rand() is an random number in the(0,1), c_1 the individual factor and c_2 the societal factor.

Proposed solution

1. In optimization algorithms the main drawback is the code complexity. The optimization is inversely proportional to the length of the code. All the algorithms optimize the code and make the results better but they have lengthy code. The solution can be we can merge the original code with the optimization code so that the length can be reduced. As the length of the code becomes shorter and the number of iterations becomes less, the complexity of the code decreases.
2. Instead of fuzzy we can apply rough sets or soft sets as they have lesser boundations and easier implementations as compared to fuzzy. Fuzzy has the drawback of multiple membership functions to work upon.

Conclusion

Data mining is explained with the important fields that are in use and have various improvement areas. The Genetic algorithm is used to find various patterns and is applied to the data to find out the processing time and makes the implementation easier. Genetic algorithms can be applied in various fields and this has been explained in the related work.

The pattern discovery and rule generation should be possible with the assistance of GA Parallel processing and biological data processing can also be done with the help of genetic algorithm. Various optimization techniques have also been discussed. One of them is particle swarm optimization, which is explained with the various works done in this field. Ant colony optimization is also a type of optimization.

References

[1]. M. B. Malik, M. A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", in proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012.

[2]. Discovery in Databases. 0738-4602-1996, AI Magazine (Fall 1996).pp: 37–53.

[3]. J. Han and M. Kamber, Data Mining: Concepts and Techniques. Second edition Morgan Kaufmann Publishers.

[4]. Lindell Y., Pinkas B.: Privacy-Preserving Data Mining. CRYPTO, 2000.

[5]. Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.

[6]. Ackley, D., and Littman, M. 1992. Interactions between learning and evolution.In C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen, eds., Artificial Life II.Addison−Wesley.

[7]. Ackley, D., and Littman, M. 1994. A case for Lamarckian evolution.In C. G. Langton, ed., Artificial Life III, Addison−Wesley.

[8]. Altenberg, L. 1995. The Schema Theorem and Price's Theorem. In L. D. Whitley and M. D. Vose, eds, Foundations of Genetic Algorithms 3. Morgan Kaufmann.

[9]. S.Deepa et al: "An Optimization of Association Rule Mining Algorithm using Weighted Quantum behaved PSO", International Journal of Power Control Signal and Computation(IJPCSC) Vol3. No1. Jan-Mar 2012 ISSN: 0976-268X.

[10]. J. Han, G. Dong, and Y. Yin.Efficient mining paritial periodic patterns in time series database. In Proc. of the15th International Conference on Data Engineering,,pages 106–115, 1999.

[11]. Yiyu Yao, "ROUGH SET APPROXIMATIONS: A CONCEPT ANALYSIS POINT OF VIEW", University of Regina, Regina, Saskatchewan, Canada, 2015.

[12]. En-Bing Lin and Yu-RuSyau, "Comparisons between Rough Set Based and Computational Applications in Data Mining", International Journal of Machine Learning and Computing, Vol. 4, No. 4, August 2014.

[13]. Türks¸en, I.B. and Tian Y. 1993. Combination of rules and their consequences in fuzzy expert systems, Fuzzy Sets and Systems, No. 58,3-40, 1993.

[14]. http://www.cs.cmu.edu/Groups/AI/html/faqs/ai/fuzzy /part1/faq-doc-4.html

[15]. Wai-HO AU, Keith C.C. Chan: An Effective Algorithm for Discovering Fuzzy Rules in Relational Databases, Fuzzy Systems Proceedings, IEEE World Congress on Computational Intelligence. Volume 2.ISSN: 1098-7584, Print ISBN: 0-7803-4863-X, Page(s):1314 – 1319, 1998.

[16]. Delgado, Miguel: Fuzzy Association Rules: an Overview. BISC Conference, 2003.

[17]. C.M. Fonseca, P.L. Fleming, An overview of evolutionary algorithms in multi-objective optimization, Evolutionary Computation 3 (1995) 1–16.

[18]. C.M. Fonseca, P.L. Fleming, Genetic algorithms for multi-objective optimization: formulation, discussion and generalization, in: Fifth International Conference on Genetic Algorithms 1993, pp. 416–423.

[19]. A.A. Freitas, A survey of evolutionary algorithms for data mining and knowledge discovery, in: A. Ghosh, S. Tsutsui (Eds.), Advances in Evolutionary Computing, Springer-Verlag, New York, 2003, pp. 819–845.

[20]. A.A. Freitas, Data Mining and Knowledge Discovery with Evolutionary Algorithms, Springer-Verlag, New York, 2002.

[21]. J Grefenstette, Optimization of control parameters for genetic algorithms, IEEE Transactions on Systems, Man and Cybernetics, v.16 n.1, p.122-128, Jan./Feb. 1986 .

[22]. E. Cantú-Paz. "A Summary of Research on Parallel Genetic Algorithms". R. 95007, July 1995. revised version, IlliGAL R. 97003. May 1997.

[23]. A. Chipperfield, P. Fleming. "Parallel Genetic Algorithms". Parallel and Distributed Computing Handbook, A. Y. H. Zomaya (ed.), MacGraw-Hill, pp. 1118-1143. 1996.

[24]. A. Grajdeanu. Parallel Models for Evolutionary Algorithms.ECLab, George Mason University, 38, 2003.