International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
Volume 5 Issue 3 May 2014

AN EFFECTIVE ALGORITHMIC APPROACH FOR ASSOCIATIVE CLASSIFICATION USING METAHEURISTICS

Anuradha Rehal, M.Tech. Scholar,
Department of Computer Science and Applications
Kurukshetra University, Kurukshetra

ABSTRACT

Data mining refers the extraction of hidden predictive information from large databases. It is a powerful technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can respond to the business questions traditionally too time consuming to resolve and search databases for hidden patterns, finding predictive information. Efficient and perfect solution to the combinatorial optimization problems in different streams have been an area of research from long time. Engineering, Industrial, Economical and Scientific problems are solved with various approaches such as Simulated Annealing, Tabu Search, Genetic Algorithms, Ant Colony Optimization, Harmony Search, Scatter Search or Iterated Local Search. These techniques known as metaheuristics presents itself as highly promising choice for nearly-optimal solutions in reasonable time where exact approaches are not applicable due to extremely large running times or other limitations. Metaheuristic is an excellent strategy that guides and modifies other heuristics to produce solutions beyond those that are normally generated in a quest for local optimality. This paper highlights the efficiency of a metaheuristic approach, Ant Colony Optimization, which is being used by the researchers now days in several Engineering, Scientific, Business and Industrial applications. For experimental and simulation purpose, the process of data farming is used using a well famed suite weka. In this manuscript, an empirical approach is used for extraction of better results against classical association rule algorithms.

Keywords – Data Mining, Associative Classification, Association Rule Mining, heuristics, metaheuristics.

I. INTRODUCTION

The data mining has attracted a great deal of attention due to the wide availability of huge number of data. It is a cooperative effort of humans and computers. Humans design databases, describe problems and set goals. Computers thoroughly examine data, looking for patterns that match these goals. Data Mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large databases [1]. The objective of this process is to sort through large quantities of data and discover new information. The benefit of data mining is to turn this newfound knowledge into actionable results, such as increasing a customer's likelihood to buy [2]. It is currently regarded as key element of much more elaborate process called Knowledge Discovery in Databases (KDD).

The term Data Mining can be used to describe a wide range of activities. A marketing company using historical

response data to build models to predict who will respond to a direct mail or telephone solicitation is using data mining. A manufacturer analyzing sensor data to isolate conditions that lead to unplanned production stoppages is also using data mining [3]. Other terms in the literature are sometimes used to describe this process in addition to Data Mining. Among these are, knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

Data mining models produces one or more output values for a given set of inputs. Analyzing data is often the process of building an appropriate model for the data. It is an abstract representation of reality. Models in Data Mining are either Predictive or Descriptive. Prediction involves using some variables in data sets in order to predict unknown values of other relevant variables (e.g. classification, regression, and anomaly detection). Description involves finding human understandable patterns and trends in the data (e.g. clustering, association rule learning, and summarization).

II. ASSORTED DATA MINING MODELS

- Classification: This model is used to classify database records into a number of predefined classes based on certain criteria. For example, a credit card company may classify customer records as a good, medium, or poor risk. A classification system may then generate a rule stating that "If a customer earns more than Rupees 40,000, is between 45 to 55 in age, and lives within a particular ZIP code, he or she is a good credit risk."[4]
- Regression: This model is used for the analysis of the dependency of some attribute values upon the values of other attributes in the same item, and the automatic prediction of these attribute values for new records. It maps a data item to a real-valued prediction variable. For example, given a data set of credit card transactions, build a model that can predict the likelihood of fraudulence for new transactions.
- Time Series: This model describes a series of values of some feature or event that are recorded over time [5]. The series may consist of only a list of measurements, giving the appearance of a single dimension, but the ordering is by the implicit variable, time. The model is used to describe the component features of a series data set so that it can be accurately and completely characterized.
- Clustering: This model is used to divide the database into subsets, or Clusters, based on a set of attributes. For example, in the process of understanding its customer base, an organization may attempt to divide the known population to discover clusters of potential customers based on attributes never before used for this kind of analysis (for example, the type of schools they attended, the number of vacation per year, and so on). Clusters can be created either statistically or by using artificial intelligence methods. Clusters can be analyzed automatically by a program or by using visualization techniques.
- Association Rule Learning / Dependency Modelling: This model is used to describe significant dependencies between variables. It identifies affinities among the collection, as reflected in the examined records. These affinities are often expressed as rules [6]. For example: "60% of all the records that contain items A and B also contain items C and D." The percentage of occurrences (in this case, 60) is the confidence factor of association. Association model is often applied to Market Basket Analysis, where it uses point-of-sale transaction data to identify product affinities.
- Sequencing: This model is used to identify patterns over time, thus allowing, for example, an analysis of customer purchases during separate visits. It could be

found, for instance, that if a customer buys engine oil and filter during one visit, he will buy gasoline additive the next time [26].

- Characterization / Summarization: This model is used to summarize the general characteristics or features of a target class of data. The data corresponding to the user-defined class are typically collected by a database query. For example, to study the characteristics of software products whose sales increased by 10% in the last year, the data related to such products can be collected [7],
- Comparison / Anomaly Detection: The model is used for comparison of the general features of target class data objects with the general features of objects from one or a set of comparative (contrasting) classes [8]. For example, the model can be used to compare the general features of software products whose sales increased by 10% in the last year with those whose sales decreased by at least 30% during the same period.

Data mining has been proved as a very basic tool in knowledge discovery and decision making process. Data mining technologies are very frequently used in a variety of applications. Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers and clusters [25]. Frequent patterns are the itemsets that are frequently visited in database transactions at least for the user defined number of times which is known as support threshold [22]. Presently, a number of algorithms have been proposed in literature to enhance the performance of Apriori Algorithm, for the purpose of determining the frequent pattern [9]. The major concern for any algorithm is to reduce the processing time. Knowledge Discovery in Databases (KDD) and Data Mining (DM) helps to extract useful information from raw data. Association rules describe how often items are purchased together. Such rules can be useful for decisions concerning product pricing, promotions, store layout and many others [10].

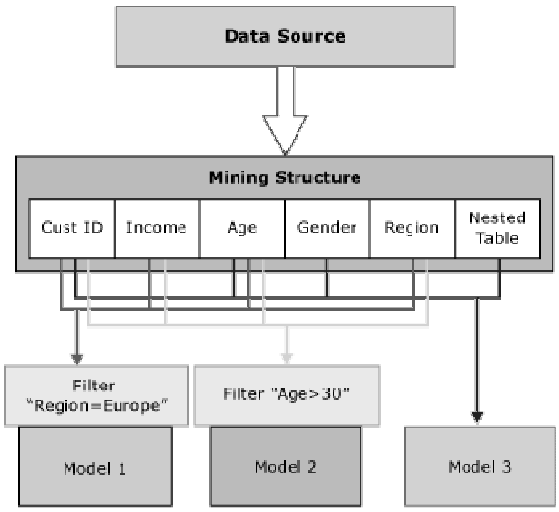

Figure 1: Collection of models from different sources

In Figure 1, it is depicted that the data is collected from multiple sources. It is integrated and finally placed in some common data store [23]. Module of it is then taken and pre-processed into a standard format. This 'prepared data' is then passed to a data mining algorithm which produces an output in the form of rules or some other kind of 'patterns' [8].

III. GENERAL METHODOLOGY TO SOLVE A PROBLEM

1. Exact Algorithms
2. Heuristic Algorithms
3. Approximation Algorithms
4. Streaming Algorithms
5. Online Algorithms

Metaheuristics are used to solve Combinatorial Optimization Problems, like Bin Packing, Network Routing, Network Design, Assignment Problem, Scheduling, or Time-Tabling Problems, Continuous

Parameter Optimization Problems, or Optimization of Non-Linear Structures like Neural Networks or Tree Structures as they often appear in Computational Intelligence [11]. Metaheuristics are generally applied to problems for which there is no satisfactory problem-specific algorithm or heuristic; or when it is not practical to implement such a method [24]. Most commonly used Metaheuristics are focused to combinatorial optimization problems, but obviously can handle any problem that can be recast in that form, such as solving Boolean equations [12].

IV. HEURISTICS AND METAHEURISTICS

Heuristic refers to "discover". A Heuristic is used when

1. Exact method are not on any help, due to execution time
2. There are errors in input data or is unreliable
3. Improvement in the performance of exact methods is required
4. There is need of a solution after a limited period of time.
5. We have to choose between addressing a more realistic model and provide an approximate solution instead of a simpler, unrealistic model that we can prove that can solve to optimality.
6. There is need of good starting points for an exact method.

V. DISADVANTAGES OF USING HEURISTICS

1. In many cases, convergence is generally guaranteed.
2. Optimality may be achieved but it is not proved.
3. In many cases, they may not be able to generate a feasible solution.

Metaheuristics are said to be high level procedures which coordinate simple heuristics such as local search, to find solutions that are of better quality than those found by simple heuristics done [13].

VI. COMMONLY USED METAHEURISTICS

- Ant Colony Optimization, Dorigo 1991
- Tabu search [Glover, 89 et 90]
- Simulated Annealing [Kirckpatrick, 83]
- Threshold accepting [Deuck, Scheuer, 90]
- Variable neighborhood [Hansen, Mladenovi´c, 98]
- Iterated local search [Loren¸co et al, 2000]
  Genetic Algorithm, Holland 1975 – Goldberg 1989
- Memetic Algorithm, Moscatto 1989
- Scatter search, Laguna, Glover, Marty 2000
- Artificial Bee Colony Algorithm, 2005

Countless variants and hybrids of these techniques have been proposed, and many more applications of Metaheuristics to specific problems have been reported. This is one of the active fields of research, with a considerable literature, a large community of researchers and users, and a wide range of applications [14] [15].

Traditional methods of search and optimization are too slow in finding a solution in a very complex search space, even implemented in supercomputers [16]. Metaheuristics consist of number of methods and theories having robust search method requiring little information to search effectively in a large or poorly-understood search space. There exists an extensive range of problems which can be formulated as obtaining the values for a vector of variables subject to some restrictions [17] [18]. The elements of this vector are denominated decision-variables, and their nature determines a classification of this kind of problems [19]. Specifically, if decision-variables are required to be discrete, the problem is said to be combinatorial. The process of finding optimal solutions (maximizing or minimizing an objective function) for such a problem is called combinatorial optimization.

VII. CONSTRUCTIVE BASED METAHEURISTIC

1. ANT COLONY OPTIMIZATION
2. CROSS ENTROPY METHOD
3. SEMI GREEDY
4. GRASP

VIII. APPLICATION OF ANT COLONY OPTIMIZATION IN COMMON PROBLEMS

- Vehicle routing problem
- Traveling salesman problem
- Minimum spanning tree problem
- Linear programming (if the solution space is the choice of which variables to make basic)
- Eight queens puzzle. (A constraint satisfaction problem. When applying standard combinatorial optimization algorithms to this problem, one would usually treat the goal function as the number of unsatisfied constraints (say number of attacks) rather than as a single boolean indicating whether the whole problem is satisfied or not.)
- Knapsack problem

IX. MAIN FEATURES OF A GOOD METAHEURISTICS

- Population intrinsic parallelism
- Indirect Coding
- Cooperation adapted crossover
- Local search in solution space
- Diversity need to be controlled
- Easy to implement the restarts
- Randomness

X. PROPOSED APPROACH

The Classical Association Rule Mining Apriori Algorithm has few drawbacks such as; the iterations involved reduce the minimum support until it finds the required number of rules with the given minimum confidence. The traditional approach can be improved by overriding some trade-off phases and discarding the unwanted objects and fields from the association analysis [20] [21]. The Apriori algorithm needs deep analysis, review as well as revision in terms of the inefficiencies or trade-offs for assorted applications.

a. CLASSICAL / EXISTING APPROACH

The Transactional Data from a web server log file is fetched and listed in Table according to the web server attributes.

Table 1 : Web Server Transactional Data

	Mailserver	HttpService	ThirdpartyAPI
		HttpService	PageNotFound
		HttpService	FTPService
	Mailserver	HttpService	PageNotFound
		Mailserver	FTPService
		HttpService	FTPService
		Mailserver	FTPService
Mailserver	HttpService	FTPService	ThirdpartyAPI
	Mailserver	HttpService	FTPService

To find frequent Itemsets, candidate generation is performed with 1-Itemset occurrences.

Table 2 : Itemset Occurrences

1-Itemset	**Support**
Mailserver	6
HttpService	7
FTPService	6
ThirdpartyAPI	2
PageNotFound	2

Since every Itemset satisfies the minimum support level, therefore none of the Itemsets will be pruned from the above candidate database. To fetch frequent patterns the

2-Itemset Combinations are generated next. The combinations are listed in Table.

Table 3 : 2-Itemset Occurrences

2-Itemset	Support
Mailserver, HttpService	4
Mailserver, ThirdpartyAPI	2
Mailserver, FTPService	4
Mailserver, PageNotFound	1
HttpService, FTPService	4
HttpService, ThirdpartyAPI	2
HttpService, PageNotFound	2
FTPService, ThirdpartyAPI	1
FTPService, PageNotFound	0
PageNotFound, ThirdpartyAPI	0

The above table consists of some combinations that do not satisfy minimum support level, therefore those Itemset – Combinations are pruned. The pruned database is listed.

Table 4: Pruned Database of 2-Itemset Combinations

2-Itemset	Support
Mailserver, HttpService	4
Mailserver, ThirdpartyAPI	2
Mailserver, FTPService	4
HttpService, FTPService	4
HttpService, ThirdpartyAPI	2
HttpService, PageNotFound	2

To fetch frequent patterns from the 2-Itemset Combinations that satisfy a minimum support combinations of 3-Itemset are generated next. The combinations are listed in Table and are checked for minimum support level for further pruning of dataset.

Table 5: 3-Itemset Occurrences

3-Itemset	Support
Mailserver, HttpService, ThirdpartyAPI	2
Mailserver, HttpService, FTPService	2
Mailserver, HttpService, PageNotFound	1

The above table consists of some combinations that do not satisfy minimum support level, therefore those Itemset – Combinations are pruned. The pruned database is listed in Table 6.

Table 6: Pruned Database of 3-Itemset Combinations

3-Itemset	Support
Mailserver, HttpService, ThirdpartyAPI	2
Mailserver, HttpService, FTPService	2

As there are no further combinations possible from Itemsets in Table, the candidate generation for 4-Itemset will not take place. Apriori terminates at this stage.

b. PROPOSED APPROACH USING ANT COLONY OPTIMIZATION

procedure I-Rule Mining (T, RSupport) { //T is the database and RSupport is the support
L_1= {frequent items};
RIs={Required Item Set Support}

for (k = 3; L_{k-1} != RIs; k++) {
C_k= candidates generated from L_{k-1}
//that is cartesian product L_{k-1} x L_{k-1} and eliminating any k-1 size itemset that is not
//frequent

for each transaction t in database do{
#increment the count of all candidates in Ck that are contained in t
L_k = candidates in Ck with RSupport
}//end for each

}//end for

return U_k L_k ;
}

To find association rules form the same database using Reverse-Apriori Transactional Database listed in Table.

To find frequent Itemsets, in Reverse Apriori the candidate generation is performed with 4-Itemset occurrences; the 5-Itemset generation combinations are found and listed in Table.

Table 7: 5-Itemset combinations

5-Itemset	**Support**
Mailserver, HttpService, FTPService, ThirdpartyAPI, PageNotFound	**0**

Since the 5-Itemset combinations generated does not satisfy the minimum support as listed in Table. The 4-Itemset combinations have to be generated.

Table 8: 4-Itemset combinations

4-Itemset	**Support**
Mailserver, HttpService, FTPService, ThirdpartyAPI	1
Mailserver, HttpService, FTPService, PageNotFound	0
Mailserver, HttpService, ThirdpartyAPI, PageNotFound	0
Mailserver, FTPService, ThirdpartyAPI, PageNotFound	0
HttpService, FTPService, ThirdpartyAPI, PageNotFound	0

As in the above Table none of the 4-Itemset combinations satisfy the minimum support so all the combinations will be pruned. The 3-Itemset combinations have to be generated.

Table 9: 3-Itemset Combinations

3-Itemset	**Support**
Mailserver, HttpService, FTPService	2
Mailserver, HttpService, PageNotFound	1
Mailserver, HttpService, ThirdpartyAPI	2
FTPService, ThirdpartyAPI, PageNotFound	0
HttpService, FTPService, ThirdpartyAPI	1
HttpService, FTPService, PageNotFound	0

There are some combinations in Table which satisfy minimum support. All those itemsets which does not satisfy minimum support are pruned and are listed.

Table 10 : Pruned Dataset

3-Itemset	**Support**
Mailserver, HttpService, FTPService	2
Mailserver, HttpService, ThirdpartyAPI	2

XI. CONCLUSION AND FUTURE WORK

This manuscript includes preprocessing, transaction identification, and data integration components. A lot of advancements are being pursued by the researchers in finding exact solutions to the combinatorial optimization problems using techniques such as integer programming, dynamic programming, cutting planes, and branch and cut methods. Still there are many hard combinatorial problems which are unsolved and needs good heuristic methods. The solutions obtained as "Optimal Solutions" is in many cases are not as per the requirements. The goal and objective of using a metaheuristics approach is to produce efficient solutions. The metaheuristics such as Ant Colony Optimization is one of the most popular approaches to explore in the field of optimization and it will bring wonder to the world of algorithms in future. In this manuscript, the server log files are analyzed for investigation of results with the proposed approach. In future this proposed metaheuristics can be further optimized to get more efficient results with different techniques of data mining..

XII. REFERENCES

[1] A Survey of Various Association Rule Mining Approaches, Jyotsana Dixit Abha Choubey, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4 Issue 3 March 2014

[2] Association Rule Mining Algorithm: A Review, NCI2TM: 2014, Shikha Dubey, Dr. Shivaji D. Mundhe, 2014, International Journal on Natural Language Computing (IJNLC) Vol. 3, No.1, February 2014, DOI : 10.5121/ijnlc.2014.3103 21

[3] An Improved Apriori Algorithm For Association Rules, Mohammed Al-Maolegi, Bassam Arkok, Computer Science, Jordan University of Science and Technology, Irbid, Jordan, 2014

[4] International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2014, A Survey on Association Rule Mining, T. Karthikeyan1 and N. Ravikumar, Associate Professor, Department of Computer Science, PSG College of Arts and Science, Coimbatore, India, Research Scholar, Department of Computer Science, Karpagam University, Coimbatore, India.

[5] Lim, T. S., Loh, W. Y., and Shih, Y. S.,. "A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms", Machine Learning, 2000.

[6] John, G. H. and Langley, P., "Estimating Continuous Distributions in Bayesian Classifiers", Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. pp. 338- 345. Morgan Kaufmann, San Mateo, 1995.

[7] Maragatham G. and Lakshmi M, "A RECENT REVIEW ON ASSOCIATION RULE MINING", International Journal of Computer Science and Engineering, ISSN : 0976-5166, Vol. 2, No. 6, pp831-836, Dec 2011-Jan 2012.

[8] Bing Liu, Wynne Hsu, Yiming Ma, "Integrating Classification and Association Rule Mining", Appeared in KDD-98, New York, Aug 27-31, 1998.

[9] A. Zemirline, L. Lecornu, B. Solaiman, and A. Echcherif, "An Efficient, and J.S. Usher, "Facility layout using swarm intelligence," in Proceedings of IEEE Swarm Intelligence Symposium, pp. 424-427, June 2005.

[10] S. Lorpunmanee, M.N. Sap, A.H. Abdullah, and C. Chompoo-inwai, "An ant colony optimization for dynamic job scheduling in grid environment," World Academy of Science, Engineering and Technology, pp. 314-321, 2007.

[11] B. Chakraborty, "Feature subset selection by particle Association Rule Mining Algorithm for Classification", ICAISC 2008, LNAI 5097, pp.

717–728, 2008, Springer-Verlag Berlin Heidelberg, 2008.

[12] Marco Dorigo and Thomas Stutzle, "Ant Colony Optimization", ISBN-81-203-2684-9and Edition 2004.

[13] A. Abraham, C. Grosan, and V. Ramos, "Swarm Intelligence in Data Mining," Studies in Computational Intelligence, Vol. 34, pp. 1-20, Springer 2006.

[14] C.T. Hardinswarm optimization with fuzzy fitness function," in 3rd International Conference on Intelligent System and Knowledge Engineering, ISKE, pp. 1038-1042, 2008.

[15] K. Mong Si, and W. Hong Sun, "Multiple ant-colony optimization for network routing," in First International Symposium on Cyber Worlds Proceedings, pp. 277-281, 2002.

[16] X. Tan, X. Luo Chen, and W.N. Jun Zhang, "Ant colony system for optimizing vehicle routing problem with time windows," International Conference on Computational Intelligence for Modeling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, pp. 209-214, 2005.

[17] E. Salari, and K. Eshghi "An ACO algorithm for graph coloring problem," Congress on Computational Intelligence Methods and Applications, pp. 659-666, 2005.

[18] M. Lee, S. Kim, W. Cho, S. Park, and J. Lim, "Segmentation of brain MR images using an ant colony optimization algorithm," Ninth IEEE International Conference on Bioinformatics and Bioengineering, pp. 366-369, 2009.

[19] Z. H. Deng and S. L. Lv. Fast mining frequent itemsets using Nodesets.[2]. Expert Systems with Applications, 41(10): 4505–4512, 2014.

[20] Z. H. Deng, Z. Wang , and J. Jiang. A New Algorithm for Fast Mining Frequent Itemsets Using N-Lists [3]. SCIENCE CHINA Information Sciences, 55 (9): 2008 - 2030, 2012.

[21] Z. H. Deng and Z. Wang. A New Fast Vertical Method for Mining Frequent Patterns [4]. International Journal of Computational Intelligence Systems, 3(6): 733 - 744, 2010.

[22] Rauch, Jan; Logical calculi for knowledge discovery in databases, in Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery, Springer, 1997, pp. 47-57

[23] Hájek, Petr; Havránek, Tomáš (1978). Mechanizing Hypothesis Formation: Mathematical Foundations for a General Theory. Springer-Verlag. ISBN 3-540-08738-9.

[24] Webb, Geoffrey I. (1995); OPUS: An Efficient Admissible Algorithm for Unordered Search, Journal of Artificial Intelligence Research 3, Menlo Park, CA: AAAI Press, pp. 431-465 online access

[25] Bayardo, Roberto J., Jr.; Agrawal, Rakesh; Gunopulos, Dimitrios (2000). "Constraint-based rule mining in large, dense databases". Data Mining and Knowledge Discovery 4 (2): 217–240. doi:10.1023/A:1009895914772.

[26] Webb, Geoffrey I. (2000); Efficient Search for Association Rules, in Ramakrishnan, Raghu; and Stolfo, Sal; eds.; Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000), Boston, MA, New York, NY: The Association for Computing Machinery, pp. 99-107 online access