DOI: 10.21917/ijsc.2020.0317

DISCOVERING POSITIVE ASSOCIATION OF ASD ATTRIBUTES WITH CLASS USING MULTI OBJECTIVE CULTURAL ALGORITHM

R. Abitha and S. Mary Vennila
Department of Computer Science, Presidency College, India

Abstract

Association rule mining (ARM) is a common and most preferable research method for bringing out the fascinating relations between the variables provided in any data set. It brings out the knowledge by satisfying the user defined values and criteria measures specified by the researcher. Frequent item set generation is a well-known method carried out by many researchers to retrieve interesting correlations among the variables that helps in decision making. The accuracy of the brought out rules by ARM is good enough to provide a conclusion on research studies. This can be improved by incorporating optimization like heuristic search techniques. In this paper cultural algorithm is used to improve the performance of rule mining by optimization which is required for categorizing the risk level of ASD individuals. Optimization is utilized in health care domain for generating optimized rules to analyze the frequently combined attributes among the patient's data. It gracefully improves the result finding process which will be tranquil to conclude the decision. The Cultural algorithm fit in to the larger course of evolutionary algorithms that is inspired by natural evolution. In this paper multi objective optimization technique is proposed by incorporating ARM and cultural algorithm by considering different objectives namely support, confidence, lift and completeness of the rule to find the positive association of Autism Spectrum Disorder (ASD) screening data features with positive class. The result of this research depicts the positive association with improved performance along with reduced number of rules.

Keywords:

Association Rule Mining, ASD, Apriori Algorithm, Rule Generation, Cultural Algorithm, Risk Level, Optimization

1. INTRODUCTION

Discovering the association between the variables in a given data source using the data mining techniques is a powerful mechanism [14] in almost all domains like health care, marketing, education biotechnology and to name a few. Data mining is a knowledge discovery step where interesting, hidden, unknown data patterns are pull out by applying intelligent techniques. In this paper Association Rule Mining is utilized with heuristic algorithm to improve the output in the search space [15]. Autism spectrum disorder is related with lots of modifiable and non-modifiable factors. Age, gender and history are non-modifiable and behavioral, genetic, environmental factors are considered as modifiable factors. This technique is applied on bench mark ASD data set and collected data, that have been collected for research purpose and applied CBARG [5] algorithm proposed by the authors to find the interesting optimized rules. The proposed algorithm results in reduced number of rule generation and reporting the strong association of attributes with class that will be discussed in the following section. And finally how these generated set of rules are used in categorizing the individual based on their risk is also shown in Fig.2 and Fig.3.

2. AUTISM SPECTRUM DISORDER (ASD)

ASD is not a disease but a disorder of composite developmental condition that has tenacious challenges in five common areas namely social interaction, communication, repetitive behavior, response and routine. In the world one in 160 children is affected with ASD [1]. Children with autism differ in their behavior and possess different symptoms. At the same time the severity of symptoms varies from mild to severe. The risk level of the children may change over a period. The symptoms can be observed by 2 to 3 years in children. But it is very hard to do the diagnosis since some children records in our collected data though they exhibit some of the qualities of ASD but they are normal. So it cannot be designated as autistic child by considering few characteristics. In this regard this paper aims to present an attributes which are to be preferably considered in diagnosing autistic traits using association rule mining with cultural algorithm (CA).

The children with ASD suffer from problems like sensory issues, cognitive abilities and emotional difficulties. But proper and periodic training helps them to some extent in their regular life style. The structured training can be given once the child is diagnosed with ASD. The first step is the screening process where the data mining and machine learning algorithms can be employed. In this study this CBARG algorithm can be useful in predicting the pattern of attributes that plays major role in the positive traits of ASD and in turn useful for categorizing risk.

3. ASSOCIATION RULE MINING (ARM)

Association Rule mining is a process of generating hidden, interesting correlations among the set of items or variables given in the data set. It aims to present the subset of items and attributes that most frequently occur in the transactions. ARM is envisioned to produce strong rules from the given transactional data base according to user defined interesting measures.

The form of association rule mining is $A \rightarrow B$ where the left hand side part 'A' is called antecedent and right hand side 'b' is called consequent and both 'A' and 'B' are separate frequent item sets occur in the data set. The expression $A \rightarrow B$ is known as "If an item set 'A' occur in a transaction then item set 'B' also will occur in the same transaction. It represents the association between the variables based on the value of "support" and "confidence" measure. Rules that satisfy the minimum support and minimum confidence value set by the user only will be generated The main issues considered with Association rule mining are: (1) ARM confines to generates more number of rules that are not easy to understand. As the number of rules grow based on the number of attributes and records considered in data base it suffers from algorithmic complexity. (2) How to excerpt the

interesting strong rules from the more repeatedly generated rules. In this paper we propose a technique by combining cultural algorithm with ARM to produce less number of optimized rules as the CA achieving global optima. And strong rules can be extracted by considering multiple objectives support, confidence and lift as measures.

- **Support:** The support of an item set *A*, sup(*A*) is the proportion of transaction in the database in which the item A appears. It signifies the popularity of an item set [6].

$$\text{Sup}(A) = (\text{No. of transaction A appears})/(\text{Total number of transactions}) \quad (1)$$

If support(A) ≥ Min_support, user defined threshold value then A is known as frequent item set.

- **Confidence:** The confidence is the proportion of transaction that includes item set *A* which also includes item set *B* the Confidence for a rule is given as follows

$$\text{Conf}(A \rightarrow B) = (\text{supp}(A \cup B))/(\text{supp}(A)) \quad (2)$$

Rules with Support($A \rightarrow B$) ≥ Min_Support and Conf($A \rightarrow B$)≥ Min_Confidence user defined threshold value then Rule ($A \rightarrow B$) is said to be strong rules [6].

- **Lift:** Lift of a rule is given by

$$\text{Lift}(A \rightarrow B)(\text{Supp}(A \cup B))/(\text{Supp}(A) * \text{Supp}(B)) \quad (3)$$

4. LITERATURE REVIEW

Abitha et al. [5] proposed CBARG to generate optimized reduced number of rules to retrieve strong rules from ASD data set. The performance over memory utilization and running time are also shown.

Son et al. [6] specified performance of the algorithm is degraded due to irrelevant rules and his proposed algorithm has minimized the running time by 39% and number of rules is reduced by 52% due to optimization.

Gupta et al. [7] designed a method using multi objective feature of genetic algorithm to generate association rules and tested on Adult, chess, zoo, wine data set and achieved maximum accuracy with four different measures support, confidence, completeness and interestingness. The efficiency of the algorithm was proved by generating reduced number of rules.

Anandhvalli et al. [8] presented a technique to find all the potential optimized rules from given data source using genetic algorithm. They designed a system which can predict the rules that contain negative rule in the generated rules along with more than one attribute in consequent body [7].

Thabtah et al. [9] designed RML algorithm to present non-redundant rule with covering learning. The empirical evaluation on different ASD dataset was done to show improved performance of RML that has useful rules needed for ASD classification.

Kuo et al. [10] proposed particle swarm optimization (PSO) based ARM algorithm (ARM–PSO) in an application of stock market to gauge speculation conduct and stock class buying [9].

Sathya et al. [11] introduced ECLAT and PSO based mining on association rule mining using éclat to improve sales in super market and proved with good accuracy and minimization of rule generation with optimization.

Wakabi-Waiswa et al. [12] derived a system known as MOGAMAR to generate high quality association rules with five quality metrics namely support, confidence, lift, J-measures and interestingness.

Kou et al. [13] proposed BPSOARM where without predefined thresholds of minimum support and confidents derived effective rules for real world industry data.

5. CULTURAL ALGORITHM (CA)

Cultural algorithm belongs to the family of an evolutionary algorithm that was introduced by Reynolds [2] in 1994 inspired by the social learning. It mimics the natural evolution and a powerful technique used for optimization problem [3]. Cultural algorithm is principally a global optimization technique which consists of an evolutionary population space whose experiences are integrated into a Belief space that in turn influences the search process to converge the problem into a solution space [4]. Cultural algorithm is an evolutionary algorithm and the mechanism is based on the strategy survival of fittest and inspired by social evolution occurring in nature [5] [15]. The original cultural algorithm has comprised of five knowledge sources in belief space, population space and the communication protocols to exchange knowledge between the macro and micro level spaces [5] [15]. In this paper CBARG algorithm proposed by the author [5] has been utilized to find out strong non-repetitive rules. Cultural algorithm uses genetic parameter such as selection cross over and mutation. A CA explores for a better solutions for a given problem by upholding updated population of candidate solution in population space and making successive generations by choosing the best solution stored in belief space and applying genetic operators to create new candidate solution. Thus CA thrives for optimal solution evolve over generations.

5.1 BELIEF SPACE

The belief space includes with five knowledge sources the Normative, Situational, Domain, Topographical, and History knowledge source [15] [5]. All these knowledge sources are used to hold different knowledge or experience collected during the process of evolution which will be useful in solving the problem. This paper suggests the modified CA where the topographical knowledge source is not used for the heuristic search [5] [15]

5.1.1 Normative Knowledge:

This knowledge source is used to store desirable range of values namely maximum and minimum list of possible values for the attributes define the population. These values are collected during the generation of rules and can be used during the cross over and mutation process [5].

5.1.2 Situational Knowledge:

Situational knowledge source stores the best instances found during the evolution process. At the end of each generation this knowledge source is updated with this exemplar and is convenient for the evolution process to search for similar pattern of the same example instead of selecting blind random population [5].

5.1.3 Domain Knowledge:

For this problem, metric values for each rule generated and fitness values are stored to evaluate each rules produced and

updated this knowledge source at the end of each generation. Here pareto optimality search principle is used by comparing the fitness values of each rule to select elite individual after each generation [5].

5.1.4 History Knowledge:

In this study of research it stores dominant individual generated at the end of each generation by comparing the fitness values stored in domain knowledge source. Since cultural algorithm is conceivable with memory by contributing these knowledge sources, but in general EAs are characterized as memory less as it does not store details about the previous generation which can be progressively changed the scenario and acquire memory in a systematic way with these different knowledge sources. Thus this history knowledge source stores elite population generated at each generation and maintain memory across the generations [5].

6. CBARG ALGORITHM

This CBARG algorithm is used to generate reduced optimized rule. It is based on cultural algorithm and association rule generation. These optimized rules are playing vital role in categorizing risk level of ASD positive children. The block diagram of CBARG is given in Fig.1.

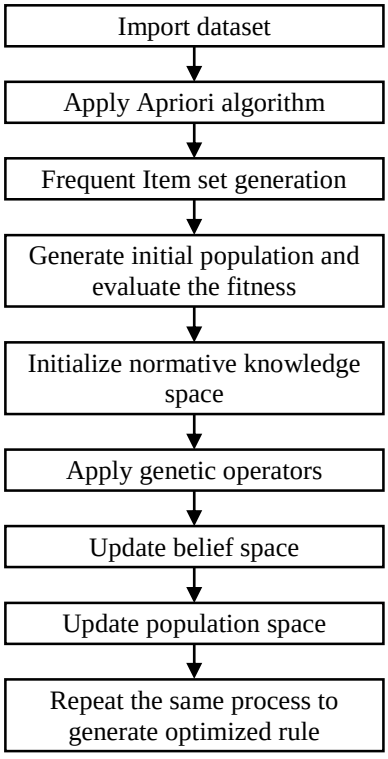

Fig.1. CBARG Block diagram

6.1 ENCODING AND INITIALIZATION

In this step antecedent and consequent part of the rules generated by Apriori algorithm are encoded using binary encoding that includes bits of strings. As found in literature, to apply evolutionary based algorithms like CA a conventional method of encoding the rules are mandatory to represent the solution in the problem space. There are two major approaches namely, Michigan's approach [16] and Pittsburg approach [16] to encode the rules. In this research Michigan's approach [16] is carried out where an individual that represent each rule is encoded. In the later approach set of rules are represented by an individual are encoded. Population initialization plays a key role in determining the overall performance of the process. Initial population is randomly generated. Elite population and other information stored in situational and normative knowledge space is used for initialization of population in subsequent generation

6.2 EVOLUTIONARY STRATEGY

6.2.1 Selection:

Based on the fitness function the selection operator selects the chromosome that has high fitness value. Based on Darwin's principle "survival of fittest" the chromosome is selected with regard of how far it is fit enough to solve the problem since this fitness is considered as a comparable measure.

6.2.2 Crossover:

In general the chromosome of the child population has pattern different from their parents to exhibit its own behavior. In this regard to show the variance from generation to generation this cross over operator is used to modify the pattern of chromosomes. In this study two point crossover is used with 80% cross over probability.

6.2.3 Mutation:

This operator operates on individuals. It mutate the specified number of gene value from '0' to '1' and '1' to '0'. And thus change the gene pattern in the individual chromosome that directs to produce better solution.

6.2.4 Fitness Function:

It is an objective function and problem dependent which is acting as a deciding criteria to choose a population to produce generations. As this study aims for multi objective optimization to extract the strong association rules, support, confidence, lift and interestingness of the rule are considered as measures for optimization. Interestingness can be mathematically formulated as follows [7].

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

where $\sigma(N)$ represents total number of transactions.

$$FF = \frac{(R_1 \times Sup) + (R_2 \times Con) + (R_3 \times lift) + (R_4 \times Interest)}{R_1 + R_2 + R_3 + R_4}$$

FF is used for fitness calculation as suggested in [7]. Based on the relative importance of support, confidence, lift and interest rank value R_1=4, R_2=3, R_3=2 and R_4=1 were assigned and ensured the obtained fitness values are between 0 to 1 [7].

7. DESCRIPTION OF THE WORK

At first CBARG algorithm is used to derive association between the attributes and the positive class. This phase of work is essential to pigeonhole the ASD affected children. By using these strong relative rules it is possible to categorize the individuals based on their risk severity. By utilizing the unsupervised version of artificial neural network categorization

will be carried out. The elementary idea is to attain rules that depicts the correlation among the attributes to predict the result.

To start with, Apriori algorithm is applied to generate rules by deriving frequent item sets. The normative knowledge source in belief space is initialized and the initial rule populations are set randomly and based on the defined measures fitness function is evaluated. The population with high fitness value are selected and the evolutionary genetic operators are applied. Then the belief space is updated that influence the population space for next generation reproduction. At the end of every generation by storing the newly created exemplar individuals the situational knowledge source in belief space is also updated.

Over a time the individuals can be substituted by their offspring that are generated by applying the genetic operators on the population. Throughout the process belief space is used as an information or knowledge repository or warehouse which stores the knowledge and experience that gained by each individuals over the generations. The process is repeated till the termination condition is reached. Here if the rules that satisfy the defined measure through the fitness function then that rule will be considered as strong rule and added in to optimized rule category.

For example if R_1 [i_1, i_4, i_5] and R_2: [i_1, i_2, i_6, i_7] are two individual rule with objective measures say [0.65,1,1.2,1] and [0.75,1,1.3,1] for rule R_1 and R_2 respectively then R_2 will be stored as strong rule based on the objective measure. Different bench mark data set ASD child, ASD toddler, ASD Adult (Table.2) and real time collected data are also used for this research.

Table.1. Sample rules generated by CBARG for collected data

Antecedent	Consequent	Lift	Interestingness	count
{i_7=y, i_{11}=y}	{*trait*=y}	1.280	0.87	545
{i_6=y, i_{13}=y, i_{12}=y, i_{10}=y}	{*trait*=y}	1.250	0.78	564
{i_4=y, i_{14}=y, i_{13}=y, i_8=y	{*trait*=y}	1.210	0.75	451
{i_7=y, i_{13}=y, i_{14}=y, i_{11}=y}	{*trait*=y}	1.308	0.70	510
{i_4=y, i_6=y, i_9=n, i_{13}=y}	{*trait*=y}	1.168	0.71	471
{i_4=y, i_7=y, i_{11}=y, i_{14}=y}	{*trait*=y}	1.328	0.69	450
{i_5=y, i_{13}=y, i_{14}=y, i_7=y}	{*trait*=y}	1.308	0.71	573
{i_5=y, i_{14}=y, i_{11}=y, i_{13}=y}	{*trait*=y}	1.000	0.70	552

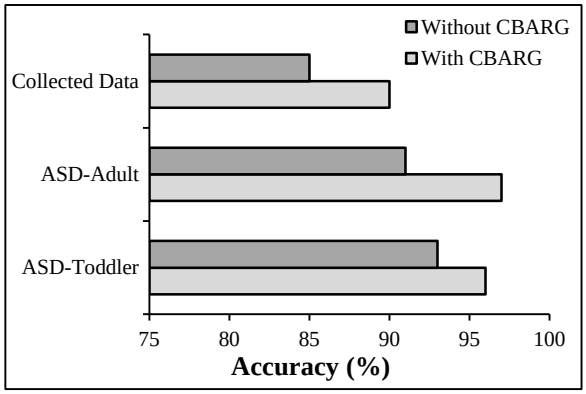

Fig.2. SOM Accuracy with and without CBARG

Table.2. Sample common rules generated by CBARG for AQ10 adult, AQ-10 child

Antecedent	Consequent	Lift	Interestingness
{A_4, A_2, A_8, A_9}	yes	1.20	.80
{A_9, A_8, A_5, A_4}	yes	1.21	.84
{A_7, A_{10}, A_2, A_9}	yes	1.03	.70
{ A_4, A_5, A_1, A_8}	yes	1.12	.76
{A_9, A_8, A_5, A_3}	yes	1.21	.78

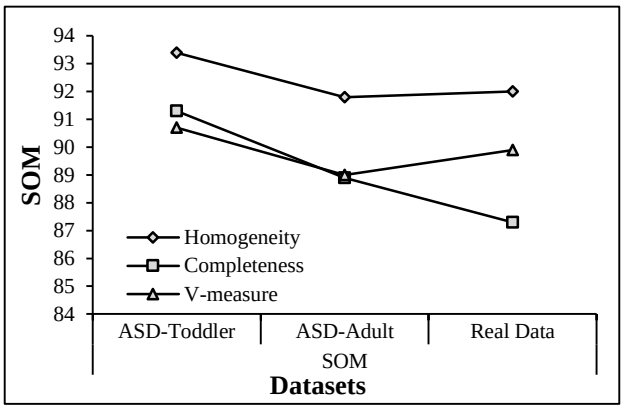

Fig.3. SOM Validation for categorization of risk severity based on the Rules generated using CBARG

8. RESULTS AND DISCUSSION

The number of rules generated from CBARG algorithm is reduced by 30% when compares with Apriori algorithm. The sample rules are given in Table.1. From this sample rules, it seems like that i_4, i_5, i_6, i_7, i_8, i_{10}, i_{11}, i_{12}, i_{13} and i_{14} have greater impact and show the positive association on the class label "Yes". These items are frequently appeared in many rules and they cover certain autistic behavior including routine, repetitive behavior, and responsiveness and communication that has to be considered while screening. The item i_9 = "n" related to intellectual function and shows though some children with positive cases does not have disability related to this attribute, but possess problem with other items. And also items i_1, i_2 and i_3 which were not included in most of the rules. This i_1, i_2, i_3 related to attributes history of the records. The very less association between the items related to history and class "Yes" is shown here. For the bench mark data the generated sample optimized rules are given in (Table.2) that include A_2, A_4, A_5, A_7, A_8, A_9 and A_{10}. These features also represent behavioral characteristics of the individual. However these predictions are made only based on the collected data and bench mark data with records of positive ASD cases. The performance of the proposed CBARG algorithm can be interpreted as good as it shows the accepted association between positive class and the mentioned autistic behavior. This work is reflected as necessary before risk level categorization of ASD positive cases using clustering technique. The Fig.2 shows the comparison of the output generated by SOM with and without the output of optimized generated rules. It is clearly reported that the output of CBARG has significant impact in categorizing the risk level of ASD.

9. CONCLUSION AND FUTURE ENHANCEMENT

In data mining, association rule mining plays vital role in bringing out the fundamental correlation between the attributes with the class value. This shows the degree of association among the features and helps in decision making. But the main limitation discussed in many research work is the vast number of generation of rules. The proposed CBARG algorithm has overcome the limitation and extracted valuable patterns that is very much useful in ASD risk level categorization. The analysis of the result shows the stronger association with the behavioral features in predicting ASD and level than features related to history. CBARG is supportive to evaluate and determine remarkable rules with the symptom patterns to find the risk level. The output of such results will be supportive for the clinicians in decision making in the field of ASD. In future these work can be extended by considering genetic and environmental features for finding the association in ASD diagnosis.

REFERENCES

[1] Autism Spectrum Disorders, Available at: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders

[2] R.G. Reynolds, "An Introduction to Cultural Algorithms", *Proceedings of 3rd Annual Conference on Evolutionary Programming*, pp 131-139, 1994.

[3] C. Chung and R.G. Reynolds, "CAEP: An Evolution-Based Tool for Real-Valued Function Optimization using Cultural Algorithms", *International Journal on Artificial Intelligence Tools*, Vol. 7, No. 3, pp. 239-291, 1998.

[4] Bidishna Bhattacharya, Kamal K. Mandal and Niladri Chakravorty, "Cultural Algorithm Based Constrained Optimization for Economic Load Dispatch Of Units Considering Different Effects", *International Journal of Soft Computing and Engineering*, Vol. 2, No. 2, pp. 1-13, 2012.

[5] R. Abitha and S. Vennila, "CBARG Cultural Based Optimized Rule Generation Method to Improve Knowledge Discovery in Autism Spectrum Disorder", *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 8, No. 6, pp. 3327-3333, 2019.

[6] Le Hoang Son, Francisco Chiclana, Raghavendra Kumar, Mamta Mittal, Manju Khari,Jyotir Moy Chatterjee and Sung Wook Baik, "ARM-AMO: An Efficient Association Rule Mining Algorithm Based on Animal Migration Optimization", *Knowledge-Based Systems*, Vol. 154, pp. 68-80, 2018.

[7] 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*, pp. 1-8, 2013.

[8] M. Anandhavalli and S. Kumar Sudhanshu, A. Kumar and M.K. Ghose, "Optimized Association Rule Mining Using Genetic Algorithm", *Advances in Information Mining*, Vol. 1, No. 2, pp. 1-4, 2009.

[9] Fadi Thabtah and David Peebles, "A New Machine Learning Model based on Induction of Rules for Autism Detection", *Health Informatics Journal*, Vol.26, No. 2, pp. 1-14, 2019.

[10] R.J. Kuo, C.M. Chao and Y.T. Chiu, "Application of Particle Swarm Optimization to Association Rule Mining", *Applied Soft Computing*, Vol. 11, No. 1, pp. 326-336, 2011.

[11] M. Sathya and K. Thangadurai, "Association Rule Generation using E-ACO Algorithm", *International Journal of Control Theory and Applications*, Vol. 27, No. 9, pp. 513-521, 2016.

[12] P. Wakabi Waiswa and V. Baryamureeba, "Mining High Quality Association Rules using Genetic Algorithms", *Proceedings of 22nd Midwest Conference on Artificial Intelligence and Cognitive Science*, pp. 73-78, 2009.

[13] Zhicong Kou and Lifeng Xi, "Binary Particle Swarm Optimization-Based Association Rule Mining for Discovering Relationships between Machine Capabilities and Product Features", *Mathematical Problems in Engineering*, Vol. 2018, pp. 1-16, 2018.

[14] J. Han and M. Kamber, "*Data Mining Concepts and Techniques*", Morgan Kaufmann, 2014.

[15] Y. Djenouri, H. Drias and Z. Habbas, "Bees Swarm Optimization using Multiple Strategies for Association Rule Mining", *International Journal of Bio-Inspired Computation*, Vol. 6, No. 4, pp. 239-249, 2014.

[16] Alex A. Freitas, "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery", Master Thesis, Department of Computer Science, Pontificia Universidade Catolica do Parana Rua Imaculada Conceicao, 2019.