Enhanced Cultural Algorithm of Data Mining for Intrusion Detection System

[1]Gurmeet Gill, [2]Prateek Gupta
[1,2]Dept. of Computer Science Engineering, SRIST Jabalpur, MP, India

Abstract
Classification rule mining is a class of problems which is the most sought out by decision makers since they produce comprehensible form of knowledge. The user should be ready to specify the properties of the foundations. The foundations discovered should have a number of these properties to render them helpful. This work proposes to enhance the cultural algorithm by application of ANT clustering mechanism to improve the performance of the system. Simple Ant clustering without cultural algorithm has also been implemented for comparison of the system.

Keywords
Cultural Algorithm, Web Data Mining, Clustering, XML, XPATH, XSL, ANT Clustering

I. Introduction

Classification rule mining could be a category of issues that is that the most sought-after out by call manufacturers since they manufacture comprehendible style of information. the principles made by the rule mining approach square measure evaluated victimization varied metrics that square measure referred to as the properties of the rule. The classification rules need to satisfy varied properties to be used as a decent classifier. The metrics typically used square measure support and confidence. But there square measure alternative properties like understandability and power of the rule that build the classifiers a lot of usable. The objectives used for analysis of rules might generally be conflicting. So the matter of constructing rules with specific properties ought to be sweet-faced as a multi-objective improvement one [1]. A cultural rule (CA) is planned for multi objective improvement of rules. a minimum of a number of the properties of the system ought to be controlled by the user to form it a lot of interactive and usable. Within the planned system the user will experiment with the system by specifying bound attributes just like the rule metrics (objectives), the rule schema, threshold for the metrics, etc. These user inputs are going to be employed by the agents within the CA for evolving optimized rules.

Cultural rule is AN organic process rule that best represents a system and consists of 2 levels of evolution: the phylogeny in a very population area and also the evolution in a very belief area. Through AN acceptance perform, the experiences of people within the population area square measure accustomed generate drawback finding information that's to be keep within the belief area. The idea area successively guides the evolution of the population area by means that of AN influence perform. Cultural algorithms are used for modeling the evolution of complicated social systems and for finding numerical improvement issues. The matter and connected work on organic process data processing for rule induction square measure mentioned in Section a pair of. Section three describes the planned cultural rule that is employed for multi objective improvement of rules. Experiments and results square measure mentioned in Section four. Section five concludes with future work.

The Problem: Given information supply and multiple objectives for improvement specifically rule metrics such by the user, the matter is presenting the user with AN optimum and novel set of rules with the desired properties.

Aims of the research:

- To require data processing generally and rule induction particularly to consecutive level by incorporating the strengths of organic process computing and Agent technology into data processing & information discovery (DM & KD).
- The sleek integration of Agent and DM for interactive DM therefore on involves users and so convert information discovered into unjust information.

II. Existing Work

Rule knowledge extraction with specific properties has received little attention in the past years, and typical rule induction algorithms tend to neglect certain desirable properties, such as the ability to induce novel knowledge [1]. Evolutionary algorithms for rule mining as a multi objective optimization problem have been proposed in the literature. Reynolds and Iglesia, [2-4] use multi-objective algorithms to induce partial classification rules and describe how the use of rule representation and modified dominance relations may increase the diversity of rules presented to the user and how clustering techniques may be used to aid in the presentation of potentially large sets of rules generated. A variety of experiments have been reported by them for partial classification. The algorithm described produce partial classification rules using misclassification cost, rule complexity, support, confidence and coverage as fitness measures for the rules which are allowed to be controlled by the users. Rafael et al. [1] report a research work that combines evolutionary algorithms and ranking composition methods for multi-objective optimization.

The candidate solutions are constructed, evaluated and ranked according to their performance in each individual objective, and then rankings are composed into a single rank to solve the multi-objective problem considering all objectives simultaneously. A Multi Objective Evolutionary Algorithm (MOEA) has been proposed by the authors in [5] and [6] for obtaining fuzzy rules for subgroup discovery taking coverage, support and confidence as measures for optimization. The rule induction problem has been considered as a multi-objective combinatorial optimization problem by Khabzaoui et al [7] for finding non frequent and interesting rules using meta-heuristic. A review of evolutionary algorithms for data mining can be found in [8] and for multi objective optimization of classification rules in [9].

Cultural algorithm is a class of evolutionary algorithms which is mostly applied for numerical optimization problems and which has a Knowledge base for representing various primitive knowledge types used by animal species. Lazar and Reynolds, [10] have used CA and rough sets for heuristic knowledge discovery. Sternberg and Reynolds [11] use an evolutionary learning approach based on cultural algorithms to learn about the behavior of a commercial rule-based system for fraud detection. The learned knowledge in the belief space of the cultural algorithm is then used to re-engineer the fraud detection system. Reynolds et al., [12] use decision trees to characterize location decisions made by early inhabitants at Monte Alban, a prehistoric urban center, and have injected these

rules into a socially motivated learning system based on cultural algorithms and inferred theories about urban site formation.

III. Extended Cultural Algorithm for Multi Objective optimization of Rules

Cultural algorithms have been used in modeling social systems to solve problems in optimization. But use of cultural algorithm for multi objective optimization of rules is hardly found in the literature. Cultural algorithms use a basic set of knowledge sources, each related to knowledge observed in various animal species. These knowledge sources are then combined to direct the decisions of the individual agents in solving optimization problems. Evolutionary Algorithms are derived from nature and works with a population of individuals in an environment. But they are memory less. Cultural algorithm is a class of evolutionary algorithm that provides a systematic and principled approach for representing knowledge through the five knowledge sources. CA's use knowledge sources to store the various Meta data during the evolution thus incorporating memory into the evolutionary process. Agents evolve using this knowledge in the KS's to produce better individuals. Integration of intelligent agents with Data mining is justified in the article by Cao [13] where the author explores the promising area of agent mining interaction and integration. The current study further integrates evolutionary computing with intelligent agents and data mining to create a complete social system to convert discovered knowledge into social intelligence. Moreover the user can experiment with the system by specifying the various attributes of the system. The following section discusses various parts of the Extended CA (ECA). Thus the proposed system can also be used as a tool kit for experimenting with the process of classification rule mining.

A. The Belief Space

The belief space comprises of the five knowledge sources namely the Normative, Situational, Domain, Topographical and the History KS. For the rule optimization problem the five knowledge sources are modified to hold different types of knowledge or Meta data used in solving the problem.

Further an additional KS has been added to hold the rules. The agents in the CA have also been given social or cognitive traits which they use in decision making, all of which are described below.

B. Normative KS

Normative Knowledge Source (NKS) contains the attributes and the possible values that the attribute can take. It gathers this information from the training data set. The normative knowledge source is used to store the maximum and minimum values for numeric attributes. For nominal or discrete attributes, a list of possible values that the attribute can take is stored in the normative KS. The normative KS is used by the agents during mutation.

C. Situational KS

Situational Knowledge Source (SKS) consists of the best exemplar found along the evolutionary process. It represents a leader for the other individuals to follow. This way, agents use the leader in-stead of a randomly chosen individual for the recombination. The user can specify a schema which can be used by agents for the search of similar or dissimilar individuals to interest the user.

D. Domain KS

Domain Knowledge Source (DKS) contains the vector of rule metrics for each rule. It is updated whenever better rules are accepted into the population at the end of each generation. The domain KS is used by the system to choose best rules for subsequent generations.

E. Topographical KS

Topographic Knowledge Source (TKS) is used to store the difference or distance between two rules for the purpose of discovering diverse set of rules to avoid local optima. Hence topographical KS can be used to create novel and interesting rules by using the dissimilarity measure of individuals. This KS is updated at the end of each generation. The topographical knowledge contains a rule pair and their dissimilarity measure.

F. History KS

History Knowledge Source (HKS) records in a list, the best individual found at the end of each generation. Evolutionary algorithms are termed as memory less since they do not retain memory of previous generations. However attempts have been made to retain elite individuals of each generation as a separate elite population to render memory to the evolutionary algorithms. Cultural algorithm renders memory to the evolutionary strategy in a systematic way by using the five knowledge sources. History knowledge is used to store best individuals of each generation chosen according to the optimization strategy, thus maintaining memory across generations.

G. The Rule KS

The cultural algorithm is extended by adding another knowledge source namely the Rule KS (RKS) in order to hold the rules. The other KS's hold a pointer which is the Rule Id to one of the rules in the RKS. The rule KS is added to the CA in order to render it to solve the problem of rule mining thus making CA as Extended CA or ECA. The representation of the rule KS is similar to that of the HKS.

H. Social Agents

The proposed ECA is also extended by adding cognitive traits to the agents. The agents are distinguished by assigning a cognitive trait namely risk taker or imitator or cautious. The agents use this trait in the selection of parents for reproduction using different knowledge sources.

I. Influence Phase

The influence function decides which knowledge sources influence individuals. In the proposed system this is left to the agents. In the proposed CA the agents use their social trait namely risk taker or imitator or cautious to choose parents for reproduction. Risk takers use knowledge from any of the five knowledge sources at random while cautious agents use only the historical knowledge source. The imitators use the situational knowledge source to create individuals which are similar to the example specified by the user. The normative knowledge source which stores the possible attribute values is used by all the agents during the mutation operation. The topographical knowledge source enables creation of a diverse set of rules. Domain knowledge stores the values of the metrics of the individuals and thus is used for choosing best individuals according to user specified metrics. Thus the five KS's guide the agents in the evolution process.

J. Acceptance Phase

The acceptance function determines which individuals and their behaviors can impact the belief space. Based on selected

parameters such as performance, for example, a percentage of the best performers (e.g., top 10%), can be accepted [12], as in the CA literature. But since the problem is one of classification rule mining, a threshold value for the rule metrics specified by the user is to be used to accept individuals for next generation. The process of agent's selection, reproduction, evaluation forms a generation. At the end of a generation (iteration), the agents return their best individuals along with a vector of rule metric values. The individuals are accepted into the belief space based on the Pareto optimization strategy using the metrics stored in the domain KS as vectors. Dominators which are obtained by the comparison of values of the user specified metrics stored as vectors are chosen for the next generation. The knowledge sources are thus updated at the end of each generation and thus evolve along with the agents. The new values in these KSs then influence the population space. Thus the macro evolution takes place by updating the KSs.

K. Evolutionary Strategy

Genetic algorithm is by far the most used evolutionary strategy which is also used in the current study. The various attributes of the GA used are discussed below.

L. Chromosome Representation

The chosen data records are converted into chromosomes and represented as a vector of attribute values. The system uses high level encoding where the attribute values are used as they appear in the data source. This reduces the cost of encoding and decoding individuals for creating rules for large data sets.

The relational operators are not included in the genotype as found in most algorithms found in the literature. Therefore they are not involved in the reproduction which further minimizes the length of the chromosome and in turn the time taken for encoding and decoding. This representation also avoids use of different types of reproduction operators for different parts of the chromosome. In the current study the class attribute is included in the chromosome. Michigan style rules as disjunction of attribute tests are created only when they are presented to the user.

M. Population initialization

Population initialization is an important aspect that decides the output of the algorithm. The initial population is created by choosing data sets from the data source at random. This process known as seeding chooses random data from the training data set to be used as seed to create the initial population.

N. Reproduction Operators

The operators used for reproduction are selection, crossover and mutation.

1. Selection Strategy

Unlike algorithms found in the literature, in the proposed CA, Agents use their social traits in choosing the individuals for reproduction as described earlier. In this way, knowledge based selection is used rather than random. This kind of selection strategy aids in creating not only good individuals but also interesting and a diverse set of individuals using the various KS's.

2. Crossover

One point crossover is used. Initially two individuals are chosen at random from the population. A crossover point is chosen at random and the contents of the chromosome after the crossover point are swapped. Thus crossover takes two parents and produces two children.

Mutation: Mutation operates on individual values of attributes in the chromosome. A mutation point is chosen similar to that of the crossover point which is a random integer. The value of the attribute at that point is replaced by another value depending upon the type of the value. For nominal attributes the value to be replaced is chosen from a list of available values which is also the case of discrete integer values. If the attribute is a continuous valued one, a random value in a specified range of minimum and maximum values is generated and used for reproduction.

O. Parameters

The parameters that are to be considered and greatly influence the algorithm performance are the crossover rate which is the probability of crossover, and the mutation rate which is the rate of mutation. Also the population size and the number of generations or the termination condition are parameters of importance. Table 2 gives a summary of the parameters used in the experiments.

P. Optimization Strategy

The optimization or multi objective optimization strategy forms the accept phase of the cultural algorithm. This also enables interactive data mining where the user can specify the attributes of the rules to be optimized along with certain threshold values for the various rule metrics. Pareto optimality and ranking composition methods are the frequently used optimization strategies. Pareto optimality has been used in the experiments. Pareto optimality is an optimization strategy that uses comparison of the metrics represented as a vector. An individual "a" is said to be better than another individual "b" if "a" is better than "b" in all the metric values or equal to "b" in all but one metric and better at least in one value. This is enabled by the use of Domain KS which stores the rule metrics as vectors. The entries in the DKS are compared with each other and the best performers in all the metrics are returned as dominators.

IV. Proposed Work

In the work done by Sujatha Srinivasan et el. an extended cultural algorithm is proposed for multi objective optimization of classification rules, where mining rules with specific properties is taken as a multi objective optimization problem. The extended CA is also improved by incorporating cognitive traits to agents so that the system can also be used as a social system to study the dynamics of an organization or any social system.

But the work done by authors have stated that the social knowledge created by the individuals in the ECA is to be converted into actionable social knowledge or collective social intelligence to be applied in various applications such as an Intrusion detection and prevention system to solve the computer security problem.

These works takes the base as the work of the authors and propose to include the actionable knowledge using the following steps:

Dataset of Social website shall be used and loaded in the system

User will input characteristics or characteristics shall be retrieved as per the Intrusion Detection Systems such as Source IP Address, Number of Hits in a day from unique IPs etc.

Information of the social dataset shall be processed using Extended Cultural Algorithm to form the rules retrieval and definitions

Rules designed will be Processed to make them actionable social knowledge for Intrusion Detection System

Rules defined by the system will be used to form the results related with time taken in processing, number of rules produced, Complexity of the algorithm etc.

V. Results & Discussion

Implementation of the proposed work has been done using C# language with dataset of Access Log dataset stored in files. The implementation uses file handling to fetch records from the dataset in RAM and processes all the records generated in results for retrieving various relevant details. Various steps used in implementation are as follows:

A. Interface Development

C#.net windows forms have been used to develop the interfaces in which have different screens. Various relevant navigational buttons have been included in it which allows traversing with validations.

B. Loading Dataset from Access Log Dataset

Access Log is the mostly widely used data set for the evaluation of systems. This is loaded in the system using C# File Handling routines, which makes it not only easier but allows for processing the read records to suit the requirements of the implementation of the proposed system.

C. Data Preprocessing

We have used File handling for loading the access log file in project and applied data pre-processing on the loaded data. In data pre-processing to filter data for important fields of the records of data has been done. Stopping has been used to remove all remove any columns with irrelevant information for the proposed implementation

D. Clustering Phase

In this phase, characteristics have been loaded and added in the system using file handling from dataset and frequencies of the words has been calculated by applying the data pre-processing phase. This phase loads the data contents one by one, filters them for various columns added using ant clustering screen. Then the messages are pre processed for clustering. Clustering is done on the basis of the column values for IP Addresses and other column details.

Results have been calculated by using the clusters and formulas for calculation of detection rate and false positive rate as specified in proposed algorithm. Graphs have been drawn using the various results as calculated for all accuracy, precision, recall and specificity.

Comparison with Existing Work: A comparative chart has been prepared for all detection rate, false positive rates etc. showing the comparison between the existing work from the base paper and evaluated results for the proposed algorithm implementation as above.

For obtaining results we have used the following decision making tabular data collection from the execution of the various steps of the proposed and ANT clustering algorithms:

Table 1: Reading of the Time Taken in Various Steps of Extended Cultural Algorithm

STEP	TIME TAKEN (ms)
Time in Data Loading	350
Time in Population	666
Time in Selection of Rules	664
Time in Rule Generation	49
Time in Rule Fitering	701

Table 2: Readings of the Time Taken in Various Steps of the ANT Clustering Algorithm

STEP	TIME TAKEN (ms)
Time in Data Loading	426
Time in Data Cleaning	1046
Ant Clustering Time	1313

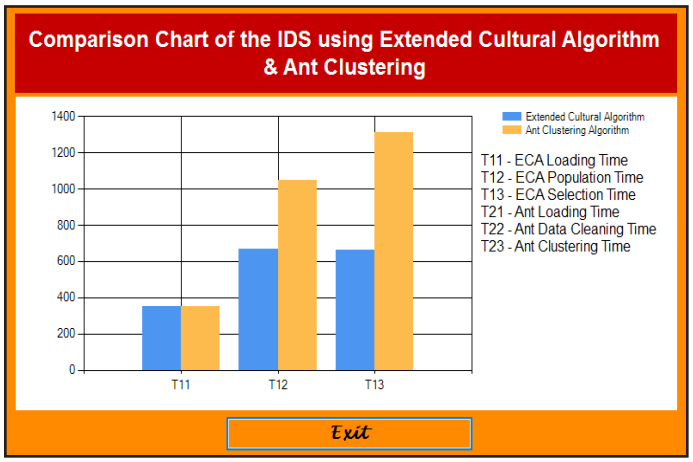

Fig. 1: Graph Created Using the C# Chart Control for Showing the Comparison Between the Proposed Algorithm and the ANT Clustering Algorithm

Inference: From the graph it is clear that the time taken in processing of the different stages of the implementation for ANT Clustering algorithm and the proposed algorithm are comparable. The time taken by the ANT clustering algorithm is equal or higher than the proposed extended cultural algorithm.

VI. Comparison

From the results obtained in the above comparison graph it is seen that the work proposed in cultural algorithm is better than the incremental clustering algorithm ANT clustering algorithm. This leads to prove that the methodologies adapted in extended cultural algorithm in spite of the genetic algorithm application takes less time in execution and provide better performance.

Also the rules generations in extended cultural algorithm are efficient and take almost no time in processing whereas the results obtained are good. The rule generation has been made flexible as the need of the data set, whereas in ANT clustering we do not have such flexibility and we directly get the clusters created at the end. For inclusion of such flexibility additional performance compromise shall be required which in turn make the ANT mechanism to be slower.

References

[1] Sujatha Srinivasan, Sivakumar Ramakrishnan,"Multi Objective Optimization of classification rules using Cultural Algorithms", International Conference on Communication Technology and System Design 2011, Procedia Engineering 30 (2012) pp. 457 – 465.

[2] Fabrizio Angiulli, Stefano Basta, Stefano Lodi, Claudio Sartori,"Distributed Strategies for Mining Outliers in Large Data Sets", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 7, July 2013 IEEE.

[3] Yinan Guo, Yun Liu, Jian Cheng,"Harmonious Color Optimization Design Based on Adaptive Interactive Cultural Algorithm", 2013 IEEE Congress on Evolutionary Computation June 20-23, Cancún, México, 2013 IEEE.

[4] Mostafa Z. Ali, Noor H. Awad, Robert G. Reynolds, "Hybrid Niche Cultural Algorithm for Numerical Global Optimization", 2013 IEEE Congress on Evolutionary Computation June 20-23, Cancún, México, 2013 IEEE.
[5] V.K.Deepa, J. Rexy R.Geetha,"Rapid Development of Applications in Data Mining", Proceedings of 2013 International Conference on Green High Performance Computing March 14-15, 2013, India, 2013 IEEE.
[6] Giusti Rafael, Gustavo E.A., Batista P.A., Prati Ronaldo Cristiano,"Evaluating Ranking Composition Methods for Multi-Objective Optimization of Knowledge Rules", Proceedings of Eighth International Conference on Hybrid Intelligent Systems, 2008, pp. 537-42.
[7] Reynolds A. P., de la Iglesia,"B. Rule induction using multi-objective metaheuristic: Encouraging rule diversity", Proceedings of IJCNN. 2006, pp. 6375-82.
[8] Reynolds A. P., de la Iglesia B.,"Rule Induction for Classification Using Multi-Objective Genetic Programming", In: Proceedings of 4th Int'l. Conf. on Evolutionary Multi-Criterion Optimization. LNCS 4403. 2007, pp. 516-30.
[9] Reynolds A. P., de la Iglesia B.," A Multi-Objective GRASP for Partial Classification", Soft Computing, 2009, 13(3), pp. 227-43.
[10] Del Jesus M. J., Gonzalez Pedro, and Herrera Francisco, " Multi-objective Genetic Algorithm for Extracting Subgroup Discovery Fuzzy Rules", Proceedings of the IEEE Symposium on Computational Intelligence in Multi-criteria Decision Making, 2007, pp.50-7.
[11] Dehuri S., Mall R.,"Predictive and comprehensible rule discovery using a multi-objective genetic algorithm Knowledge-Based Systems", 2006, 19, pp. 413–21.
[12] Khabzaoui M, Dhaenens C, Talbi EG,"Combining evolutionary algorithms and exact approaches for multi-objective knowledge discovery", RAIRO Operations Research, 2008, 42: pp. 69–83.
[13] Freitas Alex A.,"A Review of Evolutionary Algorithms for Data Mining", Soft Computing for Knowledge Discovery and Data Mining, 2007, pp. 79-111
[14] Srinivasan, S., S. Ramakrishnan,"Evolutionary multi objective optimization for rule mining: a review", Artificial Intelligence Review: 2011, 36(3): pp. 205-48.