RESEARCH ARTICLE OPEN ACCESS

Optimizing the Accuracy of CART Algorithm by Using Genetic Algorithm

Jatinder Kaur [1], Jasmeet Singh Gurm [2]
Department of Computer Science and Engineering
RIMT-IET, Mandi Gobindgarh
Punjab Technical University, Jalandhar
Punjab - India

ABSTRACT
Data mining is one of the analysis step of the "Knowledge Discovery in Databases" process, or KDD. Data mining is a logical process that is used to search through large amounts of information in order to find important data. The goal of this technique is to find patterns that were previously unknown. Once you have found these patterns, you can use them to solve a number of problems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. There are many classification algorithms in data mining but Decision Trees are commonly used in data mining with the objective of creating a model that predicts the value of target variable based on the values of several input. In this paper we will optimize the CART algorithm by using Genetic algorithm

Keywords:- Data mining, Decision Trees, WEKA, CART, Genetic Algorithm.

I. INTRODUCTION

Data mining is one of the analysis step of the "Knowledge Discovery in Databases" process, or KDD. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. It is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

There are several major data mining techniques have been developing and using in data mining projects recently including association, classification, clustering, sequential patterns and decision tree.

II. CART

It stands for classification and regression trees and was introduced by Breiman in 1984.It builds both classifications and regression trees. The classification tree construction by CART is based on binary splitting of the attributes. It is also based on Hunt‟s algorithm and can be implemented serially. It uses gini index splitting measure in selecting the splitting attribute.

CART is unique from other Hunts based algorithm as it is also use for regression analysis with the help of the regression trees (S.Anupama et al,2011). The regression analysis feature is used in forecasting a dependent variable given a set of predictor variables over a given period of time. It uses many single-variable splitting criteria like gini index, symgini etc and one multi-variable in determining the best split point and data is stored at every node to determine the best splitting point.

Decision Trees are commonly used in data mining with the objective of creating a model that predicts the value of a target (or dependent variable) based on the values of several input (or independent variables). The CART or Classification & Regression Trees methodology was introduced in 1984 by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone as an umbrella term to refer to the following types of decision trees:

Classification Trees: where the target variable is categorical and the tree is used to identify the "class" within which a target variable would likely fall into.

Regression Trees: where the target variable is continuous and tree is used to predict it's value.
The CART algorithm is structured as a sequence of questions, the answers to which determine what the next question, if any should be. The result of these questions is a tree like structure where the ends are terminal nodes at

which point there are no more questions. A simple example of a decision tree is as follows

1. Advantages of Cart Algorithm

1. CART handles missing values automatically using surrogate splits
2. Invariant to monotonic transformations of predictive variable
3. Not sensitive to outliers in predictive variables unlike regression and Great way to explore, visualize data.

2. Disadvantages of Cart Algorithm

1. Nonparametric
2. Automatically performs variable selection
3. Uses any combination of continuous/discrete variables, Very nice feature: ability to automatically bin massively categorical variables into a few categories.
4. Discovers "interactions" among variables

III. GENETIC ALGORITHM

In the computer science field of artificial intelligence, a genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. GAs is inspired by Darwin's Theory about Evolution "Survival of Fittest". GAs is adaptive heuristic search based on the evolutionary ideas of natural selection and genetics.

Genetic algorithms find application in bioinformatics, phylogenetics, computational science, engineering, economics, chemistry, manufacturing, mathematics, physics and other fields [3].

GAs simulate the survival of the fittest among individuals over consecutive generation for solving a problem. Each generation consists of a population of character strings that are analogous to the chromosome that we see in our DNA. Each individual represents a point in a search space and a possible solution. The individuals in the population are then made to go through a process of evolution.

The GA maintains a population of n chromosomes (solutions) with associated fitness values. Parents are selected to mate, on the basis of their fitness, producing offspring via a reproductive plan. Consequently highly fit solutions are given more opportunities to reproduce, so that offspring inherit characteristics from each parent. As parents mate and produce offspring, room must be made for the new arrivals since the population is kept at a static size. Individuals in the population die and are replaced by the new solutions, eventually creating a new generation once all mating opportunities in the old population have been exhausted. In this way it is hoped that over successive generations better solutions will thrive while the least fit solutions die out.

New generations of solutions are produced containing, on average, more good genes than a typical solution in a previous generation. Eventually, once the population has converged and is not producing offspring noticeably different from those in previous generations, the algorithm itself is said to have converged to a set of solutions to the problem at hand.

Outline of Basic Genetic Algorithm

1. **[Start]** Generate random population of n chromosomes (suitable solutions for the problem)
2. **[Fitness]** Evaluate the fitness f(x) of each chromosome x in the population
3. **[New population]** Create a new population by repeating following steps until the new population is complete
    a. **Selection:** Select two parent chromosomes from a population according to their fitness (the better fitness,the bigger chance to be selected)
    b. **Crossover:** With a crossover probability cross over the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents.
    c. **Mutation:** With a mutation probability mutate new offspring at each locus (position in chromosome).
    d. **Accepting:** Place new offspring in a new population
4. **[Replace]** Use new generated population for a further run of algorithm
5. [**Test]** If the end condition is satisfied, stop, and return the best solution in current population

6. **[Loop]** Go to step 2.

A typical flowchart of a genetic algorithm is shown in Figure.1

One iteration of the algorithm is referred to as a generation

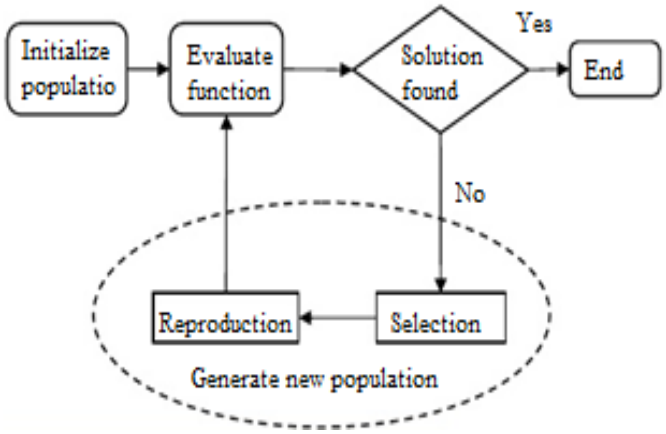

Fig.1 Flow chart of genetic algorithm

Advantages of GA

1. It can solve every optimisation problem which can be described with the chromosome encoding.
2. It solves problems with multiple solutions.
3. Since the genetic algorithm execution technique is not dependent on the error surface, we can solve multi-dimensional, non-differential, non-continuous, and even non-parametrical problems.
4. Structural genetic algorithm gives us the possibility to solve the solution structure and solution parameter problems at the same time by means of genetic algorithm.
5. Genetic algorithm is a method which is very easy to understand and it practically does not demand the knowledge of mathematics.
6. Genetic algorithms are easily transferred to existing simulations and models.

2. Disadvantages of GA

1. There is no absolute assurance that a genetic algorithm will find a global optimum. It happens very often when the populations have a lot of subjects.
2. Certain optimisation problems (they are called variant problems) cannot be solved by means of genetic algorithms. This occurs due to poorly known fitness functions which generate bad chromosome blocks in spite of the fact that only good chromosome blocks cross-over.
3. Like other artificial intelligence techniques, the genetic algorithm cannot assure constant optimisation response times.
4. It is unreasonable to use genetic algorithms for on-line controls in real systems without testing them first on a simulation model.

IV. LITERATURE SURVEY

Anuja Priyama et al. (2012) elaborates that At the present time, the amount of data stored in educational database is increasing swiftly. These databases contain hidden information for improvement of student's performance. Classification of data objects is a data mining and knowledge management technique used in grouping similar data objects together. There are many classification algorithms available in literature but decision tree is the most commonly used because of its ease of execution and easier to understand compared to other classification algorithms. The ID3, C4.5 and CART decision tree algorithms former applied on the data of students to predict their performance. But all these are used only for small data set and required that all or a portion of the entire dataset remain permanently in memory. This limits their suitability for mining over large databases. This problem is solved by SPRINT and SLIQ decision tree algorithm. In serial implementation of SPRINT and SLIQ, the training data set is recursively partitioned using breadth-first technique. In this paper, all the algorithms are explained one by one. Performance and results are compared of all algorithms and evaluation is done by already existing datasets. All the algorithms have a satisfactory performance but accuracy is more witnessed in case of SPRINT algorithm.

Rupali Haldulakar et al. (2011) explained that strong rule generation is an important area of data mining. In this paper authors design a novel method for generation of strong rule. In which a general Apriori algorithm is used to generate the rules after that authors use the optimization techniques. Genetic algorithm is one of the best ways to optimize the rules. In this direction for the optimization of the rule set they design a new fitness function that uses the concept of supervised learning then the GA will be able to generate the stronger rule set. In this direction authors optimize association rule mining using new fitness function. In which fitness function divide into two classes c1 and c2 one class for discrete rule and another class for continuous

rule. Through this direction authors get a better result. To make genetic algorithm more effective and efficient it can be incorporated with other techniques so it can provide a best result.

Alisa et al. (2006) described that Quality of life in multiple sclerosis has been often measured through the SF-36 questionnaire. In this study, validation of the SF-36 summary scores, its 'physical' component, and its 'mental' component was attempted by exploring the joint predictive power of disability, of anxiety and depression, and of disease duration, progression type, age, gender and marital status. The sample consisted of 75 patients suffering from multiple sclerosis admitted to an inpatient rehabilitation unit. The interplay between potential predictors was assessed through a particular regression model. Two main advantages of this technique are its robustness with respect to distributional assumptions and its sensitivity to high-order interactions, between independent variables, difficult to detect through conventional multiple regression. Predictive variables for physical component of the SF-36 were EDSS and HADS-D.

Alaa Al Deen et al. (2011) elaborates the concept of classification and association rule mining algorithms are discussed and demonstrated. Particularly, the problem of association rule mining, and the investigation and comparison of popular association rules algorithms. The classic problem of classification in data mining will be also discussed. The paper also considers the use of association rule mining in classification approach in which a recently proposed algorithm is demonstrated for this purpose. Finally, a comprehensive experimental study against 13 UCI data sets is presented to evaluate and compare traditional and association rule based classification techniques with regards to classification accuracy, number of derived rules, rules features and processing time.

V. SIMULATION RESULTS

To investigate the proposed method we implement it by using weka data mining tool and for testing the result we use weather.numeric dataset and breast cancer dataset. Our proposed algorithm CART with genetic algorithm is compared with simple cart algorithm .Our proposed algorithm shows better classifiers accuracy instead of simple CART algorithm.

Table 1 Classifiers accuracy for Cart + Genetic algorithm

Data Set	Correctly classified Instances	Incorrectly Classified Instances
Weather.numeric	57.1429%	42.8571%
Breast cancer	76.5734%	23.4266%

Table 2 Classifiers accuracy for Cart algorithm

Data Set	Correctly classified Instances	Incorrectly Classified Instances
Weather.numeric	50%	50%
Breast cancer	69.2308%	30.7692%

Above both tables shows the accuracy of CART algorithm and CART by using Genetic algorithm applied on some data sets. Table 1 shows that CART by using Genetic algorithm technique has higher accuracy of 57.1429% as compared to simple CART algorithm applied on weather. numeric dataset. In case of breast cancer dataset accuracy of cart using genetic algorithm is 76.5734% which is higher as compared to simple CART algorithm applied on the same dataset.

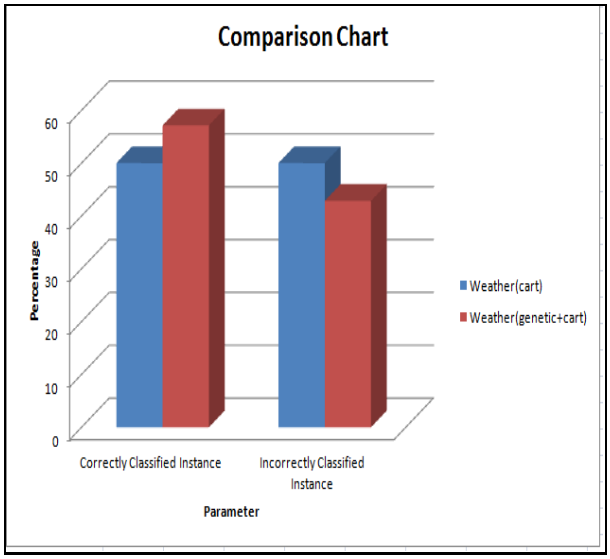

Fig.2 Comparison between cart and cart + genetic apply on weather.numeric .arff file

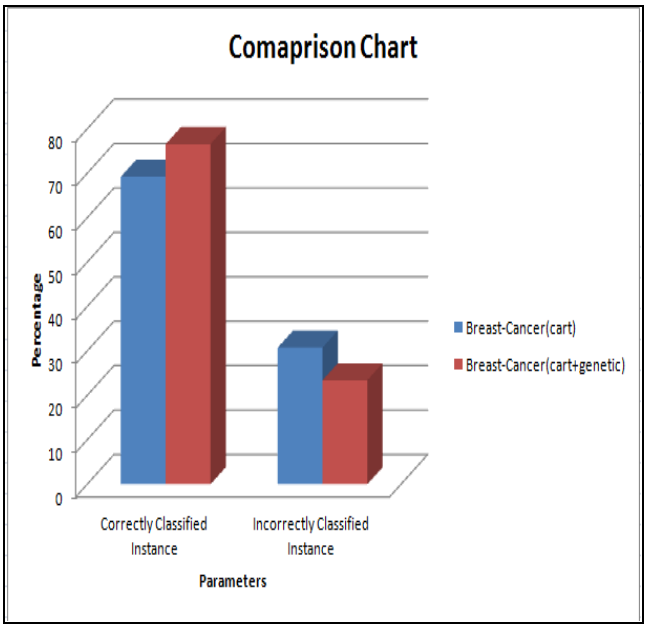

Fig.3 Comparison between cart and cart + genetic apply on Breast Cancer .arff file

VI. CONCLUSION AND FUTURE SCOPE

There are many classification algorithms in data mining but Decision Trees are commonly used in data mining with the objective of creating a model that predicts the value of target variable based on the values of several input. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. In this paper by using CART alongwith Genetic algorithm on various datasets classifiers accuracy is improved. As future work, we will process the Classification and Regression Trees algorithm with other new algorithm to refine result more accurately.

VII. REFERENCES

[1] Jatinder Kaur , Jasmeet Singh, "Review on CART and Genetic Algorithm", International Journal for Multi-Disciplinary Engineering and Business Management (IJMDEBM), ISSN: 2348 – 2249, vol. 3, Issue 2, June 2015.

[2] Anuja Priyam,"Comparative Analysis of Decision Tree Classification Algorithms", International Journal of Current Engineering and Technology, Vol.3, No.2, pp.866-883, June 2013.

[3] Rupali Haldulakar and Prof. Jitendra Agrawal, "Optimization of Association Rule Mining through Genetic Algorithm", International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 3 Mar 2011, pp. 1252-1259.

[4] Alaa Al Deen, Mustafa Nofal and Sulieman Bani-Ahmad, "Classification Based On Association-Rule Mining Techniques: A General Survey and Empirical Comparative Evaluation ", Ubiquitous Computing and Communication Journal, Vol.5, Issue.3, 2011.

[5] Agrawal R., Imielinski T. and Swami A. "Mining Association rules between sets of items in large databases", In the Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (ACM SIGMOD '93), Washington, USA, pp.207-216, 1993.

[6] Ramez Elmasri, Shamkant B.Navathe," Fundamentals of Database Systems", Pearson, fifth edition, 2009.

[7] Wei-Yin Loh," Classification and Regression trees " Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA,Vol 1, Jan-Feb 2011.

[8] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam,`` A Study of Data Mining Tools in Knowledge Discovery Process", International Journal of Soft Computing and Engineering (IJSCE),Volume-2, Issue-3, July 2012

[9] Hamidah Jantan, Mazidah Puteh, Abdul Razak Hamdan and Zulaiha Ali Othman "Applying DataMining Classification Techniques for Employee's Performance Prediction"

[10] Kuldeep Kumar, Sikander, Ramesh Sharma, Kaushal Mehta, "Genetic Algorithm Approach to Automate University Timetable", International Journal of Technology Research (IJTR) Vol 1,Issue 1,Mar-Apr 2012.

[11] Jiawei Han and Micheline Kamber,"Data Mining Concepts and Techniques", Morgan Koufmann Publishers, Second Edition 2006.

[12] Brijsh Kumar bhardwaj and Saurabh Pal (2011)"Data mining: a prediction for performance improvement using classification", International journal of computer science and information security, vol. 9, no. 4.

[13] S.Anupama Kumar and Dr. Vijayalakshmi M.N. (2011) "Efficiency of decision trees in predicting student"s academic performance", D.C. Wyld, et al. (Eds): CCSEA 2011, CS & IT.

[14] Swasti Singhal, Monika Jena "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013.

[15] Dr. Sudhir B. Jagtap, Dr. Kodge B. G. "Census Data Mining and Data Analysis using WEKA", (ICETSTM – 2013) International Conference in "Emerging Trends in Science, Technology and Management-2013, Singapore.

[16] Frawley, W., Piatetsky- Shapiro, G., Matheus, (1992). "Knowledge Discovery in Databases: An Overview", AI Magazine, fall 1992, pp. 213-228.

[17] D. Kerana Hanirex and K.P. Kaliyamurthie, "Mining Frequent Itemsets Using Genetic Algorithm" , Middle-East Journal of Scientific Research , 2014.

[18] M. Sukanya, S. Biruntha, Dr. S. Karthik and T.Kalaikumaran "Data mining: Performance Improvement in Education Sector using Classification and Clustering Algorithm", International conference on computing and control engineering (ICCCE 2012) 12 & 13 April, 2012.

[19] Shaeela Ayesha, Tasleem Mustafa, M.Inayat Khan and Ahsan Raza Sattar(2010) "Data mining model for higher education system", Europeon journal of scientific research, ISSN 1450-216X Vol. 43 no. pp.2

[20] M. Sujatha, S. Prabhakar, Dr. G. Lavanya Devi, "A Survey of Classification Techniques in Data Mining" International Journal of Innovations in Engineering and Technology (IJIET)Vol. 2 Vol. 2 Issue 4 August 2013.

[21] Anita Thengade , Rucha Dondal "Genetic Algorithm – Survey Paper ", International Journal of Computer Applications, April 2012.