



Description of Genetic and CART Algorithm using Data Mining Tool

Jatinder Kaur

Department of CSE

RIMT-IET

Punjab Technical University, Jalandhar

Punjab, India

Jasmeet Singh Gurm

Department of CSE

RIMT-IET

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 explain WEKA tool and CART and Genetic algorithm in data mining.

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

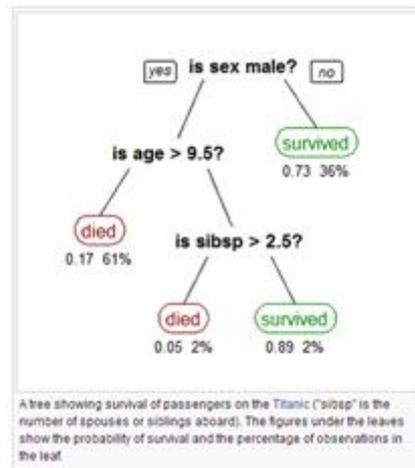


Figure 2.1 Decision Tree

II.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.

II.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.

III.1. 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.

III.2. Disadvantages of GA

1. 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.
2. 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.
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. WEKA TOOL

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering and association rules; It also includes visualization tools. The new machine learning schemes can also be developed with this package. WEKA is open source software issued. The data file normally used by Weka is in ARFF file format, which consists of special tags to indicate different things in the data file (foremost: attribute names, attribute types, and attribute values and the data). The main interface in Weka is the Explorer. It has a set of panels, each of which can be used to perform a certain task. Once a dataset has been loaded, one of the other panels in the Explorer can be used to perform further analysis [14].

IV.1. The characteristics of the WEKA system

WEKA is a free for academic license, not integer with other systems. As a typical representative of the academic data mining, it has the following characteristics:

- 1) Cross-platform, it supports Windows and Unix, and many other operating systems;
- 2) It supports the structures text file, the data mining format (C4.5), and provides database interface (JDBC);
- 3) It can handle the data types of continuous, discrete, characteristic, date types.
- 4) It provides the missing value treatment, elimination noise, standardization, data discretization, attribute structure, transform variable, split data, data balance, sample sorting, sample shuffle, data clustering, dimensional reduction, value reduction and sampling operation;
- 5) It can complete preprocessing, classification, clustering, association, visualization and other tasks;
- 6) It supports machine learning and neural networks;
- 7) It provides algorithm combinations, users embedded algorithm, algorithm parameter settings (basic, advanced)
- 8) It can generate basic reports, test reports, output format, implementation model explained, model comparison, data score function;
- 9) It achieves data visualization, mining process visualization, and the mining result visualization (comprehension, evaluation).

IV.2. File format of the WEKA system

The WEKA system supports three types of data file to open, respective that imports from the local data file, data site or database to be tested. However, whichever way the slide to open, WEKA always has a certain limit on the format of the imported data.

WEKA uses a data format called ARFF (Attribute-Relation File Format), this is an ASCII text. The ARFF file can be divided into two parts. The first part gives the Head information, including a statement of relations and attribute declarations. The second part shows the Data information, the given data in the data set.

(1) **Head information:** @ relation defines the data set name, equivalent to the data table name. @ Attribute defines the data set attribute; it contains the attribute name and possible values of attribute or the attribute type.

(2) **Data information :** @ data defines the start of data set record, the following is all the data sets record, the record is unordered, every data item between each row is separated by comma “,”. Also for the missing data items, we use “?” to express the missing value. Certainly, when we import the data file, we will find that we can also import the file form with the file extension name. csv (which may be exported by Excel or Matlab); the instance of the C4.5 original file with extension file name is .names and .data, and has been serialized the extension file name is .bsi's . That is because the WEKA system comes with three kinds of file format converters were: CSVLoader, C45Loader and SerializedInstanceLoader so when the WEKA ARFF file could not be loaded, the system will automatically call the file format converter automatically converter to the additional types of files to ARFF format for testing.

The system interface

WEKA uses a series of standard machine learning techniques that is unified graphical user interface (GUI), to combine with many pre-processing and post-processing methods, apply many different learning algorithms into data sets, and assess the corresponding results. When the user runs WEKA, the WEKA GUI Chooser interface will appear, as shown in Figure 4.1, including the Simple CLI, Explorer, Experiment, and Knowledge Flow [13].



Figure 4.1 Weka Gui Chooser

We click the Explorer button, go into the Explorer graphical user interface, as shown in Figure . 4.1

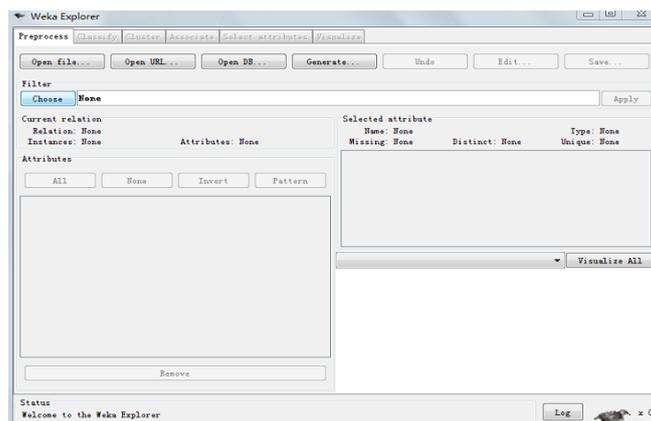


Figure 4.2 The interface of WEKA

In Figure 4.2, there are six labels at the top of the WEKA Explorer interface, separately corresponding to different data mining methods supported by WEKA. These include: Process, Classify, Cluster, Associate, Select attributes, Visualize. Through this user interface, all the WEKA functions can be completed by menu selection and form filling. This is done by changing the option into menu, setting the not applicable option as not available, and designing the user options as the form filling shape, to guide the user step by step to completely explore the algorithm in proper order. At the same time, it also gives the tools usage tips in the pop-up window, which is a great help for the users, and the reasonable default values allow the users to achieve the desired results with minimal effort.

IV.3. WEKA application interfaces

- **Explorer**
 - preprocessing, attribute selection, learning, visualization
- **Experimenter**
 - testing and evaluating machine learning algorithms.
- **Knowledge Flow**
 - visual design of KDD process
 - Explorer
- **Simple Command Line**
 - A simple interface for typing commands

V. CONCLUSION

In this paper we have discussed about the CART algorithm, Genetic algorithm and WEKA tool in data mining. 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. WEKA is open source software which implements algorithms for data preprocessing, classification, regression, clustering and association rules; It also includes visualization tools.

REFERENCES

- [1] 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.
- [2] Rupali Haldulakar and Prof. Jitendra Agrawal, "Optimization of Association Rule Mining through Genetic Algorithm", International Journal on Computer Science and Engineering (IJCSSE), Vol. 3 No. 3 Mar 2011, pp. 1252-1259.

- [3] 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.
- [4] 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.
- [5] Ramez Elmasri, Shamkant B.Navathe," Fundamentals of Database Systems", Pearson, fifth edition, 2009.
- [6] Wei-Yin Loh," Classification and Regression trees " Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA,Vol 1, Jan-Feb 2011.
- [7] 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
- [8] Hamidah Jantan, Mazidah Puteh, Abdul Razak Hamdan and Zulaiha Ali Othman "Applying DataMining Classification Techniques for Employee's Performance Prediction"
- [9] 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.
- [10] Jiawei Han and Micheline Kamber,"Data Mining Concepts and Techniques", Morgan Koufmann Publishers, Second Edition 2006.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.