ISSN No. 0976-5697

Volume 3, No. 2, March-April 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Genetic Programming in Data mining Tasks

Venkatadri. M*
Assistant Professor
Centre for Information Technology
College of Engineering Studies
University of Petroleum and Energy Studies, India
venkatadri.mr@gmail.com

Hanumat G. Sastry
Assistant Professor
Centre for Information Technology
College of Engineering Studies
University of Petroleum and Energy Studies, India
sastrygh2000@yahoo.com

Lokanatha C. Reddy
Professor Department of Computer Science
School of Science & Technology
Dravidian University, India.
lokanatha_r@yahoo.com

Abstract: Genetic Programming (GP) is a machine learning technique used to give the optimized solution for the user specified tasks from a population of computer programs based on a fitness function. Genetic Programming provides automated and optimized solutions for searching of large, poorly defined search spaces and even with the complexities of high dimensionality, multi modality and discontinuity with noise. Knowledge Discovery is an extremely complex process in the real world databases. Various data mining techniques exists for knowledge discovery process, among them, Genetic Programming data mining techniques are more efficient and suitable. Hence, this paper discus the various GP based techniques in the data mining field.

Keywords: Data mining, Classification, Clustering, Genetic Programming

I. INTRODUCTION

Data Mining is increasingly being accepted in various domains as a viable means of discovering hidden knowledge in the large volumes of data repositories. Data Mining is a multi-disciplinary field borrowing ideas from data bases, machine learning and artificial intelligence, statistics, high performance computing, signal and image processing, mathematical optimization, pattern recognition, etc [1]. Since the emergence of massive data collection technology, large volumes of data (Zeta Bytes) with complex data formats the conventional data mining techniques/algorithms are no longer sufficient [2]. To address these challenges, Evolutionary Data mining Algorithms/techniques are suitable to complement the existing approaches.

Genetic Programming is one of the better Evolutionary Algorithms, follows Darwin's theory of evolution—often paraphrased as "survival of the fittest". GP is domain independent, innovative flexible and interesting technique and has been applied to a broad range of problems with success from traditional optimization in engineering and operational research to non-traditional areas. Various Genetic Programming based Data mining techniques/ algorithms exist to handle the various data formats. Hence, this paper presents intensive study on the existing data mining algorithms based on Genetic Programming.

II. OVERVIEW ON DATA MINING TASKS

Primarily Data mining can be classified into two high level categories, such as

i. Predictive Data Mining
ii. Descriptive Data Mining

A. Predictive Data Mining:

This model of data mining techniques creates a model to predict the future values based on the past and current data values. The various Predictive Data Mining techniques are

a. Classification
b. Regression Analysis
c. Time Series Analysis
d. Prediction

B. Descriptive Data Mining:

This model of data mining techniques organizes the data, based on their general properties and transforms it into human interpretable patterns, associations or correlations. The various Descriptive Data Mining techniques are

a. Clustering
b. Summarization
c. Association Rule Mining
d. Sequence Discovery

III. GENETIC PROGRAMMING (GP)

GP is an evolutionary computation technique that automatically solves problems in a systematic, domain independent method for getting computers automatically solves problems starting from a high level statement of what needs to be done. Genetic Programming is an extension of the Genetic Algorithms and was proved, promoted and developed into a practical tool by John Koza amongst a whole range of possible evolutionary algorithms [3]. Each computer program is treated as an individual in the population of complete (or candidate) solution. Best solution is determined by computing the Fitness function on every individual. The candidate solution usually consists of data and functions. Individual solution (Computer Program) for each problem in GP works with sets of symbols namely the terminal set and function set. The terminal set typically contains variables (or attributes) and constants; whereas the function set contains functions which are believed to be appropriate to represent good solutions for the target problem. The following algorithm explains the typical GP process.

```
Start
Randomly populate an individual from the available primitives
Repeat
        Execute the each individual and determine its fitness.
        Create new individuals by applying genetic
        operations
        on one or more populated individuals, which met the
        fitness criteria.
Until an acceptable solution is found or some other stopping
condition is met
Return the best-so-far individual.
End
```

IV. GENETIC PROGRAMMING IN DATA MINING

The field of Data mining has got attracted towards the Genetic Programming, due to its advantages in automatic evolution of programs with optimized solutions without prior knowledge on data in larger search spaces with very less fatal errors. GP is also advantages over other Evolutionary approaches in construction and population of the individuals as it supports flexible variable length solution representation. Various GP techniques/algorithms are constantly expanding in the field of data mining, among these techniques Classification and Clustering are well popular and more suitable for the most of the real data mining problems.

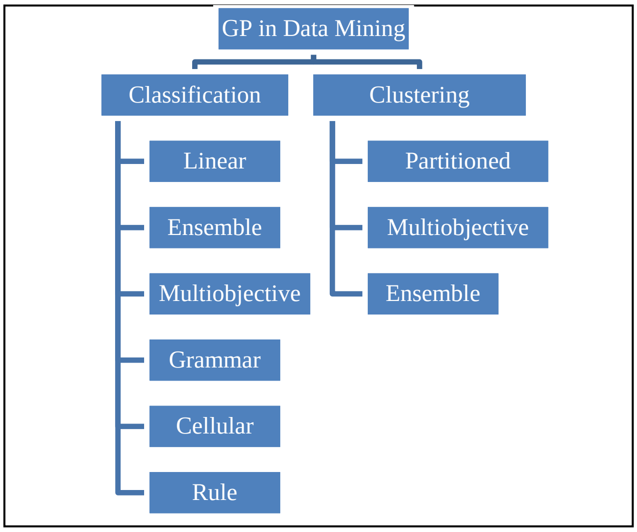

Figure 1: GP in Data Mining

V. GENETIC PROGRAMMING BASED CLASSIFICATION

GP provides optimistic solutions for the discovery of potentially important gene by generating comprehensible rules for classification [4][5][6]. Various GP based Classification techniques/algorithms are widely applied in discovering the underlying data relationships; among them the following are the popular techniques/algorithms.

a. Linear GP Classification
b. Ensemble GP Classification
c. Multi objective GP Classification
d. Grammar based GP Classification
e. Cellular GP Classification
f. Rule based GP Classification

A. *Linear GP Classification:*

This technique employs the linear function for classifying the data. Some of the Linear GP algorithms are also uses Statistical Analysis; hence it is also known as GP based Statistical Classification. Linear GP is well suitable for text classification. Genetic Parallel Programming (GPP) is a Linear Genetic Classification technique to evolve parallel programs based on a Multi-ALU Processor [7].

B. *Ensemble Genetic Classification:*

Ensemble Genetic Classification combines the heterogeneous datasets and different classification techniques. A Genetic Programming Ensemble Approach to Cancer Microarray Data Classification [8] proposed ClusSNR method by exploits the advantage of a clustering technique , namely K-Means clustering combined with feature selection and Genetic programming to construct a number of classifiers which are assembled into an ensemble by using six data sets of cancer microarray data from Bio-medical Data Analysis web site .

C. *Multi objective Genetic Classification:*

Multi objective classification deals with multiple objectives. GP based Multi objective classification employs the decision trees for individual representation. Evolutionary Multi-objective Optimization approach (EMO) with GP [9] for classification reduces the size of the trees drastically to finds out the minimum error rate for each size of decision trees through structural risk minimization approach.

D. *Grammar based Genetic Classification:*

Grammar based Genetic Classification technique employs the constraints over operators for data classification. These constraints will be represented by production rules. Grammar based GP Classification algorithm uses decision trees and rule induction method [10].

E. *Cellular Genetic Classification:*

Cellular Genetic Programming Classification technique is a decision tree classification technique based on Cellular automata framework to enable a fine grained parallel implementation GP with diffusion model [11] [12].

F. *Rule based Genetic Classification:*

Rule based Genetic Classification in GP is one type of rule extraction process from a rule set database where rules can be represented in IF<antecedent> THEN <consequent> form for classification of data. The consequent part represents a class label when the condition in the antecedent part is true. The classification rules are generated and are stored in rule set bases for each real valued attributes; each condition is defined on the input data. A Constrained Genetic Programming (CGP) introduced in [13] is a GP-based cost-sensitive classifier capable of building decision trees to minimize not only the expected number of errors, but also the expected misclassification costs through its novel constraint fitness function.

Table 1: Overview on GP Classification Techniques

Type of GP Classification Algorithm	**Technique Applied**	**Individual Representation**
Linear	Linear Discriminative/ Generative [27]	Linear Data Structure
Ensemble	Boosting/Bagging/Bayesian Network[28]	Tree/Graph Data Structure
Multiobjective	Pareto Dominance principle[29]	Tree Data Structure
Grammar	Backus Naur Form(BNF)	Linear/Tree Data Structure
Cellular	Cellular Automata[27]	Tree Data Structure
Rule	Classification Rules Set	Tree Data Structure

VI. GENETIC PROGRAMMING BASED CLUSTERING

Clustering is a popular data mining technique, determines a finite set of categories to describe a data set according to the similarities among its objects. Clustering is very popular in finding out the interesting patterns in market segmentation, image processing, document categorization and web mining. Due, to the importance of Clustering various Clustering techniques/algorithms exist. Genetic Programming techniques are also applied for better Clustering, among them the following are the popular techniques/algorithms.

a. GP based Partitioned Clustering
b. GP based Multi Objective Clustering
c. GP based Ensemble Clustering

A. GP based Partitioned Clustering

The Partitioned Clustering is a core Clustering technique, which results in a set of M clusters, each object belongs to one cluster and each cluster is represented by a cluster representative. Various, GP based techniques have implemented in Partitioned Clustering for its effective efficiency. Syntactical recognition for time series data using GP algorithm [14] has successfully demonstrated the analysis of stock prices and exchange rates. This clustering technique has two phases. One is learning phases and other is clustering phase. The individuals (syntactical expressions) are represented in a tree and the GP is implemented in learning phase to improvise the functional forms (syntactical expressions) to approximate the K-clusters used to generate the time series data. “Particle Swarm-like agents approach for Dynamically Adaptive data clustering “(PSDC) [15] is a K-means partitioned clustering technique using agent technology. In this technique the data set is represented as a set of vectors in D-dimension space. This technique has given good results on the experimental data set from Machine Learning Repository (UCI). A GP based Clustering technique is modeled by considering the populations made up of strings of Meroids with different lengths and provide a framework for evolutionary conceptual clustering that can be applied to standard relational representations for knowledge bases in the Semantic Web context [16]. This technique is a combination of evolutionary and distance based clustering approaches to improve the clusters with less noise.

B. GP based Multi Objective Clustering:

Most of the real time business applications are multi objective that described using a set of different criteria that account for conflicting properties such as the compactness of clusters and the separation between clusters. Various Genetic Programming techniques are applied for effective Multi Objective Clustering. Parallel hYbrid clusteRing using genetic progrAmming and Multi-objective fItness with Density (PYRAMID) [17] employs a novel approach to cluster a large volumes of data by combining the data parallelism with Genetic Programming and a multi objective density based fitness function. In this technique, individual are represented by tree and data parallelism is achieved by Divide and Conquer method. Evolutionary Image Segmentation Based on Multiobjective Clustering algorithm finds out various solutions (image segmentation results) through an evolutionary process. The individuals in the solution space are represented as a graph in which each pixel in the image is represented as a node, and an edge between two nodes indicates that these pixels are in the same region (cluster) [18]. Incremental Parallel hYbrid clusteRing using genetic progrAmming and Multi-objective fItness with Density (IPYRAMID) algorithms employs a combination of data parallelism, genetic programming (GP), special operators, and multi-objective density-based fitness function for efficient clustering. This algorithm employs divide and conquer technique for improving the data parallelism and the individuals in the solution space is supported by label based cluster generation mechanism [19].

C. GP based Ensemble Clustering:

Ensemble clustering improves the accuracy, stability and robustness of clustering algorithms [20][21][22]. Ensemble methods can combine both heterogeneous data and various clustering algorithms. Genetic Programming has been successfully implemented in Ensemble Clustering. The DNA-aggregator is a GP based approach to deal with the gene selection and classification tasks for multi class micro array datasets using a tree based individual representation with ensemble K-Means clustering algorithm to discover groups of similar objects from the biological micro-array datasets characterize the underlying data distribution [23].

A Decision Support System (DSS) was built on GP based ensemble clustering techniques, this system ensemble well-known genetic programming techniques namely Linear Genetic programming (LGP), Multi-Expression Programming (MEP) and Gene Expression programming (GEP) through Clustering [24].

Table 2: Overview on GP Clustering Techniques

Type of GP Clustering Algorithm	**Technique applied**	**Representation**
Partitioned	K-Means	Tree
Multiobjective	Divide and Conquer	Graph
Ensemble	Cellular Model	Tree

VII. CONCLUSION

This paper has briefly presented the survey on the Classification and Clustering techniques using Genetic Programming. In, continuation of this we are presently working on the Search Space optimization and efficient representation of the Individual other than the Decision Tree to carry out the efficient data mining tasks using GP.

VIII. REFERENCES

[1]. G.Q Wang and D.Huang, 2007. The summary of data mining technology. Micro Computer Application technology, Vol 2. 2007, pp.9-14.

[2]. Abbas H.A., Sarker R.A., Newton C.S. (eds.), 2001. Data Mining: A Heuristic Approach. Idea Group Publishing, USA.

[3]. John R. Koza. 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA, USA.

[4]. Hong,J.H. and Cho,S.B. 2006. The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming. Artificial Intelligence. Med., 36, 43–58.

[5]. Langdon,W.B. and Buxton,B.F. 2004. Genetic programming for mining DNA chip data from cancer patients. Genet. Program. Evol. Mach., 5, 251–257.

[6]. Yu,J.J. et al. 2007. Feature selection and molecular classification of cancer using genetic programming. Neoplasia, 9, 292–U216.

[7]. Sin Man Cheang, Kin Hong Lee, Kwong Sak Leung. 2003. Evolving data classification programs using genetic parallel programming.URL:http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1299582&isnumber=28874

[8]. Supoj Hengpraprohm. 2008. A Genetic Programming Ensemble Approach to Cancer Microarray Data Classification. The 3rd International Conference on Innovative Computing Information and Control (ICICIC'08), 978-0-7695-3161-8/08 © 2008 IEEE.

[9]. URL:http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4603529&isnumber=4603189.

[10]. DaeEun Kim. 2004. Structural Risk Minimization on Decision Trees using an Evolutionary Multiobjective Optimization. Proc. Seventh European Conf. Genetic Programming, pp. 338-348, 2004. http://www.ti.uni-bielefeld.de/downloads/publications/tree.pdf.

[11]. Pedro G. Espejo, Crist´obal Romero. 2005. Induction of Classification Rules with Grammar-Based Genetic Programming. International Conference on Machine Intelligence. Tozeur – Tunisia.

[12]. Aleksandra Takac. 2004. Application of Cellular Genetic Programming in Data Mining Proceedings of Conference Knowledge. Brno, Czech Republic.

[13]. Gianluigi and Clara Pizzuti etc.. 1999. A cellular genetic programming approach to classification. Proceedings of the Genetic and evolutionary computation conference.

[14]. Jin Li, Xiaoli Li and Xin Yao. 2005. Cost-Sensitive Classification with Genetic Programming. IEEE Congress on Evolutionary Computation, vol.3, no., pp. 2114- 212.

[15]. Xiaorong Chen and Shozo Tokinaga. 2005. Syntactical Recognition Systems of Time Series based on the Genetic Programming and its Applications to Clustering of Stock Prices.

[16]. https://qir.kyushuu.ac.jp/dspace/bitstream/.../1/KJ00004492593.pdf

[17]. Sherin M. Youssef, Mohamed Rizk and Mohamed El-Sherif. 2007. Particle Swarm-like agents approach for Dynamically Adaptive data clustering. International Journal of Mathematics and Computer in Simulation. Issue 2, Volume 1.pp 108-118.

[18]. Nicola Fanizzi, Claudia d'Amato and Floriana Esposito. 2007. Randomized Metric Induction and Evolutionary Conceptual Clustering for Semantic Knowledge Bases. CIKM'07, November 6–8, 2007, Lisboa, Portugal. ACM 978-1-59593-803-9/07/0011

[19]. Tout, S., Sverdlik, W., and Sun, J. 2006. Parallel hybrid clustering using genetic programming and multi-objective fitness with density (PYRAMID). Proceedings of the 2006 International Conference on Data Mining (DMIN'06), Las Vegas, NV, USA, 197-203.

[20]. Shirakawa, S., and Nagao, T. 2009. Evolutionary Image Segmentation Based on Multi objective Clustering. Congress on Evolutionary Computation (CEC '09). Norway, (2009) 2466—2473.

[21]. Alpa Reshamwala, Vijay Katkar and Mamta Ubnare. 2010. Incremental Cluster Detection using a Soft Computing Approach. International Journal of Computer Applications (0975 – 8887). Volume 11– No.8, December 2010.

[22]. Fern, X. and Brodley, C. 2003. Random projections for high dimensional data clustering: A cluster ensemble approach. In Fawcett, T., Mishra, N., eds.: Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), Washington D.C., USA, AAAI Press (2003)

[23]. Topchy, A., Jain, A., Puch, W.,2005. Clustering Ensembles: Models of Consensus and Weak Partitions. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(12) (2005) 1866–1881.

[24]. Kuncheva, L. and Vetrov, D. 2006. Evaluation of stability of k-means cluster ensembles with respect to random initialization. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(11) (2006) 1798–1808.

[25]. Zakaria Suliman Zubi and Marim Aboajela Emsaed. 2010. Using sequence DNA chips data to Mining and Diagnosing Cancer Patients. International Journal of Computers. Issue 4, Volume 4.

[26]. Ajith Abraham, Crina Grosan. 2006. Decision Support Systems Using Ensemble Genetic Programming. Journal of Information & Knowledge Management, Vol. 5, No. 4.

[27]. Pereira F., Mitchell T., Botvinick M. 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45(Suppl. 1), S199–S209. doi: 10.1016/ j.neuroimage.2008.11.007.

[28]. C. D. Stefano, F. Fontanella, G. Folino and A. S. Freca.2011.A Bayesian Approach for combining ensembles of GP classifiers. In the Proceedings of the 10th International Workshop Multiple Classifier Systems (MCS'2011), Naples, Italy, June 15 – 17, 2011.

[29]. D. Parrott, X. Li, and V. Ciesielski .2005. Multi-objective techniques in genetic programming for evolving classifiers. Proceedings of the 2005 Congress on Evolutionary Computation (CEC '05), pages 183–190, 2005.