Genetic Programming for AL goriahms

■ **Ammar .K. AL-Debr *** ■ **Alhadi mohamed yahya****

Abstract

Genetic Programming for Data we have presented two basic techniques that can be used to apply GP to DM problems and the differences between each of them. There was an emphasize on the traditional application of GP after mapping the current domain to a binary domain. Experimental results showed that this technique produces unambiguous rule sets with acceptable accuracy. Still, modified GP produced successful results on other domains but there were no results reported for the census domain (adult domain).

1- Research Genetic Programming for Data

The discovery of inherent knowledge in a large data set is an attractive but complex concept. Due t o evolution of large datasets ,computer algorithms and computer hardware, this concept starts to bloom in the early 90's researcher's suggested a model form of extracting knowledge from the dataset which become the well known association rule's algorithm[1,4].This algorithm was capable of extracting association between different attributes in the data set that fulfill a certain minimum threshold{support and confidence values). Others techniques that have been emerged from the AI and Machine Learning domains include Artificial Neural Networks (ANN's).Inductive-based learning algorithms, Decision Trees and Evolutionary Algorithms(i.e. Genetic

* Ecommerce& data analysis dept faculty of economy & political science Tripoli University

** Ecommerce& data analysis dept faculty of economy & political science Tripoli University

algorithms, Evolutionary Strategies and Genetic Programming)[5,10].Some examples of these techniques are shown in Table 1.

Mining techniques	Approach	Remarks
1.Association rules	Statistical	Discover association between attributes of a table fulfilling a certain threshold.
2. Clustering	Mathematical	Clusters "similar" instances into one class.
3.Decision Trees	Statistical	A decision tree is formed that is capable of differentiating between different classes.
4.Rule Induction	Heuristic	A set of rules that covers one class is formed using inductive algorithms.
5.Artificial Neural Networks	Mathematical	According to the ANN technique used (supervised/ unsupervised)a cluster or a mapping network is formed.
6.Evolutionary algorithms (Mainly Genetic Programming)	Heuristic	May be used to generate concepts or to extract relevant attributes.

Table 1.Examples of DM Techniques

Research in data mining has been exploring different areas like new methods/techniques/algorithms for algorithms for extracting knowledge, use of parallel architectures for DM applications p11[search space pruning(attribute selection)[12] and devising new measures for knowledge quality. One of the new and appealing ideas is researchers are focusing on

right now, namely the "interestingness" of the discovered knowledge [13-14].The mining process is in general user dependent and it should be noted that, generating a concept may not be useful if it is obvious .In fact users are more interested in "what they don to know about their data or what they can't conclude by themselves by looking into the dataset. Thus "interestingness "as measure of rule importance has been investigated. A generated concept or rule maybe considered interesting if it covers the specified class and "was not expected "by the users. The term "not expected" is usually interested as the "generated concept contains attributes that were considered insignificant". Another interpretation is that the user was focusing on other attributes that did not appear in the generated concept/rule. Thus this term "interestingness" can be defined subjectively and objectively. Note that still, objective factors are the ones discussed up till now since subjective ones are harder to define and differ between different user and domains.

Another dimension in DM is the availability of tools to accomplish the mining task. Currently there are many tools available that support different data formats and at the same time provide the user with different techniques (statistical/Al-based) for DM.

In this paper we are discussing the use of GP as a tool for the DM problem. The paper is organized as follows: Section 2 describes traditional GP as a search algoritms .Section 3 presents and discussed two different approaches for applying GP to the DM problem. Section 4 describes the domain of experiments, the census data set. Experiments and results are presented in section 5 finally the paper is concluded in section 6.

2- Genetic Programming

Genetic Programming (GP) represents one of the main area of research in evolutionary algorithms (EA).it has been proposed been [15] and since then has been applied to many applications ranging from computer programs evolution up to electronic circuit design [16].Still, GP is an evolutionary base search that algorithm that perform same general evolutionary cycle shown in fig 1.However unlike other EA systems, GP applied to symbolic representations. Thus the representations scheme, and the genetic operators have been redefined and even more domain related operators has been devised

to match the needs of definition language and application domains. In this section, a brief discussion of GP knowledge representation and operators is presented.

Knowledge representation

Traditionally GA and Evolutionary Strategy were applied to binary and numeric chromosomes respectively. These representations were adequate for many applications like optimization problems, and adaptive control. For GP, it is required to evolve computer programs, which are based on a different linguistic structure. Thus, GP was applied initially to S-expressions (LISP Symbolic expressions), where an expressions can be represented as a tree.fig 2. Shows an example of an S-expression. Notes the operators (or Functions) are represented as parent nodes within the tree and operands represent the terminal nodes.

Define domain (function set and terminal ser);
Generated initial population;
Evaluate population;
Repeat until total number of generations is exhausted
Reproduce and select new population;
Apply genetic operators;
Evaluate population;
Select best individual;
End repeat;

It should be noted that the tree representation is very flexible for problem representation and in the same time for genetic operators. The definition of the genetic operators defined in GP is discussed in the next section.

3- Genetic Operators

The basic genetic operators are crossover, mutation and inversion. They have been defined and widely applied to binary and numeric representations. However, to apply the same operators in GP they should be redefined. Crossover is redefined as follows:

Two new offspring's are formed from tow randomly selected parents as a result of exchanging sub-trees of the parents. The crossover point objected to define the root node for the sub-tree that will be subjected to crossover. The

mutation operator can be defined in different forms:

1.The replacement of a sub-tree with a randomly formed one.

2.The change of the value of a terminal node.

3.The change of the function defined in a parent node.

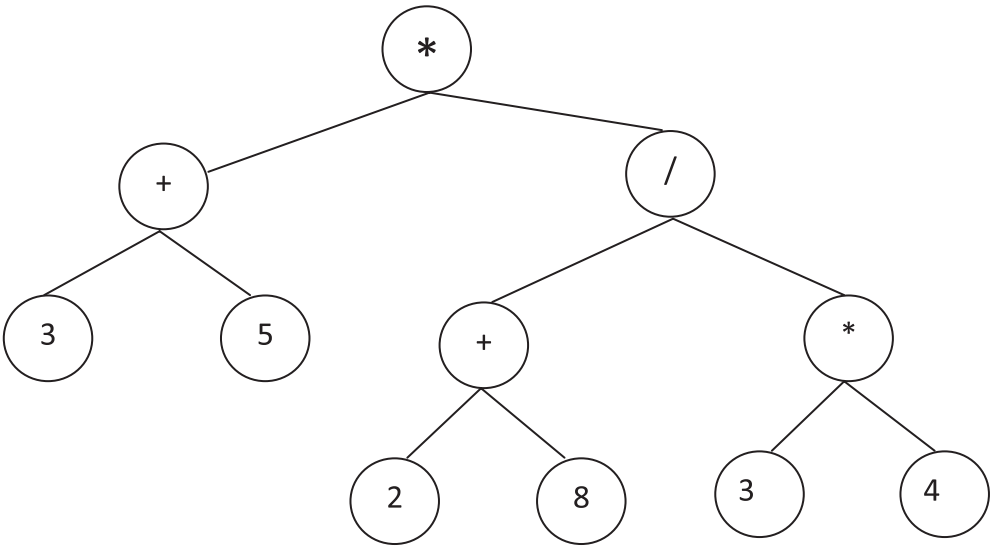
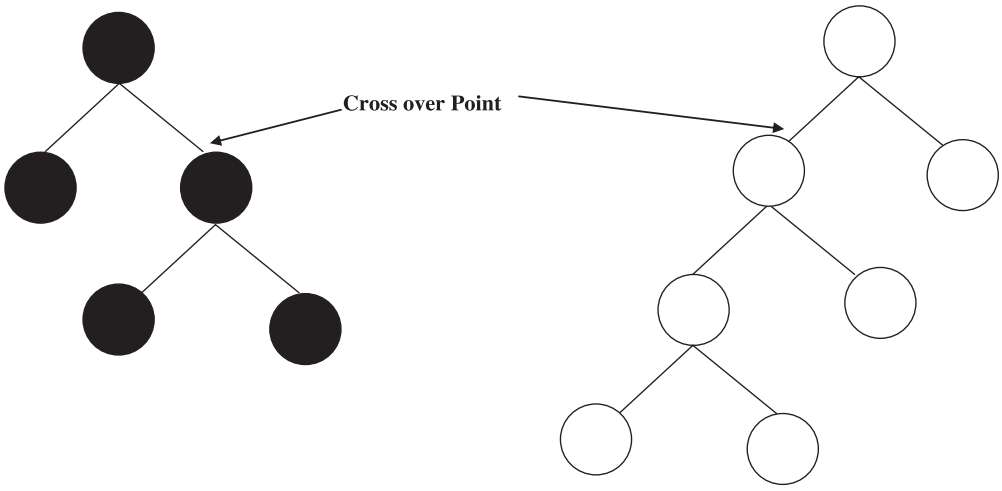


Fig 2 A simple example of S-expression (*(+ 3 5) (/ (+ 2 8) (* (3 4))) in a tree representation

The inversion operator is a unary operator (applies to one tree) and it is defined like a crossover operator. Other operator have been defined and applied to different applications but we'll restrict our discussion to the basic genetic operator's .Figure 3 shows the result of applying the crossover operator



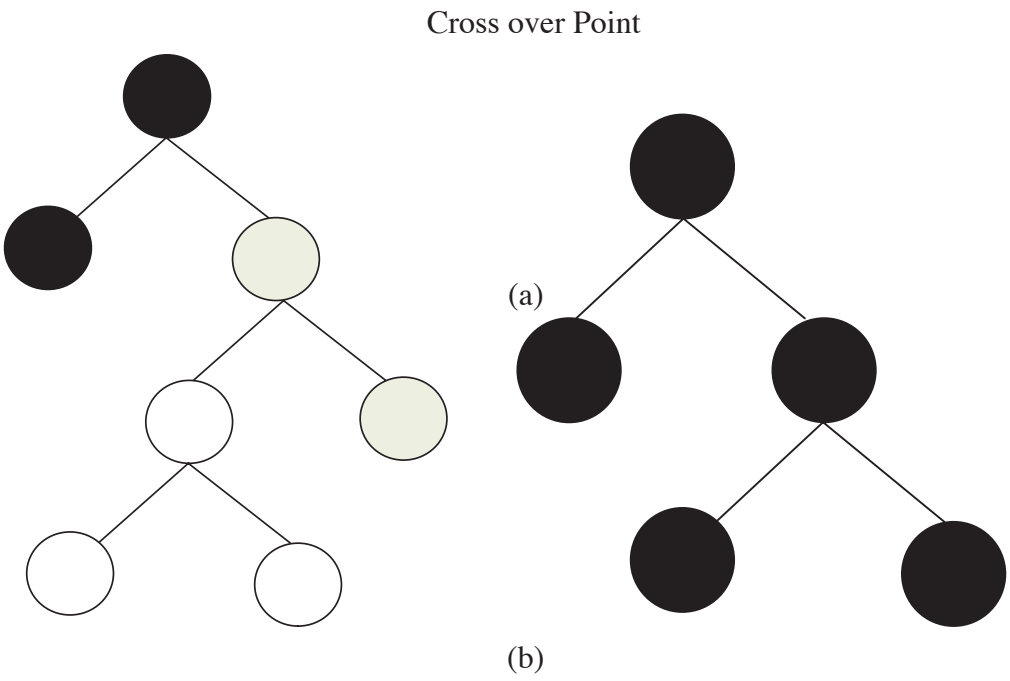


Fig 3 Two parent chromosomes (trees) and the crossover point (b) .The resulting offspring's after crossover.

GP has evolved in the last decade to include more operators that are imperative for the discovery of new programs. For example many architecture-altering have been defined for creating subroutines, adding and deleting subroutines, Automatic Defined Functions (ADF's) loop (ADL's) and iterations (ADI's).Most of these operators are background operators i.e. they operate at a very low rate as compared to the crossover operator.

4. Apply GP to Data Mining

The Data mining problem can be viewed as a search for the best hypothesis (or rule) in the space of possible hypothesis.

The dimensionality of the hypothesis space is directly proportional to the number of attributes available in the domains under consideration.

Noise is a major factor that affects the quality of the discovered rules.

The major sources of noise are:

1.Data inaccuracy.

2.Missing attribute values.

Thus, there is a need for a robust search algorithm capable for exploring the search space effectively and at the same time can be easily tuned to accommodate other features required in the discovered hypothesis (for example, simpler rules or interesting rules).GP seems to be the best candidate for this task, however there is one major problem that should be taken into consideration namely:

GP operators don not perform any semantic check on the resulting chromosomes. Thus semantically incorrect chromosomes may be generated during the GP cycle.

This problem has been encountered when Genetic Algorithms (the ancestor of GP) were applied to semantically restrict symbolic domains. In general we can specify two approaches for applying GP to DM:

1-To apply GP operators to the search space without any semantic modifications or restrictions.

2-To apply a semantically modified version of the GP operators.

Both approaches are valid. While the first one preserves the basic structure and theory of traditional GP, the second one restricts the search space to valid hypothesis / rules only.

The following section discusses how to apply each of these approaches and their advantages and disadvantages.

4.1. Applying a restricted form of GP to DM

Restricted forms of GP are systems that apply genetic operators to programs (GP chromosomes) to produce only semantically valid offspring's. Recently a constrained-based GP system has been proposed as a DM tool for discovering interesting rules [17]. In this case, allowing crossover and mutation operators to generate only legal trees restricted the structure of the generated tree using GP. A legal tree is one that fulfills the following restrictions:

1. The child of an AND function is either an AND function or an operator from the set of operators {>, <, =, <>}.
2. A child node of a comparison operator should be either an attribute or an attribute value.

3. Any attribute can occur only once in the rule tree antecedent.
4. The input and output data types are pre-specified and only parameters following these are allowed.

These rules are enforced during GP generations and maintained throughout the applications of the GP operators. The proposed model was applied to three diseases drawn from the medical domain:

1.Breast cancer with two classes, 138 records and 7 attributes.

2.Dermatology with 6 classes 366 records and 34 attributes.

3.Chest pain with 12 classes 138 records and 161 attributes.

It has produced acceptable results (simpler rules with a little bit higher accuracy) as compared to the well known decision tree algorithm ID3 (the tool used was C4.5).

The obvious advantage of such system is its ability to limit the search space to valid rules only, i.e. rules that are not ambiguous and are well structured. On the other hand there is a book keeping cost associated with these constraints. It is required to perform checks after each application of generic operators to make sure that the resulting offspring follows the pre-set constraints.

4.2. Applying traditional GP to DM

To apply traditional GP to DM problems, we must take into consideration an important aspect of GP, namely the closure property. The closure property implies that an output of a function can be input to another one with no restriction. This property although acceptably in many situations, is not acceptable for DM since different attributes may have different data types and can be used only with a specific set of function. This is one of the main reasons for having the above mentioned constraints when applying GB to DM problems.

However, this aspect to be overcome by using a mapping function that maps non-binary attributes to binary domain. The mapping can be described as follows:

- Logical attributes: not effected.
- Categorical attributes: each value of the attributes is considered as a new attribute that may have one or two values, either true (exists) or false (does not exist).

● Numerical values: transformed into ranges and each range is then treated as a new attribute that may have one of two values (True or False).

As an example, consider the attribute race in the census domain. Initially it may have one of the following values:

{White, Asian-Pac- Islander, Amer-Indian-eskimo, Other, black}

After applying the mapping function to this attribute, it will be transformed to five binary attributes of the form:

{IS white,Is Asian-Pac-Islander, Is Amer-Indian-eskimo, Is Other, Is black}

Each of these new attributes may only have one value, zero or one (or true or false). Although the number of the attributes in the search space has increased, the formation and the evaluation of the rule tree became much easier with no need for semantic constraints. One should notice that the binary transformation did not affect the number of examples (tuples) in the data set.

In fact it has re-distributed these tuples in a different pattern. Hence, the problem is converted to a binary optimization problem.

After the binary mapping phase traditional GP is applied and trees are evaluated based on their completeness and consistency given a specific training set. A test set formed of unseen examples is used to check the accuracy of the best discovered tree.

The proposed cycle (mapping then application GP) is applied to the census domain to find out the possibility of using unconstrained traditional GP for DM applications.

5. The Census domain (Adult domain)

The census domain (know now as the adult domain) consist of 14 attributes (6 continuous numeric and 8 nominal) and 48842 records. The domain has been studied as a good example for DM techniques and tools since:

1.Number of records is quite large.

2.Attributes values are continuous and nominal.

3.Many attributes may be considered as irrelevant to the mining task, thus they represent some overheads (noise) for the mining algorithm.

4.Another source of noise that exists within the data set itself is that 7% of the data set has missing values.

A sample results of applying DM techniques has been reported in [8] and it is shown in table 2, the complete environment can be downloaded from the UCI Machine Learning repository web site: http://www.ics.uci.edu/mlearn / Machine-Learning.html

Table 2. Sample results of different DM techniques to the census domain

No.	Techniques/Algorithm	Error Value
1	C4.5 rules	14.94
2	Voted ID3 (0.6)	15.64
3	HOODG	14.82
4	FSS Naïve Bayes	14.05
5	NB Tree	14.10
6	C4.5-auto	14.46
7	IDTM(Decision Table)	14.46
8	OCI	15.04
9	C4.5	15.54
10	Naïve-Bayes	16.12
11	Voted ID3(0.8)	16.47
12	Nearest-neighbor(3)	20.35
13	Nearest-neighbor(1)	21.42

The goal of the mining task is to find the hypothesis that classifies those who earn more than $50K/year and those who don't.

6. Experiments

In this research GP has been applied to the census data after mapping the domain attributes to binary domain. Experiments have been conducted to extract complete and consistent rules while maintaining the simplicity of the

discovered rules themselves. A sample of the resulted rule se(s) is shown in table 3 together with the corresponding criteria used for evolving the rule set, achieved accuracy on the test set number of hits and number of misses.

Table 3 Discovered rules

Rule Set	Discovered Rules	Criteria	Accuracy	Hits
Rule 1	AND (Has B.Sc., capital Gain Greater Than Zero Hoursperweek>40>)	Normal No(bias)	82.7%	12416
Rule2	Has Ph.D.			
Rule3	AND (CapitalGainGreaterThanZero, OR (HasMs. IsSelfEmployedInc.)			
Rule4	AND (IsMarried,Or(HasB. Sc., HasM.Sc.,IsSelfEmpInc, USAmerican CapitalLossGreaterThanZero))			
Rule1	CaptialGainGreaterThanZero	Simpler	78%	11714
Rule1	HasPh.D	Normal (different Seed)	82.6%	12395
Rule2	AND (IsMarried,(OR (CapitalGainGreaterThanZero, HasM. Sc.,IsFederalGov.,HasB.Sc., CapitalLossGreaterThanZero))			

By inspecting the outcomes of the experiments it is quite interesting to

notice that all discovered rules share the same attributes (sometimes the same expressions). There are only 7 attributes out of 14 that appear in the resulting rules. Another interesting remark is the accuracy of the discovered rules which ranges from 78% to around 83%.

As compared with previous techniques applied to this domain, traditional GP is comparable with the nearest Neighbor algorithms and Naïve Bayesian classifiers in terms of accuracy. Interestingly, the attribute CapitalLoss Greater ThanZero (these are two different attributes in the original data set with continuous numeric values).Although conflicting but this kind of conflict cannot be detected even if a constraint-based version of the GP system is used. The reason is that these attributes are treated as two different attributes and there is no relation that describes how they should be treated if both of them appeared in the discovered rule. One final note, the resulting rule sets are kind of clear and readable.

For the purpose of this research the test and training sets have been set to 30,000 and 15000 respectively. The function set used is composed of {AND, OR, NOT} and the terminal set is {True, False}. The Process was executed for 50 generations and the population size was set to 100. The initialization is done using the Half builder technique (half the time the initialized trees are GROWN to a randomly selected tree size, usually between 2 and 6 while the rest of the time, trees are built to the actual full tree size). Crossover probability is set to 0.9 and the selection of the parents uses Tournament selection. Mutation probability 0.1 is used and the max. Tree depth is set to 17. The GP simulator used is JAVA evolutionary computation simulator release 8 by Sean Luke.

7. Conclusion

In this paper, we have presented two basic techniques that can be used to apply GP to DM problems and the differences between each of them. There was an emphasize on the traditional application of GP after mapping the current domain to a binary domain. Experimental results showed that this technique produces unambiguous rule sets with acceptable accuracy. Still, modified GP produced successful results on other domains but there were no results reported for the census domain (adult domain).

References

[1.] R.Srikant, R.Agrawal, "Mining Generalized Association Rules", Proc.Of the 21st Int'l Conference on Very Large Databases,Switzerland September,1995.

[2.] R.Srikant, R.Agrawal, "Mining Quantitive Association Rule" in large relational tables' Proc. Of the ACM-SIGMOD Conference on knowledge discovery in database and data mining, Portland, Oregon 1996.

[3.] R.Srikant, R.Agrawal, "Mining sequential patterns: Generalizations and Performance Improvements" Proc. Of the fifth Int'l conf. On extending Database Technology (EDBT), France, March 1996.

[4.] Usama Fayyad, George Peteteski-Shapiro, Padhairc Smyth, and Ramasamy Thurusamy, "advances in knowledge discovery and Data minding", AAAI Press/The MIT Press,1996.

[5.] A.A Freitas, A survey of evolutionary algorithms for data mining and knowledge discovery. To appear in: A Ghosh and S.Tsutsui. (Eds)Advances in Evolutionary computation springer-verlag,2002

[6.] Christopher J. Matheus, Philip K. Chan and Gregory Piatetsky Shapiro, "Systems for knowledge discovery in databases", IEEE Transactions on knowledge and Data Engineering, Vol., 5, No.6, Dec.1993.

[7.] Quinlan,J.R. (1986a), "Introduction of Decision Trees", Machine Learning Journal (1),81-106.

[8.] J.Han, Y.Fu. "Discovery of Multi-Level Association Rules from Large Databases". Proc. Of VLDB conf. Switzerland, September 1995.

[9.] Kaufman, K.A., Michaski,R.S. and Kerschberg, L."Mining for knowledge in Database: Goals and General Description of the INLEN System", in knowledge Discovery in Databases, Piatetski-Shapiro, G. and Frawley, W.J. (Eds). AAAI Press/The MIT, Menlo Park, CA 1991.

[10.] H.Han and Y.Fu. "Exploration of the Power of Attribute-Oriented Induction in Data Mining", U.M. Fayyad, G.Piatetski-Shapiro, P.Smyth, and R. Uthurusamy (eds.), advance in knowledge Discovery and Data mining AAAI/ MIT Press.1996.

[11.] A.A. Freitas, Simon H.Lavington, "Mining Very Large Database with Parallel Processing" Kluwer Academic Publisher, Boston.1999.

[12.] A.A. Freitas, "A summary of the Papers Presented at the AAAI-99 & GECCO-99 Workshop on Data Mininig with evolutionary Algorithms: Research directions. (1-page extended abstract). Proc. Of the GECCO-99 workshop Program, 226. Orlando, FL, USA. July 1999.

[13.] Freitas, A.A. "On rule interestingness measures", knowledge-based Systems 12 (1999) 309-315

[14.]E. Noda, A.A.Freitas, H.s. Lopes. "Discovering interesting prediction rules with a genetic algorithm." Proc. Congress on Evolutionary Computation (CEC-99) 1322-1329. Washington D.C. USA, July 1999.

[15.]Koza J.Genetic "Programming: On the Programming of Computers by Means of Natural selection", the MIT Press 1992.

[16.]Koza J., Bennett III F., Andre D., and Keane M. Genetic Programming III: Darwinian Invention and Problem solving, Morgan Kaufman 1999.

[17.] C.C. Bojarczuk, II S. lopes, and A.A. Freitas, "Data Mining with constrained-syntax genetic programming applications in medical data sets". Proc. Intelligent Data analysis in Medicine and Pharmacology (IDAMAP-2001), a Workshop at Medinfo -2001. London, UK, SEP-2001.

[18.] R.Kohavi, "Scaling up the accuracy of Naïve-Bayes Classifiers: A Decision-Tree Hybrid" Proceedings of the Second International conference on knowledge Discovery and Data mining, 1996.