5-18-2013

Heuristic Guided Evolution

Hayden Wimmer
UMBC, hwimmer@bloomu.edu

Roy Rada
UMBC, rada@umbc.edu

Heuristic Guided Evolution

Hayden Wimmer
UMBC
hwimmer1@umbc.edu

Roy Rada
UMBC
rada@umbc.edu

ABSTRACT

Exploiting knowledge to guide the evolutionary process in evolutionary computing is a concept that has the potential to increase the performance of evolutionary algorithms. The research question of this paper is "*Can heuristics derived from past experiences be incorporated into evolutionary computing in order to increase the performance?"* In order to answer the research question the following hypothesis is developed: *"A heuristically-guided mutation of decision trees will outperform randomly mutated decision trees in terms of classification accuracy."*The methodology for answering the hypothesis is an experiment that tests a knowledge-guided mutation of a decision tree using heuristics created from prior decision trees as a form of knowledge. This is compared with a random mutation of the same decision tree. This experiment supports the theory that using knowledge in the form of heuristics to guide mutation will produce a difference in the performance of the classification of data instances. This supports the need for further research into knowledge guided evolutionary algorithms.

Keywords

Evolutionary Computing, Genetic Algorithms, Knowledge. Heuristics

INTRODUCTION AND RELATED WORKS

Stock market portfolio optimization is a challenge that has had many unique approaches. Data mining approaches are common in order to investigate patterns in large datasets. Commonly invested large data sets include historical stock market and publically available fundamental financial data such as balance sheets and income statements. One such approach in data mining is the use of decision trees and decision rules (Apté et al. 1997). Due to dynamic and changing stock market patterns predicting future patterns is a difficult task. Employing evolutionary computing is an accepted technique when data and patterns frequently change thereby making models of the data obsolete as time progresses. There is research into using genetic algorithms in stock selection where high performing stocks are identified (Zhou et al. 2006). Genetic algorithms and evolutionary computing have been applied to data mining techniques. One such approach is to implement genetic algorithms upon decision trees in order to "build a better mousetrap." This has been implemented with statistical methodologies (Fu 2002). Additionally, employing cross-over mutation between decision trees created by the C4.5 algorithm has been studied and tested against standard decision trees (Fu et al. 2003). Merging genetic algorithms with financial portfolio optimization is a logical next step. Decision trees generated by the C4.5 algorithm may be mutated using different techniques. One technique is to employ a random mutation to the resulting decision tree. Another approach is to use knowledge-guided mutation. Knowledge guided mutation may arise in many forms. One such form would be to examine past C4.5 generated decision trees as knowledge structures and employ a cross-over technique such as employed in (Fu et al. 2003).

This work begins for describing the general framework employed in this experiment which includes terminology and background information on decision trees and genetic algorithms and ends with our hypothesis. Next, the methodology of the experiment is discussed which includes information about the dataset, the methods for analyzing the data, and how the genetic process was conducted. Following the methodology section, results are presented in the tests of random versus heuristic guided tests. The work is finalized with conclusions and future directions.

FRAMEWORK

Decision trees are directed graphs which contain nodes and edges that may be utilized in order to support the decision making process. A node is an attribute of the underlying dataset where the decision tree splits based on certain criteria– or creates new branches. Nodes are also referred to as vertices in graph theory. An edge is defined as a set of two nodes. In decision trees the edges are directed edges or arcs. Terminal nodes – nodes which contain no more splits or branches – are referred to as leaves. The depth of a decision tree is determined by the maximum number of nodes that are traversed before a leaf – of terminal node – is reached. Nodes may be references via their depth within the tree. The node where the initial split occurs is referred to as a depth of 0 or the root node. The level – or depth – of a node within the tree is computed as its distance to the root node. The tree contains leaves – or nodes – which are an attribute from a set of data. Decision trees will split on the attributes. A split is where a node then creates new branches based on some criteria of the attribute in which it split. For example, if we were determining if we should purchase a product one attribute we may consider is price. We may split our decision to purchase based on a high or a low price. If the price is low we decide to purchase the product and if the price is high we do not purchase the product. Decision trees are frequently converted into rules for decision making. The above scenario may be represented as rule set or a decision tree. The example in *Table 1* shows a decision tree with only one level of depth and the associated rule set.

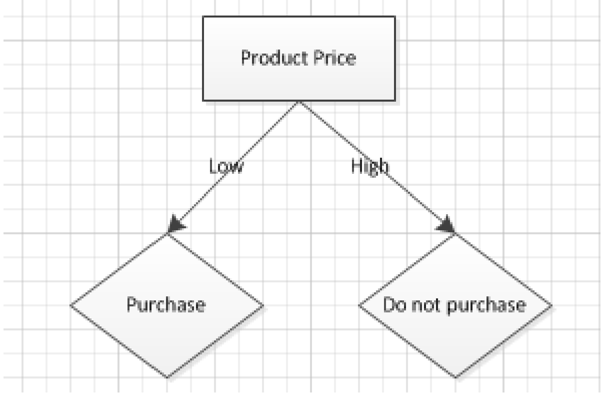
Rule Set	Decision Tree
IF Price = LOW THEN Purchase Product Else Do Not Purchase Product End IF	

Table 1 – Rule Extraction from a Decision Tree

One such example is utilizing a genetic algorithm to determine which attribute would be ideal for a split.

Genetic algorithms may be employed in order to modify decision trees. A genetic algorithm is an algorithm that follows a path of evolution in order to achieve a goal – typically and ideal solution which is referred to as fitness. A genetic algorithm works on populations of organisms. A population of organisms is considered a group of organisms. An organism is a collection of genes. Genes are also referred to as chromosomes and are attributes of the organism related to the ability to solve the given optimization problem. Organisms within the population are altered using different methods which include cross-over where two organisms are combined to form a new organism. Another method is mutation where organisms' genes are changed. Organisms are evolved and the organisms with a high level of fitness remain in the population and less-fit organisms are discarded. Each occurrence of this process is referred to as a generation. Genetic algorithms randomly mutate an organism and then test the fitness or how well the organism performs or solves the problem. Random mutation is when the organism's genes are modified randomly or without knowledge. In the case of a decision tree, modifying which attributes constitute splits without any form of knowledge or guidance is a random mutation. Another method of mutation is knowledge guided mutation. Knowledge guided mutation occurs when human knowledge is codified and integrated into the process of mutation within the genetic algorithm. Examples of codified knowledge are semantic networks, ontologies, domain knowledge, and literature. In addition to codified human knowledge, heuristics may be employed in the mutation process. In the mutation phase heuristics from past examples may be incorporated to guide the mutation process. This leads us to the following hypothesis:

Hypothesis: A heuristically-guided mutation of decision trees will outperform randomly mutated decision trees in terms of classification accuracy.

METHODOLOGY

Data for the preliminary experiment was retrieved from the COMPUSTAT global database. The COMPUTSTAT database follows a relational database model and contains attributes extracted from fundamental data such as an organization's financial statements. Attributes are grouped by the primary key (GVKEY) and year. Attributes include fields such as Cost of Goods Sold, EBITA, and Common Shares Outstanding. Data from the years 2000 through 2006 was retrieved. The data was discretized by dividing the data into quintiles in the following set {1, 2... 5}. The total data set size was 11,828 records.

In addition to discretization, the dataset was restricted to only data from GBR and companies that remained active throughout the period 2000 through 2006. There were two reasons for only selecting this subset of data. First, currency fluctuations between countries pose difficulties accurately standardizing the currency - especially over time. Second, different markets exhibit different data trends and patters. Factors for the difference in trends and patterns include macroeconomic factors such as political environment and local economy. Finally, some attributes were removed from the dataset. Each attribute was counted based on the number of instances in which it was present. Any attribute that was not present for 90% of the data instances from 2000-2006 was removed. A total of 68 data attributes remained. The attribute P1PriceDiv was selected to build a target. This attribute is the geometric mean of the past two years values of the PriceDiv which is defined as the price attribute added to the dividend attribute (price + dividend). This target attribute was named direction and was computed as a nominal data type and was based on data from the subsequent year and had the values {UP, DOWN}. The P1PriceDiv from the subsequent year was utilized to determine if the P1PriceDiv would increase or decrease. In *Equation 1* let P = P1PriceDiv.

$$\boxed{Direction = \begin{Bmatrix} \text{" }UP\text{ " if } P_{year} < P_{year+1} \\ \text{" }DOWN\text{ " if } P_{year} \geq (P)_{year+1} \end{Bmatrix}}$$

Equation 1 – Calculation of Direction

For example, if GVkey(X) had a P1PriceDiv of 1 in year 2000 and 2 in 2001 then the attribute direction for the year 2000 dataset would be listed as "UP."

The completed data set was then imported into WEKA to process with J48 which is WEKA's implementation of the C4.5 algorithm. This scenario is replicated four times employing decision trees created from the C4.5 decision tree algorithm as implemented in WEKA (Holmes et al. 1994; Mark et al. 2009).The decision trees were trained and tested with 10 fold cross-validation with a minimum of 10 instances to constitute a split. Evaluation is based upon the accuracy of correct classifications. There were three sample exercises conducted for four total years using the aforementioned parameters. The test tests will be referred to as actual, random, and heuristic-guided. The exercise is based on a single decision tree organism with branches representing attributes. The branches may be weighted or numbered in order to represent their level of depth or proximity to the root node of the decision tree.

The actual test is when a decision tree is constructed with data from year n and trained to predict the direction of the p1pricediv in year n+1. An example is 2001 data used to train the decision tree in order to predict the direction of the 2002 p1pricediv. This is considered the standard by which the remaining two models may be evaluated.

In order to perform the following random mutation and heuristic-guided tests WEKA was utilized with separate training and testing data sets. A training set was constructed in order to force the resulting decision tree to take on the structure – or directed graph - that was predetermined for the test. Then, this structure was tested with the same data set as the actual test above.

The next test is the random mutation test. In this test, the decision tree created from training on year n to predict the direction of p1pricediv in year n+1 is randomly mutated. Any node that splits on depth level three is randomly changed to any other attribute or potential node. Therefore, any random attribute is applied to depth level 3 of the actual decision tree and the resulting decision tree is then trained with year n data to predict price direction for year n+1.

The final testing scenario is the heuristic-guided test. This test uses knowledge in the form heuristics from a decision tree from a prior year. The decision tree from the test actual from year n and n-1 are combined. The decision tree from year n is mutated using cross-over with year n-1. A node from tree n-1 on the depth level 2 is then replaces a node of depth level 3 of the decision tree from year n. The resulting tree is then utilized to predict the direction of p1pricediv for year n+1. The figure below details the mutation process for the heuristic-guided mutation. An example of this process is shown in *Figure 1.*

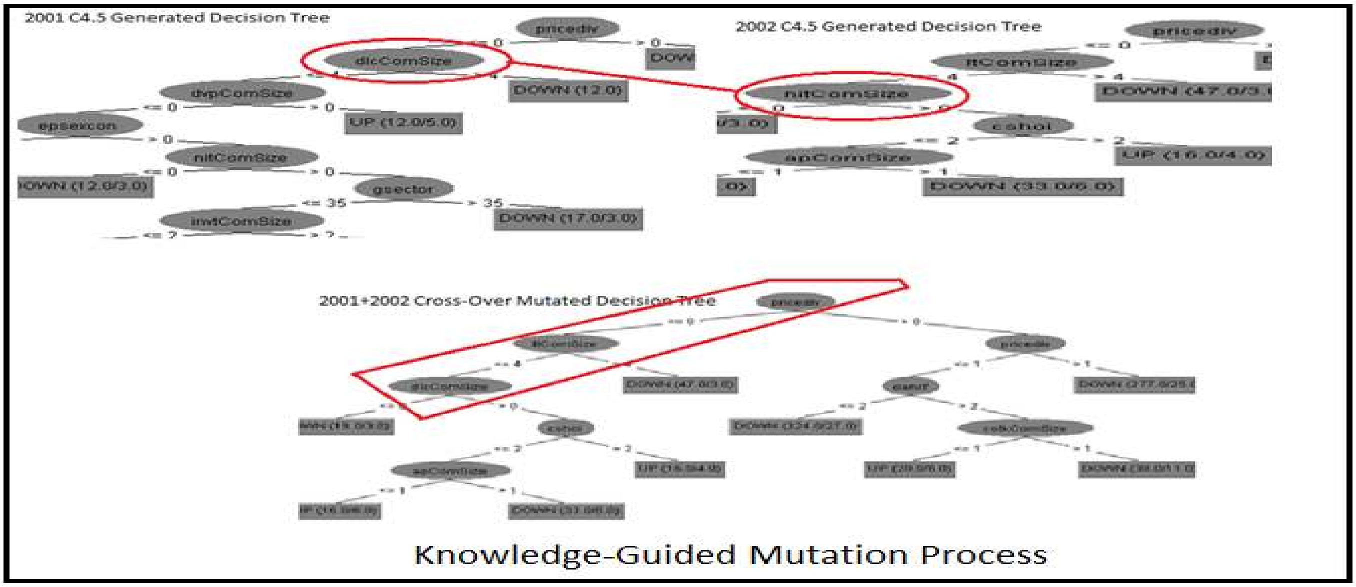


Figure 1 – Using past decision tree organisms to form a new organism via controlled cross-over

RESULTS

The following table and chart illustrates the success rates as a percentage of correct classifications. This leads us to conclude there is a difference between the heuristic-guided and random mutation's performance. The hypothesis "*A heuristically-guided mutation of decision trees will outperform randomly mutated decision trees in terms of classification accuracy*" has been answered by demonstrating that heuristic guided mutation can outperform random mutation. The results for the experiments are detailed in *Table 2.*Based on the results one may conclude the research question has been answered and heuristics from past experiences can have an effect on the evolutionary mutation process. The performance as measured by classification accuracy is illustrated in *Figure 2.* Currently, sufficient data exists for an applicable statistical test; however, based on the difference in performances of classifications between the models additional research is necessary to investigate this phenomenon. All decision trees and WEKA output as produced from the above experiments are available by contacting the primary author.

Data Year	Prediction Year	Heuristic-Guided	Random	Actual
2001	2002	77.34	74.36	82.63
2002	2003	81.06	80.54	85.31
2003	2004	80.56	80.45	82.44
2004	2005	65.84	63.97	78.40
Average		76.20	74.83	82.19

Table 2 – Comparison of Random mutation and Heuristic-Guided mutation for Classification Accuracy

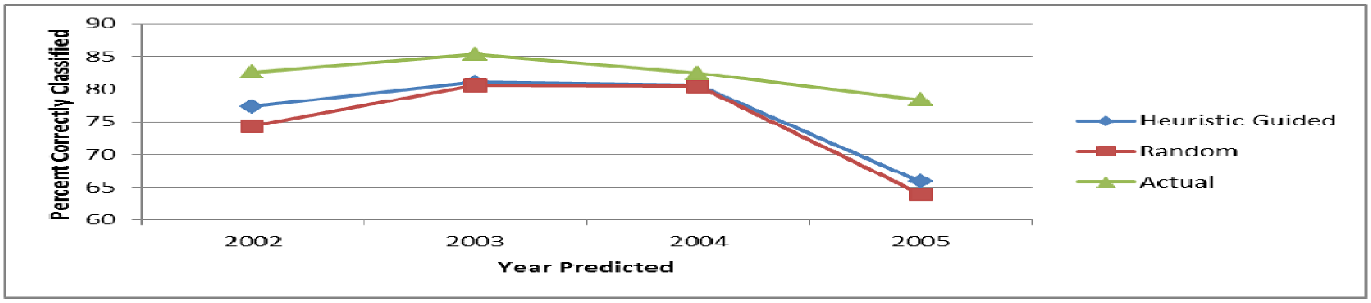


Figure 2 – Performance of Random, Heuristic-Guided, and Actual Decision Tree Performance as Measured by Classification Accuracy

CONCLUSIONS AND FUTURE WORK

In this research we examined the role of knowledge in the evolutionary process. Decision trees were created from financial data. We posed the research question: "*Can heuristics derived from past experiences be incorporated into evolutionary computing in order to increase the performance?"* Next, the hypothesis was developed as: *"A heuristically-guided mutation of decision trees will outperform randomly mutated decision trees in terms of classification accuracy."*Heuristics were extracted from the decision trees and applied to the evolutionary mutation process. This was compared with a random mutation of the decision trees. Our results indicate there is an advantage to employing knowledge in the mutation process. Future work will expand upon this study to incorporate knowledge structures such as semantic networks and ontologies into the evolutionary process. Additional future work will incorporate a cost benefit analysis to determine if the performance gains are worth the additional computational overhead. Besides the cost benefit analysis, a larger and more extensive data set will be employed and a greater number of generations will be included in the testing processes. The technical implications of this research have the potential to increase the efficiency of evolutionary algorithms. The managerial implications of this research aim to create high quality knowledge management systems to aid in the decision making process.

WORKS CITED

Apté, C., and Weiss, S. 1997. "Data mining with decision trees and decision rules," *Future Generation Computer Systems* (13:2–3), pp 197-210.

Fu, Z. 2002. "Using Genetic Algorithms-Based Approach for Better Decision Trees: A Computational Study Discovery Science," S. Lange, K. Satoh and C. Smith (eds.), Springer Berlin / Heidelberg, pp. 195-198.

Fu, Z., Golden, B. L., Lele, S., Raghavan, S., and Wasil, E. A. 2003. "A Genetic Algorithm-Based Approach for Building Accurate Decision Trees," *INFORMS J. on Computing* (15:1), pp 3-22.

Holmes, G., Donkin, A., and Witten, I. H. Year. "WEKA: a machine learning workbench," Intelligent Information Systems,1994. Proceedings of the 1994 Second Australian and New Zealand Conference on1994, pp. 357-361.

Mark, H., Eibe, F., Geoffrey, H., Bernhard, P., Peter, R., and Ian, H. W. 2009. "The WEKA Data Mining Software: An Update."

Zhou, C., Yu, L., Huang, T., Wang, S., and Lai, K. 2006. "Selecting Valuable Stock Using Genetic Algorithm Simulated Evolution and Learning," T.-D. Wang, X. Li, S.-H. Chen, X. Wang, H. Abbass, H. Iba, G.-L. Chen and X. Yao (eds.), Springer Berlin / Heidelberg, pp. 688-694.