Discovering The Classification Rules For Egyptian Stock Market Using Genetic Programming

Samah Refat[1], Mohammed El-Telbany[2], Hisham Hefny[3], Ahmed Bahnasawi[4]

[1] Girls College for Arts, Science and Education, Ain Shams University, Cairo, Egypt
[2] Computers and Systems Dept., Electronics Research Institute, Dokki, Egypt.
[3] Computers Science Dept., Statistical Studies and Research Institute, Cairo University, Egypt.
[4] Electronics and Communications Engineering Dept., Faculty of Engineering, Cairo University, Egypt

***Abstract**-Applications of learning algorithms in knowledge discovery are promising and relevant area of research. It is offering new possibilities and benefits in real-world applications, helping us understand better mechanisms of our own methods of knowledge acquisition. Genetic programming (GP) posses certain advantages that make it suitable for discovering the classification rule for data mining applications, such as convenient structure for rule generation. This paper, intended to discover classification rules for the Egyptian sock market return by applying GP. Since the Egyptian stock market return data have a large number of specific properties that together makes the generalized classification rules unusual. The process behaves very much like a random-walk process and regime shift in the sense that the underlying process is time varying. These reasons cause greats problems for the traditional classification algorithms. Experiments presenting a preliminary result to demonstrate the capability of GP to mine accurate classification rules suitable for prediction, comparable to traditional machine learning algorithms i.e., C4.5.*

I. INTRODUCTION

Recently, there has been upsurge of interest in the area of *data mining* in knowledge discovery in financial data and discovering the classification rules or trading rules is one of the fundamental activities in data mining which are unconcerned with continual price prediction and wait for a certain conditions to trigger by or sell actions. The use of GP for discovering *accurate*, *comprehensible* and *predictive* classification rules is relatively under-explored area. It is a promising approach due to their effectiveness in searching very large spaces and the ability to perform global search for candidate rules. In comparison with the traditional decision tree techniques (e.g., CART [13]; C4.5 [16]) which provided relatively accurate and understandable rules, they generated rules are more complex than necessary. Moreover is local or greedy search technique since it selects only one attribute at a time based on the entropy measure, and therefore the feature space is approximated by a set of hyper-cubes [8]. GP classifiers have been proposed alternative methods. They posses certain advantages that make it suitable for classification, such as convenient structure for rule generation, for instance. Further, they perform a global search in which genetic operators can modify many different combinations of attributes - using the several different functions available in the function set. Hence, even if the original attributes do not have much predictive power by themselves, the system can effectively create "*derived attributes*" with greater predictive power, by applying the function set to the original attributes [3][1]. Finally, it offers wide range of possibilities of variations and modifications of algorithm that may lead to improve overall performance of the application. In this paper, we are interested in using GP technique for automatic rule discovery and the evaluation of its performance over real world data of the Egyptian stock market represented in the Commercial International Bank (CIB) is presented. Especially, the Egyptian stock market return data have a large number of specific properties that together makes the generalized classification rules unusual. The process behaves very much like a random-walk process and regime shift in the sense that the underlying process is time varying. These reasons cause greats problems for the traditional classification algorithms. This paper is organized as follows. Section 2 presents an overview of GP. Section 3 describes the related work of using GP system for discovering classification rules. Section 4 examines the data used to assess the Egyptian stock market, represents the methodology and the classification results. Section 5 concludes the paper.

II. GENETIC PROGRAMMING

GP [11][12], belong to a class of optimization techniques broadly called evolutionary algorithms. GP is modification of genetic algorithms with one major difference. The population consists of individuals represented by specific data structure – trees. Inner nodes of the trees can represent functions (e.g. arithmetic operators, conditional operators or problem specific functions) and leaf nodes would be terminals - external inputs, constants, and zero argument functions. Both the function set and the terminal set must contain symbols appropriate for the target problem. Each individual of the population is evaluated with respect to its ability to solving the target problem can grow in size and shape in a very dynamical way. This evaluation is performed by a fitness function, which is problem dependent function and the population evolved through the actions of genetic operators such as reproduction,

0-7803-8294-3/04/$20.00 ©2004 IEEE

mutation and crossover. Mutation means the evolution of a completely new structure determined at random at selected node (i.e., subtree in the chosen node is deleted and new one is randomly created, by "grown" in the chosen mutation point). Crossover involves the branches from two parent structures being swapped as shown in Figure 1.

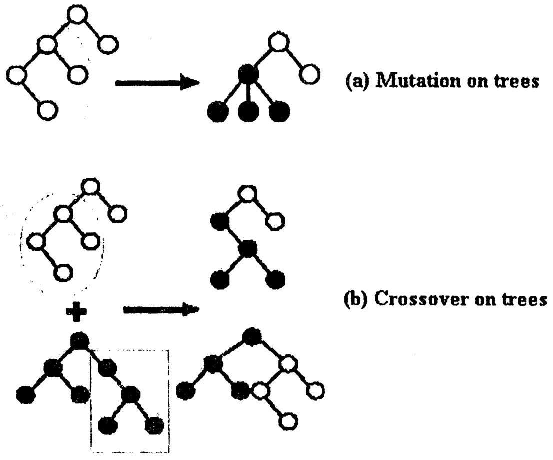

Figure 1. Genetic programming operators (a) mutation (b) crossover.

Unlike most data mining algorithms, GP can automatically discover any logical, mathematical combination of different input attributes. To apply GP to data mining classification problem one must define a terminal set, containing the attributes and their corresponding domain values, and a function set, containing the operators to be applied to the terminal set elements.

III. Genetic Programming in Classification

In this section, we briefly review related approaches using GP for classification in the context of data mining. In this context, the discovered knowledge is often expressed in the form of *IF-THEN* classification rule. GP has bee used to evolve decision trees for classification [11][12] but its role as an inductive leaner remain elusive. Applying GP for classification task is relatively straightforward as long as all the attributes are continuous numeric data and these methods have shown to be capable of producing accurate classifiers. Since the continuous attribute satisfy the *closure property* of GP. Closure of the search space assumes that the return value of any subtree is a valid argument for any function, which means that only admissible offspring is produced. In certain cases this could be a problem, for example if we combine Boolean and numerical functions. In order to maintain this property with nominal or logical attributes so that the classification program still has the ability to performing accurate classification, several solutions have been proposed to cope with the closure property of GP. The *first* approach is based on the use of constrained-syntax GP. The key idea is that, for each function available in the function set, the user specifies the type of its arguments and the type of its result [17][19][15]. In this case, GP can naturally discover rules expressed in a kind of *first-order logic*. The *second* approach based on modifying the data being mined - rather than modifying the standard GP algorithm. An example of this approach consists of booleanizing all the attributes being mined by converting its possible values into binary value (zero or one) and then using logical operators (AND, OR, etc.) in the function set. Hence, the output of any node in the tree will be a Boolean value, which can be used as input to any logical operator in the corresponding parent node. In this case, the resulted classification rule is expressesed in *proportional logic*. Systems based on this approach are described in References [10][5][6][7][18] and our system adopted this approach. The *third* approach is convert the nominal attributes and logical attributes into continuous data and then using mathematical operators (ADD, SUB, etc.) in the function set. The conversion based on labeling each distinct nominal or logical value for a given attribute with an integer value starting at zero reflects its values [18].

IV. Discovering the Classification Rules for Egyptian Stock Market

A. Task and Data

The Egyptian Stock Exchange (ESE) comprises of two exchanges, respectively the Alexandria stock exchange and Cairo. The two exchanges were integrated and govern by the same board of directors (i.e., Capital market authority) as an independent regulatory agency. The two exchanges share the same trade, clearing and settlement systems, so that market participates have access to stock listed in both exchanges. The behavior of the ESE stock return is evaluated using the many daily aggregated indices. These are CASE30 daily index and the Hermes Financial Index (HFI), which are the most widely known and acknowledged performance indicator that measures the return on investment from the change in value of the stocks and other companies also calculate their own indices [14][2]. One of the key banks that contribute to Egyptian market is the Commercial International Bank (CIB); the data for CIB stock price is used for the period between January 2001 and September 2003 for discovering classification rules that predict the stock price of CIB given the prices of the previous closed prices. This gave us 663 data point; each data point comprises the stock prices of the previous five days stock price and the classification of the future stock price. Since our technique learns concept description, we define the stock price as two disjoint concepts: *stock-will-increase* and *stock-will-decrease*. We somewhat arbitrarily took the first 440 data points for training, the next 223 points for validation and test.

B. Classification using Genetic Programming

The system follows the standard GP framework, where the GP expresses the classification rule in proportional logic, which allows one to get human-readable forecasting rule that is sufficient to predict whether a given instance belongs to that class or not as a two-class problem. GP in general has a five components, i.e., the function set, terminal set, fitness function, control parameters and stop condition, must be determined to solve a problem. For a given classification problem, the genetic programming algorithm is executed on population of trees that represent Boolean functions. The function set consists of Boolean functions AND, OR and NOT. The terminal set consists of all input prices of the last week. For the fineness function, there are several formulas have been used to measure the quality of the classification rule that integrating *completeness* (i.e., $\mathrm{com} = \mathrm{p/P}$) and *consistency* (i.e., $\mathrm{con} = \mathrm{p/p+n}$), where p and n are the number of positive and negative examples covered by the rule and P is the total number of positive examples in the training set [9]. The fitness function calculated using a more quality metric is consistency gain, which defined as:

$$fitness = \left(\frac{p+n}{P+N}\right) * exp\left(\frac{p+n}{P+N} - 1\right) \qquad \ldots(1)$$

The consistency gain takes into account the distribution of positive and negative examples in the training set. Where N is the total number of negative examples in the training set. This fitness function has the advantages of being simple and returning a normalized value in the range [0..1].

4.3 Experiments results

The GP is tested for discovering a classification rule for predicting the change in the CIB price. The algorithm was trained and tested with the following parameters we use are the population size is 2000; the number of generations is 51 (including the first random population). The initial population is generated using the *ramped-half-and-half* method. The maximum depth of the new individual is 6. The maximum depth of individual after crossover is 17. Fitness proportionate selection is used. 90% of selected individuals undergo crossover, with 70% internal and 20% external crossover points being selected. The remaining 10% of selected individuals undergoes reproduction. The productivity of discovered rules evaluated using the *accuracy rate* (i.e. $acc = (p+n)/(P+N)$). In order to compare the performance of GP with traditional classification methods, we also conducted experiments using C4.5 [16] results using WEAK which a library of Machine Learning Algorithms in Java. The parameters for those classifiers were chosen to be the default one used by WEKA. The mean accuracy of results of training and test data is presented in table 1.

Table 1. The Classification Accuracy Comparison.

	GP algorithm		C4.5 algorithm	
	Training	Test	Training	Test
Accuracy	65.0%	59.6%	57.9%	52.4%

The C4.5 are able to get a prediction error about 42.1% and 47.6% (on 66%-34% training and test set data on *stock-will-increase/stock-will-decrease*). Using GP can effectively create comprehensive tree with greater predictive power (i.e., GP achieved an accuracy rate of 65%.), by applying the function set to the original attributes this reduce the error to 35% and 40.4% for training and test data sets. However, the basic drawback of GP, compared with C4.5 is the speed.

V. Conclusion and Discussion

This paper explores the synergy of GP in comparison with decision tree classifications such as C4.5 [16] for discovering a classification rule from the data of the Commercial International Bank (CIB) as a key bank at the Egyptian stock market. The C4.5 are able to get a prediction error about 47.6% on data of test set however, using GP can effectively create comprehensive tree with greater predictive power, by applying the function set to the original attributes this reduce the error to 40.4%.

Acknowledgments

We are indebted to Mr. Tarek El-Showarby from EFG-Hermes for fruitful discussion and for providing us with the data of Commercial International bank.

VI. References

[1] A., Freitas, "*A survey of evolutionary algorithms for data mining and knowledge discovery*", In: Ghosh A. and Tsutsui S. (Eds.) Advances in Evolutionary Computation. Springer-Verlag, 2002.

[2] B., Azab, *The Performance of the Egyptian Stock Market*, MSc. Thesis, Business School, University of Birmingham, 2002.

[3] B., Masand, and G., Piatetsky-Shapiro, "Discovering Time Oriented Abstractions in Historical Data to Optimize Decision Tree Classification", In Angeline P., and Kinnear K.. (eds.) Advances in Genetic Programming II, MIT, 1996.

[4] B., Mulloy, R., Riolo, and R., Savit, "Dynamics of genetic Programming and Chaotic Time Series Prediction", In Proceedings of the Genetic Programming First Annual Conference, 1996.

[5] C., Bojarczuk, H., Lopes, and A., Freitas, "Data Mining with Constrained-Syntax Genetic Programming: Application in Medical Data Set", In Proceedings of Intelligent Data Analysis in Medicine and Pharmacology, a Workshop at Medinfo-2001. London, UK, 2001.

[6] C., Bojarczuk, H., Lopes, and A., Freitas, "Discovering comprehensible classification rules using genetic programming: a case study in a medical domain", *Proc. Genetic and Evolutionary Computation Conference (GECCO-99)*, 953-958, 1999.

[7] C., Bojarczuk, H., Lopes, and A., Freitas, "Genetic programming for knowledge discovery in chest pain diagnosis" IEEE Engineering in Medicine and Biology magazine - special issue on data mining and knowledge discovery, 19(4), 38-44, 2000.

[8] C., Zhou, W., Xiao T., Tirpak P., and Nelson, "Discovery of Classification Rules by Using Gene Expression Programming",

In Proceeding of International Conference In Artificial Intelligence.

[9] I., Bruha, "Quality of Decision Rules: Definitions and Classifications Schemes for Multiple Rules", In Nakhaeizadeh, G.; and Taylor, C., (ed.) *Machine Learning and Statistics, The interface*, pp. 107-131, Johan Wiley & Sons, Inc., 1997.

[10] J., Eggermont, A., Eiben, and J., van Hemert, "A Comparison of Genetic Programming Variants for Data Classification", *Proc. Intelligent Data Analysis (IDA-99)*, 1999.

[11] J., Koza, *Genetic Programming II: Automatic Discovery of Reusable Programs*. MIT Press, 1994.

[12] J., Koza, *Genetic Programming: on the Programming of Computers by Means of Natural Selection*. Cambridge, MIT Press, 1992.

[13] L., Breiman, J., Friedman R., Olsen C., and Stone "Classification and Regression Trees", Wadsworth International Group, 1984.

[14] M., Mecagni, and M., Shawkey, "Efficiency and Risk-Return Analysis for the Egyptian Stock Exchange", Working Paper 37, the Egyptian Center for Economic Studies, 1999.

[15] P., Ngan, K., Leung, M., Wong, J., and Cheng, J., "Using Grammar Based Genetic Programming for Data Mining of Medical Knowledge", In Koza, J.; Banzhaf, W.; Chellapilla, K.; Deb, K.; Dorigo, M.; Fogel, D.; Garzon, M.; Goldberg, D.; Iba, H.; Riolo, R., (Eds.) *Genetic Programming 1998: Proc. 3rd Annual Conf.*, 146-151. Morgan Kaufmann, 1998.

[16] R., Quinlan, "C4.5: Programs for machine learning", Morgan Kaufmann, 1993.

[17] S., Bhattacharyya, O., Pictet, G., Zumbach, "Representational semantics for genetic programming based learning in high-frequency financial data", *Genetic Programming 1998: Proc. 3rd Annual Conf.*, 11-16. Morgan Kaufmann, 1998.

[18] T., Loveard, and V., Ciesielski, "Evolving Nominal Attributes In Classification Using Genetic Programming", In Proceedings of the 4th Asia-Pacific conference on Simulated Evolution and Learning, 2002.

[19] Y., Hu, "A Genetic Programming Approach to Constructive Induction", In Koza, J.; Banzhaf, W.; Chellapilla, K.; Deb, K.; Dorigo, M.; Fogel, D.; Garzon, M.; Goldberg, D.; Iba, H.; Riolo, R., (Eds.) *Genetic Programming 1998: Proc. 3rd Annual Conf.*, 146-151. Morgan Kaufmann, 1998.