National Research University Higher School of Economics
Faculty of Computer Science
School of Applied Mathematics and Informatics

Using Machine Learning Techniques for Trading Agent Modeling

Student: Dushatskiy Arkadiy
Group: BPMI 123

Language Consultant: Associate professor Tarusina S. A.

Word count: 2496

Moscow
2016

This project is intended to investigate the applications of machine learning in field of automated stock trading. Considered class of machine learning algorithms is genetic algorithms, they are examined in general and the focus is made on evolutionary ways of constructing decision trees and ensembles of decision trees. These machine learning algorithms are used as classifiers to predict the most profitable actions in different time prediction horizons, it is assumed that actions are executed by trading agent without human involvement. The practical part of project consists of implementation of such algorithms in Python programming language with Cython extension and their performance evaluation on real market data through trading simulations. The project is divided into two main parts: first part is devoted to constructing single decision trees and at the second part they will be united in ensembles. It is expected, that results gained at the first stage will be satisfactory, compared to benchmark 'buy-and-hold' trading strategy and the second part of work will be able to significantly improve them.

Contents

Introduction

Background. The study area of this research is application of machine learning algorithms to automated trading.

Machine learning has become recently one of the most cutting-edge field of computer science, now it is used in wide range of problems: from contextual advertising and fraud detection in on-line payments to various medical applications. Many methods of machine learning are well-studied nowadays and it is important enough that there already exist algorithms which are multipurpose, have good scalability and can show a good performance out of the box in different types of problems.

Genetic algorithms are a general class of machine learning algorithms, they seem to be an interesting area for research, because the core idea of this type of algorithms is borrowed from evolution process which takes place in nature. In complex tasks it is often impossible to find an optimal solution straightly, but genetic algorithms can manage to find a solution which is quite optimal, what is achieved by gradual evolution of initial feasible solution. It should be noticed that genetic algorithm is a very general framework that gives an opportunity to apply it to a wide range of tasks. The well-known areas, in which genetic algorithms managed to significantly improve the state of the art results are the following: robotics, complicated optimization tasks such as designing telecommunication networks, artificial intelligent agents for computer games, computer-aided molecular design. The current project is intended to apply genetic algorithms to area of automated stock trading.

Automated stock trading is a type of trading which prevalence is growing now, because in rapidly changing environment it is very difficult and often impossible to make efficient trading strategies manually. Furthermore, nowadays there exist a type of trading, called High Frequency Trading (HFT), which peculiarity is conducting numerous deals in a very short time periods that cause the fact that

even milliseconds may matter. Naturally, only automated trading systems (also called trading robots) can take part in such kind of trading. Moreover, it is rather difficult to create a successful trading strategy by programming it directly because of complexity and latency of market behavior patterns. Thus, the combination of algorithmic trading and machine learning, usage of machine learning algorithms for creating trading strategies is now a subject of growing concern of both specialists in machine learning and quantitative researchers.

Problem statement. The current investigation will be focused on algorithms which predict the most valuable action for automated trading system at the current moment. The objective of the research can be formulated as follows: investigate the techniques of combining genetic algorithms with decision trees in application to automated trading, model trading agents based on different variations of such algorithms and conduct the experiments which consist of applying algorithms to real market data and evaluating their performance by executing trading simulations.

Decision trees. Definitions. Decision trees is a general class of machine learning algorithms. In current project decision trees will be used as classifiers: the target variable is an action, which should be executed by trading robot at the current moment.

Generally, decision tree is a type of tree data structure which is common in computer science. The structure of decision tree can be described by following:

1. In general case, the tree may be not binary, but for the most practical purposes binary decision trees are considered.
2. The vertex consists of three components: the variable, the inequality sign and the threshold value.
3. The predictable variables are presented in leaves of tree.

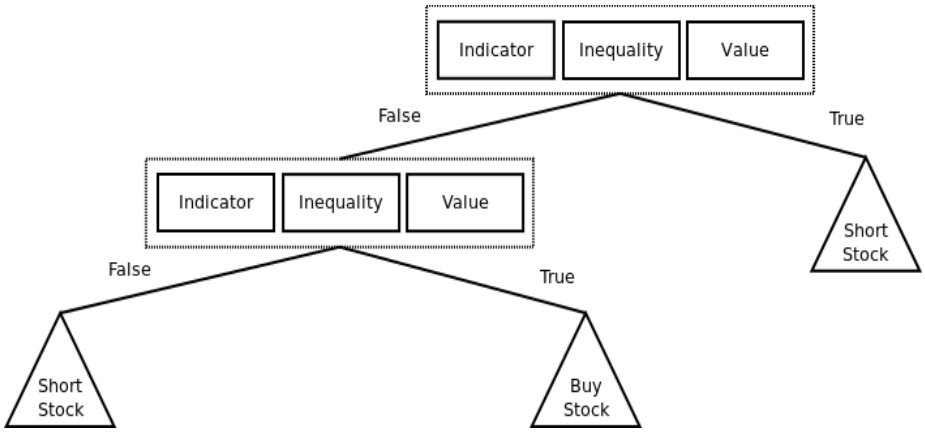
4. The process of prediction is executed as follows.

   4.1. Tree traversal is started in root.

   4.2. While traversing the tree, the movement down to next node from the current one is made with respect to variable value: if the inequality (which can be written formally as *value<sign>threshold*) is true then the next node of traversal is the right child node of the current one, otherwise it is the left child node.

   4.3. If the leave is reached then its value is declared as prediction.

Example 1. This example illustrates a simple binary decision tree in context of predicting action for trading according to Tiwari (2012). 'Buy stock' and 'short stock' are predictable actions, by indicator a technical market indicator is meant.



Genetic algorithms. Definitions. The most general genetic algorithms framework is suitable for all types of tasks in which some metrics should be optimized. There are several basic terms related to genetic algorithms. First of them is *fitness function*: it is the considered metrics, achieving the maximum value of fitness function is the goal of algorithm. *Individual* is an object which performance is evaluated by fitness function. The third main concept in evolutionary algorithms is *population*: it is a set of individuals which is

gradually evolving. *Threshold of evolution* is the ratio of population which remains after selection of individuals with the highest values of fitness function**.** There may be differences in implementations of genetic algorithms due to specific properties of application areas, but there are main parts, which are included in quite all genetic algorithms:

1. Creating initial random population.
2. Iterative part:
   2.1. Selecting individuals based on fitness. It is made in a such way, that only individuals with the highest values of fitness function remain in population.
   2.2. Applying operators (crossover and mutation) to create new individuals.
   2.3. Computation of fitness values of new individuals.
   2.4. Checking if stopping condition is reached.

The meaning of crossover operator is creating a better individual by taking some parts from two good enough individuals. The role of mutation process is further improvement of individuals performance by slightly changing the current best ones. Due to specialty of trees data structure, these basic genetic operators are defined for trees in a peculiar way.

1. *Crossover*

   Crossover operator is performed with respect to tree structure, it is meant that new tree is created by taking a tree and replacing a subtree of it by corresponding subtree from another tree.

2. *Mutation*

   In mutation operator the tree structure is also considered and mutation is processed in the following way: a random vertex is selected and a subtree with root in this vertex is substituted by randomly generated new subtree.

Literature Review

There are numerous research papers dedicated to decision trees and genetic algorithms, but the present review is focused on investigations which combine the both approaches.

In an investigation conducted by Barros, Basgalupp, Carvalho (2012) genetic algorithms are used for constructing decision trees. In the work the general approach of constructing decision trees in an evolutionary way is described. Wide range of different types of decision trees, including decision trees for regression and classification are considered. The work is focused on comparing performance demonstrated by decision trees built with genetic algorithms with trees constructed in traditional way using greedy approach. In the conclusion of research it is stated that in several application domains genetic algorithms are valuable for creating decision trees.

The efficiency of genetic method in task of finding vector of weights to combine learning agents in an ensemble is shown in research work by Sylvester and N. V. Chawla (2005). This investigation is focused on generalizing the procedure of using genetic algorithm for building ensembles consisted of more trivial classifiers. It should be noticed, that the way of constructing ensemble of decision trees using genetic algorithm is a particular case of general method described in the paper.

In a research conducted by Potvin, Soriano and Vallée (2004) the very general types of decision trees are used for generating complex trading rules. It is propounded to apply real-valued functions (exponent, logarithm, trigonometric functions etc.) to object values in nodes of decision tree while traversing it. Though this approach allows to generate more sophisticated rules than traditional decision trees, the multiplicity of functions enlarges the search space for genetic algorithm dramatically, what makes a probability of finding a good trading rule significantly lower. It should be noticed, that usage of this approach

with restriction of the number of different functions in nodes may be reasonable in some cases.

The practical experiments of applying genetic algorithm in combination with decision trees to trading are conducted in work by Myszkowski and Rachwalski (2009). Experiments consist of using decision trees built in evolutionary way in the Warsaw Stock Exchange. In this research results of experiments are provided, which demonstrate, that evolutionary algorithm is an effective technique for generating satisfactory predictions for stock market data. In the investigation it is highlighted that with such method it is possible to generate a more profitable trade strategy than traditional 'buy- and-hold' strategy, which is considered as a benchmark in many papers related to subject.

All the above review can indicate that with the approach of using decision trees combined with genetic algorithms in field of automated trading advance can be achieved. Nevertheless, this area is not entirely explored and there exist opportunities for further investigations in both theoretical and practical parts of considered issue.

Methodology

The project consists of two main stages. On the first stage it is planned to build single decision trees using evolutionary approach. The features in nodes of trees are preliminarily calculated market technical indicators such as Moving Average (MA), Relative Strength Index (RSI) and etc. for different time periods. The method of evaluating the strategy performance, which in terms of genetic algorithms, is called fitness function, is trading simulation: some period of time in the past is fixed and the simulation of trading is executed, the result of it is the gained profit (or loss) for this period of time. It is supposed that in the beginning of simulation as well as at the ending of simulation, the robot has no stocks – in terms of trading, its position is zero. Such assumptions are made to provide a

reliable calculation of fitness function and reduce a probability of situation, when success in simulation was achieved by chance.

On the second stage of project it is proposed to build the ensembles of decision trees. Genetic algorithm will be used to find optimal vector of real numbers which will represent the weights of trees in an ensemble. To speak more specifically, in considered method, ensemble of N decision trees is vector $W=(w_{1,\dots},w_N), w_1 \in \mathbb{N}$ and the predicted action by ensemble is calculated as majority of weighted predicted actions by single trees. Such approach of using ensembles of decision trees gives a chance to improve results from the first stage.

The possible field for experimental work is very wide because of variety of properties which can be modified during experiments, parameters of both algorithms and data extraction can be adjusted. To talk in more detail, data specifications which can vary in experiments are the following ones: periodicity of market situation observations, considered time period, type and name of financial instrument. Algorithms parameters which may have significant impact on performance include decision trees depth, mutation probability, population size, threshold of evolution. It should be mentioned that detailed market data is owned by stock exchanges and is not public, but in experiments in current project free-access market data will be used. The market features presented in these datasets are limited but they provide enough information for the goals of research.

Results Anticipated

It is expected that the approach of constructing decision trees in evolutionary way can achieve satisfactory results in automated trading tasks. It should be highlighted, that the evaluation of algorithms will be processed on the real market data. The results achieved by ensembles of decision trees are anticipated

to be significantly better than results of the single classifiers. Performance of algorithms will be compared to common benchmarks in algorithmic trading, such as 'buy-and-hold' strategy. Experiments will be conducted for wide range of stocks and markets, it is possible that results on different markets and different stocks may vary significantly, comparison of them may also provide some nontrivial insights of market data. Finally, the structure of decision trees will be examined in order to reveal the most informative technical indicators.

Conclusion

The issue of using machine learning techniques in automated stock trading is now an active area of research. In the framework of the project it is proposed to use decision trees and ensembles of decision trees constructed by evolutionary techniques as the methods of predicting the most valuable action on stock market in real time. It is anticipated that the considered approach can achieve satisfactory results in automated trading tasks, evaluation of performance will be processed on the real market data. If the hypothesis, that ensembles of decision trees constructed with genetic algorithm, can achieve significantly better results than single trees, will be supported by experimental results, then the extension of this technique to other machine learning algorithms may be an issue of future investigations.

References

[1] Barros, R.C.; Basgalupp; M.P., de Carvalho, A.C.P.L.F; Freitas, A.A. (2012). A Survey of Evolutionary Algorithms for Decision-Tree Induction. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol.42, no.3, pp.291-312.

[2] Myszkowski, P.B., Rachwalski L. (2009). Trading rule discovery on

Warsaw Stock Exchange using coevolutionary algorithms. *Computer Science and Information Technology, IMCSIT '09,* pp.81-88.

[3] Potvin, J.-Y.; Soriano, P., ; Vallée M. (2004). Generating trading rules on the stock markets with genetic programming. *Computers and Operations Research*, vol.31, no.7, pp. 1033-1047.

[4] Sylvester, J.; Chawla, N. V. . (2005). Evolutionary ensembles: Combining learning agents using genetic algorithms. *AAAI Workshop on Multiagent Learning*, pp. 46–51.

[5] Tiwari, S.; Pandit, R.; Richhariya, V. (2012). Predicting future trends in stock market by decision tree rough-set based hybrid system with HHMM. *International Journal of Electronics and Computer Science Engineering*, vol. 1, no.7 , pp. 1578 – 1587.