An approach using Artificial Neural Network and Genetic Algorithm for Day Trade Portfolio Selection

Paula Campigotto Department of Computer Science Universidade do Estado de Santa Catarina Joinville, Brazil paula.campigotto@edu.udesc.br Omir Correia Alves Junior Department of Computer Science Universidade do Estado de Santa Catarina Joinville, Brazil omir.alves@udesc.br

Abstract—In the financial market there are several types of investors, from the most conservative to the most daring, who are subject to greater risks in the expectation of greater returns on their investments. However, the concept of risk, in investment portfolios, makes it possible to measure it in different ways. This paper aims to present a method created to select portfolios for Day Trade financial investments using different metric risks, such as CVaR, EWMA and GARCH, and the ensemble of Genetic Algorithm NSGA-II and LSTM Artificial Neural Network, comparing it's selected portfolios' performance with another method which uses only NSGA-II and Buy and Hold financial strategy. The results show that the proposed method, with LSTM ANN achieved better returns in the year of 2019.

Index Terms—portfolio selection, genetic algorithm, artificial neural network

I. INTRODUCTION

The selection of portfolios is a strategy used to invest in the financial market and consists of selecting assets based on several factors, among them the history of quotations of such assets. The Modern Portfolio Theory, proposed by [1], aims to maximize the expected return of a portfolio while minimizing its risk. Such values are measured using the average rate of return and covariance, respectiveley, which are obtained by the historical quotations of the assets present in the portfolio.

In the financial context, the asset expected return is defined as the gain, or loss, of an investment over a given period of time. Beyond the covariance, used by [1], the risk can be measured by several strategies, such as standard deviation, semi-variance, Conditional Value at Risk (CVaR) and many others, which have in common the characteristic of considering greater risks for less consistent portfolios [2]. Thus, one of the objectives of this work is to compare the performance of portfolios using 5 different risk measures, with the same sample of quotes, in order to identify correlations between the risk metrics and the types of investors.

In this paper, only stock-type assets will be used, which represent portions of the Net Equity of a company. In other words, they are equity securities that may fluctuate in value as they are conditioned to the performance of that company. In addition, to perform the selection and optimization of the portfolio, the meta-heuristic Non-Dominated Sorting Genetic Algorithm 2 (NSGA-II) will be used, which proved to be efficient in papers related to the optimization of investment portfolios, such as in [3], [4], [5] and [6].

Furthermore, this paper will compare the results about the accumulated return of the portfolios selected using only NSGA-II method and the portfolios selected using the Day Trade strategy with NSGA-II and Long Short Term Memory (LSTM) Artificial Neural Network (ANN).

This paper is structured as follows: in the section II, the research problem is presented. The risk models used in the simulations are conceptualized in the Section III. The algorithms used for the experiments are explained in Section IV. Section V explains which experiments were carried out, and their results and analysis are exposed in Section VI. Finally, in the Section VII the final considerations of the work are presented, as well as suggestions for future works researches.

II. PORTFOLIO SELECTION PROBLEM

The Modern Portfolio Theory proposed by [1] analyzes a set of assets in order to compose a great investment portfolio. This theory aims to optimize a multi-objective function, based on the expected return and risk of the portfolio.

The return and risk of a portfolio can be calculated using the quotations of the portfolio's assets. For the return value of an asset, its daily rates of return are averaged, while for the calculation of risk different metrics such as variance, standard deviation, volatility and others can be used. The calculation of the risk and return of a portfolio is given by the sum of these same measures of each asset multiplied by the proportion of the asset in the portfolio, according to Equation 1, where R refers to risk or return on portfolio, k is its cardinality, or quantity of assets, r_i is the risk or return of the asset i(depending on the R being calculated) and w_i is the proportion of the asset i in the portfolio.

$$R = \sum_{i=1}^{k} r_i w_i \tag{1}$$

In order to minimize risk, it is necessary to diversify the portfolio, since risk and diversification are inversely proportional variables. Thus, it is necessary to analyze the ideal size for the cardinality of the portfolio, based on criteria that determine a limited and balanced value of assets. [7].

One of the restrictions, addressed by the portfolio selection problem, is the cardinality constraint, which limits the value of assets in a portfolio, classifying the problem as *NP-Hard*. The portfolio problem is in the class of combinatorial optimization problems and exact methods cannot find optimal solutions in polynomial time, according to [8]. All of these factors motivate the use of a multi-objective meta-heuristic algorithm to optimize the problem and obtain solutions close to the optimal solution.

III. RISK METRICS

In the context of investment portfolios and in order to achieve the expected return, it is necessary to define the time periods in which asset purchase and sale movements should occur, but, due to the uncertainties inherent in the process, it is not possible to guarantee that the expected return will be reached. Risk can be defined as the ability to measure such uncertainty for a decision, based on the probability of certain values or scenarios occurring.

The risk is usually associated with the possibility of loss, but likewise, there is a possibility of gain, which encourages investors to increase their risk exposures. Most of the time, the risk is measured based on the standard deviation, thus, the further the expected return value deviates from the interval defined by the average and standard deviation of the previously observed values, the greater the risk of this investment. Likewise, if the expected return has a value that does not differ from the historical behavior of the series of values, the risk will be low [2]. There are different ways to measure the risk, in addition to the standard deviation, which will be presented next.

A. GARCH (1,1)

The Generalized Autoregressive Conditional Heteroskedastic (GARCH), proposed by [9], is a statistical technique that aims to measure the volatility of the returns on financial assets of a given time series, being a generalization of the ARCH technique, proposed by [10]. Volatility in the financial market measures the dispersion of the returns on an asset or index. Thus, the greater the volatility of an investment portfolio, the greater the risk of gain or loss. The most common GARCH model is GARCH (1,1) and is expressed according to Equation 2.

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{2}$$

The value of σ_t^2 represents the variance at time t, while r_t^2 refers to the rate of return at time t. In addition, $\omega = \gamma V_L$, which represents the long-term weighted variance. The following restrictions must be observed: $\omega > 0, \alpha \ge 0, \beta \ge 0$ and $\gamma + \alpha + \beta = 1$.

B. EWMA

The *Exponentially Weighted Moving Average* (EWMA) was exposed for the first time by [11] and calculates the variance of a given time sequence, but giving more weight to the most recent observations, as [12]. In this way, the risk, based on the EWMA, can be calculated according to Equation 3.

$$\sigma_t^2 = (1 - \lambda)r_{t-1}^2 + \lambda \sigma_{t-1}^2$$
(3)

Similarly to the values of GARCH, the value of σ_t^2 represents the variance at the time t and r_{t-1}^2 refers to the rate of return at the time t-1. The value of λ belongs to the interval [0.1] and models the short-term memory of the stock, being called the decay factor, as it is responsible for carrying out the weighting that gives greater weight to the most recent values.

C. LPM

The way to measure risk based on *Lower Partial Moments* (LPM) uses semi-variance as a criterion to determine risk, [13]. The LPM risk measure can be calculated from a target return ξ which can be represented by the average of the asset or by a market *benchmarking*. The formulation of the LPM is represented in the equation 4:

$$LPM_{s,\xi}(r) = E\{[\min(r-\xi, 0]^s]^{\frac{1}{s}},$$
(4)

where r is the return on the asset, ξ is the minimum acceptable or expected return, and s is the order of the model (degree of risk aversion to the investor).

D. VaR

The Value at Risk (VaR) is a statistical method that corresponds to the maximum potential loss, at a specific confidence level β , in a given time and can be translated as the amount in which the losses are not exceed $(1 - \beta)$ % of scenarios, [14]. The VaR formula was defined by [15] according to Equation 5:

$$VaR = \alpha_{\beta}(x) = \min\{\alpha \in \mathbb{R} : \Psi(x, \alpha) \ge \beta\}$$
(5)

$$\Psi(x,\alpha) = \int_{f(x,y) \le \alpha} p(y) dy \tag{6}$$

$$f(x,y) = -[x_1y_1 + \dots + x_ny_n] = x^T y,$$
(7)

where x is the portfolio in question and $\alpha_{\beta}(x)$ is the lowest value for α when $\Psi(x, \alpha) = \beta$. The function $\Psi(x, \alpha)$ can be viewed in Equation 6, where x represents the vector of asset weights in a portfolio, and y is an uncertain vector

A. NSGA-II

The Nondominated Sorting Genetic Algorithm II (NSGA-II), proposed by [16], is a multi-objective meta-heuristic of the genetic algorithms class. In short, it uses concepts of the Evolution of the Species from *Charles Darwin* and Genetics from *Gregor Mendel* to optimize a problem. NSGA-II is an evolution of its first version - NSGA, proposed by [17]. The improvement of NSGA-II over NSGA is on its complexity, which has been reduced from $O(N^3m)$ to $O(N^2m)$, where N is the number of individuals and m is the number of objective functions. For the problem in question, the individuals are the selected portfolios and the objective functions are the minimization of the risk and the maximization of the expected return.

NSGA-II operation begins with a random creation of the first population P_t , which has a size equal to N. After this creation, a new population, Q_t , also of size N, is generated from the selection, recombination and mutation operations over P_t . The union of these two populations forms the population $R_t = P_t \cup Q_t$, of size 2N, which will be subjected to ordering according to non-domination criteria.

The concept of non-domination is related to Pareto Boundary and to the objective functions of the problem, thus, one solution dominates another when its return to the fitness function, which measures the efficiency of the solution, is better than the return of another.

The ordering segregates individuals from the population, associating solutions to different classes $-F_1, F_2, F_3, ..., F_n$ – according to their elitism (dominance over other solutions). Thus, individuals belonging to the class F_1 are the best solutions, which will be present, in full, in the population P_{t+1} and the higher the index of the class, the worse the solutions contained therein. Therefore, for the formation of the population P_{t+1} , of size N, the individuals contained in the first classes are allocated while the others are rejected. Finally, if P_{t+1} meets the stop criterion, the solution will be returned, otherwise P_{t+1} will act as the starting population for the next iteration. [16]

B. Long-Short Term Memory Artificial Neural Network

Recurrent Neural Networks (RNN) are those that feed on their own output data, so that such information persists throughout learning. As a subset of the RNNs, there is the Long-Short Term Memory architecture, which allows the storage of data for a longer period of time within a Neural Network. Such a Neural Network model can remember past information in the following way: the ANN cells are responsibles for storing past data, which are filtered through the forgetting and input modulation gates. The oblivion gate decides whether data will be reused or discarded. The input modulation gate makes adjustments to the stored input data. [18]

C. Portfolio Selection and GA

In order to adapt the portfolio selection problem to the Genetic Algorithm, it is necessary to map some concepts from



Fig. 1. GA portfolio selection model

a context to another. Figure 1 presents an example of the population, chromossomes, genes and alleles used to create the GA adapted algorithm.

Also, some changes and definitions needed to be done to guarantee the validity of the solutions, such as using the arithmetic crossover strategy and balancing the mutation of the portfolio proportions, in order to maintain the sum of 100% to it's total proportion.

V. PROPOSED METHOD AND PARAMETERS

The bio-inspired optimization methods, often metaheuristic, or also the Artificial Neural Network, are classified as computational intelligence algorithms and are approximate, that is, their execution cannot be guaranteed individually. Thus, it is necessary that their performances be evaluated from an average of executions together with the observed standard deviation. In addition, such methods have many parameterizable variables. Table I presents the values used for such parameters in the performed experiments.

TABLE I PARAMETERS USED IN THE EXPERIMENT

Parameter	Value	ANN	NSGA-II
Executions	10		\checkmark
Iterations	50		\checkmark
Population size	50		\checkmark
Mutation (%)	0.1		\checkmark
λ : ewma	0.94		\checkmark
ω : garch	0.001		\checkmark
α : garch	0.75		\checkmark
β : garch	0.1		\checkmark
τ : lpm	0		\checkmark
k: lpm	0.001		\checkmark
Cardinality	9	\checkmark	\checkmark
Optimizer	Adam	\checkmark	
Loss	Mean squared	\checkmark	
	error	•	
Cells	100	\checkmark	

The chosen values for the parameters were based on previous experiments and executions of the algorithm, but they can also be optimized by another system. Our solutions usually hit the convergency with 50 iterations, as shown in Figure 2, that is why this number was pinned. Furthermore, 10 executions gave us a satisfactory standard deviation, as the execution time for this type of problem tends to be longer, we opted for the smallest number of executions.



Fig. 2. Convergency Graph for NSGA-II and risk metrics

The experiments were performed on a Dell Inspiron 14 7000 series machine with the following specifications: Linux operating system (Ubuntu 20.04), 16 GiB of memory, Intel Core i7 processor (7th generation) and 1TB of hard disk.

The experiments were carried out using the Yahoo Finance database for the years between 2015 and 2019, with such data being the daily quotations of all Ibovespa¹ shares.

The investment strategy used to carry out the tests using a *LSTM ANN* model, as initially proposed, will be the *Day Trade* Strategy, which executes daily exchange of portfolios. It will be compared with another approach using only NSGA-II algorithm and Buy and Hold investment strategy, which maintain the same portfolio during a long period of time. For this, in the Day Trade method, an ANN model will be created for each stock.

Figure 3 ilustrates an example of how the proposal takes place. It shows that the Neural Network models will be responsible for, each day, informing the 9 assets that will have the highest return on the following day, based on its prediction made by the history of the last 30 days, and 30 is an execution parameter. The number 9 for the portfolio's cardinality was defined based on the results of previous works, as mentioned by [5] and [19].

After the *ANN* information about the best potential assets for the next day, the *NSGA*-II will select, also based on the history of the previous 30 days, which are the best proportions for the 9 assets selected, considering the average rate of return and also the 5 different risk metrics mentioned before.

After the ANN models perform the predictions for each day n, the real values for the day n-1 and the predicted value for the day n will be compared, thus obtaining the simple rate of return for the day n. Having the expected rate of return for each asset and for each day, a daily descending order of the expected return of each asset will be performed and, thus, the 9 potential best assets will be chosen to compose the portfolio of the day n.





Fig. 3. Proposed method using LSTM ANN aproach

It was created a LSTM ANN model for each asset with a training set of 4 years (2015-2018) and the testing set was the last year (2019). Figure 4 shows the prediction of one of the generated models used to compose each portfolio. It is shown this the prediction and real value are very similar for that VALE3 asset, but on the other hand, for MGLU3, Figure 5, the results were not as similar as VALE3.



Fig. 4. Prediction and real values for VALE3 asset

It is important to mention that this project did not focus on optimizing an ANN model for the assets, but focused on optimizing the selection of the portfolios using NSGA-II and Day Trade Strategy. That is why there is no numbers or comparisons for our ANN models' result, since it was used a model with only two LSTM layers and two Dense layers. The optimization of the model can be used to improve the strategy's success in future works.

Afer the ANN use, which returned all the selected portfolios for each day of 2019, NSGA-II was used in order to return



Fig. 5. Prediction and real values for MGLU3 asset

the distribution of the assets on each of the portfolios, in other words, it returns what percentage each asset would have in the portfolio. Figure 6 shows the example of the selected assets (stocks) from Figure 3 with the proportions, after the execution of NSGA-II, creating the portfolios.

NSGA-II provides the portfolios for each day of 2019 with the best distribution of proportion possible. It was executed 10 times, so it was possible to measure the mean and the standard deviation of the accumulated returns of each selected portfolio.



Fig. 6. Selected Stocks (portfolios) after NSGA-II execution

VI. RESULTS

Figure 7 presents the accumulated return of the NSGA-II selected portfolios, expressed by it's mean and standard deviation, using the Buy and Hold financial strategy. The last position of the lines (december) are in the same order of the legend. As we can see, the results were really good, as expected, since all the selected portfolios had better results than Ibovespa, a brazilian stock exchange index.

Figure 8 illustrates the Pareto Boundary for all risk metrics when running NSGA-II. It is possible to observe that the result of the average accumulated return of the portfolios is consistent with the Pareto frontier, since the CVaR and VaR frontiers are dominant and as the LPM and EWMA dominated, evidencing the concepts of Pareto optimality. It is also possible



Fig. 7. Accumulated Return of the NSGA-II selected portfolios (2019)

to notice that, the greater the risk and return, the smaller the diversity of portfolios present in the solutions for all metrics, except the EWMA, which has the opposite behavior. This fact is verified from the density of portfolios displayed in the graph.



Fig. 8. Pareto Boundary of the NSGA-II selected portfolios (2019)

Figure 9 shows the same results of Figure 7, but using Day Trade financial strategy and also the ANN to selected the assets before NSGA-II distributes their proportions. The mean accumulated returns obtained by this method were much better than the ones obtained only by NSGA-II.

Table II presents the accumulated return at the end of 2019 of the two experiments, which shows that, making the choice of portfolios with the *Day Trade* method and the daily filtering of assets by ANN, the results are considerably better for all risk metrics, especially for VaR, which achieved a cumulative return 23.24% higher than the same metric in NSGA-II-only experiment and also GARCH, which had an increase of 17.82%.

Figure 10 shows the Pareto boundaries observed for Day



Fig. 9. Accumulated Return of the LSTM ANN and NSGA-II selected portfolios (2019)

TABLE II Final mean accumulated return of the two mentioned methods: NSGA-II and ANN + NSGA-II (2019)

	NSGA-II	ANN + NSGA-II
CVaR	43.40%	58.85%
VaR	44.50%	67.74%
EWMA	28.99%	34.45%
GARCH	35.79%	53.61%
LPM	30.27%	37.93%
Ibovespa	25.49%	25.49%

Trade executions, after ANN pre-filtering. As the mentioned strategy is daily, more elements were used to find the total average of the solutions. In comparison with the values observed in the experiment that uses only NSGA-II, it is possible to notice that the CVaR presented the dominant frontier in most of the graph, despite the fact that, in almost the entire period, the accumulated return of VaR was greater. The dominated frontiers, as seen in Figure 8, were also those of GARCH and LPM, which had the lowest accumulated return among the analyzed metrics.

VII. CONCLUSION

The asset selection problem in a cardinality-constrained investment portfolio, as noted, is a combinatorial problem of the NP-Hard class, which makes its optimal solution practically impossible to find. For this reason, portfolios, or solutions, can be found using heuristic methods, which find an approximate solution to the problem. It was possible to observe that the NSGA-II algorithm, by itself, already obtained very good results, but with the use of ANN's asset pre-filtering, the results were even better.

Given the results presented, it can be concluded that the method proposed with the joint use of the NSGA-II and ANN LSTM algorithms, proved that it selects investment portfolios with superior quality, thus providing returns that exceed, for example, the Ibovespa, this index was used as *benchmark* for the analysis of the results obtained.



Fig. 10. Pareto Boundary of the NSGA-II and LSTM ANN selected portfolios (2019)

As future work, we suggest the implementation of a model similar to the one presented in this work, but which filters assets, daily, with different ANNs models, so, for each day, the quotations until the day before are used as a training value to be predicted and that these last values have greater weight than old observations, such as the EWMA concept, so that the ANN can more accurately predict the possible assets with better returns for the day. Furthermore, it is also suggested the use of ANN predictions in the modeling of the *fitness* function, which currently uses the ratio between the return and risk of portfolios. It can also be used more data as input to the ANN, like the own Ibovespa and other economic indexes.

REFERENCES

- H. Markowitz, "Portfolio selection," *The journal of finance*, vol. 7, no. 1, pp. 77–91, 1952.
- [2] A. N. Assaf, *Mercado financeiro*, 12th ed. São Paulo: Editora Atlas S.A., 2014.
- [3] C. M. Quinzani, "Otimização multiobjetivo de portfolios utilizando algoritmos evolutivos," Universidade Estadual de Campinas - UNICAMP, Campinas, Tese de mestrado, 2010.
- [4] M. d. S. d. Oliveira, "Análise comparativa entre medidas de risco na otimização multiobjetivo de carteira de ações com restrição de cardinalidade," Centro Federal de Educação Tecnológica de Minas Gerais -CEFET, Belo Horizonte, Tese de mestrado, Março 2016.
- [5] A. Heiden, F. M. Ramos, and O. C. Alves Junior, "Otimização multiobjetivo de portfólios de investimento utilizando algoritmos evolutivos," *LI Simpósio Brasileiro de Pesquisa Operacional*, 2019.
- [6] M. Kaucic, M. Moradi, and M. Mirzazadeh, "Portfolio optimization by improved nsga-ii and spea 2 based on different risk measures," *Financial Innovation*, vol. 5, no. 1, Jun. 2019. [Online]. Available: https://doi.org/10.1186/s40854-019-0140-6
- [7] C. Blum and A. Roli, "Metaheuristics in combinatorial optimization: Overview and conceptual comparison," ACM computing surveys (CSUR), vol. 35, no. 3, pp. 268–308, 2003.
- [8] G. Hanoka, R. Cardoso, and F. Paiva, "Modelo multiobjetivo para seleção de portfólios com restrição de cardinalidade, custo de transação e valor em risco condicional," *TEMA (São Carlos)*, vol. 17, pp. 353 – 365, 12 2016.
- T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, Apr. 1986.
 [Online]. Available: https://doi.org/10.1016/0304-4076(86)90063-1

- [10] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation," *Econometrica*, vol. 50, no. 4, p. 987, Jul. 1982. [Online]. Available: https://doi.org/10.2307/1912773
- [11] S. W. Roberts, "Control chart tests based on geometric moving averages," *Technometrics*, vol. 1, no. 3, pp. 239–250, 1959.
- [12] J. S. Hunter, "The exponentially weighted moving average," *Journal of Quality Technology*, vol. 18, no. 4, pp. 203–210, Oct. 1986. [Online]. Available: https://doi.org/10.1080/00224065.1986.11979014
- [13] T. Foo and S. Eng, "Asset allocation in a downside risk framework," *Journal of Real Estate Portfolio Management*, vol. 6, no. 3, pp. 213–223, 2000.
- [14] R. Mansini, W. Ogryczak, and M. G. Speranza, *Linear and Mixed Integer Programming for Portfolio Optimization*. Springer International Publishing, 2015. [Online]. Available: https://doi.org/10.1007/978-3-319-18482-1
- [15] R. T. Rockafellar, S. Uryasev *et al.*, "Optimization of conditional valueat-risk," *Journal of risk*, vol. 2, pp. 21–42, 2000.
- [16] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.
- [17] N. Srinivas and K. Deb, "Multiobjective optmization using nondominated sorting in genetic algorithms," *Evolutionary Computation*, 1994.
 [18] D. S. Academy, "Deep learning book,"
- [18] D. S. Academy, "Deep learning book," ihttp://www.deeplearningbook.com.br/¿, 2019, accessed: 2021-02-06.
- [19] P. Campigotto, G. M. Utiama, and O. C. Alves Junior, "Análise da eficiência de diferentes métricas de risco e estratégias de investimento na otimização de portfólios," *LII Simpósio Brasileiro de Pesquisa Operacional*, 2020.