Journal of Autonomous Intelligence (2019) Volume 2 Issue 2.
doi:10.32629/jai.v2i2.44

Original Article

A Method to Identify Anomalies in Stock Market Trading Based on Probabilistic Machine Learning

Anderson Rodrigo Barretto Teodoro, Paulo André Lima de Castro*

Autonomous Computational Systems Lab, Technological Institute of Aeronautics (ITA), São José dos Campos - SP, Brazil; Email: barretto@ita.br, pauloac@ita.br

ABSTRACT

Financial operations involve a significant amount of resources and can directly or indirectly affect the lives of virtually all people. For the efficiency and transparency in this context, it is essential to identify financial crimes and to punish the responsible. However, the large number of operations makes it infeasible for analyzes made exclusively by humans. Thus, the application of automated data analysis techniques is essential. Within this scenario, this work presents a method that identifies anomalies that may be associated with operations in the stock exchange market prohibited by law. Specifically, we seek to find patterns related to insider trading. These types of operations can generate big losses for investors. In this paper, we use the public available information from SEC and CVM, based on real cases on BOVESPA, NYSE and NASDAQ stock exchanges, that it was used as a training base. The method includes the creation of several candidate variables and the identification of which are the most relevant. With this definition, classifiers based on decision trees and Bayesian networks are constructed, evaluated and then selected. The computational cost of performing such tasks can be quite significant, and it grows quickly with the amount of analyzed data. For this reason, the method considers the use of machine learning algorithms distributed in a computational cluster. In order to perform such tasks, we use the WEKA framework with modules that allows the distribution of the processing load in a Hadoop cluster. The use of a computational cluster to execute learning algorithms in a large amount of data has been an active area of research, and this work contributes to the analysis of data in the specific context of financial operations. The obtained results show the feasibility of the approach, although the quality of the results is limited by the exclusive use of publicly available data.

Keywords: *Data Mining; Machine Learning; Anomaly Detection*

ARTICLE INFO
Received: July 8, 2019
Accepted: Aug 5, 2019
Available online: Aug 20, 2019

***CORRESPONDING AUTHOR**
Paulo André Lima de Castro, Autonomous Computational Systems Lab, Technological Institute of Aeronautics (ITA), São José dos Campos - SP, Brazil; pauloac@ita.br;

CITATION
Anderson Rodrigo Barretto Teodoro, Paulo André Lima de Castro. A Method to identify anomalies in stock market trading based on Probabilistic Machine Learning. Journal of Autonomous Intelligence 2019; 2(2): 42-52. doi: 10.32629/jai.v2i2.44

1. Introduction

Financial fraud is a problem that generates big losses for the corporate sector, the government, the financial industry and consumers in general. Anomaly detection is a big challenge in the financial market. As new detection and prevention techniques are achieved, new methods to crack the guards are developed by fraudsters. So, it's a job that requires constant improvement (West and Bhattacharya, 2016).

There are many challenges involving data mining applied to fraud detection in the stock market. These challenges include, for example, the dataset involving the transactions, because it is massive and arranged in different ways. Data collection for this type of problem is also another big challenge because the available labeled data are very rare. The process to label the data is costly and requires expert investigation. Moreover, the number of positive fraud cases is only a small percentage of the total sample (Golmohammadi and Zaiane, 2012). Other challenges for using autonomous intelligence are described in more detail in

(Cotsaftis, 2018). An interesting recent approach was described in (Saldanha and Omar, 2019) where it was proposed to use machine learning to identify disruptive movements in the negotiation from the use of spoofing and layering but using first an unsupervised algorithms to create a learning database and after using this database to detect patterns.

In this work, we used insider trading data from cases in Brazil, obtained from CVM (Securities commission in Brazil) (Prado and Vilela, 2015) and in the United States of America, from SEC (Securities and Exchange Commission) through (SEC, 2016)(Gorman, 2016). The case study considers the financial transactions of the companies involved in anomalies at BOVESPA, NYSE and NASDAQ stock exchange, which is related to insider trading anomalies. We have used some supervised learning techniques trained to detect new anomalous cases. The dataset was built using the cited data and we also create some variable that helped to identify possible relevant events.

The final dataset includes many different variables from many distinct subjects, (109 variables in total). For example, we created variables related to elections (presidential election, midterm elections and so on) and others related to financial aspects currency variation, stock index and others. We present a detailed explanation about those variables in section 3.

Our goal is to build a model that is capable to identify possible anomalies (like insider trading) in the stock market data. This task is performed by using data mining techniques and machine learning. We have used decision trees with C4.5 algorithm and Bayesian networks. Other techniques like neural networks have been tested, but the best results were achieved using decision trees and Bayes Networks in preliminary tests. Furthermore, it is easy to gain insight about the more relevant variable when using Probabilistic machine learning approaches rather traditional neural networks.

Market manipulation with criminal behaviors may affect the lives of virtually all people. However, the large number of operations makes it infeasible for human experts analyze them all. To address this problem, we present a model that it is able to identify automatically anomalies that may be associated with insider trading operations in the stock exchange market, which is prohibited by law.

The main contribution of this work is the creation of a model to detect anomalies by using publicly available information, transforming these no-value data in a signal that points to probable anomalies in the stock market. The remaining of this paper is organized as follows. In section 2, we present the main points of this work background and on the Section 3 the methodology used to build the model. We also tested several variables to verify their relevance for this scenario. These results are presented in Section 4. Finally in section 5, we present some conclusions, discussions and point out some future research alternatives.

2. Background

In this section, we briefly present the problem domain and the machine learning techniques used in this work. We also discuss aspects of using machine learning algorithms in distributed systems, since we have done that in the context of this project.

2.1 Machine learning

In machine learning process we use algorithms to build Bayesian networks (Bouckaert, 2001) and decision trees (Russell and Norvig, 2010). We use C4.5 algorithm (Mitchell, 1997) for the implementation of decision trees. On the Bayesian networks creation, various network configurations are possible. We use search algorithms to find the best model. Each algorithm has a feature, and works in a particular way. In this work, we use some generic search algorithms, such as Hill Climber (Russell and Norvig, 2010). Some specific algorithms for building Bayesian networks have also been used, such as: NaiveBayes (John and Langley, 1995), K2 (Cooper and Herskovits, 1992), Tan (Friedman *et al*., 1997) and TabuSearch (Hertz and de Werra, 1987).

Each algorithm together with Bayesian networks has various setting parameters for their execution (Witten and Frank, 2005). However, it has some common parameters. The main common parameters considered, in most cases, are the estimator (Hall *et al*., 2009),

maximum number of parents (Bouckaert, 2004), the number of folds, minimum number of instances for leaves and confidence factor (Drazin and Montag, 2012).

We use the cross validation technique to perform the training and test on the database, from the machine learning algorithms (Cabena *et al*., 1998). We use the following metrics to check the quality of classification models: the kappa index (Landis and Koch, 1977), accuracy, precision, sensitivity, sensibility (Mitchell, 1997) and the confusion matrix to check the amount of correct and wrong classifications (Russell and Norvig, 2010).

2.2 Software Tools

The main tools used were WEKA (Witten and Frank, 2005). framework and Hadoop (Konstantin *et al*, 2010). We used WEKA to perform data mining, through machine learning and Hadoop to perform distributed processing of the data in cluster. Both were used in the Ubuntu Linux operating system. For more information about, we suggest the reader to address the following papers: (Abualigah, 2019), (Abualigah *et al*., 2018), (Abualigah *et al*., 2017) and (Abualigah and Khader, 2017).

Hadoop is a framework targeted for applications that use clusters to process large volumes of data. The main elements of Hadoop are MapReduce programming models (West and Bhattacharya, 2016). Currently this platform is considered one of the best to carry out the processing of high demand for data and has numerous benefits, as discussed in (Shvachko *et al*., 2010).

WEKA is a free software used for data mining and currently is consolidated as the data mining tool most used in academic environments. One of its great advantages is the amount of available algorithms and the provided resources to work with large volumes of data, like distributed WEKA plugin. This resource allows performing the data mining in a distributed way using a Hadoop cluster. WEKA provides some specific commands to accomplish this task. The main ones are the definition of the maximum size of split, number of chunks, number of nodes on the cluster and the command for changing Hadoop settings (Hall, 2013).

2.3 Problem domain

Stock market is a place that offers the necessary means to carry out purchase and sale of marketable securities (Linton and Mahmoodzadeh, 2018). There are regulatory agencies that are responsible for ensuring the order and smooth progress of the negotiations and to ensure the proper functioning of those involved in the transactions at the stock exchange. Moreover, these agencies have the power to apply penalties to those responsible when they commit some violation and generally supervise, regulate and discipline the securities market. In Brazil, this agency is called CVM (Brasil, 1976) and in EUA is called SEC (SEC, 2016).

There are a variety of anomalies that occur in the context of financial transactions on the stock exchange, but we are focused only on insider trading cases. This crime usually occurs when employees of a company operates on the stock market, based on inside information. In most cases, it occurs with managers or employees of administrative areas that have high professional position within a company (Carlton and Fischel, 1983).

3. Methodology

In this work, we used a cluster with Hadoop, composed by eight computers, to run the jobs in a distributed way on the data processing. Together with Hadoop and running on this distributed environment, the WEKA was used to perform the data mining tasks. In parallel to the use of the cluster, a personal computer was used, running the sequential version of WEKA to compare the results obtained in the Hadoop cluster. **Figure 1** illustrates the architecture of the distributed WEKA with Hadoop.

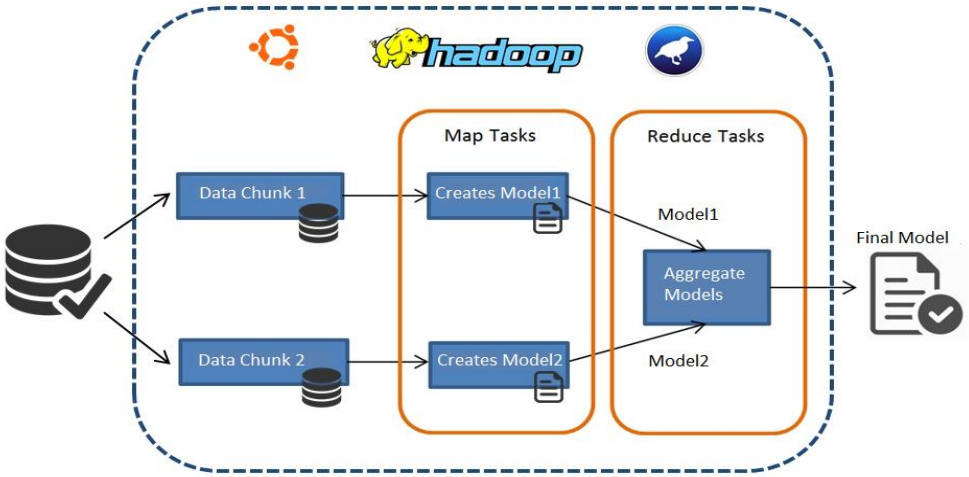

Figure 1. Distributed WEKA Architecture With Hadoop

The database was created using four steps. First, it was the collection of anomalous data labeled by the SEC and CVM. On the second step, we have collected the respective historical financial operations on the stock market for each company involved in anomalous cases. On the third step, we have created candidate variables, which could be related, or not, to anomalous occurrences, such as: elections, holidays, currency quotation on the stock exchange and some variables related to the price and volume variations in a certain period of time. Thus, we have stored all databases used to create the candidate variables within a data lake. Finally, the complete database was created, consisted of financial transactions records of companies operating on the BOVESPA, NYSE and NASDAQ stock market ex- changes. **Table 1** illustrates a description of the variables used in this work.

Variable	**Value**
RelevantInformation	The time since the last publication of the relevant fact.
Var StockIndex X Days Before *ranging from 1 to 20 days	Variables describing the variation of the stock market index X days before
VarVolumeAverageXDays *ranging from 1 to 20 days	Describe the change in volume relative to the moving average of the last X days
VarPriceLastXDays *ranging from 1 to 20 days	Describe price change from previous X days
VarVolumeLastXDays *ranging from 1 to 20 days	Describe the change in volume from the previous X days
Var QuotationOfCurrencies 2 Days Before *ranging from 1 to 20 days	Currency quote in the last X days
DayOfTheWeek	Define the day of the week
NameOfTheMonth	Define the name of the month
NameOfTheQuarter.	Define the name of the quarter
Holiday	Indicates a holiday at the time of the financial operation.
Presidential	Indicates if there was a presidential election
PossessionOutlet	Swearing in of President-elect
OffYear	Year when small elections are held, but no significant elections.
Midterm	These are the midterm elections, which take place midway through the president's four-year term.

Table 1. Description of the Evaluated Variables

In order to set the number of domain elements of each candidate variable and their respective range of values during the discretization process, we have used the clustering algorithm, called canopy (McCallum *et al*., 2000). The main goal of this stage is to find out the number of groups that best describes each numeric variable to facilitate the process of the discretization. We compare the results of a numeric database with the results obtained with a discretized database in order to analyse the performance of the clustering (Abualigah *et al*., 2017); (Abualigah *et al*., 2018).

We calculate the Lift, based on probabilities of Bayes rule (Pearl, 2014), to identify the most important variables in the context of the problem addressed. The first step is to get the apriori values of the anomalies. This mathematical notation used on the first step is presented below on Equation 1.

$$P(Anomaly = 'Y' \mid CandidateVariable) = \left(\frac{P(CandidateVariable \mid Anomaly = 'Y')}{P(CandidateVariable)}\right) \quad (1)$$

As illustrated in Equation 1, the parameter CandidateVariable identifies each variable available on the database for evaluation. After calculating this values using Bayesian rules it was applied the lift calculation, as presented on Equation 2.

$$Lift = \left(\frac{P(Anomaly='Y' \mid CandidateVariable)}{P(Anomaly='Y')}\right) \quad (2)$$

Thus, for each candidate variable in the database, each probability was checked from the values of the Anomaly variable. Through the results obtained, we selected the variables that have the highest value of Lift; in other words, the variables that have the highest dependence on the cases in which there are anomalies. We have used this criterion for the database construction.

Pre-processing of data was one of the most labor-intensive stages, involving cleaning, in- tegration and data transformation. With the database created, we have started the analysis first using decision trees and after with probability calculations using Bayes rule. We use the C4.5 algorithm for decision trees and Bayesian networks for the probability calculation. The next step was checking the behavior of the common parameters of the algorithms and understanding the variation reflection of these parameters on the results. We have created several simulations for this task, in which the values of only one parameter were varied and all other parameter values were fixed. The database was initially set up using 109 variables, considering the same database for training and validation. After that, we have created a new database with only the most relevant variables, identified by the Lift results. Finally, we began the final stage, looking for a good classification model, using cross-validation on a database with only the selected variables.

4. Results

We identified the most relevant variables, for the problem addressed, from the calculation of the lift values. In general, the variables that have the greatest lift, and consequently, that are considered the most important, are those related to the relevant fact and the price variation. We use these results as the basis for creating the database. **Table 2** illustrates the top 10 lift values of the most relevant candidates variables.

Variable	**Value**	**Lift**
RelevantInformation	LastDay	246,767
RelevantInformation	1WeekAgo	159,49
RelevantInformation	2WeekAgo	104,01
RelevantInformation	LastHours	89,82
RelevantInformation	4WeekAgo	41,75
RelevantInformation	3WeekAgo	36,61
RelevantInformation	3MonthsAgo	18,32
VarPriceLast2Days	Low	15,39
VarPriceLast3Days	High	12,83
VarPriceLast5Days	StrongHigh	8,11

Table 2. The top 10 lift values of the most relevant candidates variables.

We use a numeric database and other discretized to compare the performance of the discretization process. We applied the C4.5 algorithm with the same configuration for each database. **Figure 2** illustrates one of the simulations, comparing the performance of the numeric and discrete databases with the variation in the number of folds, using non-binary trees with two chunks, confidence factor equal to 0.9 and a maximum split size equal to 5000000.

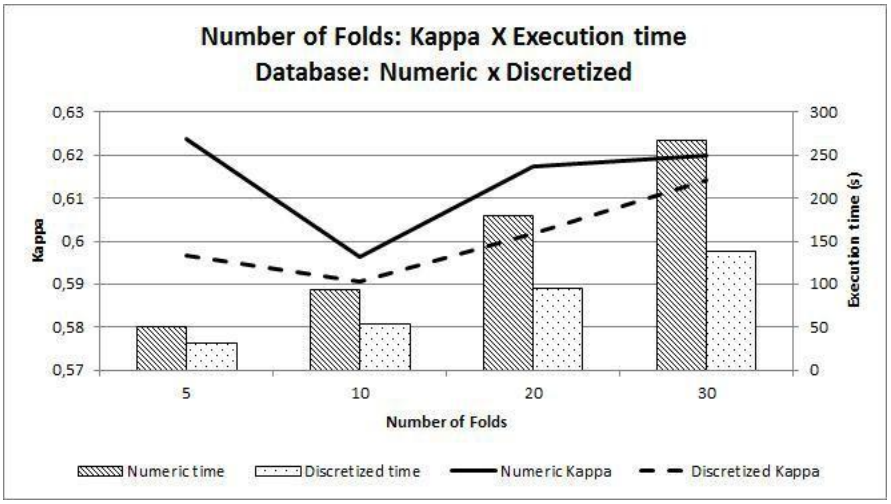

Figure 2. Performance achieved with non-binary decision trees built with C4.5 algorithm comparing the numeric and discrete databases.

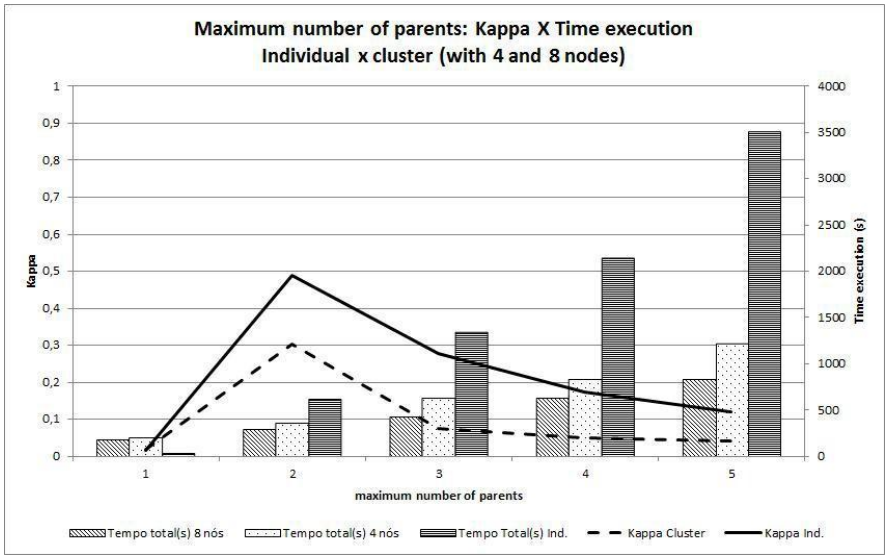

Figure 3. Performance comparison obtained with Bayesian networks to individual computer and the cluster using 4 and 8 nodes.

In general, all simulations had the same behavior. Using the database with numeric values, the processing time gets bigger and there is a slight improvement in the quality of the models, like in the kappa metric. For this reason, we considered the use of numeric database to search for a better model for classification.

We observed, from the analysis of the parameters that are common in several algorithms, that the main ones that influence the quality of the classifier on Bayesian networks are mainly the maximum number of parents and the estimator. On decision trees, the main parameters are the minimum number of instances for leaves and confidence factor. The process to search for a model with better quality was the last step, we varied only those parameters and fixed all other typical values, as the best results obtained in the analysis of each parameter. The main typical parameters are the number of folds and number of chunks equals to 2.

Figure 3 illustrates a performance comparison between the cluster and the individual computer, comparing the processing time and kappa, using cross-validation. For this simulation, we use Bayesian networks and K2 algorithm, with two chunks, maximum split size equal to 5000000, estimator equal to 0.1 and 10 folds as fixed parameters, varying only the maximum number of allowed parents.

As illustrated in **Figure 3**, despite the improvement on the cluster performance, on pro- cessing time, compared to individual computer, the quality of the model in the cluster was worse in most executions. The

worsening on the kappa, observed in most of the executions between the cluster and the individual computer, occurred by the way the cluster works in distributed WEKA, by dividing the complete database in chunks. As the model training is performed in each chunk, the model consequently has less anomalies cases to perform the training, producing a decline in the model quality on the cluster, in practically all simulations. Thus, for the final step of search by models with better quality, we use only the individual machine.

Table 3 illustrates the results of each metric to the best models found using the full database. The last column illustrates the best result, on average, obtained with the Bayesian networks, using the Hill Climber as the search algorithm. We use the following configuration: estimator equal to 0.1, the maximum number of parents equal to 5 and 30 folds. The second column shows the best result using the C4.5 algorithm, considering a binary tree with confidence factor equals to 0.9, minimum number of instances of leaves equal to 2 and 35 folds.

Metrics	**Algorithms**	
	Decision trees using C4.5	**Bayesian Net. using HillClimber**
Kappa	0.6944	0.5382
Accuracy	0.9988	0.9986
Precision	0.7531	0.8345
Sensibility	0.6451	0.3978
Specificity	0.9995	0.9998

Table 3. Results of the main models obtained using the complete database.

By observing the decision trees and Bayesian networks that generated the results il- lustrated in Table 3, it was found that occurred over-fitting on the models. By using the complete database, several variables could be removed for simplification and generalization of the model. Thus, on this stage we maintained only the most relevant variables as iden- tified using the lift calculation. **Table 4** illustrates the results of each metric for the best models found using the database with only the most relevant variables.

Metrics	**Algorithms**	
	Decision trees using C4.5	**Bayesian Net. using K2**
Kappa	0.5188	0.5122
Accuracy	0.9986	0.9984
Precision	0.8045	0.6432
Sensibility	0.3835	0.4265
Specificity	0.9998	0.9995

Table 4. Results of the main models obtained using the database with only the most relevant variables.

The second column of Table 4 illustrates the best result using decision trees. For this model, we used the C4.5 algorithm, considering non binary trees with confidence factor equals to 0.1, minimum number of instances per leaves equals to 2 and 10 folds. **Figure 4** illustrates a model obtained through the use of decision trees.

The accuracy and specificity for both models was very good, as illustrated in Table 4. For the precision, the best model was obtained using decision trees. For sensibility, the best model was obtained using Bayesian networks. Therefore, we observed an interesting behavior at the performance, wherein the use of decision trees complemented the use of Bayesian networks, and vice versa. **Table 5** illustrates the confusion matrix of the results illustrated in Table 4.

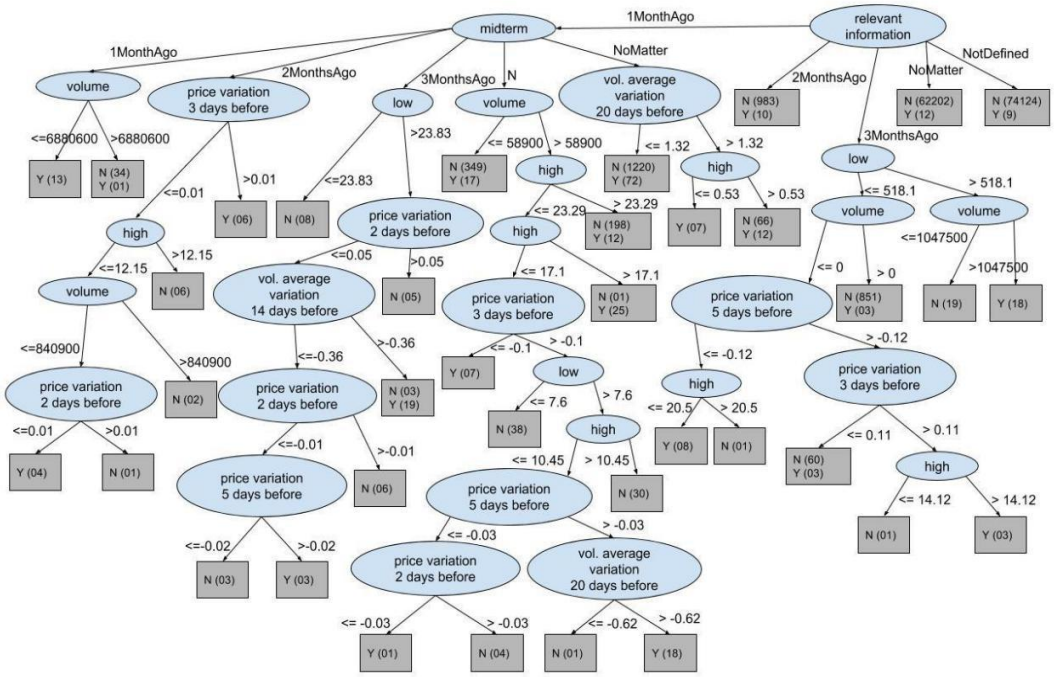

Figure 4. Best model obtained using decision trees.

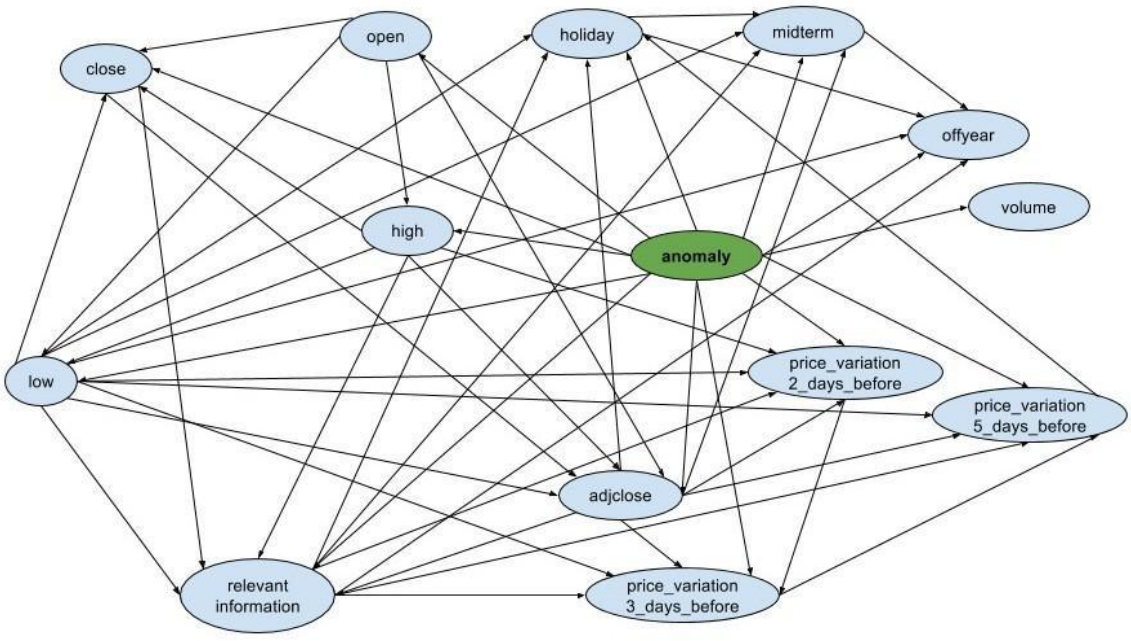

Figure 5. Best model obtained using Bayesian networks.

Decision Trees using C4.5			**Bayesian Net. using K2**		
N	Y		N	Y	
140039	26	N	139999	66	N
172	107	Y	160	119	Y

Table 5. Confusion matrix of the best models obtained using the database with only the relevant variables.

As shown in Table 5, the confusion matrix obtained by referring to the model using Bayesian networks, obtained a greater amount of true positives. This means that this model is better to classify anomalous cases. In contrast, the model obtained through the use of decision trees achieved a greater number of true negatives. Thus, this model is better to classify cases that are not anomalies.

5. Conclusions

In this work, the main contribution was the construction of a method to identify patterns, with the creation of classifiers to detect possible insider trading anomaly. These classifiers can be used within detection stock market anomalies procedures in order to list possible cases of insider trading and to subsidize further investigations. One of the advantages of the classifiers created in this work is that we use only publicly available information. However, this brings a cost in terms of classifier quality. We believe that if individualized data per investor were used the quality of the classifiers would significantly increase as the discrepancy between the insiders and the others investors would be sharper.

From the available financial transactions data on the stock exchange, together with the anomalous data collected, it was possible to acquire a sufficient mass of data to perform machine learning. We have created several candidate variables and analyzed their relevance to the problem addressed. Among these variables, it was concluded that the most important ones are those related to the relevant facts, the price and volume variation and some other variables related to political elections, as midterm and off-year.

We noticed that through the use of the numerical database, it was possible to produce models with better quality compared to the discretized database. Moreover, when performing the truncation of models treating the overfitting, the numerical database gets better results. By performing the same procedure in discretized database, we observed a further loss in the quality of the models.

We have also observed the difference in the results obtained in the individual machine and the cluster in almost all simulations. We concluded from these observations, that these differences occurred mainly because of the way both of them process the data. In the cluster, the input data are divided into chunks and distributed in data nodes by the masternode. The cross-validation process is performed on each chunk, whereas the single computer is performed on the entire database. For this reason, the kappa ends up being different.

Thus, we observed that, as the number of chunks increases, the kappa quality tends to decrease. We concluded that this happens due to the influence of this parameter on the size of the training database, since the bigger the number of chunks, the smaller the amount of data available for training the models in each chunk. This makes the parameter number of chunks, in most cases, inversely proportional to the quality of the kappa. Therefore, when using the cluster, the best kappa we obtained was using 2 chunks. However, most of these results are worse than the results achieved using a single machine, working with the complete database. For this reason, we obtained all final models using only the individual machine.

5.1 Discussions

One of the major difficulties encountered in this work was to obtain data anomalies, par- ticularly those related to foreign exchanges data. A large part of the available news about the crimes in the stock market does not label the type of crime. Therefore, the reading of each one of them and the correct interpretation to label the types of abnormalities in each case and discard those with incomplete information are required.

Several anomalous cases found occurred in companies that have merged or do not currently exist. In such cases, we searched for the whereabouts of the shares held by these companies at the time of occurrence of the anomaly. In many cases, it was not possible to find the records of these companies. Thus, these companies were removed from the analysis, decreasing the size of the database.

Although the results obtained with the integration of distributed WEKA with Hadoop, we have encountered some difficulties with the use of the plugin. The lack of a detailed documentation relating to the functioning and the commands that are used for running the algorithms and the data in a distributed way was one of the big challenges.

The use of the distributed environment, in this particular work, has not brought benefits to the quality of the models. Despite the performance improvement at the processing time, the quality of the models on the cluster was worse compared with the quality of the models obtained on an individual computer. Therefore, we concluded that the cluster could be a good alternative for works that have a really big database, with a large volume, in the order of terabytes, and that, in fact, could not be resolved using a single machine.

5.2 Future research

One of the main aspects that could be addressed in future projects is the inclusion of more data of anomalies. A possible idea is to incorporate anomalous data from the Asian and European financial market. Thus, with this information, it is possible to verify the performance of the classifier generated in this work on stock exchanges in different continents. This would also allow the comparison between similarities and differences in the behavior of crimes committed in different stock exchanges and use these results to create a classifier which may cover a global behavior.

Several types of crimes happen on the stock market, in addition to insider trading. Thus, it would be interesting to build models for other types of anomalies, such as the market prices manipulation, also known as pump-dump. Furthermore, it would be possible to make a comparison between the generated models to determine if there are similarities between these different anomalies.

Concerning the technical aspect, it would be interesting to accomplish a comparative study using Hadoop and SPARK together with WEKA, in order to discover if there is a significant difference of the computational performance at processing time. Besides that, another interesting test would be to use other data mining tools, such as R and KNIME, comparing the performance of these tools with the distributed WEKA.

Acknowledgments

We are grateful to CNPQ for financial support and the Technological Institute of Aeronautics (ITA), for providing the space and facilities.

References

1. COTSAFTIS, Michel. The autonomous intelligence challenge. *Journal of Autonomous Intelligence* 2018; 1(1): 1.
2. Saldanha, M., & Omar, N. Real-time manipulation in the Financial Market of Stocks and Derivatives by Spoofing and Layering, 2019.
3. Abualigah, L. M. Q. Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering. *Studies in Computational Intelligence*. 2019
4. Abualigah, L. M., &Khader, A. T. Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. *The Journal of Supercomputing* 2017; 73(11): 4773-4795.
5. Abualigah, L. M., Khader, A. T., &Hanandeh, E. S.. Hybrid clustering analysis using improved krill herd algorithm. *Applied Intelligence*. 2018
6. Abualigah, L. M., Khader, A. T., &Hanandeh, E. S. A Combination of Objective Functions and Hybrid Krill Herd Algorithm for Text Document Clustering Analysis. *Engineering Applications of Artificial Intelligence*. 2018
7. Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *Journal of Computational Science* 2017; 25: 456-466. doi: 10.1016/j.jocs.2017.07.018
8. Abualigah, L. M., Khader, A. T., Hanandeh, E. S., &Gandomi, A. H. (). A novel hybridization strategy for krill herd algorithm applied to clustering techniques. *Applied Soft Computing* 2017; 60: 423-435.
9. Remco R Bouckaert. Bayesian network classifiers in weka. Department of Computer Science, University of Waikato Hamilton, 2004.
10. Remco Ronaldus Bouckaert. Bayesian belief networks: from construction to inference. 2001.
11. Brasil. Law no. 6,385, of December 7, 1976, 1976. Available from: http://www.planalto.gov.br/ccivil_03/LEIS/L6385original.htm.
12. Peter Cabena, Pablo Hadjinian, Rolf Stadler, *et al*. *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc., 1998.
13. Dennis W Carlton and Daniel R Fischel. The regulation of insider trading. *Stanford Law Review*. 1983; 857–895.
14. LINTON, Oliver; MAHMOODZADEH, Soheil. Implications of high-frequency trading for security markets. *Annual Review of Economics* 2018; 10: 237-259.
15. Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of proba- bilistic networks from data. *Machine learning* 1992; 9(4):309–347.

16. Lawrence Davis. *Handbook of genetic algorithms*. 1991.
17. Sam Drazin and Matt Montag. Decision tree analysis using weka. *Machine Learning-Project II, University of Miami* 2012; 1–3.
18. Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters* 2006; 27(8):861–874.
19. Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine learning* 1997; 29(2-3):131–163.
20. Koosha Golmohammadi and Osmar R Zaiane. Data mining applications for fraud detection in securities market. In I*ntelligence and Security Informatics Conference (EISIC)*, European. IEEE; 2012. p. 107–114.
21. Thomas Gorman. Sec actions, 2016. Available from: http://www.secactions.com/.
22. Mark Hall. Mark hall on data mining weka. 2013. Available from: https://markahall.blogspot.com.br/2013/10/weka-and-hadoop-part-1.html.
23. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 2009; 11(1):10–18.
24. Alain Hertz and Dominique de Werra. Using tabu search techniques for graph coloring. *Computing* 1987; 39(4):345–351.
25. George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995. p. 338–345.
26. J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics* 1977; 159–174.
27. Andrew McCallum, Kamal Nigam, and Lyle H Ungar. Efficient clustering of high- dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000. p. 169–178.
28. Tom M Mitchell. Machine learning. Burr Ridge, IL: McGraw Hill 1997; 45:37. Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.
29. Viviane Muller Prado and Renato Vilela. Insider trading x-ray in the brazilian securities commission (cvm) 2002-2014, 2015.
30. Stuart Jonathan Russell and Peter Norvig. *Artificial intelligence: a modern approach. Pearson*, 3 edition, 2010.
31. SEC. Sec enforcement actions: Insider trading cases, 2016. Available from: https://www.sec.gov.
32. Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*, IEEE; 2010. p. 1–10.
33. Jarrod West and Maumita Bhattacharya. Intelligent financial fraud detection: a comprehensive review. *Computers & Security* 2016; 57:47–66.
34. Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and tech- niques*. Morgan Kaufmann, 2005.