Comparison of Classification Algorithms on Financial data

NURSEL SELVER RUZGAR
Ted Rogers School of Management
Ryerson University
350 Victoria Street, Toronto, ON, M5B 2K3
CANADA
nruzgar@ryerson.ca

Abstract: - Today's life, big data can be seen in many fields. There are many computer-based methods developed and continuing to be developed to assess the big data more efficiently. Data mining is one of them. In this paper, two Canadian banks' daily stock market price changes are examined by ten data mining algorithms to see which algorithm or algorithms classify the financial data well. For this purpose, thirty-seven years of daily stock price changes for two Canadian banks with 21 independent variables and one dependent variable, price, were obtained from NASDAQ. Ten data mining algorithms were applied to two datasets separately and the performances of the algorithms were compared and tested based on accuracy, kappa statistic, process time and confusion matrix. It was observed that tree algorithm, J48, and meta-analysis algorithms, Meta-Attribute Selected Classifier, Meta-Classification via Regression and Meta-Logitboost, classified the financial data with high accuracy. The results show that tree algorithm, J48, and the meta-analysis algorithms, Meta-Attribute Selected Classifier, Meta-Classification via Regression and Meta-Logitboost, are promising alternative to the conventional methods for financial prediction.

Key-Words: - Classification, Logistic Regression, Fuzzyrough-NN, Genetic Programming, J48, Random Forest, Navie Bayes, Navie Net, Meta-Analysis, Weka, Data mining

1 Introduction

In today's digital age, large volume of data can be seen in many fields. The essential goal of scientists and researchers to extract valuable information from the big data utilizing the appropriate methods. Data mining (DM) is one of them. DM is a data analysis technique based on statistical application; it aims to extract information that could previously not be determined, from massive quantities of data [1].

DM and knowledge discovery are a family of computational methods that aim at collecting and analysing data related to the function of a system of interest to gain a better understanding of the system [2]. DM attempts to formulate, analyse and implement basic induction processes that help extract meaningful information and knowledge from unstructured data. DM that aims to reveal valuable information from the overwhelming volume of data and achieve better strategic management and customer satisfaction is the process of using statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and knowledge assembled from large databases [3, 4]. DM, also referred to as knowledge discovery is the science of extracting critical information from large amount of existing raw data and deploying that information across the organization [5, 6]. DM can be used in different disciplines, such as finance [7], business and banking [8, 9], engineering [10], medicine [11-13] and science [14].

There are many DM methods to perform the analysis, such as clustering, classification, and association. Classification is of the widely used DM method to extract information from various high-dimensional data sets [4]. The classification includes the following algorithms; Logistic regression (LR), J48, Discriminant analysis (DA), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Decision Tree (DT), Rough sets (RS), Fuzzy Rough (FR), Fuzzy Rough -NN (FRNN), Random Forest (RF), Genetic Programming (GP), Associative Classification (AC), Neural Network (NN) and Support Vector Machine [15]. In literature, it is seen that LR have been applied to show a relationship between banking sector development and the financing of businesses with loans [16, 17], RS were used to classify price movements [18], financial data [19] and also to classify credit ratings in the global banking industry [20]. The literature on classification models is vast and offers a myriad of techniques that approach the classification problem

from multiple angles [21, 22]. Fuzzy Rough Set (FRS) has been extended to cast the RCN approach withing the framework of FRS to eliminate the need for a user-specified similarity threshold while retaining the model's discriminatory power [23, 24]. GP has been successfully applied in economic and financial prediction [25]. These algorithms encode a potential solution for a specific problem into a simple chromosome-like data structure and apply recombination operators to these structures to preserve critical information [26]. Similarly, J48 has been used to classify banking data [27].

Classification is one of the commonly used DM method. Many algorithms, including LR, cluster analysis, RS, FR, FRNN, GP, Meta-Attribute Selected Classifier (MAS) algorithms, and several other techniques, have been used for the classification of price changes [4]. In this paper, ten classification algorithms of WEKA software as a tool are used to classify the daily stock market closing price change of two Canadian banks: BN, NB, RF, J48, FRNN, GP, LR, MAS, Meta-Classification via Regression (MR) and Meta-Logitboost (ML). Weka allows the user to analyse the data from various perspectives and angles, in order to derive meaningful relationships. To determine which algorithm or algorithms classify the two datasets more effectively and efficiently with high accuracy, WEKA 3.7.2 and 3.9.3 are used as a tool. The classifier performances are tested based on accuracy, kappa statistic, running (processing) time and confusion matrix.

The rest of the paper is organized as follows. In section 2, classification algorithms used in this paper will be discussed briefly. Then, in Section 3, the purpose and methodology will be presented. In Section 4, findings will be discussed, and finally conclusion will be given in Section 5.

2 Classification Algorithms

In literature, there are different methods to assess the big data. Classification is one of the most used DM method. Classification is process of partitioning data into different classes or groups and collect the items into target classes. The main purpose of classification is to predict the target class for the data [28]. There are many different classifiers or algorithms. It is not exactly known that which will perform most efficiently and accurately in any given case. To find out which algorithm or algorithms classify the data more accurately, at least some of widely used one should be run. In this research, WEKA software will be used for the classification as a tool. In the rest of this section, some properties of WEKA software and ten classification algorithms used in this paper are summarized.

Weka is a DM software that implements data mining algorithms using a java language. It is an open source program developed by the University of New Zealand. Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection. All techniques of Weka software are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes. This software has many important advantages, so that we use it in our work: 1) It is fully implemented in the Java programming language, therefore runs on almost any architecture; 2) it is easy to use due to its graphical user interface; 3) It is a huge collection of data pre-processing and modelling techniques. [29] There are three steps to follow for classification by Weka: 1) preparing the data, 2) selecting and applying appropriate algorithm, 3) analysing the results. In the first step, the data should be converted to special dataset format. It supports multiple dataset format like csv data files, Json Instance files, libsvm data files, Matlab ASCII files etc., with the default being ARFF. After running the suitable algorithm, using performance measurements of classifier, analyse the results. These measurements are kappa statistic, accuracy, root mean square error, ROC, confusion matrix and so on. The classification accuracy is the percent ratio of the number of correctly predicted data points to the total number of data points. In literature, 80% is assumed as the threshold point [30] for financial data. The second performance measurement of classifier is kappa statistic which is used to indicate the agreement between the model's prediction and true values. Kappa statistic has a range from −1 to +1 [31]. It has been suggested the kappa result can be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement [32]. To measure the classifiers' prediction capability, the Kappa coefficient is used. Cohen's Kappa coefficient measures the inter-rater agreement for categorical items. It is usually deemed a more robust measure than the standard accuracy since this coefficient considers the agreement occurring by chance [33]. The other one is confusion matrix which presents a visualization of the classification performance based on a table that contains columns representing the instances in a predicted class and rows representing the instances

in an actual class. In this paper, processing (running) time was used as another performance measure.
Weka has various classification algorithms. Classification contains seven different types of classifiers: Bayes, Functions, Lazy, Meta, Misc, Rules and Trees [34]. Each classifier contains different number of algorithms. Ten of them were used in this study and illustrated in Fig. 1.

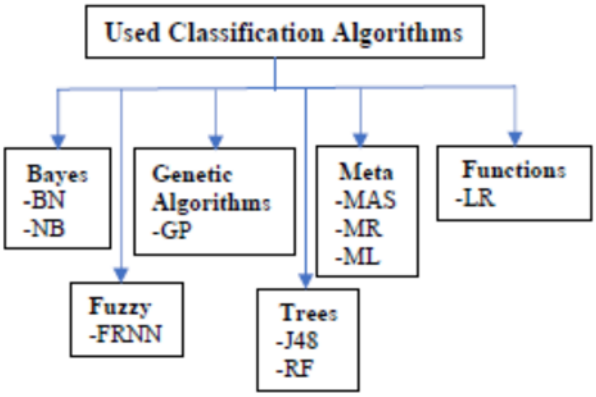

Fig. 1. Used classification algorithms

Bayesian classifiers, BN and NB, are both probabilistic algorithms. BN are directed acyclic graphs (DAG) whose nodes represent random variables. The nodes can be any observable quantities, variables, unknown parameters or hypotheses [34, 35]. Edges are the conditional dependencies. Nodes which are not connected represent the independent variables. Each node is associated with a probability function that takes as input a set of values for the node's parent variables and gives the probability of the variable represented by the node [35]. NB is a probabilistic classification algorithm using estimator classes, where numeric estimator precision values are chosen based on the analysis of the training data [36].
Fuzzy classifier, FRNN, is a new approach to fuzzy-rough nearest neighbour classification. Nearest neighbour model that utilizes the lower and upper approximations from fuzzy rough set theory to classify test instances [37].
GP algorithms encode a potential solution to a specific problem on a simple chromosome like data structure and apply recombination operators to these structures to preserve the critical information [38].
J48 Induces classification rules in the form of a pruned/unpruned decision tree. case J48 creates a decision node higher up in the tree using the expected value of the class. J48 can handle both continuous and discrete attributes, training data with missing attribute values and attributes with differing costs. Further it provides an option for pruning trees after creation [39], while RF is bagging of Random Trees [40].
Meta classification indicates the usage of combination of multiple classifiers. This combination is carried out within three steps: In first step, multiple training subsets are constructed from a training set. In second step, each classifier is solely constructed according to both the algorithm and data training subset. In third step, the results of base classifiers are integrated, and results are obtained in a higher-level step called Meta classifier. There is also a Multiclass Classifier Meta classifier that does this for any binary class classifier [41]. With MAS Algorithm, the range of the training data and testing data is lessened by this algorithm before being departed onto the classifier. The classifier is raised, so various search approaches are used during the phase of attribute selection. ML is a boosting algorithm is an extension of Adaboost algorithm. It replaces the exponential loss of Adaboost algorithm to conditional Bernoulli likelihood loss. This Class is used for performing additive logistic regression. [41]. MR uses regression approaches for classification.
Finally, LR is a classifier building linear logistic regression models. LogitBoost with simple regression functions as base learners is used for fitting the logistic models [42].

3 Methodology

The paper aims to explore which classification algorithm or algorithms out of ten DM algorithms classify the stock price changes for two Canadian banks effectively, accurately and fast. In this research, thirty-seven years of data for the period of 1980 to 2017 for two major banks in Canada were used to analyse by ten classification algorithms. The purpose of this paper is

- to determine the best classification algorithm or algorithms to classify the data based on process time
- to determine the best classification algorithm or algorithms to classify the data based on performance parameters of classifiers, accuracy, kappa statistic and confusion matrix
- to classify the daily stock price changes for two banks for future predictions
- to find out which method gives the most accurate results when the method applied for each bank's data separately.

Stock market prices are very important parameters for the investors. They would like to invest on any financial instrument which gives more profit on the stock market. The price changes affect their

investments. For this purpose, two large Canadian banks' stock market daily price changes, over the period from 1980 to 2017, were examined by ten different data mining classification algorithms, BN, NB, RF, J48, FRNN, GP, LR, MAS, MR and ML by using Weka 3.7.2 and 3.9.3 as a tool. Data for two banks were obtained from NASDAQ [43].

Each data set has twenty-one independent variables, and one dependent variable. Independent variables are Daily Opening price, Daily Opening bid, Daily Opening ask, Daily Closing price, Daily Closing bid, Daily Closing ask, Daily High, Daily Low, Daily Transactions, Daily Volume, Daily Quotes, Daily Quote changes, Daily Return, S&P/TSX Composite Price Index, S&P/TSX Composite Total Return Index, Sector 40 (Financials) Price Index, Sector 40 (Financials) Total Return Index, S&P/TSX 60 Price Index, S&P/TSX 60 Total Return Index, Call Loan Interest Rate and Foreign Exchange Rate (CA$/US$) and the dependent variable is the change in daily closing price according to the previous day's closing price. The dependent variable is the daily closing price change, which is grouped as "up", "down" and "same" according to the previous day's stock market daily closing price. If the closing price increased relative to the previous day's closing price, "up" is assigned as the new variable component, if the closing price is decreased relative to the previous day's closing price, "down" is assigned as the new variable component, and similarly, if the closing price remained the same, "same" is assigned as the new variable component

4 Findings

In this paper, ten algorithms, BN, NB, RF, J48, FRNN, GP, LR, MAS, MR and ML, were applied to the datasets for two major banks, TD and RBC, in Canada from 1980 to 2017 using Weka 3.7.2 and 3.9.3 to determine the best classification with the best prediction accuracies. For this purpose, the performance of classification algorithms was analyzed regarding the four commonly used parameters, accuracy, kappa statistics, process time and confusion matrix. Table 1. represents accuracies, kappa statistics and the process times of each algorithms used in this paper for TD and RBC datasets.

The first classification parameter is accuracy in Table 1. The classification accuracy is the proportion of the total number of predictions that were correct. The accuracy was evaluated using 10-fold cross-validation. The correctly classified numbers and accuracy rates of the classifiers for the both banks were summarized second and fifth columns. For example, accuracy of BN algorithm with real instances for TD was 64.6032% which means 6178 out of 9563 stock prices were classified correctly. For TD data set, the highest same accuracy, 92.3455 %, is given by the algorithms J48, MAS, MR and ML. Similarly, the highest accuracy for RBC dataset is given by J48 (92.1259%), MAS (92.1468%), MR (92.1259%) and ML (92.1364%). The accuracy of RF algorithm for both datasets followed the others. This shows the algorithms of meta-analysis and trees classify this type of financial data more accurately. The algorithms BN, NB and FRNN did not classify both datasets well based on accuracy.

Table 1. Accuracies, Kappa statistics and process time of Algorithms

	TD			RBC		
	Accuracy (%)	Kappa statistic	Time (seconds)	Accuracy (%)	Kappa statistic	Time (seconds)
BN	(6178/9563) 64.6032%	0.4932	0.33	(6239/9563) 65.241%	0.498	0.28
NB	(5058/9563) 52.8914%	0.341	0.13	(5283/9563) 55.2442%	0.3671	0.08
RF	(8780/9563) 91.8122%	0.852	3.95	(8772/9563) 91.7285%	0.8496	4.28
J48	(8831/9563) 92.3455%	0.861	0.19	(8810/9563) 92.1259%	0.8562	0.19
FRNN	(6945/9563) 72.6237%	0.5289	0.13	(6497/9563) 67.9389%	0.4487	0.13
GP	(8571/9563) 89.6267%	0.8104	31.67	(8522/9563) 89.1143%	0.8001	27.09
LR	(8603/9563) 89.9613%	0.8182	1.69	(8556/9563) 89.4698%	0.8081	1.45
MAS	(8831/9563) 92.3455%	0.861	0.52	(8812/9563) 92.1468%	0.8566	0.42
MR	(8831/9563) 92.3455%	0.861	2.47	(8810/9563) 92.1259%	0.8562	1.95
ML	(8831/9563) 92.3455%	0.861	7.36	(8811/9563) 92.1364%	0.8564	7.33

The second parameter is kappa statistic which measures the classifiers' prediction capability or the interrater agreement for categorical items. The highest kapa statistic for TD data set was obtained from J48, MAS, MR and ML. Although J48, MAS, MR and ML algorithms were having the highest kappa statistics, RF, GP and LR algorithms gave reasonable kappa statistic, more than 80% agreement, which is the recommended agreement by many texts as the minimum acceptable interrater agreement [44]. Similarly, the highest kappa statistics for RBC dataset were seen on J48, MAS, MR and ML algorithms. Besides these, RF, GP and LR algorithms also produced the valid kappa statistics. Kapa statistics of BN, NB and FRNN are in the range of 0.34 to 0.53 for TD dataset, and 0.36 to 0.50 for RBC dataset. Since these are under 80% threshold, they are not good classifiers for this kind of financial data based on kappa statistic.

The classification process time is the third parameter for classification. The running times of the all classifiers were measured in seconds and depicted in Table 1 for TD and RBC dataset. 0.13 sec. is the shortest running time of the NB and FRNN algorithms for TD dataset whereas NB has the shortest running time with 0.08 seconds for RBC data. Then with 0.19 sec. running time, J48 algorithm is the second-best classifier for TD dataset and with 0.13 sec. FRNN is the second-best algorithm for RBC dataset. On the other hand, GP is with the longest process time for both datasets. TD dataset, the classified time of GP algorithm in case of parameter used will takes more time. It took 31.67 sec. compared with other used algorithms. Similarly, for RBC dataset, GP algorithm running time took 27.09 sec. Based on the running time of the classifiers, NB and FRNN algorithms run in shortest time for both data sets.

The overall comparison based on accuracy, kappa statistic and time shows that first J48, then the meta-analysis algorithms, MAS, MR ad ML, classify the financial data well.

Table 2. Confusion Matrices

Algorithms		TD				RBC			
		Up	Down	Same	TCC*	Up	Down	Same	TCC*
BN	Up	2849	0	1680	6178	2922	0	1606	6239
	Down	26	2562	1589		31	2618	1563	
	Same	48	42	767		61	63	699	
NB	Up	2565	62	1902	5058	2663	41	1824	5283
	Down	658	1696	1823		609	1877	1726	
	Same	39	21	797		48	32	743	
RF	Up	4509	1	19	8780	4500	0	27	8772
	Down	35	4094	48		57	4130	26	
	Same	339	341	177		351	330	142	
J48	Up	4527	2	0	8831	4527	0	0	8810
	Down	40	4134	3		60	4149	4	
	Same	344	343	170		356	333	134	
FRNN	Up	3586	564	379	6945	3323	755	449	6497
	Down	567	3189	421		757	3036	420	
	Same	329	358	170		346	339	138	
GP	Up	4427	102	0	8571	4372	149	6	8522
	Down	46	4105	26		61	4133	19	
	Same	346	472	39		368	438	17	
LR	Up	4406	106	17	8603	4391	113	23	8556
	Down	126	4028	23		167	4035	11	
	Same	342	346	169		358	335	130	
MAS	Up	4528	1	0	8831	4527	0	0	8812
	Down	42	4133	2		60	4149	4	
	Same	346	341	170		356	331	136	
MR	Up	4528	1	0	8831	4526	1	0	8810
	Down	41	4134	2		56	4150	7	
	Same	346	342	169		356	333	134	
ML	Up	4528	1	0	8831	4527	0	0	8811
	Down	42	4133	2		60	4149	4	
	Same	346	341	170		356	332	135	

*TCC: Total correctly classified instances.

Table 2 shows the confusion matrix containing statistical measures used to describe the ability of the classifier to discriminate among the cases with "up", "down" and "remains same" classes. Confusion matrix presents a visualization of the classification performance based on a table that contains columns representing the instances in a predicted class and rows represent the instances in an actual class. The classification accuracy is the proportion of the total number of correct predictions. Confusion matrix presents a visualization of the classification performance based on a table that contains columns representing the instances in a predicted class and rows representing the instances in an actual class. Table 2 demonstrates the confusion matrices obtained from different algorithms for two data sets, TD and RBC banks, containing statistical measures used to describe the ability of the classifier to discriminate among the cases with "up", "down" and "remains same" classes.

In the actual case for TD with BN algorithm, out of 9563 data points, 6178 stock prices were classified correctly. Where stock prices increased in the previous day, 2849 of them increased; where stock prices decreased in the previous day, 2562 of them decreased; and where stock prices remained same in the previous day, 767 of them remained the same with 64.6032% ((2849+2562+767)/9563%) accuracy. In the actual case for TD with NB algorithm, out of 9563 prices 5058 stock prices were classified correctly. Where the stock prices increased in the previous day, 2565 of them increased; where stock prices decreased in the previous day, 1696 of them decreased; and where stock prices remained the same in the previous day, 797 of them remained the same with 52.8914% accuracy. Similarly, for TD bank with RF algorithm, 8780 real instances out of 9563 with 91.8122% accuracy; with J48 algorithm, 8831 real instances out of 9563 with 92.3455% accuracy; with FRNN algorithm, 6945 real instances out of 9563 with 72.6237% accuracy; with GP algorithm, 8571 real instances out of 9563 with 89.6267% accuracy; with LR algorithm, 8603 real instances out of 9563 with 89.9613% accuracy were classified correctly. In the actual case for TD bank with all meta-analysis algorithms, MAS, MR and ML, give the same number of correctly classified items, 8831 out of 9563 and the same accuracy rates of 92.3455%. From Table 2, it can be concluded that according to the correctly classified real instances the data of TD bank were best classified by the algorithms J48, MAS, MR and ML.

In the actual case for RBC with BN algorithm, out of 9563 data points, 6239 stock prices were classified correctly. Where stock prices increased in the previous day, 2922 of them increased; where stock prices decreased in the previous day, 2618 of them decreased; and where stock prices remained same in the previous day, 699 of them remained the same with 65.241% ((2922+2618+699)/9563 %) accuracy. In the actual case for TD with NB

algorithm, out of 9563 prices 5058 stock prices were classified correctly. Similarly, for RBC bank with NB algorithm, 5283 real instances out of 9563 with 55.2442% accuracy; with RF algorithm, 8772 real instances out of 9563 with 91.7285% accuracy; with J48 algorithm, 8810 real instances out of 9563 with 92.1259% accuracy; with FRNN algorithm, 6497 real instances out of 9563 with 67.9389% accuracy; with GP algorithm, 8522 real instances out of 9563 with 89.1143% accuracy; with LR algorithm, 8556 real instances out of 9563 with 89.4698% accuracy were classified correctly. In the actual case for RBC bank with meta-analysis algorithms, 8812 real instances out of 9563 with 92.1468% accuracy with MAS algorithm, 8810 real instances out of 9563 with 92.1259% accuracy with MR algorithm and 8811 real instances out of 9563 with 92.1364% accuracy with ML algorithm were correctly classified. From Table 2, it can be concluded that according to the correctly classified real instances the data of RBC bank were best classified by the algorithms MAS, J48, MR and ML.

When the algorithms are compared according to the number of correctly classified instances for both TD and RBC datasets, Meta-Analysis algorithms, MAS, MR and ML, and J48 correctly classify this kind of financial data. From Table 2, it can be concluded that NB and BN Algorithms are not good classifiers. It seems that RF, GP and LR algorithms are moderate classifiers.

5 Conclusion

In the digital age, the vast data should be analyzed properly. There exist several DM techniques and tools. In this paper, ten different DM classification algorithms were selected and run to classify the daily stock market price changes of two big Canadian banks in the period from 1980 to 2017. WEKA was selected as a tool and 3.7.2 and 3.9.3 versions were used for the analysis. The test mode used for analysis is 10-fold cross validation and full training set. The analysis is based on accuracy, kappa statistic, process time and confusion matrix.

The analysis results based on accuracy showed that J48 and meta-analysis algorithms, MAS, MR and ML classified both datasets well whereas the algorithms BN, NB and FRNN did not classify.

Similarly, the analysis results based on the second parameter, kappa statistics, and the third parameter, confusion matrix, gave the same name of algorithms, J48, MAS, MR and MS, as the best classifiers for both datasets. Unlike to classifications based on accuracy, kappa statistics and confusion matrix, based on the running time of the classifiers, NB and FRNN algorithms run in shortest time and GP is the longest time for both data sets.

The results have shown that, overall, the J48, MAS, MR and ML algorithms proposed in this paper best classify the price changes when compared with the other algorithms, so these algorithms are promising alternative to the conventional methods for financial prediction.

Many other parameters are left for future research such as more test modes can be considered, more datasets can be taken, and other data mining tools can also be compared. For further studies, it is recommended to use different algorithms to further clarify the best classification algorithm.

References:

[1] Inmon, W. H., Building Data Warehouse, QED/Wiley, Hoboken, NJ, USA, 2005.

[2] Triantaphyllou, E., Data Mining and Knowledge Discovery via Logic-Based Methods. New York: Springer, 2010.

[3] Kusrini, dan L.E.T. Algoritma, Data Mining, Andi Publishing, 2009, Yogyakarta. Indonesia.

[4] Ruzgar, N. S., Classification of Stock Market Price Change by Data Mining, The Journal of American Academy of Business, Cambridge, Vol. 25(2), 2020, pp.1-9.

[5] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufmann Publishers, ISBN 978-0123814791, July 2011.

[6] Sharma, N., Om, H., Early Detection and Prevention of Oral Cancer: Association Rule Mining on Investigations, WSEAS Transactions on Computers, Vol. 13, 2014, E-ISSN: 2224-2872, pp: 1-8.

[7] Cheng, C. H., Chen, T. L., Wei, L.Y., A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting, Information Sciences, Vol. 180, 2010, pp. 1610–1629.

[8] Dalloshi, P., Badivuku-Pantina, M., Empirical assessment of the impact of banking sector development on firm external financing, using the MELR model, WSEAS Transactions on Business and Economics, Vol. 15, 2018, pp. 512-521.

[9] Ferreira L., Borenstein D., Righi, M. B., Filho D. Teixeira, A., A fuzzy hybrid integrated framework for portfolio optimization in private banking, Expert Systems with Applications, Vol. 92, 2018, pp. 350–362.

[10] Kumar, A., Kumar, S., Decision Tree based Learning Approach for Identification of

Operating System Processes, WSEAS Transactions on Computers, Volume 13, 2014, pp. 277-288.
[11] Zeynu, S., Patil, S., Prediction of Chronic Kidney Disease Using Data Mining Feature Selection and Ensemble Method, WSEAS Transactions on Information Science and Applications, Vol. 15, 2018, pp. 168-176.
[12] Ivasic-Kos, M., Ipsic, I., Ribaric, S., Multi-level Image Annotation Using Bayes Classifier and Fuzzy Knowledge Representation Scheme, WSEAS Transactions on Computers, Vol. 13, 2014, pp. 635-644.
[13] Ramamurthy, B., Chandran, K.R., Shape-BasedImage Retrieval Using Canny Edge Detection and K-Means Clustering Algorithms for Medical Images. International Journal of Engineering Science and Technology, Vol. 3, 2011, pp. 1870–1877.
[14] Zeffora, J., Shobarani. A, R., Statistical Analysis of Random Forest on Real Estate Prediction, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Vol. 8 (8S), June 2019, pp. 640-644.
[15] Andriansah, Jl. R. C. and Achmad S.Jl. R. C., Comparative Accuracy of Regression Logistic Algorithm and C4.5 Based Chi Squared and Practical Swarm Optimization for Prediction Feasibility of Credit Giving, International Journal of Advance Studies in Computer Science and Engineering, IJASCSE, Vol. 7(7), 2018, pp. 1-7.
[16] Dalloshi, P., Badivuku-Pantina, M., Empirical assessment of the impact of banking sector development on firm external financing, using the MELR model, WSEAS Transactions on Business and Economics, Vol. 15, 2018, pp. 512-521.
[17] Ruzgar, B., Ruzgar, N. S., Classification of the Insurance sector with logistic regression, International Journal of Mathematical Models and Methods in Applied Sciences, Vol. 1(1), 2007, pp. 168-174, ISSN: 1998-0140, http://www.naun.org/journals/m3as/
[18] Ruzgar, N. S., Ruzgar B., Unsal, F., An Analysis of Price Movements Using the Rough Set Theory Approach, 19th International Conference on Applied Mathematics (AMATH '14), Mathematics and Computers in Science and Engineering Series, Vol. 38, 2014, pp. 91-98.
[19] Ruzgar, N. S., Ruzgar, B., Unsal, F., Rough set theory and discriminant analysis to classify financial data, International Journal of Economics and Statistics, Vol. 3, 2015, pp. 110-116.
[20] Chen, Y-S., Cheng, C.-H., Hybrid models based on rough set classifiers for setting credit rating decision rules in the global banking industry, Knowledge-Based Systems, Vol. 39, 2013, pp. 224–239.
[21] Witten, I. H., & Frank, E., Data mining: Practical machine learning tools and techniques. 2nd ed., San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.
[22] Nápoles, G., Mosquera, C., Falcon, R., Grau, I., Bello, R., Vanhoof, K., Fuzzy-Rough Cognitive Networks, Neural Networks, Vol. 97, 2018, pp. 19–27.
[23] Cornelis, C., De Cock, M., Radzikowska, A. M., Fuzzy rough sets: from theory into practice. In Handbook of granular computing, 2008, pp. 533–552
[24] Inuiguchi, M., Wu, W.-Z., Cornelis, C., Verbiest, NFuzzy-rough hybridization, Springer Berlin Heidelberg, 2015, pp. 425–451,
[25] Kim, M.J., Min, S.H., Han, I., An evolutionary approach to the combination of multiple classifiers to predict a stock price index, Expert Systems with Applications, Vol. 31, 2006, pp. 241–247
[26] Cheng, C-H., Chen, T-L., Liang-Ying W., A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting, Information Sciences, Vol.180, 2010, pp. 1610–1629
[27] Vanitha, K., Libia Rani, G., Analysis of Classification and Clustering Algorithms using Weka For Banking Data, International Journal of Advanced Research in Computer Science, Vol. 1 (4), 2010, pp. 104-107
[28] Fan, W., Bifet, A., Mining Big Data: Current Status, and Forecast to the Future, SIGKDD Explorations, Vol. 14(2), 2012.
[29] Kasperczul, A., Dardzinska, A., Comparative Evaluation of the different data Mining Techniques used for the Medical Database, acta mechanica et automatica, vol.10(3), 2016) DOI 10.1515/ama-2016-0036
[30] Laurier, C., Meyers, O., Serra, J., Blech, M., Herrera, P., Serra, X., Indexing music by mood: design and integration of an automatic content-based annotator. Multimedia Tools Applications, Vol. 48, 2010, pp. 161–184.
[31] McHugh, M. L., Interrater reliability: the kappa statistic, Biochem Med, Zagreb, Oct; 22(3), 2012, pp. 276–282.

[32] Cohen, W.W., Fast effective rule induction, in: Proceedings of the 12th International Conference on Machine Learning, 1995, pp. 115–123.

[33] Eugenio, B. D., Glass, M., The kappa statistic: a second look. Computational Linguistics, Vol. 30(1), 2004, pp. 95–101.

[34] Hemlata, Comprehensive Analysis of Data Mining Classifiers Using Weka, International Journal of Advanced Research in Computer Science (0976-5697), Vol. 9 (2), March-April 2018, pp. 718-723.

[35] Hussain, N. I., Choudhury, B., Rakshit, S., A Novel Method for Preserving Privacy in Big-Data Mining, International Journal of Computer Applications, (0975-8887) Vol. 103(16), October 2014,

[36] John, G. H., & Langley, P., Estimating continuous distributions in Bayesian classifiers, In Proceedings of the eleventh conference on uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc, 1995, pp. 338–345.

[37] Jensen, R., & Cornelis, C. In Proceedings of the 6th international conference on rough sets and current trends in computing, Vol. 5, 2008, pp. 310–319.

[38] Elmi, Z., Faez, K., Goodarzi, M., Goudarzi, N., Feature selection method based on fuzzy entropy for regression in QSAR studies, Research Article, Molecular Physics, Vol. 107(17), 2009, pp. 1787–1798.

[39] Quinlan, J. R., C4.5: programs for machine learning. Morgan Kauffman Publishers, 1993.

[40] Breiman, L., Random forests. Machine Learning, Vol. 45(1), 2001, pp. 5–32

[41] Devi, T. S., Sundaram, K. M., A Comparative Analysis of Meta and Tree Classification Algorithms Using Weka, International Research Journal of Engineering and Technology(IRJET), www.irjet.net, Vol.3(11) Nov-2016, pp. 77-83.

[42] Sumner, M., Frank, E., Hall, M., Speeding up logistic model tree induction. In Knowledge discovery in databases: PKDD, Springer, 2005, pp. 675–683.

[43] http://clouddc.chass.utoronto.ca.ezproxy.lib.ryerson.ca/ds/cfmrc/displayTSX.do?ed=2018&t=ts&f=daily &lang=en#v2, Accessed: May 4, 2019.

[44] Laurier, C., Meyers, O., Serra, J., Blech, M., Herrera, P. and Serra, X., Indexing music by mood: design and integration of an automatic content-based annotator. Multimedia Tools Applications, Vol. 48, 2010, pp. 161–184.