Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 122 (2017) 788–795

Procedia Computer Science

www.elsevier.com/locate/procedia

Information Technology and Quantitative Management (ITQM 2017)

Market Manipulation Detection Based on Classification Methods

Aihua Li[a,*], Jiede Wu[a], Zhidong Liu[a]

[a] *School of Management Science and Engineering, Central University of Finance and Economics, Beijing, 100081,China*

Abstract

In this paper, we use supervised machine learning methods to detect the market manipulation in China based on the information released by China Securities Regulation Commission (CSRC) and data in the security market. Among the supervised machine learning, we mainly use classification methods to detect the anomaly from the daily and tick trading data of manipulated stocks. As a result, we find that the supervised machine learning methods are good at detecting market manipulation from daily trading data and have poor performance on tick data, based the measure method of accuracy, sensitivity, specificity and area under the curve (AUC). The best used supervised machine learning models are K-Nearest Neighbor (KNN) and Decision Tree (DT) which have over 99% of them.

Keywords: Market manipulation, Classification methods, Anomaly detection;

1. Introduction

Generally, market manipulation was defined that a deliberate attempt to intervene in the free and fair operation of the market which can create artificial, false or misleading appearances with respect to price of security [1]. Market manipulation, broadly defined, has existed since the infancy of financial markets [2], which has become an important issue of emerging and developed market. Some researchers regarded that market manipulated could enhance the mobility. More focused on the disadvantage aspect such as that manipulated security market not only distorted the prices and transactions in the security market, but also undermined the function of security market. What's more, many investors would lose significantly because of most manipulators' illegal profit making. Thus, it is meaningful to detect the action before or after it happened.

There has been increasing research on data mining techniques in detecting market manipulation. Allen and Gale [3] classified three kinds of types, which are manipulation as action-based manipulation, information based manipulation, and trade-based manipulation. Here, we studied how to detect the trade-based manipulation that a trader attempts to manipulate a stock price simply by buying and then selling (or vice versa), without releasing any false information or taking any other publicly observable actions designed to alter the value of security, by using supervised and unsupervised machine learning methods based on information about the manipulated stocks from China Securities Regulation Commission (CSRC). To obtain this goal, daily stock data and tick stock data of 64 manipulated stocks published in CSRC website were selected in this paper. After data cleaning and transformation,

[*] Corresponding author. Tel: +86-10-62288622 ; fax: +86-10-62288622.
Email address: aihuali@cufe.edu.cn (Aihua Li)

10.1016/j.procs.2017.11.438

supervised machine learning models and rules were built, such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Tree (DT), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR), Artificial Neural Networks (ANN) and so on. Whats more, K-folds cross validation methods were used here to test the robust of these supervised methods.

The main contributions of our work are as follows. First, currently, most scholars' study of detecting market manipulation are theoretical and pattern description which is still hard to accurately and fast detect market manipulation. Here, in order to increase the possibility and efficiency of detecting market manipulation, we apply different supervised machine learning models to detect market manipulation in real time stocks data and find that the most effective supervised machine learning methods are K-Nearest Neighbor and Decision Tree. Second, we use daily and tick real time trading stock data to evaluate those supervised machine. Third, it would be benefit to the investors and regulators on stock market because of the high performance of market manipulation detection which is over 99% accuracy, sensitivity and specificity for some supervised machine learning models.

The remainder of the article is organized as follows. In section 2, we summarized the main result of prior study on market manipulation detection. In section 3, we introduced 7 kinds of supervised machine learning models that was used on detecting the market manipulation. Section 4 is the main section, in which we used these supervised machine learning models to detect market manipulation and evaluate the result of these models. Section 5 gave the conclusion and some future works.

2. Literature Review

Allen and Gale [3] and Jarrow [4] were the first researchers to study manipulation. After studying the history of stock-price manipulation, Allen and Gale classified the manipulation as action-based manipulation, information-based manipulation and trade-based manipulation[3]. They defined trade-based manipulation as manipulating stocks through actual trading or trading orders by distorting stock market prices, rather than changing company values or issuing false information. In a dynamic model of asset markets, by investigating large traders' manipulation of trading strategies in the securities market, Jarrow found that large investors had a greater impact on stock prices [4]. Carhart and Reed [5] found that a large trader can manipulated the market or lure the market to "manipulate" their own trades. And he developed a theory for option pricing in an economy with a larger trader showing that the standard option pricing models holds, but with a random volatility price process. Hanson and Oprea [6] developed an experimental model to study whether the manipulators could distort the prices in a prediction market.

Some researchers have attempted to detect manipulation in different methods. Aggarwal and Wu [7] studied the US stock market from 1990 to 2001 to detect market manipulation and found that the manipulated stock had abnormal stock prices, liquidity, volatility and return. H.Öğüt and Aktaş [8] compared the performance of Artificial Neural Networks (ANN) and Support Vector Machine (SVM) with discriminant analysis and logistics regression on detecting the market manipulation and found that data mining techniques were better than multivariate techniques. According to Mongkolnavin and Tirapat [9], association rules were applied to detect mark-the-close in intraday trades from the Thai Bond Market Association. Price variation in the market and behavior of investors were integrated to analyze warning signals in real time. And the method showed that a list of investors were in the market, who perhaps are manipulators. Fallh and Kordlouie [10] used logit model, artificial neural network, and multiple discriminant analysis to create stock price manipulation models in Tehran stock exchange, and the performances of three aforesaid models were effective. The selected data were thoroughly analyzed by runs test, skewness test, and duration correlative test and the selected samples were divided into two sets: manipulated and non-manipulated companies. The factors that were related to stock price manipulation were defined such as: size of company, P/E ratio, liquidity of stock, status of information clarity, and structure of shareholders. In Yangs paper [11], logistic regression model was chosen to detect stock price manipulation activities in Shanghai and Shenzhen market. They analyzed independent variables based on primary component analysis, which increased performance for forecasting the model. And the model was better than the linear regression model. Cao and McGinnity [12] proposed the Adaptive Hidden Markov Model with Anomaly States (AHMMAS) to detect intraday stock price manipulation activities. The stock tick data were level 2 data from NASDAQ and London stock exchange and the model was tested with simulated data and real market data. The performance evaluation of AHMMAS outperforms other benchmark algorithms such as Gaussian Mixture Models (GMM), K-Nearest

Neighbors Algorithm (kNN), and One Class Support Vector Machines (OCSVM). Leangarun and Thajchayapong [8] investigated two popular scenarios of stock price manipulations where are pump-and-dump and spoof trading. They defined and used level 2 data to train the neural network models which achieved high accuracy for detecting pump-and-dump, and two dimensional Gaussian model which showed that it could detect spoof trading.

The study above showed that market manipulation detection has drawn more attention from different countries researchers, and some classification methods played some roles for the detection process. Whats more, classification method is also called supervised learning which includes many models and algorithms. However, most of the researchers were focused on one or two method for market manipulation detection. And the experiment data were low frequency.

3. Methodology

Supervised machine learning is the search for algorithms that reason from externally supplied samples to produce general hypotheses, which then can predict the future samples [13]. In other words, the goal of supervised learning is to build a concise model, rule or pattern of the distribution of class labels in terms of predictor attributes. The resulting classifier is then used to assign class labels to the testing where the values of the predictor attributes are known, but the value of the class label is unknown. Supervised machine learning mainly consists of classification and regression. Classification is learning a function that maps (classifies) a data item into one of several predefined classes [14]. And similarly regression is learning a function that maps a data item to a real-valued prediction variable. In this paper, differnct supervised learning method including K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Tree (DT), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR), Artificial Neural Network (ANN) were used for security market manipulation detection.

KNN is an instance based classifier method. The parameter units consist of samples that are used in the method and this algorithm then assumes that all instances relate to the points in the n-dimensional space R^N[15]. The method is labor intensive when given large training sets, and did not gain popularity until when computing power became available.

SVM is a classier based on optimal method, which performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM method provides an optimally separating hyperplane and the margin between two groups is maximized. It is proven to be advantageous in handling classification tasks with excellent generalization performance [16].

Decision Tree is one of the most popular techniques for prediction. An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. Most of researchers have used this technique because of its simplicity and comprehensibility to uncover small or large data structure and predict the value. And the results are easily understood because of their reasoning process and can be directly converted into set of IF-THEN rules [17].

LDA and QDA could be called statistical learning method, and the fromer is widely used in discriminant analysis to predict the class label based on a given set of measurements on new unlabeled samples [14]. LDA is with the ability to capture statistical dependencies among the predictor variables indicates that it would be suitable to explore the linear classification problems. QDA is closely related to linear discriminant analysis, and it is assumed that the measurements from each class are normally distributed. Unlike LDA however, in QDA there is no assumption that the covariance of each of the classes is identical.

LR is a method that would use the given set of features either continuous, discrete, or a mixture of both types and the binary target, LR then computes a linear combination of the inputs and passes through the logistic function[18]. This method is commonly used because it is easy to implement and it provides competitive results.

ANN is another popular technique used in data mining. The advantage of neural network is that it has an ability to detect all possible interactions between predictors' variables[19]. Artificial neural network could also do a complete detection without having any doubt even in complex nonlinear relationship between dependent and independent variables[20]. Therefore, artificial neural network technique is selected as one of the best prediction method.

Above all, popular supervised learning methods such as KNN, SVM, DT, LDA, QDA, LR, and ANN were described, which were used for security market manipulation detection.

4. Empirical Analysis

We used supervised machine learning models to detect the daily trading data and tick trading data of the manipulated stock to find out the manipulated time and evaluate the supervised machine learning model. The experiment showed that supervised machine learning models are suitable to daily trading data and have high accuracy to detect the manipulated data, while have poor performance of detecting the tick trading data.

4.1. Data Description and Data Preprocessing

In this paper, we study the market manipulation cases released by China Security Regulatory Commission (CSRC). Once CSRC officials discovered someone has manipulated the stock market, they would punish the manipulators and release this case on the website regularly. Based on that, we can download the daily trading data and tick trading data of these manipulated stocks. We get 64 manipulated stocks of daily trading data, which include daily open price, daily highest price, daily lowest price, daily close price and daily trading volume and tick trading data, which include tick price, tick price change volume, tick trading volume, tick trading amount and type released by CSRC from 2013 to 2016. For each stock's analysis period from T_{i1} to T_{i2}, which make sure the rate of manipulated data about 20%.

$$T_{1i} = t_{1i} - 2 * p_i, \quad i = 1, 2, \cdots, n \tag{1}$$

$$T_{2i} = t_{2i} + 2 * p_i, \quad i = 1, 2, \cdots, n \tag{2}$$

In the (1) and (2), T_{1i} and T_{2i} represent the start day and the end day of the analysis of each stock respectively, p_i represents the manipulated period of each stock, t_{1i} and t_{2i} represent the start day and the end day of manipulation of each stocks respectively, and the n represents the number of stocks.

Table 2 showed that we merged all of the daily trading data and tick trading data, and standardize them by zero-mean normalization. We get 4, 593 daily trading data including 919 market manipulated points, and 8, 986, 466 tick trading data including 1,985,096 market manipulated points , which were obtained by data cleaning and integration. 1% of the tick trading data are randomly selected in order to compute faster . And we transformed the nonnumerical data into numerical data. Then, labeled the data as 1 if it was manipulated at the time according to CSRC, and others as 0. So the input data for the supervised machine learning was the preprocessed daily trading data and tick trading data, while the output is binary, marked 0 and 1.

Table 1. Part samples of manipulated stock cases

Stock name	Stock code	Start day of manipulation	End day of manipulation	Start day of analysis	End day of analysis
Jingyi Co., Ltd	002295	1/19/2015	1/19/2015	1/15/2015	1/21/2015
Fuda Co., Ltd	603166	7/5/2016	7/18/2016	5/26/2016	8/25/2016
Zhongxing Commerce Co., Ltd	000715	1/4/2013	5/26/2014	2/12/2009	4/17/2018
Shibeigaoxin Co., Ltd	600604	9/8/2015	9/9/2015	9/2/2015	9/15/2015
Shuangxin Co., Ltd	300100	11/26/2014	11/28/2014	11/18/2014	12/8/2014

Table 2. Analysis stock data

	Daily	Tick	Simplified tick
Total point	4,593	8,986,466	89,864
Out point	919	1,985,096	19,851
Out rate	20.27%	22.09%	22.09%

4.2. Performance Evaluation

Measures of the quality of classification are built from a confusion matrix which records correctly and incorrectly recognized examples for each class. Table 3 presents a confusion matrix for binary classification, where TP are true positive, TF are false positive, FN are false negative, and TN are true negative counts. In this paper, for example, TN means this data are predicted as being manipulated and they are truly being manipulated according to CSRC.

Table 3. Confusion Matrix

		Predicted class	
		success	failure
Actual class	success	TP	FN
	failure	FP	TN

Classifying performance without focusing on a class is the most general way of comparing algorithms. It does not favor any particular application. The most used empirical measure, accuracy, does not distinguish between the number of correct labels of different classes. Accuracy approximates how effective the algorithm is by showing the probability of the true value of the class label. In other words it assesses the overall effectiveness of the algorithm.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

Corresponding to it, two measures that separately estimate a classifiers performance on different classes are sensitivity and specificity which approximates the probability of the positive (negative) label being true; in other words, it assesses the effectiveness of the algorithm on a single class.

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{FP + TN} \tag{5}$$

A comprehensive evaluation of classifier performance can be obtained by the ROC which shows a relation between the sensitivity and the specificity of the algorithm ROC curves, which plot sensitivity as a function of specificity for all possible thresholds [21], and illustrate a classifiers trade-off between true positives and false negatives. A higher value of sensitivity for a given value of specificity indicates better performance. The area under the ROC curve (AUC) is a commonly used metric for evaluating a classifiers performance.

4.3. Experimental Results

We use 5-fold cross-validation to evaluate predictive models by partitioning the daily trading data and tick trading data into a training set to train the models, and testing set to evaluate them. For the supervised machine learning models, we use K-nearest neighbors (KNN), Decision tree classifier (DTC), Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), Logistic regression (LR), Artificial neural networks (ANN), Support vector machine(SVM) to build models.

Table 4 and Table 5 show the counting number of TP and FN in the confusion matrix and specificity and sensitivity for each supervised machine learning model of daily training data and tick trading data. As can be seen, most of the specificity and sensitivity have a rather high rate of daily trading data which means excellent performance of detecting manipulated data, while the sensitivity of tick trading data are low meaning poor performance of that.

Fig 1 presents the result of different models of daily and tick trading data based on above evaluation indexe, accuracy, sensitivity, specificity and AUC. From the fig 1(a), most of models of daily trading have a good accuracy, sensitivity and AUC which are over 90%, and some of them even reach 100%, while the result of specificity is not good as other index for some models of daily trading data. It is a remarkable fact that KNN and DTC perform excellent in the four indexes that all exceed 99%.

Table 4. Confusing Matrix of daily trading data

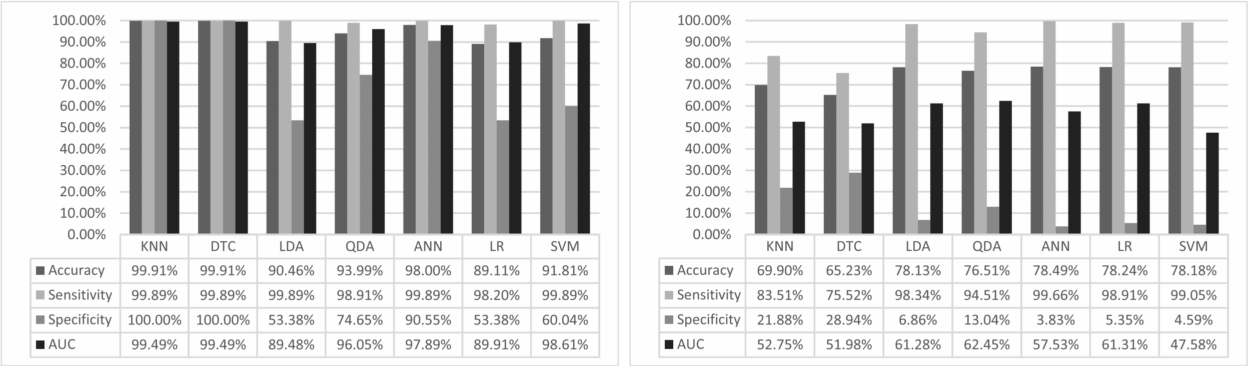
	KNN	DTC	LDA	QDA	ANN	LGR	SVM
Counting number	3658 931	3658 931	3658 497	3622 695	3658 843	3596 497	3658 559
Specificity	99.9%	99.9%	99.9%	98.9%	99.9%	98.2%	99.9%
Sensitivity	100.0%	100.0%	53.4%	74.7%	90.5%	53.4%	60.0%

Table 5. Confusion Matrix of tick trading data

	KNN	DTC	LDA	QDA	ANN	LGR	SVM
Counting number	58468 4344	52871 5745	68854 1361	66168 2589	69772 760	69250 1063	69348 911
Specificity	83.5%	75.5%	98.3%	94.5%	99.7%	98.9%	99.1%
Sensitivity	21.9%	28.9%	6.9%	13.0%	3.8%	5.4%	4.6%

The result is not ideal for the supervised machine learning models to train tick trading data according to the fig 1(b) showing all of the models have high score of sensitivity and rather low specificity. Our mainly aim is to accurately detect the manipulated data which pursue a high specificity.

Fig 2 shows Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) for different models of daily and tick trading data. The AUC of all the models of daily trading data is greater than 90% which means excellent estimate result, while the AUC of all the models of tick trading data is nearly 50% which is slightly better than random estimate result.



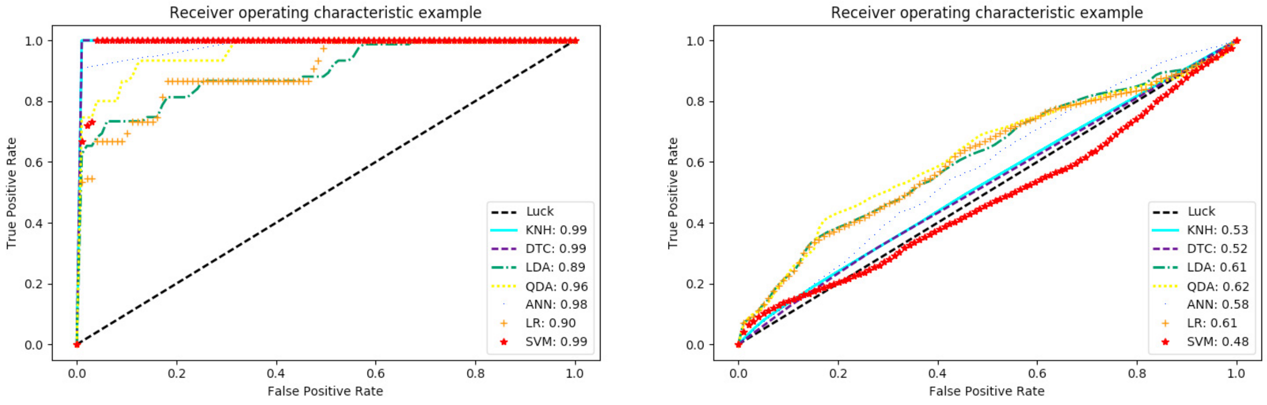
	KNN	DTC	LDA	QDA	ANN	LR	SVM
■ Accuracy	99.91%	99.91%	90.46%	93.99%	98.00%	89.11%	91.81%
■ Sensitivity	99.89%	99.89%	99.89%	98.91%	99.89%	98.20%	99.89%
■ Specificity	100.00%	100.00%	53.38%	74.65%	90.55%	53.38%	60.04%
■ AUC	99.49%	99.49%	89.48%	96.05%	97.89%	89.91%	98.61%

(a) daily trading data

	KNN	DTC	LDA	QDA	ANN	LR	SVM
■ Accuracy	69.90%	65.23%	78.13%	76.51%	78.49%	78.24%	78.18%
■ Sensitivity	83.51%	75.52%	98.34%	94.51%	99.66%	98.91%	99.05%
■ Specificity	21.88%	28.94%	6.86%	13.04%	3.83%	5.35%	4.59%
■ AUC	52.75%	51.98%	61.28%	62.45%	57.53%	61.31%	47.58%

(b) tick trading data

Fig. 1. Comparative result of different models of daily trading data and tick trading data



(a) daily trading data

(b) tick trading data

Fig. 2. ROC for different models of daily trading data and tick trading data

5. Conclusion and Future Work

This paper presents a comparative machine learning method for market manipulation detecting, especially for stock-price market manipulation. Based on manipulated information released by China Securities Regulation Commission (CSRC), we use supervised and unsupervised machine learning models to detect the anomaly from daily and tick trading data from the manipulated stocks. The supervised machine learning models, which are mainly classify machine learning models, have excellent performance for detecting the anomaly from the daily data, while in the performance of tick trading data is poor. Among the used classify machine learning models, KNN and DTC are the best, which exceed 99% of all the indexes, including accuracy, sensitivity, specificity and AUC.

For the supervised machine learning models to detecting the anomaly from the tick trading data, one of the reason for its poor performance is hard to exactly label the tick trading data as normal or abnormal, because it is almost impossible to know exactly which specific time or specific tick trading data was manipulated. As the same, the evaluation of clustering tick trading data are inaccurate as the difficulty of deciding accurate anomalies. So a more suitable way to process the tick trading data is worth to be further studied.

Acknowledgement

This paper is partly supported by the National Natural Science Foundation (71401188), Beijing Social Science Foundation (15SHB017) and Supported by Program for Innovation Research in Central University of Finance and Economics.

References

[1] T. C. Lin, The new market manipulation, Emory LJ 66 (2016) 1253.

[2] J. W. Markham, Law enforcement and the history of financial market manipulation, ME Sharpe, 2013.

[3] F. Allen, D. Gale, Stock-price manipulation, The Review of Financial Studies (1992) 503–529.

[4] R. A. Jarrow, Market manipulation, bubbles, corners, and short squeezes, Journal of financial and Quantitative Analysis 27 (3) (1992) 311–336.

[5] M. M. Carhart, R. Kaniel, D. K. Musto, A. V. Reed, Leaning for the tape: Evidence of gaming behavior in equity mutual funds, The Journal of Finance 57 (2) (2002) 661–693.

[6] R. Hanson, R. Oprea, A manipulator can aid prediction market accuracy 76 (302) 304–314.

[7] R. K. Aggarwal, G. Wu, Stock market manipulation - theory and evidence, Ssrn Electronic Journal (2003) 20–28.

[8] H. Öğüt, M. M. Doğanay, R. Aktaş, Detecting stock-price manipulation in an emerging market: The case of Turkey, Expert Systems with Applications 36 (9) (2009) 11944–11949.

[9] J. Mongkolnavin, S. Tirapat, Marking the Close analysis in Thai Bond Market Surveillance using association rules, Expert Systems with Applications 36 (4) (2009) 8523–8527.

[10] F. R. Roodposhti, M. F. Shams, H. Kordlouie, Forecasting stock price manipulation in capital market (80) 151.

[11] F. Yang, H. Yang, M. Yang, Discrimination of China's stock price manipulation based on primary component analysis, in: Behavior, Economic and Social Computing (BESC), 2014 International Conference on, IEEE, 2014, pp. 1–5.

[12] Y. Cao, Y. Li, S. Coleman, A. Belatreche, T. M. McGinnity, Adaptive hidden Markov model with anomaly states for price manipulation detection, IEEE transactions on neural networks and learning systems 26 (2) (2015) 318–330.

[13] S. B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised Machine Learning: A Review of Classification Techniques, 2007.

[14] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI magazine 17 (3) (1996) 37.

[15] R. S. Michalski, J. G. Carbonell, T. M. Mitchell, Machine Learning: An Artificial Intelligence Approach, Springer Science & Business Media, 2013.

[16] B. Zheng, S. W. Yoon, S. S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, Expert Systems with Applications 41 (4) (2014) 1476–1482.

[17] C. Romero, S. Ventura, P. G. Espejo, C. Hervás, Data mining algorithms to classify students, in: Educational Data Mining 2008, 2008.

[18] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, N. Afonoso, Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values, Computers in biology and medicine 59 (2015) 125–133.

[19] G. Gray, C. McGuinness, P. Owende, An application of classification models to predict learner progression in tertiary education, in: Advance Computing Conference (IACC), 2014 IEEE International, IEEE, 2014, pp. 549–554.

[20] P. M. Arsad, N. Buniyamin, others, A neural network students' performance prediction model (NNSPPM), in: Smart Instrumentation, Measurement and Applications (ICSIMA), 2013 IEEE International Conference On, IEEE, 2013, pp. 1–5.

[21] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems 47 (2009) 547–553.