March - April 2020
ISSN: 0193-4120 Page No. 4574 - 4583

A Bigdata Approach to Auditing Documents

Geun-Won Kim[1], Seong-Taek Park[2], Jinhwa Kim[*3]

[1]Graduate Student, Sogang University, 35 Baekbeom-ro Mapo-gu Seoul, 04107, Korea
[2]Professor, Sunmoon University, 70, Sunmoon-ro 221 beon-gil, Tangjeong-myeon, Asan-si, Chungnam, 31460, Korea
[*3]Professor, Business, Sogang University, 35 Baekbeom-ro Mapo-gu Seoul, 04107, Korea
dark_kgw@hanmail.net[1], solpherd@nate.com[2], jinhwakim@sogang.ac.kr[*3]

Article Info
Volume 83
Page Number: 4574 - 4583
Publication Issue:
March - April 2020

Abstract

Establishment and focus: The purpose of this study is to suggest an automatic document audit system using techniques in big data. In this paper, 200 documents on budget request are collected as test data. Text mining technique is used to analyze the documents. Major keywords regarding the requested budgets are induced. The documents are decided into training data and test data. A neural network, support vector machine, and regression analysis are used to test its own model. Finally, the performances of these three methods are compared to find the best model. The test confirms that techniques in big data can be applied to document auditing. This study can also be applied to similar problems such as lie detection and defect findings

System: This study suggests models for document auditing using techniques in big data such as text mining and data mining. A problem predicting costs of bill or budget is used as an example for a test problem. Documents containing cost of bill or budget are analyzed with techniques in text mining. Three data mining techniques such as neural network, logistic regression, and support vector machine are used to predict the output values as target values. The performance of these three methods are measured and compared. Among these three methods, support vector machine shows the best performance compared to other two methods of regression and neural network.

Keywords: *Document Audit, Value Prediction, Text Mining, Neural Networks, Logistic Regression, Support Vector Machine*

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 26 March 2020*

1. Introduction

One of the most important and difficult jobs in accounting is auditing, and one of the recent and rising techniques in auditing is big data. Frauds in financial areas, especially in accounting, have been problems continuously for a long time. Software in automatic document auditing has been developed in industries. Its functionalities and applications still remain in fundamental level with simple applications such as checking mistakes in documents and finding inconsistencies of contents in documents[1].

Automatic document auditing has long been studied in both industry and academia. Its complexity, however, has always been a big problem in this area as handling documents needs new advanced techniques. With the development of big data techniques such as text mining and data mining, the study on audit can now challenge new applications such as a fraud detection in documents. One of the popular techniques used in this area is expected to be auditing using big data techniques.

This study has significance in testing better prediction models in classifying documents like budget requests or bills of costs with big data

techniques for an automatic document audit system[2]. The study finds a solution to big data application to auditing documents in accounting. The study suggests a model that evaluates a requested budget or cost, whether it is correct and reasonable. The requested amount can sometimes be miscalculated, over exaggerated, or even manipulated. The suggested model can be used to find these problems in reasonable error range.

The study uses actual documents on monthly budget reports, which are used to reimburse costs from diverse purchases in an organization. A technique in text mining is used to analyze the documents and to find patterns related to the amount in budget or cost. Data mining techniques such as neural network, logistic regression, and support vector machine are used to predict the amount using the patterns from text mining. The prediction/classification performance of these three methods are evaluated and compared. Tests shows that support vector machine shows the best prediction performance consistently among these three methods at three different error ranges. This study shows that big data techniques such as text mining and data mining can be applied to document auditing in accounting.

2. Related Works

Big data techniques have been used in various fields ranging from political campaign to businesses applications. Frauds in accounting has been problems for long time until now[3]. Companies have already been using their own automated systems for monitoring frauds in documents[4]. Commercial software for automated document audit is under being developed, and its accuracy is also under improvement[5]. Modern techniques in big data are expected to improve them. Therefore, this study investigates an automatic document audit system using big data.

Text mining techniques are applied to diverse applications such as customer review analysis, news analysis, and social data analysis[6]. Fashion industry uses text mining in analyzing claim data such as customer complaints[7]. Text mining is also used in sport games. Real time strategies in soccer game is induced from the analysis of texts from broadcasting on the games[8]. Environmental scanning has been used to find social issues using text mining and topic analysis [9]. Research trends are identified from the analysis of papers on a specific topics using text mining techniques such as keywords analysis and association analysis[10]. Text mining is used to analyze political speeches to find major political issues they want to emphasize[11]. Customer reviews on a specific product or service are analyzed for target marketing. Text mining and data mining are used to find associated reviews for a customer group in this study[12]. Text mining is applied to future's wheel, which is a future prediction tool, to predict future. The study shows how text mining can be used to predict future of a new government policy[13]. There is a research on forecasting of the number of tourists using text mining and issue analysis[14].

Neural networks have been used in diverse areas such as prediction, classification, and inference. One of the popular applications of neural networks is corporate bankruptcy prediction. There are researches that compare the prediction performance of neural network and discriminant analysis using corporate financial data[15]. A hybrid model combining neural network and case-based reasoning system is suggested to improve the prediction accuracy[16]. Neural network is used to predict house prices in a region. The study suggests that neural networks show high prediction performance in predicting house prices compared to tradition methods[17]. Neural networks have been used for weather forecasting such as a predicting precipitation[18]. There is a study on prediction of route and intensity of typhoons in Japan. The study combined neural networks model and multi linear regression model together for prediction[19].

March - April 2020
ISSN: 0193-4120 Page No. 4574 - 4583

There is a study on a model which evaluates internal accounting system using neural networks and logistic regression model. Neural network model shows better performance compared to logistic regression model[20].

Support vector machine is used in diverse prediction and classification problems. There is a study on intruder detection system. Normalized image patterns are analyzed with support vector machine and factor analysis[21]. Support vector machine is used in semiconductor yield analysis. The study suggests an advanced support vector machine called steps-support vector machine, and compares its performance with that of other models[22]. There is a study on application of support vector machine to multifactor dimension reduction. The study uses times series data on agricultural product such as a beef[23]. Support vector machine is also used in manufacturing. It is used to find defects in manufacturing ball bearings[24].

3. Methods

This study suggests an automatic document audit system using big data techniques such as text mining and data mining. Figure 1 shows the processes in this research. Table 1 shows the documents file represented in keywords and their frequencies. For example, word 2 is shown 7 times and word 5 is shown 2 times in document 1. On the other hand, word 3 is shown 1 time and word 6 is shown 6 times in document 2.

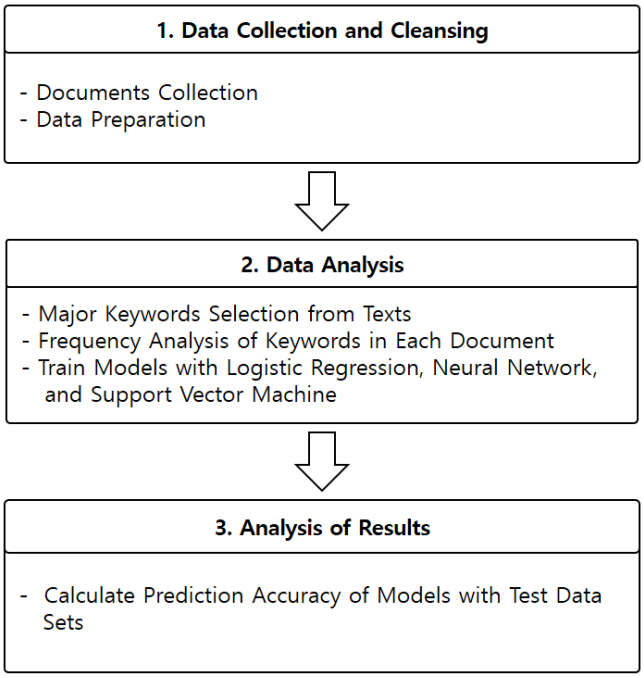

Figure 1. Research Processes in This Study

A data set of 200 bills of costs is collected for the study in step 1. The data set is divided into two groups of training data set and test data set. The documents on text with total budgets are collected as a data. As documents normally have lots of noises in them, data cleansing is necessary part in text analysis processes.

Unnecessary words are cleaned from the documents and only meaningful nouns are saved in the documents.

Table 1. Documents Represented in Keywords and Their Frequencies

	word 1	word 2	word 3	word 4	word 5	word 6	.	.	word n
document 1	0	7	0	0	2	0	.	.	0
document 2	0	0	1	0	0	6	.	.	0
document 3	0	1	0	3	0	0	.	.	0
document 4	0	0	6	0	1	0	.	.	2
document 5	0	2	0	1	0	2	.	.	0
document 6	0	0	0	0	0	0	.	.	1
document 7	0	0	1	0	3	0	.	.	0
.	.	.	.	.	.	.	.	.	.

March - April 2020
ISSN: 0193-4120 Page No. 4574 - 4583

.	.	.	.	.	.	.	.	.	.
document m	0	1	0	3	0	2	.	.	2

Similar words such as buy and purchase are named as one unified name.

In step 2, major keywords in bills of costs are extracted using text mining. The major keywords regarding the budgets are induced from these documents. These induced keywords in each document are words related to the total amount of budget or reimbursement.

Each document is represented with these keywords and their frequencies in each document as is shown in Table 1. Therefore, each document is represented as a words vector. These keywords with frequency are used to predict the output, the budget in this case, using neural network, logistic regression, and support vector machine. The amounts of the budget are normalized into 0 to 1 range for better performance in prediction. The prediction performance of these three prediction methods are evaluated with test data. The tests are performed 10 times with randomly permutated test data sets and the average performance of these three methods are compared.

4. Experimental Results

This study tests the models with 10-fold cross validation methods. It means the prediction accuracy of a model is tested 10 times with 10 different test data sets, which are selected randomly from original data set. Table 2, table 3, and table 3 show how prediction performance of three methods is calculated.

Table 2 shows prediction accuracy of neural network model using a test data set. The first column shows actual value, which is an original value to be predicted. The original values are normalized into 0 to 1 scale. Predicted value shows value predicted by neural network model. Error in % is calculated as ((actual value – predicted value)/actual value)/100. Positive error rate means predicted value is less than actual value. Negative error rate means predicted value is greater than greater than actual value.

As it is difficult to predict numeric number such as 0.514, there should be tolerable reasonable error rages. The tests have been done with three different error rages of ±5%, ±10%, and ±15%. The smaller the target, the more difficult to hit it right, and the bigger the target, the easier to hit it right. If the target it too big, prediction accuracy loses its credibility. The number 0 or 1 under these three error rate means hit or no hit. The number 0 means neural networks fail in predicting right target value, and the number 1 means success in predicting right target value.

The prediction performance of neural network at error rate ±5% is 10%. The prediction performance of neural network at error rate ±10% is 30%. The prediction performance of neural network at error rate ±15% is 40%. This is test result from one set of test data.

Table 2. Prediction Accuracy of Neural Network Using a Test Data Set

Actual Value	Predicted Value	Error in %	±5%	±10%	±15%
0.514735	0.803773	-56.1528	0	0	0
0.2294	0.277661	-21.0376	0	0	0
0.283009	0.27837	1.638965	1	1	1
0.420279	0.41135	2.124504	1	1	1
0.127372	0.273875	-115.02	0	0	0

0.298572	0.274378	8.10333	0	1	1
0.391656	0.40941	-4.53296	1	1	1
0.892187	0.740593	16.99132	0	0	0
0.228625	0.280933	-22.8794	0	0	0
0.208947	0.317537	-51.97	0	0	0
0.445324	0.41135	7.629017	0	1	1
0.253432	0.278024	-9.70379	0	1	1
0.435783	0.196954	54.80467	0	0	0
0.452122	0.803718	-77.7657	0	0	0
0.925616	0.798502	13.73299	0	0	1
0.160765	0.273883	-70.362	0	0	0
0.226479	0.273616	-20.8131	0	0	0
0.473947	0.41135	13.20754	0	0	1
0.057854	0.270471	-367.507	0	0	0
0.878353	0.803215	8.554359	0	1	1
0.563036	0.803769	-42.7562	0	0	0
0.645327	0.803772	-24.5528	0	0	0
0.18451	0.055895	69.7063	0	0	0
0.768047	0.672768	12.40534	0	0	1
0.239717	0.280933	-17.194	0	0	0
0.480507	0.569076	-18.4325	0	0	0
0.862849	0.802682	6.973044	0	1	1
0.435664	0.397309	8.803837	0	1	1
0.81933	0.419978	48.74133	0	0	0
0.597145	0.803771	-34.6024	0	0	0
			10%	30%	40%

Table 3 shows prediction accuracy of logistic regression using a test data set. The prediction performance of logistic regression at error rate ±5% is 16.67%. The prediction performance of logistic regression at error rate ±10% is 33.33%. The prediction performance of logistic regression at error rate ±15% is 43.33%. This is test result from one set of test data.

Table 3. Prediction Accuracy of Logistic Regression Using a Test Data Set

Actual Value	Predicted Value	Error in %	±5%	±10%	±15%
0.514735	0.648961	-26.0768	0	0	0
0.2294	0.258775	-12.8051	0	0	1
0.283009	0.28768	-1.65049	1	1	1
0.420279	0.322797	23.19464	0	0	0
0.127372	0.132149	-3.75063	1	1	1
0.298572	0.381381	-27.7349	0	0	0
0.391656	0.399095	-1.89938	1	1	1
0.892187	0.64692	27.49054	0	0	0
0.228625	0.292425	-27.9056	0	0	0

0.208947	0.423904	-102.876	0	0	0
0.445324	0.322797	27.51417	0	0	0
0.253432	0.138039	45.53209	0	0	0
0.435783	0.460071	-5.57342	0	1	1
0.452122	0.82762	-83.0523	0	0	0
0.925616	0.807992	12.70768	0	0	1
0.160765	0.126243	21.47354	0	0	0
0.226479	0.301851	-33.28	0	0	0
0.473947	0.322797	31.89177	0	0	0
0.057854	0.187831	-224.664	0	0	0
0.878353	0.923265	-5.11318	0	1	1
0.563036	0.746624	-32.6069	0	0	0
0.645327	0.639472	0.907294	1	1	1
0.18451	0.195798	-6.11749	0	1	1
0.768047	0.829901	-8.05331	0	1	1
0.239717	0.292425	-21.9876	0	0	0
0.480507	0.481324	-0.17021	1	1	1
0.862849	0.929061	-7.67363	0	1	1
0.435664	0.369796	15.11891	0	0	0
0.81933	0.733762	10.44367	0	0	1
0.597145	0.736398	-23.3199	0	0	0
			16.67%	33.33%	43.33%

Table 4 shows prediction accuracy of support vector machine using a test data set. The prediction performance of support vector machine at error rate ±5% is 26.67%. The prediction performance of support vector machine at error rate ±10% is 43.33%. The prediction performance of support vector machine at error rate ±15% is 56.67%. This is test result from one set of test data.

Table 4. Prediction Accuracy of Support Vector Machine Using a Test Data Set

Actual Value	Predicted Value	Error in %	±5%	±10%	±15%
0.514735	0.529641	-2.89584	1	1	1
0.2294	0.326765	-42.4431	0	0	0
0.283009	0.287	-1.4104	1	1	1
0.420279	0.40086	4.620674	1	1	1
0.127372	0.349673	-174.53	0	0	0
0.298572	0.361903	-21.2112	0	0	0
0.391656	0.406945	-3.90366	1	1	1
0.892187	0.787974	11.68059	0	0	1
0.228625	0.285967	-25.0809	0	0	0
0.208947	0.489915	-134.468	0	0	0
0.445324	0.40086	9.984804	0	1	1
0.253432	0.256571	-1.23869	1	1	1
0.435783	0.383304	12.04262	0	0	1

0.452122	0.704048	-55.7207	0	0	0
0.925616	0.834125	9.884394	0	1	1
0.160765	0.352405	-119.205	0	0	0
0.226479	0.346681	-53.0747	0	0	0
0.473947	0.40086	15.42105	0	0	0
0.057854	0.236588	-308.939	0	0	0
0.878353	0.867315	1.256662	1	1	1
0.563036	0.634431	-12.6804	0	0	1
0.645327	0.527257	18.29614	0	0	0
0.18451	0.25862	-40.1657	0	0	0
0.768047	0.677965	11.72869	0	0	1
0.239717	0.285967	-19.2936	0	0	0
0.480507	0.452265	5.877404	0	1	1
0.862849	0.881034	-2.10764	1	1	1
0.435664	0.409808	5.934921	0	1	1
0.81933	0.795182	2.947327	1	1	1
0.597145	0.630506	-5.58685	0	1	1
			26.67%	43.33%	56.67%

Table 5, Table 6, and Table 7 shows the prediction performance of three prediction techniques of logistic neural network, regression, and support vector machine at three different error ranges of ±5%, ±10% and ±15%. 10 fold cross validation is carried out to have stable performance of the models.

Table 5 shows comparison of prediction performance of three different method using 10-fold cross validation at error range of ±5%. Among three methods, support vector machine shows the best performance of 21%. Neural network shows the second best performance of 18% and logistic regression shows the third best performance of 14.67%.

Table 5. Comparison of Prediction Performance at Error Range of ± 5%

10 Fold Cross Validation	Error Range: ± 5%		
	Neural Network	Logistic Regression	Support Vector Machine
1	10%	16.67%	26.67%
2	26.67%	26.67%	6.67%
3	16.67%	16.67%	23.33%
4	16.67%	16.67%	3.33%
5	16.67%	6.67%	23.33%
6	10%	13.33%	20%
7	16.67%	10%	30%
8	20%	23.33%	26.67%
9	30%	6.67%	30%

10	16.67%	10%	20%
AVG	**18%**	**14.67%**	**21%**

Table 6 shows comparison of prediction performance of three different method using 10 fold cross validation at error range of ±10%. Among three methods, support vector machine shows the best performance of 41.67%. Neural network shows the second best performance of 37.67% and logistic regression shows the third best performance of 32.33%.

Table 7 shows comparison of prediction performance of three different method using 10 fold cross validation at error range of ±15%. Among three methods, support vector machine shows the best performance of 59%. Neural network shows the second best performance of 49.67 and logistic regression shows the third best performance of 44%.

At all error range of ±5%, ±10% and ±15%, support vector machine shows the best performance. Neural network shows the second best performance and logistic regression shows the third best performance. As show in comparative analysis of these three methods, the support vector machine is the best method when it is applied to document auditing. Figure 2 shows graphical representation of prediction performances of three different methods.

Table 6. Comparison of Prediction Performance at Error Range of 10%

10 Fold Cross Validation	Error Range: ± 10%		
	Neural Network	Logistic Regression	Support Vector Machine
1	30%	33.33%	43.33%
2	46.67%	40%	36.67%
3	36.67%	33.33%	40%
4	40%	33.33%	20%
5	43.33%	23.33%	43.33%
6	40%	26.67%	43.33%
7	30%	33.33%	56.67%
8	33.33%	40%	40%
9	46.67%	26.67%	46.67%
10	30%	33.33%	46.67%
AVG	**37.67%**	**32.33%**	**41.67%**

Table 7. Comparison of Prediction Performance at Error Range of 15%

10 Fold Cross Validation	Error Range: ± 15%		
	Neural Network	Logistic Regression	Support Vector Machine
1	40%	43.33%	56.67%
2	50%	50%	56.67%
3	56.67%	43.33%	60%
4	46.67%	40%	40%
5	53.33%	40%	60%
6	50%	36.67%	70%
7	40%	50%	60%
8	50%	46.67%	63.33%

March - April 2020
ISSN: 0193-4120 Page No. 4574 - 4583

9	53.33%	50%	63.33%
10	56.67%	40%	60%
AVG	**49.67%**	**44%**	**59%**

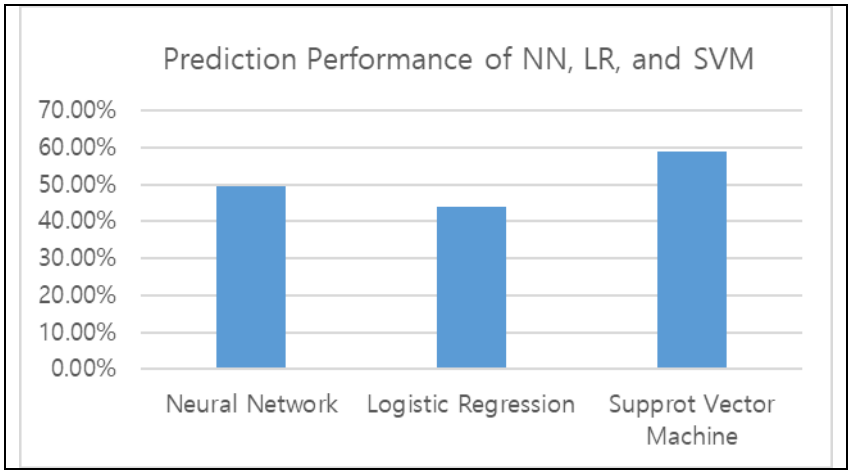

Figure 2. The Comparison of Prediction Performance of Three Methods at Error Rate ±15%

5. Conclusion

This study suggests models for document auditing using techniques in big data such as text mining and data mining. A problem predicting costs of bill or budget is used as an example for a test problem. Documents containing cost of bill or budget are analyzed with techniques in text mining. Three data mining techniques such as neural network, logistic regression, and support vector machine are used to predict the output values as target values. The performance of these three methods are measured and compared. Among these three methods, support vector machine shows the best performance compared to other two methods of regression and neural network. The test confirms that techniques in big data can be applied to document auditing. This study can also be applied to similar problems such as lie detection and defect findings.

One of the limitations of this study is the sample size. The sample size in this study is 200 due to limited availability of actual bills of costs or budgets. Diverse real world problems in auditing can be challenged with the suggested approach in this study. This study is using techniques in text mining and data mining. Future study can use move advanced techniques in text mining and machine learning. Techniques in artificial intelligence such as a deep can be applied to document auditing in the future.

References

[1] Kim KH. A Method for The application of Text Mining to Market Segmentation Using Online Customer Review. Journal of Korea Contents Association. 2009;9(8):272-284.

[2] Kim RS. Bankruptcy Prediction with Neural Networks. Korea East-West Economic Review. 2004;16(1):65-80.

[3] Chu HS, Min JK, Lee IH. A Study on Classification of Corporate Data and Bankruptcy Prediction Using Multiple Neural Network Models. Yonsei Business Review. 2004;41(2):513-539.

[4] Kim MJ. A Comparative Study on Performance of Internal Audit System Using Regression Analysis and Neural Network. Korean International Auditing Review. 2012;46:1-30.

[5] Kim SK, Cho HJ, Kang JY. Application of Text Mining in Academic Researches and Analysis of Major Techniques. Journal of Information Technology and Architecture. 2016;13(2):317-329.

[6] Kim KH, Oh SY. A Method for Market Segmentation Using Text Mining with Customer

March - April 2020
ISSN: 0193-4120 Page No. 4574 - 4583

Reviews in Online. The Journal of the Korea Contents Association. 2009;9(8):272-284.

[7] Oh HS, Jo SK, Kang CW, Lim DS. An Analysis of Claim Data Using Text Mining. Journal of the Korean Data Analysis Society. 2010;12(1):297-305.

[8] Jung HS, Lee JW, Yoo JH, Lee HS, Park DH. Analysis of Soccer Games Using Web Casts and Text Mining. 2011;11(10):59-68.

[9] Won JY, Kim DG. Inducing Social Risk Issues Using Text Mining. Crisisonomy. 2014;10(7):33-52.

[10] Lim SY, Lim YM, Lee JY. A Study on Research Trend in U-City and Smart City using Text Mining. Journal of the Korean society for geospatial information science. 2014;22(3):87-97.

[11] No HN. Big Data Text Mining on Political Speeches. Journal of Speech Communication. 2014;26:289-325.

[12] Kim JY, Kim DS. A Study on Models for Customized Information Using Text Mining with Customer Reviews in On-line. The Journal of Society for E-Business. 2016;21(2):151-161.

[13] Kim HJ, Kim JH. A Prediction on Success of Government Policies Using Text Mining and Future's Wheels. The Society of Digital Policy & Management. 2016;14(12):141-153.

[14] Park SJ, Shin JO, Song SH, Jung C. Demand Forecast on Tourists Using Text Mining and Search Engine. International Journal of Tourism Sciences. 2017;41(1):13-27.

[15] Kim SR. A Bankruptcy Prediction Using Neural Networks. Studies on Eastern Western Economy. 2004;16(1):65-80.

[16] Lee KJ, Ahn BY, Kim MH. A Diagnosis System Using Neural Networks and Case-Based Reasoning System. 2006;16(1):130-133.

[17] Jung WG, Lee SY. A Study on Predicting Price Index of Apartments Using Neural Networks. Housing Studies. 2007;15(3):39-64.

[18] Kang BC, Lee BK. Predicting Precipitations Using Neural Networks with Mid-Size Water Use Prediction. Proceeding of Korean Society of Civil Engineers. 2008;28(5B):485-493.

[19] Choi KS, Kang KR, Kim DW, Kim TR. Prediction of Moving Routs and Intensity of Typhoons Using Neural Networks. Journal of the association of Korean geographers. 2009;30(3):294-304.

[20] Kim MJ. Comparison of Evaluation Models on Internal Accounting Systems Using Logistic Regression and Neural Networks. International Studies on Accounting. 2012;46:1-30.

[21] Jung SY, Kang BD, Kim SK. An Intrusion Detection System Using Principle Component Analysis and Support Vector Machines. Proceeding of Korea Multimedia Society. 2013;314-317.

[22] Ahn DW, Ko HH, Kim JH, Baik JG, Kim SS. A Yields Prediction in the Semiconductor Manufacturing Process Using Stepwise Support Vector Machine. IE Interfaces. 2009;22(3):252-262.

[23] Lee JY, Chang JE, Oh DY. Major Gene Identification for SREBPs and FABP4 Gene in Korean Cattle. Journal of Korean Data & Information Science Society. 2015;26(3):677-685.

[24] Kim YS, Lee DH, Kim DW. Fault Severity Diagnosis of Ball Bearing with Support Vector Machine. Proceeding of The Korean Society of Mechanical Engineers. 2012;57-64.