Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 1681 – 1690

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2016, 5-7 September 2016, York, United Kingdom

Model selection for financial statement analysis: Variable selection with data mining technique

Ken Ishibashi[a]*, Takuya Iwasaki[a], Shota Otomasa[a] and Katsutoshi Yada[a]

[a] *Data Science Laboratory, Kansai University, 3-3-35 Yamate, Suita, Osaka 564-8680, Japan*

Abstract

The purpose of this study is to verify the effectiveness of a data-driven approach for financial statement analysis. In the area of accounting, variable selection for construction of models to predict firm's earnings based on financial statement data has been addressed from perspectives of corporate valuation theory, etc., but there has not been enough verification based on data mining techniques. In this paper, an attempt was made to verify the applicability of variable selection for the construction of an earnings prediction model by using recent data mining techniques. From analysis results, a method that considers the interaction among variables and the redundancy of model could be effective for financial statement data.

Peer-review under responsibility of KES International

Keywords: Financial statement analysis; earnings prediction model; model selection; variable selection; data mining

1. Introduction

Recent advancement in information and communication technology is dramatically improving computational speeds. Under the circumstances, researchers have addressed studies focused on big data accumulated in various areas. Data mining techniques play an important role in data-driven analysis and modeling. Various methods related to data mining have been developed until now, and software such as SPSS and Weka has been developed to enable us to use them easily. However, for these applications, we generally need to select a method appropriate to data.

The purpose of this study is to verify the effectiveness of a data-driven approach for the financial statement analysis. In the area of accounting, Ou and Penman (1989)[1)] addressed the construction of an earnings prediction

* Corresponding author. Tel.: +81-6-6368-1121.
E-mail address: r108047@kansai-u.ac.jp

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.216

model focused on financial statement data. They constructed a prediction model for the probability of a firm's earnings increase in the subsequent fiscal year by using stepwise logistic regression analysis. By introducing variable selection, their prediction model used variables' interactions that have not been proved theoretically. That is, it is possible that they constructed an earnings prediction model using unusual information that other people do not have.

The result of Ou and Penman (1989)[1)] has various problems related to the practical use of their method. In that research[1)], they did not state the reason why they applied logistic regression analysis to the model construction. Furthermore, follow-up studies[2), 3)] pointed out various problems through additional verifications of the model of Ou and Penman (1989)[1)]. For example, Holthausen and Larcker (1992)[2)] applied the strategy of Ou and Penman (1989)[1)] to another fiscal period, but could not obtain anomalies of the probability of earnings on investment. One reason of this result is over-fitting caused by the data dependence of variables adopted by the model. In the area of accounting, many researchers have mainly addressed the over-fitting problems by constructing models based on theoretical concepts such as the corporate valuation theory and behavioral economics (Gow 2011, p.119)[4)]. However, an approach using data mining techniques to financial statement data has not been addressed enough since Ou and Penman (1989)[1)]. Thus, the applicability of data mining techniques has not been verified enough.

In this study, an attempt is made to verify the effectiveness of the construction of an earnings prediction model based on financial statement data, by using data mining techniques. Financial statement data contains many variables. Thus, two stages-analysis that consists of variable selection and model selection is expected to be suitable for the construction of a prediction model. In this paper, focusing on the variable selection as the first stage, the verification of the applicability of data mining techniques is performed by applying methods used in the benchmarking of Hall and Holmes (2003)[5)]. We construct logit models[1)] for the prediction of the probability of firm's earnings increase by using multiple datasets for prediction of different fiscal periods, and compare each method based on their characteristics. In this way, this study verifies the applicability of variable selection using data mining techniques for financial statement data.

2. Application of variable selection

In this study, an attempt is made to construct a model of earnings prediction based on financial statement data using data mining techniques. Verification of the applicability of this approach is performed by a two-stage investigation: variable selection and model selection, because the financial statement data is composed of a large number of variables. In the modeling, increase in the number of variables tends to prevent construction of an accurate prediction model due to the complication of model. Therefore, the variable selection supports construction of accurate model by removing variables that are not related to the prediction previously. In this paper, focusing on variable selection, we verify the applicability of variable subsets obtained by several methods.

2.1. Variable selection with data mining technique

Generally, variable selection with data mining technique is categorized into two types: "filter" and "wrapper". Filters evaluate variables by using appropriate alternative criteria. On the other hand, wrappers use results obtained by applied a learning algorithm. Therefore, results of wrappers have relatively high accuracy but they tend to be computationally expensive because they contain learning algorithms.

Recently, various methods for variable selection have been developed. For the verification, this study applies several methods used in the benchmark test of variable selection[5)]. Firstly, Relief[6)], Correlation-based feature selection (CFS)[7)] and Consistency-based subset evaluation (CNS)[8)] are applied as filters. Existing research[5)] showed that these methods had different characteristics regarding variable selection. Secondly, this study uses C4.5 decision tree learner[9)] as a wrapper. In addition, stepwise methods are used for comparison with previous research[1)]. Relaying on Ou and Penman (1989)[1)], this study verifies the applicability of variable selection through the construction of a prediction model with logistic regression.

2.2. Relief

Relief is an instance-based attribute ranking scheme proposed by Kira and Rendell (1992)[6], and later improved by Kononenko (1994)[10]. This method is applied to the estimation of a variable's importance for the classification. In a classification of certain class, Relief decides a variable's importance by focusing on instances located around the border of the class. From these instances, two instances are selected as near-miss and near-hit. The near-miss is an instance that is the closest to randomly selected samples but is not the same class as them. On the other hand, an instance selected as near-hit is the closest to them and is the same class. In Relief, the importance of a variable is decided based on the effectiveness for the classification of near-miss. Existing research[5] showed that this method had large tolerance to noise but low redundancy.

In the application of Relief to variable selection, variables to adopt are generally decided by setting a threshold to their estimated ranks. In this study, the importance of variables is decided by 10-fold cross-validation, and we adopt variables for which the "Merit" criterion for the classification is more than 0 are adopted.

2.3. Correlation-based feature selection

CFS is a method that evaluates subsets of variables, not individual variables[7]. This method searches subsets containing variables that are highly correlated with the class and have low inter-correlation with each other. CFS tends to be computationally cheap and choose small variables' subsets, but it is difficult to search solutions if there are strong variable interactions[5].

In this study, we use a Greedy algorithm to search for a subset that has the best CFS's evaluation.

2.4. Consistency-based subset evaluation

CNS evaluates variables' subsets by using class consistency[8]. This method searches for combinations of variables which divide the data into subsets containing strong single class majority. Thus, this search tends to be biased in favor of small variable subsets with high-class consistency. Compared with CFS, CNS is useful if there are strong variable interactions, but the size of subset tends to be large[5].

In this study, CNS searches for subsets by using a Greedy algorithm like in CFS.

2.5. C4.5 decision tree learner

C4.5 is a learning algorithm that constructs a decision tree by selecting variables appropriate to maximize the mutual information for classification[9]. This method can avoid over-training to data by the function called "branch pruning", which removes branches that have little mutual information or classify few instances. In the variable selection, variables contained in the decision tree are adopted as a subset of variables.

In this study, a decision tree is constructed by using all training data for modeling, and then branches of which the number of classifying data is less than 50 are removed by the pruning. In this way, we obtain a subset with a size equivalent to CFS's subsets.

2.6. Stepwise method

In existing research, Ou and Penman (1989)[1] constructed an earnings prediction model by using stepwise logistic regression. Stepwise method is a conventional method that sequentially chooses variables to enhance evaluation criteria. In this method, the process of variable selection is very clear. However, because the effect of each variable is sequentially evaluated, this method is computationally expensive and it is difficult to take account of the interaction among variables.

In this study, we construct a logit model by using the stepwise forward selection method with all variables. In addition, an attempt is made to apply the same method as Ou and Penman (1989)[1]. The previous method constructs a logit model through three stages. In the first stage, logit models with each variable are constructed respectively, and then variables are removed if their p-value in each model is more than 10%. In the second stage, all variables

that remained in the first stage are entered into a logit model, and then some of them are removed in the same way as the first stage. Finally, a subset of variables is decided by applying stepwise backward selection method to the remaining variables.

3. Variable selection for financial statement data

In order to verify the applicability of variable selection using data mining techniques, we apply the methods described in Chapter 2. In addition, the dependency on data is verified by applying the obtained subset of variables to multiple datasets.

3.1. Data set

In this study, an earnings prediction model is constructed by using consolidated financial statement data of firms listed in the first section of the Tokyo Stock Exchange during 2007-2014, obtained from NIKKEI NEEDS financial data. We construct a prediction model with the same 65 variables as the previous research[1]. The previous research used 68 variables. However, this study excepted 3 variables because it is difficult to calculate them from the data described above.

Relying on Ou and Penman (1989)[1], this study estimates the probability of earnings increase (or decrease) in the subsequent fiscal year that is indicated by descriptors in the financial statements and the prediction model. We denote current profit earnings per share for a given firm i in fiscal year t as $e.p.s._{it}$, and define Pr_{it+1} (the earnings change from that year) as follows:

$$\mathrm{Pr}_{it+1} = \begin{cases} 0 & \left(e.p.s._{it+1} - e.p.s._{it} - drift_{it+1} < 0\right) \\ 1 & \left(e.p.s._{it+1} - e.p.s._{it} - drift_{it+1} > 0\right) \end{cases} \tag{1}$$

where, $drift_{it+1}$ represents the average change of $e.p.s.$ in the last 4 years from fiscal year t+1 for firm i. In this study, a firm of which the amount of increase in fiscal year t+1 from the previous year's $e.p.s.$ is larger than the average change during the last 4 years is defined as an earnings increase firm (Pr_{it+1} = 1), and a firm of which the amount of increase is smaller is defined as an earnings decrease firm (Pr_{it+1} = 0). Pr_{it+1} indicates the relative ability of firm i to generate earnings in the subsequent fiscal year. Thus, it has the character of a 'future earning power' attribute referred to by traditional fundamental analysts[1].

In this analysis, the prediction of earnings change in the subsequent fiscal year is performed by using datasets for 2012, 2013 and 2014. An earnings prediction model is constructed by using data in the 5 years period before the predicting year. For example, the prediction of 2014 uses financial statement data from 2009 to 2013 as training data, and Pr_{it+1} (earnings change from 2014 to 2015) is predicted with data of 2014 as test data. Here, the training data for model construction includes only each firm's financial statement data disclosed to public in the same months during 5 years. Under this condition, the number of extracted firms in each dataset is shown in Fig. 1. In Fig. 1, the number of firms whose earnings in the subsequent year decreased (Pr_{it+1} = 0) is labeled as "Decrease", and the number of firms whose earnings increased (Pr_{it+1} = 1) is labeled as "Increase".

All firms' financial statement data in the predicting fiscal year is used as the test data for prediction. The number of firms in each predicting fiscal year is shown in Table 1. From Table 1, it is found that the number of earnings increase firms is equivalent to earnings decrease firms in the datasets for 2012 and 2014. On the other hand, a large part of the dataset for 2013 is data of earnings increase firms.

3.2. Construction of earnings prediction model

Relying on Ou and Penman (1989)[1], this study constructs a model for earnings prediction by using logistic regression. The dependent variable of this model is the probability of an earnings increase in the subsequent fiscal year Pr_{it+1} described in Section 3.1. The explanatory variables are selected from 65 variables by using methods

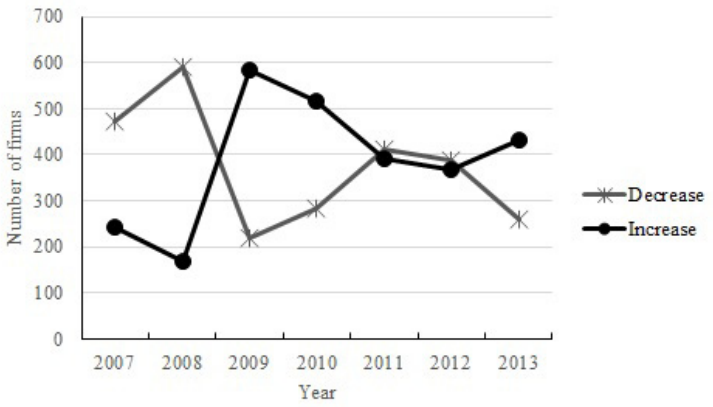

Fig. 1. Number of firms in training data sets over time.

Table 1. Number of firms in each test dataset

Prediction year of dataset	2012	2013	2014
Decrease	601	457	450
Increase	506	635	461
Total	1107	1092	911

Table 2. Results of variable selection (1)

Variable		Relief	CFS	CNS	C4.5	Stepwise	Previous
cr	Current ratio					A	
c_cr	%⊿ in current ratio	**A, B, C**		C			
qr	Quick ratio				B		
c_qr	%⊿ in quick ratio	**B, C**				C	
ds_ar	Days sales in accs. receivable	**A, B, C**			**A, B, C**		
c_ds_ar	% ⊿ in days sales in accs. receivable	**A, B, C**	**A, B, C**	**A, B, C**	**A, B, C**	**A, B, C**	**A, B, C**
it	Inventory turnover	A			A	B	
ia	Inventory / total assets	A					A
c_ia	%⊿ in inventory / total assets	**A, B, C**		**A, B**	C		
c_inv	%⊿ in inventory	**A, B, C**	**A, B**	**B, C**			
c_sales	%⊿ in sales	**A, B, C**	**A, B**	**A, C**			
c_dep	%⊿ in depreciation	B	B	**A, B**	B	B	B
c_dps	%⊿ in dividend per share	**A, B, C**		**A, B, C**		A	A
dep_fa	Depreciation / plant assets					B	
c_dep_fa	%⊿ in depreciation / plant assets	**B, C**					
roe	Return on operating equity	**A, B, C**	**A, B, C**	**A, C**	**B, C**	**B, C**	
c_roe	%⊿ in return on operating equity		B	B			**A, B**
c_cap_ex	%⊿ in (capital expenditure / total assets)	C	**A, B**	**A, B, C**		C	**A, C**
l1_c_cap_ex	%⊿ in (capital expenditure / total assets), one year lag	**A, C**			A		A

⊿ indicates changes. In calculating %⊿, observations with zero denominators are excluded and absolute values are used in all denominators.

Table 3. Results of variable selection (2)

Variable		Relief	CFS	CNS	C4.5	Stepwise	Previous
debt_equity	Debt-equity ratio				B		
c_debt_equity	%⊿ in debt-equity ratio	**A, B, C**	**A, B, C**	**A, B, C**		**A, B**	
LTdebt_equity	Long term debt to equity						
c_LTdebt_equity	%⊿ in long term debt to equity			**A, B, C**	B		
c_equity_fa	%⊿ in equity to fixes assets	**A, B, C**		**A, B, C**	A		
tie	Times interest earned		**A, B, C**	**B, C**			
c_tie	%⊿ in times interest earned			**A, C**			
sa_ta	Sales / total assets			A			
c_sa_ta	%⊿ in Sales / total assets	**A, B, C**	**A, B, C**	C			C
rot	Return on total assets	**A, B, C**	**A, B, C**	**A, B, C**	C	**A, B, C**	**A, B, C**
roc	Return on closing equity	**A, B, C**	**A, B, C**	**A, B**	**A, B, C**	A	**A, C**
gross_margin	Gross margin ratio	**A, B, C**			**A, B**	**A, B**	A
ops	Op. profit (before dep.) to sales	**A, B, C**		C		**B, C**	**A, B**
c_ops	%⊿ in Op. profit (before dep.) to sales			A			
pis	Pretax income to sales	**A, B, C**	A	C		C	
c_pis	%⊿ in pretax income to sales		C	B	C		
npm	Net profit margin	**A, B, C**	**A, B, C**	**A, B, C**	C	B	**A, B**
c_npm	%⊿ in net profit margin		**A, C**				**A, B**
stc	Sales to total cash				C	**A, B, C**	
c_pcosts	%⊿ in production	**A, B, C**		**A, B, C**			
c_rd	%⊿ in R&D		A	**A, B, C**		**A, B, C**	**A, B, C**
c_ade	%⊿ in advertising expense	**A, B, C**	**A, B**	**A, B, C**	A	**A, B**	C
c_ads	%⊿ in (advertising /sales)	**A, B, C**		**A, B, C**			C
c_ta	%⊿ in total assets	**A, B, C**		C			C
ctd	Cash flow to total debt	B			C	**A, B, C**	**A, B**
wct	Working capital / total assets	A			B		
c_wct	%⊿ in working capital / total assets	**A, B, C**		C			
oit	Operating income / total assets	**A, B, C**	B	**A, B, C**	C	**A, B, C**	**A, B, C**
c_oit	%⊿ in operating income / total assets		**A, B, C**		B		
rlt	%⊿ in total uses of funds				A		
c_funds	%⊿ in funds				B		
c_wcp	%⊿ in working capital			**A, B, C**			
noc	Net income over cash flows		**A, B, C**	**A, B, C**	**A, B, C**	**B, C**	**B, C**

⊿ indicates changes. In calculating %⊿, observations with zero denominators are excluded and absolute values are used in all denominators.

Table 4. Results of prediction

			All	Relief	CFS	CNS	C4.5	Stepwise	Previous
2012	Number of variables		65	27	17	23	10	12	15
	AUC	Decrease	0.655	0.661	0.654	0.660	0.647	0.663	**0.677**
		Increase	0.655	0.661	0.654	0.660	0.647	0.663	**0.677**
		Average	0.655	0.661	0.654	0.660	0.647	0.663	**0.677**
	Recall	Decrease	**0.672**	0.646	0.642	0.666	0.587	0.661	0.666
		Increase	0.557	0.563	0.591	0.585	**0.599**	0.575	0.571
		Average	0.620	0.608	0.619	**0.629**	0.593	0.621	0.622
	Precision	Decrease	0.643	0.637	0.651	**0.656**	0.635	0.649	0.648
		Increase	0.589	0.572	0.582	**0.596**	0.550	0.588	0.590
		Average	0.618	0.607	0.619	**0.628**	0.596	0.621	0.622
	F-Measure	Decrease	0.657	0.641	0.647	**0.661**	0.610	0.655	0.657
		Increase	0.573	0.568	0.586	**0.590**	0.573	0.581	0.580
		Average	0.619	0.608	0.619	**0.628**	0.593	0.621	0.622
2013	Number of variables		65	27	17	22	13	16	11
	AUC	Decrease	0.685	0.685	0.686	0.676	0.660	0.688	**0.690**
		Increase	0.685	0.685	0.686	0.676	0.660	0.688	**0.690**
		Average	0.685	0.685	0.686	0.676	0.660	0.688	**0.690**
	Recall	Decrease	0.764	0.753	**0.768**	0.740	0.676	0.764	0.753
		Increase	0.469	0.476	0.482	0.510	**0.554**	0.477	0.517
		Average	0.592	0.592	0.602	0.606	0.605	0.597	**0.615**
	Precision	Decrease	0.509	0.508	0.516	0.521	0.522	0.512	**0.528**
		Increase	0.734	0.728	0.743	0.731	0.704	0.737	**0.744**
		Average	0.640	0.636	0.648	0.643	0.628	0.643	**0.654**
	F-Measure	Decrease	0.611	0.607	0.617	0.611	0.589	0.613	**0.621**
		Increase	0.573	0.575	0.585	0.601	0.620	0.579	**0.610**
		Average	0.588	0.588	0.598	0.605	0.607	0.594	**0.614**
2014	Number of variables		65	27	12	26	12	12	11
	AUC	Decrease	0.666	0.675	0.666	0.675	0.679	0.677	**0.684**
		Increase	0.666	0.675	0.666	0.675	0.679	0.677	**0.684**
		Average	0.666	0.675	0.666	0.675	0.679	0.677	**0.684**
	Recall	Decrease	0.651	0.638	0.600	0.631	**0.653**	**0.653**	0.620
		Increase	0.584	0.614	**0.668**	0.605	0.586	0.599	0.627
		Average	0.617	0.626	**0.634**	0.618	0.619	0.626	0.623
	Precision	Decrease	0.604	0.617	**0.638**	0.609	0.606	0.614	0.619
		Increase	0.631	0.635	0.631	0.627	0.634	**0.639**	0.628
		Average	0.618	0.626	**0.635**	0.618	0.620	0.626	0.624
	F-Measure	Decrease	0.627	0.627	0.619	0.620	0.629	**0.633**	0.619
		Increase	0.607	0.624	**0.649**	0.616	0.609	0.618	0.628
		Average	0.617	0.626	**0.634**	0.618	0.619	0.625	0.623

described in Chapter 2. Tables 2 and 3 shows the results of variables selections. In addition, setting 0.5 as the threshold of the probability of an earnings increase (Pr = 0.5), Table 4 shows the prediction results of logit models constructed using variable subsets.

Tables 2 and 3 represent selected variables obtained by applying each method to 3 datasets. If a variable is chosen in the dataset for 2012, the value of its row is represented as A. A selected variable in the dataset for 2013 is B, and C represents a variable chosen in the dataset for the prediction of 2014. The columns are applied methods. "Previous" is the method of Ou and Penman (1989)[1] described in Section 2.6. For example, in the row of variable "cr", this means that the variable was chosen by "Stepwise" when that method was applied to the dataset for 2012. In addition, the result of "c_cr" with "Relief" was represented as "**A, B, C**" because "Relief" selected "c_cr" in all datasets. In this way, a variable chosen in multiple datasets is represented in bold font.

In the results of variable selection, it is found that Relief and CNS choose many variables compared to other methods. The result obtained by stepwise forward selection method tended to be similar to the previous method[1] because they followed the same criteria, although their procedures are different. In comparison with result obtained from each dataset, every method except for C4.5 tended to choose similar variables regardless of datasets. For example, 15 variables were selected by CFS in 2 or more datasets, while it totally chose 21 variables in all datasets. Thus, it is expected that each method has a preference of choosing variables independent of datasets.

In Table 4, prediction accuracies of logit models constructed by using variable selection results were assessed with 4 evaluation criteria (Witten and Frank 2005, pp.168-173)[11]. Area under the curve (AUC) is an area under the receiver operating characteristic (ROC) curve. In this study, we assess the prediction accuracy by using F-Measure, because this paper does not discuss the importance of Recall and Precision in the earnings prediction. Furthermore, in Table 4, "Decrease" represents the prediction accuracy for earnings decrease firms, the accuracy for earnings increase firms is "Increase", and "Average" is the total prediction accuracy for test data calculated by averaging "Decrease" and "Increase". Values represented in bold font represent the best results among prediction accuracy of each dataset. On the other hand, underlined values mean that the applied subset of variables decreased the prediction accuracy compared with logit model "All", which is constructed by using all variables.

From prediction results, it was shown that methods obtaining high accuracy differ depending on the dataset. Compared with a model using all variables, results of Relief could improve AUC but it was difficult to enhance F-Measure. The reason may be that adopted variables were decided based on estimated importance without considering influence on prediction. CFS obtained high accuracy for the predictions of 2014, but prediction accuracy for 2012 was decreased. As referred to by existing research[5], it is expected that this was caused by the interaction among variables. That is, it is indicated that there is strong interaction among variables in financial statement data. On the other hand, the effect of CNS on the improvement of prediction accuracy was relatively small, but the method had high accuracy for prediction of 2012. Therefore, it is showed that CNS is useful for a dataset with large interaction among variables.

C4.5 wrapper method had relatively small effects on the improvement of prediction accuracy. We conjecture that this was caused by applying the selection result to the logit model. Variable selection of C4.5 is effective in the prediction with constructed decision tree. Thus it seems that compared with other methods, the bias of prediction by C4.5 between "Increase" and "Decrease" is small, but it is difficult to obtain high accuracy. Finally, in the variable selection with stepwise methods, the method of Ou and Penman (1989)[1] obtained larger AUC than another stepwise method. This previous method decides variables' subset through more steps than "Stepwise", which applies the stepwise forward selection method to all variables directly. In this way, the previous method may previously remove variables not related to prediction accuracy and explanation power of model. In addition, both methods using stepwise method had larger AUC than other methods. It is expected that stepwise logistic regressions have a similar characteristic to wrapper because their variable selection was performed with considering the criteria of logit model.

From these results, it was shown that variable selections using data mining techniques could improve the accuracy of earnings prediction model by using financial statement data. However, the prediction for each dataset obtained the best F-Measure by different methods. This indicates that the influence of variables on earnings prediction is not stable due to prediction periods. Therefore, it is necessary to consider a measure against over-fitting referred to in existing studies[4], by investigating the data dependence of each method.

Table 5. Results of additional experiment

		All	CNS 2012	CNS common	Prev. 2013	Prev. common	CFS 2014	CFS common
	Number of variables	65	22	23	11	12	12	15
2012	AUC	0.655	0.660	0.656	0.668	**0.678**	0.636	0.642
	Recall	0.620	**0.629**	0.625	0.628	0.628	0.594	0.603
	Precision	0.618	**0.628**	0.625	**0.628**	**0.628**	0.599	0.604
	F-Measure	0.619	**0.628**	0.625	**0.628**	**0.628**	0.595	0.604
2013	AUC	0.685	0.684	0.686	**0.690**	0.696	0.662	0.666
	Recall	0.592	0.592	0.599	**0.615**	0.607	**0.615**	0.614
	Precision	0.640	0.644	0.643	**0.654**	0.649	0.637	0.641
	F-Measure	0.588	0.616	0.596	0.614	0.605	**0.617**	0.616
2014	AUC	0.666	0.677	**0.678**	0.664	0.675	0.666	0.666
	Recall	0.617	0.626	0.625	0.618	0.615	**0.634**	0.629
	Precision	0.618	0.626	0.625	0.618	0.615	**0.635**	0.629
	F-Measure	0.617	0.626	0.625	0.618	0.615	**0.634**	0.629

3.3. Verification of data dependence

Based on the results of Section 3.2, we verify the data dependence of each method. Firstly, variables' subsets that obtained the best F-Measure for each dataset (CNS 2012, Prev. 2013 and CFS 2014) are applied to all datasets. Next, from results of three methods that obtained subsets described above, they respectively make new subsets by adopting variables that were chosen for multiple datasets in common. In this way, based on the selection tendency of each method, we verify the effectiveness of methods independent of certain dataset. In this study, a variable selected for 2 or more datasets from the results of each method shown in Tables 2 and 3 is defined as a common variable for the method (CNS common, Prev. common and CFS common). For example, common variables for CFS are "c_ds_ar", "c_inv", "c_sales", "roe", "c_cap_ex", "c_debt_equity", "tie", "c_sa_ta", "rot", "roc" and "npm". Table 5 shows the prediction results of constructed logit models by applying 6 subsets to each dataset.

In Table 5, the value of prediction accuracy of each variable subset is the average of predictions for earnings increase firms and decrease firms. Firstly, it was shown that the prediction accuracy of variables' subset decided by considering one dataset (CNS 2012, Prev. 2013 and CFS 2014) decreased when they are applied to datasets of a different prediction fiscal year. In particular, the subset determined by CFS decreased its prediction accuracy on the dataset for 2012. This indicates that the result of variable selection tends to depend on the training data.

Next, it was shown that subsets that consisted of common variables of each method had smaller effect on the improvement of prediction accuracy than subsets decided by considering one dataset, but their variation of prediction accuracy was small regardless of datasets. In addition, even if methods used common variables for subsets, the prediction accuracy and tendency of each method had similarities. With CNS common it was difficult to obtain high prediction accuracy, but it enhanced the accuracy for all datasets compared with "All". CNS performs variable selection in consideration of interaction among variables. On the other hand, CNS common did not consider the interaction. Thus, we conjecture that CNS common has small benefits for improving accuracy. Prev. common decreased its accuracy in the prediction for 2014. The previous method decided a minimized subset of variables by using the stepwise method. For this reason, we conjecture that the number of variables contained by Prev. common is small and the method tends to be strongly dependent on data. CFS common decreased the accuracy of prediction for 2012. Thus, that is considered as a variable subset that has method characteristics not appropriate for data which has interaction among variables.

These additional experiments suggest that variable selections with data mining techniques have data dependence. However, it showed that it is possible to create subsets that decrease dependence on data and maintain their characteristics simultaneously, by using variables selected for multiple datasets in common.

4. Conclusions

In this study, variable selections with data mining techniques were applied to the construction of an earnings prediction model, in order to verify the effectiveness of a data-driven approach for financial statement analysis. From results applying variable subsets selected by various methods to several datasets, it was shown that variable selection method enhanced the prediction accuracy compared with a model with all variables. However, there were differences in effect of each method, depending on the applied datasets. In addition, this paper did not consider influence of model construction method. Therefore, it is necessary to also verify the effectiveness of this approach.

Methods useful for datasets with interaction among variables, such as CNS and C4.5, are expected to be effective in variable selection for financial statement data. In this analysis, CFS, which is said to have decreased accuracy when there is interaction among variables[5)], obtained much lower accuracy for certain datasets. If stepwise methods do not consider the interaction, they also have possibilities of decreasing prediction accuracy.

In order to achieve our purpose, many issues remain in this paper because this is the first stage for the verification of effectiveness of a data-driven approach. In this study, we constructed linear models with logistic regression. Thus, it was difficult for a subset with variable selection methods like C4.5 to obtain sufficient improvement effect on accurate prediction. In addition, as was also pointed out for the study of Ou and Penman (1989)[1)], the grounds have not been shown for the validity of logit model construction for earnings prediction[4)]. In future works, it is necessary to verify the effectiveness of this approach by performing model selections using data mining and machine learning methods such as neural network and naive Bayes. Furthermore, we need to verify the significance and applicability of this study through the analysis of total results obtained by both variable selection and model selection.

Acknowledgements

This work was supported by “Strategic Project to Support the Formation of Research Bases at Private Universities”: Matching Fund Subsidy from MEXT (Ministry of Education, Culture, Sport, Science and Technology), 2014-2018.

References

1. Ou J. A. and Penman S. H. Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics* 11 (4): 295-329, 1989.
2. Holthausen R. W. and Larcker D. F. The prediction of stock returns using financial statement information. *Journal of Accounting and Economics* 15 (2-3): 297-331, 1992.
3. Abarbanell J. S. and Bushee B. J. Abnormal returns to a fundamental analysis strategy. *The Accounting Review* 73-1, pp. 19-45, 1998.
4. Gow I. Fundamental data anomalies. In L. Zacks (Ed.), *The handbook of equity market anomalies* (pp. 117-128), New-Jersey: John Wiley and Sons, Inc., 2011.
5. Hall M. A. and Holmes G. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transaction on Knowledge and Data Engineering*, Vol. 15, No. 6, pp. 1437-1447, 2003.
6. Kira K. and Rendell L. A practical approach to feature selection. *Proc. Ninth Int’l Conf. Machine Learning*, pp. 249-256, 1992.
7. Hall M. Correlation-based feature selection for discrete and numeric class machine learning. *Proc. 17th Int’l Conf. Artificial Intelligence* (*ICML2000*), 2000.
8. Liu H. and Setiono R. A probabilistic approach to feature selection: A filter solution. *Proc. 13th Int’l Conf. Machine Learning*, pp. 319-327, 1996.
9. Quinlan J. R. C4.5: Programs for machine learning. *San Mateo*, *Calif.: Morgan Kaufmann*, 1993.
10. Kononenko I. Estimating attributes: Analysis and extensions of Relief. *Proc. Seventh European Conf. Machine Learning*, pp. 171-182, 1994.
11. Witten I. H. and Frank E. Data mining: Practical machine learning tools and techniques. *Morgan Kaufman*, 2005 (2nd edition).