African Journal of Business Management Vol. 5(29), pp. 11599-11606, 23 November, 2011
Available online at http://www.academicjournals.org/AJBM
DOI: 10.5897/AJBM11.802
ISSN 1993-8233 ©2011 Academic Journals

Full Length Research Paper

Information disclosure prediction using a combined rough set theory and random forests approach

Der-Jang Chi[1] and Ching-Chiang Yeh[2]*

[1]Department of Accounting, Chinese Culture University, Taipei, Taiwan.
[2]Department of Business Administration, National Taipei College of Business, Taipei, Taiwan.

Accepted 17 August, 2011

In recent years, corporate disclosure and transparency analysis has been of interest in the academic and business community. The objective of this study is to increase the accuracy of information disclosure prediction by combining rough set theory (RST) and random forests (RF) technique, while adopting corporate governance as predictive variables. The effectiveness of this methodology has been verified by experiments comparing RF model. The sample is based on 580 Taiwan information technology (IT) firm in 2007. The results show that the proposed model provides better prediction results and corporate governance does provide valuable information in information disclosure prediction model.

Key words: Information disclosure, rough set theory, random forests.

INTRODUCTION

A series of corporate scandals including that of Enron and WorldCom as well as increasing high-profile financial fraud have served to impede corporate growth. Most of the problems were caused by the asymmetric nature of information between the insiders and the outsiders of the firms. As the world's economy has been experiencing severe challenges during the past decade, more and more companies, no matter how large or small, are facing the problem of filing bankruptcy. Thus, accurate information disclosure prediction models are of critical importance in terms of the decision support system for investors and governments, provide timely warnings of a company's real situation.

The various techniques for the evaluation of companies' disclosure of information have attracted a good deal research interest in the academic and business community. However, these are well-established techniques to solve disclosure of information prediction problems, which include two main problems.

Firstly, the most common method consists of calculating a firm-based disclosure level and then running a multivariate linear regression, with various characteristics of the related firm (firm size, leverage, financial performance, etc.) as independent variables. Lang and Lundholm (2003) claim that the disclosure of governance practice can reduce information symmetry and, enable shareholders and investors to effectively monitor management decisions. Hence, the governance characteristic of a corporation is generally acknowledged a key factor to its disclosure level, but it is usually excluded from early studies.

Secondly, the more independent variables the model contains, the more interesting it will be and the easier it will be to find explanations of a particular disclosure behavior. However, the inclusion of too many variables may create a multicollinearity difficulty (Wallace et al., 1994). Several solutions have been proposed, including the regression of several separate models based on a selection of independent variables or a factor analysis of all independent variables (Bah and Dumoniter, 2001). These conventional statistical methods, however, have some restrictive assumptions such as the linearity, normality

*Corresponding author. E-mail: ychinc@webmail.ntcb.edu.tw. Tel: +886-2- 2322-6161. Fax: +886-2- 2322-6405.

Abbreviations: RST, Rough set theory; **RF,** random forests; **OLS,** ordinary least square; **SFI,** Securities and Futures Institute; **TSEC,** Taiwan Stock Exchange Corporation; **ITDRS,** Information Disclosure and Transparency Rankings System; **GTSM,** Gre Tai Securities Market; ROA, return on assets.

and independence of predictor or input variables. Considering that the violation of these assumptions for independent variables frequently occurs within financial data (Breiman, 1996), these methods can have limitations in terms of effectiveness and validity.

Recently, artificially intelligent approaches have proved less vulnerable to these assumptions, such as random forests (RF), a tree-based classification and regression method, developed by the late Breiman and Cutler (2003). Currently, RF is unsurpassable in accuracy in comparison to other current artificially intelligent approaches (Breiman, 2001). RF has been used extensively in different applications, such as modelling (Xu and Jelink, 2007), prediction (Guo et al., 2004; Lariviere and Van Den Poel, 2005), and pattern analysis in multimedia information retrieval, intrusion detection system (Dong et al., 2006; Zhang and Zulkernine, 2006) and machine fault diagnosis (Yang et al., 2008). Unfortunately, to the best of our knowledge, RF has not been applied in the prediction problem of information disclosure. Moreover, there are several arguments which stat that variable selection, also called feature selection, is a fundamental problem that has a significant impact on the prediction accuracy of the models. Many methods have been developed to best prepare for data inputs, such as rough set theory (RST), developed by Pawlak (1991).

Numerous RST-based reduction and feature selection algorithms have been proposed. Consistency of data (Pawlak, 1991), dependency of attributes (Wang et al., 2002), mutual information (Skowron and Rauszer, 1992), discernibility matrix (Wang and Miao, 1998) and genetic algorithm are all employed to find reducts of an information system (Moradi et al., 1998). In addition, these techniques are applied to text classification (Swiniarski and Larry, 2001), face recognition (Liu and Setiono, 1998), texture analysis (Swiniarski and Skowron, 2003), process monitoring (Dubois and Prade, 1992), machinery diagnosis (Wang and Chen, 2008) and model of system (Pavel et al., 2008). An extensive review of RST-based feature selection is given in (Thangavel and Pethalakshmi, 2009).

The objective of this study is to increase the accuracy of information disclosure prediction and propose a novel model to combine RST and RF techniques, called RST+RF, while adopting corporate governance as predictive variables.

Firstly, RST is used to perform variable selection because of its reliability in obtaining the significant independent variables. Secondly, this study will use the obtained significant independent variables from RST as inputs for the RF models. The obtained results can then be compared to see whether the one including corporate governance characteristics will give better classification accuracy or not. The effectiveness of the methodology has been verified by experiments comparing the RF model.

The paper proceeds with the literature review providing an overview of some relevant prior studies that have investigated the information disclosure assessment. This is followed by descriptions of the analytical methods while describing the methods used in the paper: RST and RF, respectively. Our proposed approach is discussed in, followed by an analysis of our results. Finally, conclusions and suggestions are contained in the study.

LITERATURE REVIEW

Information disclosure

Bushman et al. (2004) define corporate transparency as the availability of firm-specific information to outside investors and stakeholders. Furthermore, they argue that the availability of information is critical to resource allocation decisions and economic growth. Apparently, the levels of corporate transparency depend on the levels of corporate disclosure and transparency exhibited by the firm. As a result, corporate disclosure and transparency are the twin cornerstones that protect shareholders' rights. Shareholders should be treated equally, should be able to participate in the decisions affecting the firm, and should be able to elect directors to represent them. Finally, outside investors need to be assured that no individual shareholder (or group of shareholders) receives preferential treatment or has influence greater than their respective share of ownership.

Additionally, shareholders should also be able to exert their influence over the board of directors and hold directors liable for breaches of their fiduciary duty. Only through full and complete disclosure and transparent management practices can shareholders feel confident that the firm to which they have give their funds is being operated with their best interests in mind.

Determinants of corporate disclosure

Prior studies have found that firm characteristics found to be associated with extent of disclosure in Singhvi and Desai (1971) include listing status and earnings margin. Companies that have high financial leverage should have higher degrees of transparency because creditors require them to disclose more information (Khanna et al., 2004). Khanna et al. (2004) find a positive relation between market capitalization and overall transparency scores. It is possible that past performance can affect the degrees of corporate disclosure. Jensen and Meckling (1976) posit that collateral assets can reduce agency conflicts because lenders can take possession of fixed assets in case of bankruptcy. The reduction in agency conflicts may reduce the need to disclose information so it is possible that there is a negative relation between collateral values and the degrees of disclosure. Cheung

et al. (2009) argue companies have to disclose more relevant information to outside investors, which, in turn, leads to high levels of corporate disclosure and transparency for companies with high levels of assets utilization. Lang and Lundholm (2003) find that disclosure is associated with return variability, firm size and need for financing. Skinner (1994) examines earnings-related disclosures. Skinner (1994) finds that large negative earnings surprises are more often pre-empted by corporate disclosures.

Assessment of corporate disclosure

The use of statistical methods for information disclosure prediction can be traced back to 1968, when Singhvi utilized univariate statistical analysis in an attempt to explain the extent of disclosure. Many subsequent studies used ordinary least square (OLS) to predict extent of disclosure (Tian and Chen, 2009; Wallace and Naser, 1995). Lang and Lundholm (2003), a few authors have applied the rank regression in the context of disclosure studies (Tian and Chen, 2009; Wallace and Naser, 1995). Some have even used both procedures (with unranked and ranked data) in order to compare the results (Wallace and Naser, 1995). Independently of the issue of the nature of the relationship between dependent and independent variables, several authors have used the stepwise procedure (Ahmed, 1994; Depoers, 2000; Giner, 1997; Malone et al., 1993; Raffoumier, 1995). Healy and Palepu (2001), Core (2001), Chavent et al. (2006) and a discussion by Tian and Chen (2009) provide a broad overview of the empirical information disclosure literature and introduced the new trends in this area.

The general conclusion from these efforts in information disclosure prediction using statistical methods was that a simple model with a small list of financial variables could explain the extent of disclosure. These statistical models were succinct and were easy to explain.

However, the problem with applying these methods to the information disclosure prediction problem is that the multivariate normality assumptions for independent variables are frequently violated in financial data sets (Core, 2001), which make these methods theoretically invalid for finite samples.

METHODS

Rough sets theory

RST is a machine learning method, which is introduced by Pawlak (1991) in the early 1980s. It has proven to be a powerful tool for uncertainty and is usually applied to data reduction, rule extraction, data mining and granularity computation. Here, we illustrate only the basic ideas of RST that are relevant to contemporary work.

Let I= (U, A) be an information system, where U is the universe, a non-empty finite set of objects. A is a non-empty finite set of attributes. For $\forall a \in A$ determines a function $f_a : U \to V_a$. If $P \subseteq A$, there is an associated equivalence relation:

$$IND(P) = \{(x,y) \in U \times U : f(x,a) = f(y,a), \forall\, a \in P\} \quad (1)$$

The partition of U, generated by IND (P) is denoted U/P. If $f(x,y) \in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]_P$. The indiscernibility relation is the mathematical basis of rough set theory.

Let $X \subseteq U$. The P-lower approximation of *x* (denoted by $P_*(x)$ and the P-upper approximation of *x* (denoted by $P^*(x)$) are defined as follows:

$$P_*(x) = \{x \in U : [x]_P \subseteq X\},$$
$$P^*(x) = \{x \in U : [x]_P \cap X \neq \phi\}. \quad (2)$$

Let $P, Q \subseteq A$ be equivalence relations over U, then the positive, negative and boundary regions can be defined as:

$$POS_P(Q) = \bigcup_{X \in U/Q} P_*(x)$$
$$NEG_P(Q) = U - \bigcup_{X \in U/Q} P^*(x)$$
$$BND_P(Q) = \bigcup_{X \in U/Q} P^*(x) - \bigcup_{X \in U/Q} P_*(x) \quad (3)$$

The positive region of the partition U/Q with respect to P, $POS_P(Q)$ is the set of all objects of U that can be certainly classified to blocks of the partition U/Q by means of P. A set is rough (imprecise) if it has a non-empty boundary region.

An important issue in data analysis is discovering dependencies between attributes. Dependency can be defined in the following way. For *P*, $Q \subseteq A$, P depends totally on Q, if and only if $IND(P) \subseteq IND(Q)$. That means that the partition generated by P is finer than the partition generated by Q. We say that Q depends on P in a degree k (0 ≤ k ≤ 1), denoted $P \Rightarrow_k Q$, if

$$k = \gamma_P(Q) = |POS_P(Q)| / |U| \quad (4)$$

If k=1, Q depends totally on P, if 0<k<1, Q depends partially on P, and if k=0 then Q does not depend on P.

In other words, Q depends totally (partially) on P, if all (some) objects of the universe U can be certainly classified to blocks of the partition U/Q, employing P.

In a decision system, the attribute set contains the condition attribute set C and decision attribute set D, that is $A = C \cup D.$ The degree of dependency between condition and decision attributes, $\gamma_C(D)$ is called the quality of approximation of classification, induced by the set of decision attributes (Pawlak, 1991).

The goal of attribute reduction is to remove redundant attributes so that the reduced set provides the same quality of classification as the original. A reduct is define as a subset *R* of the conditional attribute set C such that $\gamma_P(D) = \gamma_C(D)$. Any given decision table may have many attribute reducts. The set of all reducts are defined as:

$$\mathrm{Red} = \left\{ R \subseteq C \middle| \gamma_R(D) = \gamma_C(D) \ \forall B \subset R, \gamma_B(D) \neq \gamma_C(D) \right\} \quad (5)$$

In rough set attribute reduction, a reduct with minimal cardinality is searched for. An attempt is made to locate a single element of the minimal reduct set $\mathrm{Red}_{\min} \subset \mathrm{Red}$:

$$\mathrm{Red}_{\min} = \left\{ R \in \mathrm{Red} \middle| \ \vee R' \in \ \mathrm{Red}, \ |R| \ \leq \ |R'| \right\} \quad (6)$$

The intersection of all reducts is called the core, the elements which are those attributes that cannot be eliminated. The core is defined as:

$$Core(C) = \cap \mathrm{Re}\, d \quad (7)$$

Random forests

Recently, there has been a lot of interest in "ensemble learning"-methods that generate many classifiers and aggregate their results. Two well-known methods are boosting (Schapire et al., 1998) and bagging (Breiman, 2003). In boosting, successive classifiers give extra weight to points incorrectly predicted by earlier predictors. In the end, a weighted vote is taken for prediction. In bagging, successive classifiers do not depend on earlier classifiers-each is independently constructed using a bootstrap sample of the data set. In the end, a simple majority vote is taken for prediction. Starting from bagging, Breiman proposed RF (Breiman, 2001), which add an additional layer of randomness. The RF is ensembles of trees (classification or regression trees). However, in addition to constructing each tree using a different bootstrap sample of the data, random forests also change how the trees are constructed. In standard trees, each node is split using the best split amongst all variables. In a RF, each node is split using the best among a subset of predictors (that is, features) randomly chosen at that node. This somewhat counterintuitive strategy turns out to perform very well compared to many other classifiers, including discriminate analysis, support vector machines and neural networks, and is robust against over fitting (Breiman, 2001). In addition, RF has only two hyperparameters (the number of variables in the random subset at each node and the number of trees in the forest), and is usually not very sensitive to their values. Thirdly, RF automatically gives variable ranking. The RF algorithm estimates the importance of a variable by looking at how much prediction error increases when data not in the bootstrap sample (what Breiman calls "out-of-bag" data) for that variable is permuted while all others are left unchanged. The necessary calculations are carried out tree by tree as the RF is constructed.

RESEARCH DESIGN

Proposed approach

The objective of this study is to propose a novel model, RST+RF, to increase the accuracy of the prediction of information disclosure. Firstly, firm characteristics and corporate governance characteristics are used as predictive (independent) variables, and secondly, RST is used to perform variable selection because of its reliability in obtaining the significant independent variables. Next, we use the obtained significant independent variables from RST as inputs of RF models. To test if the corporate governance characteristics are helpful in information disclosure prediction, we take the corporate governance characteristics before and after the information disclosure into consideration. The obtained results can then be compared to see whether the one including corporate governance characteristics will give better classification accuracy or not. Finally, for verifying the applicability of methodology, we also use the RF model as the benchmark.

Table 1. Description of the sample.

Rating	Number	%	Number of samples used in	
			Training set	Testing set
A	106	18.28	85	21
B	355	61.21	273	82
C	119	20.51	106	13
Sum	580	100	464	116

Data set

Information disclosure and transparency rating has a relatively short history in Taiwan. Taiwan's Securities and Futures Institute (SFI), entrusted by the Taiwan Stock Exchange Corporation (TSEC) and the Gre Tai Securities Market (GTSM) (formerly known as the Over-the-Counter Securities Market), recently launched the "Information disclosure and transparency rankings system" (ITDRS) to evaluate the level of transparency for all listed companies in Taiwan since 2003. The ITDRS ratings are A+, A, B, C and C- , with the grade "A+" representing the highest scores. After matching and filtering data with missing values, we obtain a data set of 580 samples with IT firms in 2007. As the number of samples in extreme-rating classes is too small, we combine both A+ and A as A, while C and C- are merged as C. In this classification model, we design an output variable to represent the three classes, namely: A, B and C. We randomly partition the data set into two parts in a proportion of 4:1. The first part is used for training and validation to select optimal parameters for the RF model. The second part is used for testing, as shown in Table 1.

Potential predictor variables

According to the studies of Chen and Jaggi (2001); Ho and Wang (2001); Eng and Mak (2003); Chen and Courtenay (2006) and Chavent et al. (2006), we summarized 19 attributes (including 14 firm characteristics and 5 corporate governance characteristics) which probably have high relationship with transparency are listed in Table 2.

RESULTS AND ANALYSES

Variable selection

To pick out the significant independent variables that are informative and closely related to the corporate condition,

Table 2. Definition and measurement of variables.

Firm characteristics		
Definition	**Measurement**	**Reference**
Debt ratio	Total liability/ total assets	Eng and Mark (2003)
Current assets over current liabilities	Current asset/ current liability	Chen and Courtenay (2006)
Stock return	Change in stock price over the year	Eng and Mark (2003)
Firm size	Evaluated by the total assets	Chen and Jaggi (2000), Chen and Courtenay (2006), Eng and Mark (2003), Ho and Wong (2001)
Dividend yield	Dividends/market share price	Chavent et al. (2006)
Price-earning ratio	Year-end price of ordinary shares/earnings per share	Eng and Mark (2003)
Leverage ratio	Total liability/ total equity	Chen and Jaggi (2000), Chen and Courtenay (2006), Eng and Mark (2003), Khanna et al. (2004)
Asset-in-place	Fixed assets/total assets	Dubois and Prade (1992)
ROE	Return on equity as a percentage	Chavent et al. (2006), Chen and Jaggi (2000), Eng and Mark (2003), Ho and Wong (2001)
ROA	Return on assets	Chen and Courtenay (2006), Eng and Mark (2003)
Market to book value of assets	Market value of firm/book value of total assets	Chen and Courtenay (2006), Eng and Mark (2003)
Market to book value of equity	Market value of ordinary shares/book value of ordinary shareholder's equity	Chen and Jaggi (2000), Eng and Mark (2003)
Growth	The 3-year average growth in total assets prior to year	Chen and Courtenay (2006), Eng and Mark (2003)
Corporate governance characteristics		
Manager holding	CEO holding/outstanding stock	Eng and Mark (2003)
Director holding	Director holding/outstanding stock	Eng and Mark (2003)
Independent director holding	Independent director holding/outstanding stock	Chen and Jaggi (2000), Chen and Courtenay (2006), Eng and Mark (2003), Ho and Wong (2001)
Institutional investor holding	Institutional investor holding/outstanding stock	Eng and Mark (2003)
Block holding	Block holder (more than 10%) holding/outstanding stock	Eng and Mark (2003)

in this study the RST-based application RSES (a collection of algorithms and data structures for rough set computations, developed at the Group of Logic, Institution of Mathematics, University of Warsaw, Poland) and also the genetic reduction algorithm (Komorowski et al., 2003) were used. In order to find the relative importance of the independent variables, the frequencies of occurrence of the independent variables in the reducts generated from the samples are computed. Table 3 gives the independent variables' average frequency of occurrence in reducts generated. As described in Table 3, independent variables with the higher average frequency of occurrence in the reducts generated is debt ratio, stock return, firm size, price-earning ratio, leverage ratio, return on assets (ROA), market to book value of assets, growth, manager holding, director holding, independent director holding and block holding. The remaining independent variables give an average of below 2 reducts and are eliminated in the subsequent experiment. Therefore, the reduced 13 variables (including 9 firm characteristics and 4 corporate governance characteristics) are selected as potential predictor variables. The selected variables are taken as input of the classifier of RF.

Table 3. The independent variables' average frequencies of occurrence in reducts generated.

Firm characteristics	
Definition	**Average**
Debt ratio	2.5
Current assets over current liabilities	1.0
Stock return	4.7
Firm size	3.2
Dividend yield	1.0
Price-earning ratio	4.8
Leverage ratio	3.5
Asset-in-place	1.0
ROE	1.2
ROA	4.6
Market to book value of assets	5.4
Market to book value of equity	1.3
Growth	5.3
Corporate governance characteristics	
Manager holding	4.3
Director holding	3.0
Independent director holding	2.2
Institutional investor holding	1.4
Block holding	2.6

Combined rough set theory (RST) and random forests (RF) approach

After the significant independent variables are picked out, RF classifiers are implemented. We employ the RF available in the R package random Forest (Liaw and Wiener, 2002). This implementation is based on the original FORTRAN code authored by Leo Breiman, the inventor of RF. Following the suggestions of (Breiman, 2003; Liaw and Wiener, 2002) and http://www.stat.berkeley.edu/breiman/RandomForests/, in preliminary tests we found that the performance of RF does not depend much on the actual value of its hyper parameters inside a large interval-as already reported in some studies (Breiman, 2001). To speed up the training, we hence use the preset values of the package random forest: Number of trees (n_{tree}) = 1300 and number of features at each split (m_{try}) = sqrt (total feature number), that is sqrt (13).

To ascertain whether corporate governance characterristics will be helpful in information disclosure predictions, we tested these two possible hybrid models. RST+RF model including both firm characteristics and corporate governance characteristics is model 1, and RST+RF model using only firm characteristics as independent variables is model 2. The results of the confusion matrix using the two obtained models are summarized in Tables 4 and 5 respectively. Table 4 shows the confusion matrix of a full-variable model, RST+RF, while Table 5 shows the confusion matrix consisting of the "only firm characteristics" model. We can observe that the average correct classification rate is 92.24% for model 1, and 83.62% for model 2. From the improved correct classification rate of the model, we can conclude that corporate governance characteristics should be helpful in improving the classification accuracy of the prediction model. In order to evaluate the classification capabilities of the proposed information disclosure models, we also compared the pure RF model. From the results shown in Table 6, we can observe that the average correct classification rate is 84.25% for model 3 and 82.13% for model 4. Again, the improved correct classification rate of the model considering both firm and corporate governance characteristics shows that corporate governance does provide valuable information in information disclosure prediction. By examining the input variables and accuracies of the models, we can provide useful information about the information disclosure process. Firstly, the models including both firm characteristics and corporate governance characteristics provide better classification results than the models only using firm characteristics. The above phenomenon implies that corporate governance does provide valuable information in predicting information disclosure. Secondly, we proposed RST+RF model provides better classification results than the RF model, even when only considering firm characteristics or the model including

Table 4. RST+RF model (model 1) classification results with both firm and corporate governance characteristics.

Actual class	Total cases	Percent correct	Class A	Class B	Class C
A	21	71.43	15	5	1
B	82	91.46	5	75	2
C	13	53.85	3	3	7

Average correct classification rate: 83.62%.

Table 5. RST+RF model (model 2) classification results with only firm characteristics.

Actual class	Total cases	Percent correct	Class A	Class B	Class C
A	21	85.71	18	1	2
B	82	96.34	2	79	1
C	13	76.90	1	2	10

Average correct classification rate: 92.24%.

Table 6. Predictive accuracies of the constructed model.

Model	Average accuracy (%)	
	Training set	Testing set
RST + RF (model 1) classification results with both firm and corporate governance characteristics	91.37	92.24
RST + RF (model 2) classification results with only firm characteristics	82.35	83.62
RF (mode 3) classification results with both firm and corporate governance characteristics	83.12	84.25
RF (model 4) classification results with only firm characteristics	81.47	82.35

RST+RF model is a better alternative since it exhibits the capability of identifying important independent variables, which may provide valuable information for further information disclosure purposes.

CONCLUSIONS AND SUGGESTIONS

The objective of this study is to increase the accuracy of information disclosure prediction by combining RST and RF techniques, while adopting corporate governance characteristics as predictive variables. In order to verify the applicability of this methodology, we also use RF model as the benchmark, and applied to the Taiwan IT industry.

The results of this study can be summarized as follows. Firstly, the corporate governance characteristics are useful ex-ante determinants of information disclosure. The new input variables, that is, non-financial factors are intended to enhance the classification effectiveness of information disclosure. We have done this by emphasizing the links between corporate governance and financial decisions. In particular, we have shown that corporate governance is an important ex-ante indicator of the information disclosure rating. Secondly, the proposed RST+RF model provides better classification results than RF model, even when only considering firm characteristics or the model including both firm and corporate governance characteristics. Hence, the RST+RF model appears to be an efficient alternative for investors and government.

However, several problems are worthy of further research. Firstly, we used corporate governance characteristics as the predictive variable is related to the classification accuracy of the information disclosure prediction. Future research may use potential non-financial factors as part of the firm's overall information disclosure. Secondly, before RF is implemented, the (*mtry*, *ntree*) parameters have to be optimized in order to construct a first-class classifier. Consequently, extracting the optimal parameters is crucial when implementing RF model. Finally, we will continue to compare our proposed feature selection approach with state-of-the-art approaches in the future.

REFERENCES

Ahmed K (1994). The impact of non-financial company characteristics on mandatory disclosure compliance in developing countries: The case of Bangladesh. Int. J. Account., 29(1): 62- 77.

Bah R, Dumontier P (2001). R&D intensity and corporate financial policy: some international evidence. J. Bus. Financ. Account., 28(5-6): 671-692.

Breiman L (1996). Bagging predictors. Mach. Learn., 24: 123-140.

Breiman L (2003). Manual on setting up, using, and understanding Random Forests v4.0. ftp://ftp.stat.berkeley.edu/pub/users/ breiman/.

Breiman L (2001). Random forests. Mach. Learn., 45: 5-32.

Breiman L, Cutler A (2003). Random forests, http://www. berkeley.edu/users/breiman/RandomForests/cc_home.htm, University of California, Berkeley, CA, USA.

Bushman RM, Piotroski JD, Smith AJ (2004). What determines corporate transparency? J. Account. Res., 42: 207-252.

Chavent M, Ding Y, Fu L, Stolowy H, Wang H (2006). Disclosure and determinants studies: an extension using the divisive clustering method (DIV). Eur. Account. Rev., 15(2): 181-218.

Chen CJP, Jaggi B (2000). Association between independent non-executive directors, family control and financial disclosures in Hong Kong. J. Account., Pub. Pol., 19: 285-310.

Cheng ECM, Courtenay SM (2006). Board composition, regulatory regime and voluntary disclosure. Int. J. Account., 41: 262-289.

Cheung SY, Connely JT, Limpaphayom P, Zhou L (2009). Determinants of corporate disclosure and transparency: evidence from Hong Kong and Thailand. 2006 China Int. Conf. Financ. Available at http://www.ccfr.org.cn/cicf2006/enrc.php.

Core JE (2001). A review of the empirical disclosure literature: Discussion. J. Account. Econ., 31: 441-456.

Deakin EB (1972). A discriminant analysis of predictors of business failure. J. Account. Res., 10: 167-179.

Depoers F (2000). A cost benefit study of voluntary disclosure: Some empirical evidence from French listed companies. Eur. Account. Rev., 9(2): 245-263.

Dong SK, Sang ML, Jong SP (2006). Building lightweight intrusion detection system based on random forest. LNCS 3973. Berlin Heidelberg: Springer-Verlag, pp. 224-230.

Dubois D, Prade H (1992). Putting fuzzy sets and rough sets together, intelligent decision support: Handbook of Applications and Advances of the Rough Sets Theory. Slowiniski, R. (Ed.). Boston: Kluwer Academic Publishers, pp. 203–232.

Eng LL, Mak YT (2003). Corporate governance and voluntary disclosure. J. Account. Public Pol. 22: 325-345.

Giner B (1997). The influence of company characteristics and accounting regulation on information disclosed by Spanish firms. Eur. Account. Rev., 16(1): 45-68.

Guo L, Ma Y, Cukic B, Singh H (2004). Robust prediction of fault-proneness by random forests. Proc. 15th Int. Symposium Software Reliability Engineer. (ISSRE'04), France: Brittany. pp. 417-428.

Healy PM, Palepu KG (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. J. Account. Econ., 31: 405-440.

Ho S, Wong KS (2001). A study of the relationship between corporate governance structures and the extent of voluntary disclosure. J. Int. Account. Audit. Tax., 10(2): 139-156.

Jensen MC, Meckling MH (1976). Theory of the firm: managerial behavior, agency costs and ownership structure. J. Financ. Econ., 20: 305-360.

Khanna T, Palepu KG, Srinivasan S (2004). Disclosure practices of foreign companies interacting with U.S. markets. J. Account. Res., 42: 475-508.

Komorowski K, Øhrn A, Skowron A (2003). The ROSETTA rough set software system, In Handbook of Data Mining and Knowledge Discovery. Klösgen W, Zytkow J (eds.), Oxford University Press.

Lang M, Lundholm R (2003). Cross-sectional determinants of analyst ratings of corporate disclosures. J. Account. Res., 31(2): 246-271.

Lariviere B, Van Den Poel D (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Syst. Appl., 29(2): 472-482.

Liaw A, Wiener M (2002). Classification and regression by random forest, R News 2: 18-22.

Liu H, Setiono R (1998). Some issues on scalable feature selection. Expert. Syst. Appl., 15: 333-339.

Malone D, Fries C, Jones T (1993). An empirical investigation of the extent of corporate financial disclosure in the oil and gas industry. J. Account. Audit. Financ., 3(3): 249-273.

Moradi H, Grzymala-Busse JW, Roberts JA (1998). Entropy of English text: experiments with humans and a machine learning system based on rough sets. Inform. Sci., 104(1): 31-47.

Pawlak Z (1997). Rough set approach to knowledge-based decision support. Euro. J. Oper. Res., 99: 48-57.

Pawlak Z (1991). Rough Sets: Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer Academic Publisher.

Pavel J, Jiří K, Miroslava K (2008). Systems modelling on the basis of rough and rough-fuzzy approach. WSEAS Trans. Inform. Sci. Appl., 5(10): 1448-1457.

Raffournier B (1995). The determinants of voluntary financial disclosure by Swiss listed companies. Eur. Account. Rev., 4(2): 261-280.

Schapire RE, Freund Y, Bartlett P, Lee WS (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. Ann. Stat., 26: 1651-1686.

Singhvi SS, Desai HB (1971). An empirical analysis of the quality of corporate financial disclosure. Account. Rev., 46(1): 129-138.

Skinner DJ (1994). Why firms voluntarily disclose bad news. J. Account. Res., 32(1): 38-60.

Swiniarski RW, Larry H (2001). Rough sets as a front end of neural networks texture classifier. Nurocomputating, 36: 85-102.

Swiniarski RW, Skowron A (2003). Rough set methods in feature selection and recognition. Pattern Recognit. Lett., 24(6): 833-849.

Skowron A, Rauszer C (1992). The discernibility matrices and functions in information systems. Intelligent Decision Support: Handbook of Applications and Advances of Rough Set Theory, pp.331-362.

Thangavel K, Pethalakshmi A (2009). Dimensionality reduction based on rough set theory: a review. Appl. Softw. Comput., 9(1): 1-12.

Tian Y, Chen J (2009). Concept of voluntary information disclosure and a review of relevant studies. Int. J. Econ. Financ., 1(2): 55-59.

Wallace RSO, Naser K (1995). Firm-specific determinants of comprehensiveness of mandatory disclosure in the corporate annual reports of firms listed on the stock exchange of Hong Kong. J. Account. Pub. Pol., 14: 311-368.

Wallace RSO, Naser K, Mora A (1994). The relationship between the comprehensiveness of corporate annual reports and firm characteristics in Spain. Account. Bus. Res., 25(97): 41-53.

Wang G, Hu H, Yang D (2002). Decision table reduction based on conditional information entropy. Chinese J. Comput., 25(7): 1-8.

Wang J, Miao DQ (1998). Analysis on attribute reduction strategies of rough set. J. Comput. Sci. Tech., 13(2): 189-193.

Wang H, Chen P (2008). Condition diagnosis of blower system using rough sets and a fuzzy neural network. WSEAS Trans. Bus. Econ., 5(3): 66-71.

Xu P, Jelinek F (2007). Random forests and the data sparseness problem in language modelling. J. Comput. Speech Lang., 21(1): 105-152.

Yang BS, Di X, Han T (2008). Random forests classifier for machine fault diagnosis. J. Mech. Sci. Tech., 22: 1716-1725.

Zhang J, Zulkernine M (2006). A hybrid network intrusion detection technique using random forests. Proc. IEEE 1st Int. Conf. Availability, Reliability and Security (ARES'06).