Neural Comput & Applic (2016) 27:749–760
DOI 10.1007/s00521-015-1893-0

ORIGINAL ARTICLE

A hybrid data mining model in analyzing corporate social responsibility

Ping-Feng Pai[1] · Lei-Chun Chen[1] · Kuo-Ping Lin[2]

Received: 19 May 2013 / Accepted: 20 March 2015 / Published online: 1 April 2015

Abstract Over the past two decades, corporate social responsibility (CSR) has received worldwide attention. Publication of CSR reports has become the trend for domestic and foreign enterprises. In the constantly changing and competitive corporate environment, public attention has come to be focused on how enterprises play the role of corporate citizen, and how they achieve a balance of profitable, environmental and charitable activities. However, most quantitative CSR studies to date have concentrated on traditional statistical approaches. The data mining technique has not been widely explored in this area. Thus, this investigation proposes a hybrid data mining CSFSC model, which stands for the first letters of CFS, SMOTE, FCM, SVMOAO and C5.0, integrating data-preprocessing approaches, a classification method and a rule generation mechanism for analyzing CSR data. The data-preprocessing approaches include correlation-based feature selection (CFS), the synthetic minority over-sampling technique (SMOTE) and the fuzzy c-means (FCM) clustering algorithm. The support vector machine one-against-one (SVMOAO) method was employed as a classifier for performing multiclassification, and the C5.0 decision tree algorithm was utilized to generate rules from the results of the SVMOAO model. In this study, CSR data collected from China's listed firms in 2010 were used to test the performance of the proposed model. The empirical results showed that the designed CSFSC model yields satisfactory classification accuracy, and can provide rules for decision makers. Therefore, the presented CSFSC model is a feasible and effective alternative in analyzing CSR data.

Keywords Support vector machines · Classification · Rule generation · Corporate social responsibility

1 Introduction

The concept of CSR is that enterprises create profit while upholding responsibility in regard to shareholders, employees, communities, the ecological environment and consumers. Many studies found that 80 % of enterprises actively published reports related to CSR activities in their annual reports [4]. CSR has gradually been integrated into organizations and become an important part of enterprise operation. Therefore, engaging in positive CSR activities has become an important way for enterprises to improve their competitiveness. Enterprises not only pursue the maximization of shareholders' profit, but also take into account employees, consumers, the environment and community. Bearing this in mind, for the sake of sustainable management, the incorporation of CSR will become a future development trend of business strategy.

There is a growing trend of related researches on CSR in both industry and academia. As opposed to traditional investment, investors now attempt to apply the concept of CSR to investments, and look forward to obtaining higher interest. While a variety of previous quantitative studies have focused on statistical analysis to explore the effect of CSR on the financial performance of, or its importance to, enterprises, data mining techniques have rarely been

✉ Ping-Feng Pai
paipf@ncnu.edu.tw

[1] Department of Information Management, National Chi Nan University, 1, University Rd., Puli, Nantou 545, Taiwan, ROC

[2] Department of Information Management, Lunghwa University of Science and Technology, 300, Sec. 1, Wanshou Rd., Guishan, Taoyuan 33306, Taiwan, ROC

applied to the analysis of CSR issues. In recent years, a wave of CSR has also arisen in China. According to Zairi and Peters' [56] report, when enterprises fulfill their social responsibility, this will have a positive impact on their corporate image and help to maintain their competitive advantage. The emergence and development of CSR in China are especially important [50]. Thus, the RANKINS CSR RATINGS (RKS, http://www.rksratings.com/), which were collected by a third-party rating agency in China, were employed for the study in this investigation.

According to the broad definition by the World Business Council for Sustainability and Development, CSR is the ethical behavior of enterprises and its effect on society. Business Social Responsibility states that CSR signifies an enterprise upholding ethical values and respecting human rights, communities and the natural environment. Moreover, Weber [55] believed that CSR could enhance the corporate image and reputation, and help organizations to improve sales and the positive attitude of staff, while saving costs and improving financial performance. However, as CSR may depend on differences in culture or industries, the practical applications of CSR all over the world are different. Bowen [5] stated that the obligations of corporations must be established in order for business activities to conform to societal expectations. However, over time, several studies have offered alternative definitions of CSR. Friedman [17] argued that public activities for solving social problems are not duties of corporations, but the prerogative and obligation of governments. Additionally, Davis and Blomstrom [13] indicated that in pursuing the shareholders' profit, decision makers have to act to protect and promote social well-being. Nevertheless, the most widely accepted definition is Carroll's [7]; it stated that CSR includes four levels: economic, legal, ethical and discretionary. Freeman [16] stated that CSR was the performance of a company not only to satisfy shareholders as a standard, but also to offer value to customers, employees, suppliers, communities and society. Michael Porter [43] considered that combining social responsibility with operating strategies was the new method of increasing the competitiveness of a company. Thus, the emergence of CSR helps to improve corporate image, enterprise value and brand value.

Some quantitative studies of CSR are investigated as follows. Huseynov and Klamm [27] used multivariate regression analysis to deal with S&P audit fee data from KLD STATS, and to study tax management and effective tax rates. The results showed that CSR significantly impacted corporate tax avoidance. Kong [34] used regression analysis to explore cumulative abnormal returns on CSR level and to analyze the influences of CSR on investors' behaviors in the food industry in China. The findings revealed that CSR can influence investors' trading behaviors in a short period of time. Garay and Font [19] investigated CSR reasons, practices and impacts in small and medium accommodation enterprises by using correlation analysis. Their results revealed that the main reason for engaging in CSR is altruistic and that a high correlation exists between CSR and corporate financial performance (CFP). Galbreath and Shum [18] employed confirmatory factor analysis, principal component analysis and factor analysis to study the relationship between CSR and firm performance. The results indicated that CSR is related to reputation and customer satisfaction and that reputation mediates the CSR and firm performance relationship. Inoue and Lee [30] applied the ordinary least-squares regression technique to investigate the effects of five dimensions (community, diversity, employees, natural environment and product) of CSR on CFP in tourism-related industries from the KLD STATS database. The experiment results showed that the four tourism-related industries (airline, casino, hotel and restaurant industries) could improve financial performance through each CSR dimension. Smith and Langford [47] employed confirmatory factor analysis to explore the impact between CSR, employee engagement and traditional human resource practices. They reported that CSR and employee engagement have a significant positive correlation, but that there was no obvious correlation between CSR and traditional human resource practices. Rahim et al. [45] applied multiple linear regression analysis to investigate the influence of CSR on consumer behavior in Malaysia. The results showed an obvious positive relationship between CSR and the buying behavior of consumers. Ghoul et al. [22] used descriptive statistics and the Pearson pairwise correlation to study whether CSR affected the cost of capital. They reported that the better the CSR scores of firms, the cheaper the equity finance they have. Chiu and Sharfman [10] employed correlation analysis, Kruskal–Wallis analysis and hierarchical regression analysis to examine the perspectives of legitimacy, visibility and the antecedents of corporate social performance (CSP). The results indicated that more profitable enterprises may not actively engage in CSP unless they are subject to a more rigorous review by various corporation stakeholders. Usunier et al. [51] applied confirmatory factor analysis, multivariate analysis of covariance and hierarchical regression analysis to study CSR compatibility for multinational corporations. They found that for multinational corporations with great power gaps between countries, loose corporate governance and complete business training, social responsibility is considered to be relatively incompatible with economic responsibility. Choi et al. [11] employed cross-sectional regression analysis and two-stage least-squares regression analysis to investigate the relationship between CSR and CFP in Korea. The results showed that stakeholder-weighted CSR index and CFP

have a significantly positive correlation, but the equal-weighted CSR index does not. Tang and Li [50] employed factor analysis to analyze CSR communication of Chinese and global corporations in China. They reported that companies usually take one of the following three major approaches in their CSR communication: CSR as *ad hoc* public philanthropy, CSR as strategic philanthropy and CSR as ethical business practices. Baden et al. [1] applied co-relational analysis to study the effect of buyer pressure on suppliers in small- and medium-sized enterprises in the UK to display CSR practices. They found that most respondents agreed that CSR can act as an incentive to make suppliers more responsible, particularly in environmental criteria, and concluded that encouraging small- and medium-sized enterprises to engage in CSR would be an effective strategy in supply chains. Consolandi et al. [12] employed probability distributions and portfolio allocation to explore the rationality and ethical preferences of investors. Their findings indicated that subjects' behavior was based not only on their individual payoffs, but also on the information of the ethical standards. Husted and Allen [28] analyzed the strategy and value creation of CSR among large firms in Spain and found that CSR can be managed for value creation. Siegel and Vitaliano [46] employed descriptive statistics and correlation matrix to study the importance of product type or service to managers when investing in CSR. The results showed that the firms selling durable experience goods or credence services are much more socially responsible than other firms. Maignan and Ferrell [37] used factor analysis to compare how consumers in the USA, France and Germany evaluate CSR toward both society and organizational stakeholders. They reported that there are significant differences between the USA and the two European nations and that guidance was provided to build the image of a responsible organization internationally. McWilliams and Siegel [38] employed descriptive statistics and correlation matrix to estimate the effect of CSR on CSP. The results indicated that CSP, research and design were highly correlated. Moreover, when research and design intensity is included in their estimated model, CSP is shown to have a neutral effect on profitability. Some assumptions of conventional statistical approaches such as linearity, normality and independence among independent variables and predefined functional forms of dependent variables and independent variables restrict applications of statistical techniques in the real world [25].

In this study, several techniques were integrated into the developed CSFSC model in order to analyze CSR data. First, in data preprocessing, CFS [23], SMOTE [8] and FCM clustering algorithm [3] were employed to extract essential attributes, cope with imbalanced data distribution and exclude extreme outliers, respectively. Second, the SVMOAO classifier was used to deal with the multiclassification problem. Third, the C5.0 decision tree algorithm [44] was utilized to yield rules for decision makers. The rest of this study is organized as follows. Section 2 introduces methodologies used in this study. Section 3 describes the CSFCS model. A numerical example and experiment results are discussed in Sect. 4. Finally, conclusions are drawn in Sect. 5.

2 Methodology

2.1 Correlation-based feature selection

Feature selection, as one of the preprocessing steps in data mining, is an important issue in reducing irrelevant attributes. CFS has been applied in various fields, such as fault detection of automobile engines [21], biomedical engineering [24] and the study of diabetes [32]. In this investigation, the CFSBFS method, a combination of CFS and best-first search algorithms, was used to identify important attributes. The CFS algorithm is a heuristic for evaluating the importance of features. This heuristic technique takes account of the usefulness of individual features for predicting class labels. The CFS is expressed as Eq. (1):

$$A_{\mathcal{F}} = \frac{\rho\overline{\mathcal{W}\beta f}}{\sqrt{\rho + \rho(\rho-1)\overline{\mathcal{W}ff}}} \tag{1}$$

where $A_{\mathcal{F}}$ is the merit of a feature subset $\mathcal{F}$ including ρ features; $\overline{\mathcal{W}\beta f}$ is the average feature and class correlation average; and $\overline{\mathcal{W}ff}$ is the average feature to feature correlations, respectively, where $f \in \mathcal{F}$. The heuristic handles irrelevant features, as they are bad predictors of the class. Redundant attributes are eliminated when they are highly correlated with the other features [23].

2.2 Synthetic minority over-sampling technique

Generating synthetic data by operating in the feature space rather than the data space, SMOTE generates real data by introducing synthetic examples, not by replacing minority classes. Furthermore, it can also deal with the over-fitting problem, and the classification performance is much higher than random over-sampling [2, 8]. SMOTE became a commonly used in the classification of imbalanced data and was used in several fields, such as medical image recognition [54] and the UCI database [36, 48]. The SMOTE algorithm is expressed as follows. First, choose the sample O and its nearest neighbors Õ randomly, and take the difference between them. Then, the difference is multiplied by a random number between 0

and 1, and the result is added to the sample. The new sample is defined as:

$$O_{\text{new}} = O + \text{rand}(0,1) \times (\ O - O). \tag{2}$$

2.3 Fuzzy C-means clustering algorithm

FCM applies fuzzy logic to enhance the stability and accuracy of clustering. The FCM method is an unsupervised technique that has been successfully applied in many fields, such as image segmentation [14, 26, 49], medical image recognition [31], knowledge systems in forest fires [29], audio signal classification [42] and semiconductor fabrication [9]. The FCM algorithm applies a membership function to calculate a membership value between 0 and 1 for each sample. Each sample has different membership values for different clusters. A sample belongs to a certain cluster when that sample's membership value to a certain cluster is higher than the others. Through the previous process, the cluster boundary becomes smooth or overlapping. The FCM performs this procedure recursively until all samples are assigned to a cluster. Let $Y = \{y_{1,}y_{2,\ldots,}y_M\}$ denote a dataset containing M samples. The expected number of clustering is $G = \{G_{1,}G_{2,\ldots,}G_C\}$. The aim of FCM is to partition the sample Y to all G clusters using the weighted distance. Let J_F be the objective function. The purpose of FCM is to minimize the following objective function:

$$\text{Min}\, J_F = \sum_{v=1}^{C}\sum_{j=1}^{M} W_{jv}^{q}\left\|y_j - u_v\right\|^2 \tag{3}$$

s.t.

$$\sum_{v=1}^{C} W_{jv} = 1, \quad j = 1,2,\ldots,M \tag{4}$$

where q presents the exponent on each fuzzy membership, u_v is the center of the vth cluster and W_{jv} is the membership degree of y_j in cluster v. Then, the membership degree and the center of the cluster are computed as follows:

$$W_{jv} = 1 \Bigg/ \sum_{k=1}^{C}\left(\frac{y_j - u_v}{y_j - u_i}\right)^{2/(q-1)} \tag{5}$$

and

$$u_v = \frac{\sum_{k=1}^{M} W_{kv}^{q}\cdot y_k}{\sum_{k=1}^{M} W_{kv}^{q}} \tag{6}$$

By finding the center of the cluster and seeking the membership function repeatedly, this procedure stops when the value of the objective function is less than a predetermined value or the procedure reaches the maximum number of iterations.

2.4 Support vector machine one-against-one

Proposed by Vapnik [52], SVM is an efficient method which mainly deals with binary classification. However, in the real-world, multiclass problems are more than two class problems. In this investigation, the support vector machine one-against-one (SVMOAO) method [33] was employed. A number of investigations have applied SVMOAO in many fields, such as machine condition monitoring and fault diagnosis [53], medical image recognition [15] and UCI databases [35]. Generally, SVM is based on structure risk minimization and is utilized to deal with linear and nonlinear problems of binary classification. The goal of SVM is to find an optimal hyperplane that can separate the classes and maintain maximum distance from both classes. If the data are nonlinearly separated, the input vectors are mapped into a high-dimensional vector space. Through the nonlinear mapping, SVM can construct a linear model to estimate a decision function by the largest margin between the hyperplane and the support vectors. The basic form of the SVM classifier is formulated as follows.

Let $\{x_k, y_k\}$, $k = 1,2,\ldots,s$ be the training dataset with input $x_k \in R^m$, and $y_k \in \{\pm 1\}$ represent the corresponding class label. SVM finds the optimal separating hyperplane with the largest margin. For separating two classes, the maximum separating margin is a constrained optimization problem represented by Eqs. (7) and (8):

$$\text{Minimize}\ \frac{1}{2}w^2 + C\sum_{k=1}^{s}\varphi_k \tag{7}$$

$$\text{subject to}\ y_k(wx_k + b) \geq 1 - \varphi_k, \varphi_k \geq 0 \quad \text{for}\, k = 1,2,\ldots,s \tag{8}$$

where w is a weight vector, φ is a slack parameter used for fault tolerance, C is the penalty parameter that controls the trade-off between decision function and wrongly classified testing samples, and b is a bias. The dual problem is transformed into quadratic programming by Lagrange multipliers. The solution to Eqs. (7) and (8) can be expressed as:

$$\text{Maximize}\ \sum_{k=1}^{s}\pi_k - \frac{1}{2}\sum_{k=1}^{s}\sum_{j=1}^{s} y_k y_j \pi_k \pi_j K(x_k, x_j) \tag{9}$$

$$\text{subject to}\ \sum_{k=1}^{s}\pi_k y_k = 0,\ 0 \leq \pi_k \leq C \quad \text{for}\, k = 1,2,\ldots,s \tag{10}$$

where π is a Lagrange multiplier, and $K(x_k, x_j) = \Gamma(x_k)\cdot \Gamma(x_j)$ is the kernel function which is applied to map the input vector space to higher-dimensional vector space from nonlinearly separable data, so that the problem can become

linearly separable. By the previous equations, we can obtain the location of support vectors, and the hyperplane is defined by the support vectors. The decision function can be represented as Eq. (11):

$$D(x_k) = \text{sign}\left[\sum_{k=1}^{s} \pi_k y_k K(x, x_k) + b\right] \quad k = 1, 2, \ldots, s \tag{11}$$

In this investigation, the radial basis function was used as a kernel function and shown as Eq. (12):

$$K(x, x_k) = \exp\left(-x - x_k^2/2\sigma^2\right) \tag{12}$$

However, the above discussion deals with only the binary classification, and not multiclass problems. The SVMOAO method can solve multiclass classification problems. The SVMOAO method constructs $k(k-1)/2$ classifiers, using all the binary pairwise combinations of k classes, and the ith SVM is trained on data from two classes. We deal with the binary classification problem by training data from the ith and the jth classes, expressed as Eqs. (13) to (16):

$$\text{Minimize } \frac{1}{2}\left\|w^{Kj}\right\|^2 + C\sum_{d=1}^{s} \varphi_d^{kj}\left(w^{kj}\right)^T \tag{13}$$

$$\text{Subject to}\left(w^{kj}\right)^T \theta(x_d) + b^{kj} \geq 1 - \varphi_d^{kj} \quad \text{if} \quad y_d = k, \tag{14}$$

$$\left(w^{kj}\right)^T \theta(x_d) + b^{kj} \leq -1 + \varphi_d^{kj} \quad \text{if} \quad y_d = j, \tag{15}$$

$$\varphi_d^{kj} \geq 0, \quad j = 1, \ldots, \zeta \tag{16}$$

The super index T denotes the matrix transpose. We employ the largest vote to predict x in the class. After calculating sign $\left[(w^{ij})^T \theta(x) + b^{kj}\right]$, if x in the ith class, then the kth class increases by one; if not, the jth class increases by one.

2.5 C5.0 decision tree algorithms

In order to find the best attribute with classification ability, the C5.0 decision tree calculates Information Gain and leaves threshold. Then, the dataset is divided into multiple subsets by the best attribute. For each subset, this process is conducted recursively and finds other subsets until all subsets contain only one attribute or the number of samples is smaller than a certain threshold. Moreover, by carrying out the merger, or pruning each leaf of the decision tree, C5.0 algorithm improves the precision of the decision tree classification. Suppose set α contains n samples. The decision attribute D and each sample belong to different classes D_i ($i = 1, 2 \ldots u$). Let r_i be the number of class D_i samples. The Information Entropy is then defined as Eq. (17):

$$E(\alpha) = -\sum_{i=1}^{u} \theta_i \log_2(\theta_i) \tag{17}$$

where $\theta_i = r_i/|\alpha|$ represents the proportion of the number of class D_i samples to the number of all samples. Let attribute f have τ different values which divide the set α into τ subsets. Then, calculate the dataset α under the attribute f as Conditional Information Entropy. The equation is shown as follows:

$$E(\alpha|f) = -\sum_{j=1}^{\tau} \theta_j' \sum_{i=1}^{u} \theta_{ij} \log_2\left(\theta_{ij}\right) \tag{18}$$

where θ_j' denotes the proportion of the sample number with attribute value f_j to the number of all samples. $\theta_{ij} = r_{ij}/\left|\alpha_j\right|$; $\left|\alpha_j\right|$ is the number with attribute value f_j; and r_{ij} is the number of samples with attribute value f_j belonging to class D_i. The next step is to calculate the Information Gain Ratio of attribute f. The functions are illustrated as follows:

$$\text{Gain}(f) = E(\alpha) - E(\alpha|f) \tag{19}$$

$$\text{Split}\,(f) = -\sum_{j=1}^{\tau} \theta_j' \log_2(\theta_j') \tag{20}$$

$$\text{Gain Ratio}\,(f) = \text{Gain}\,(f)/\text{Split}\,(f) \tag{21}$$

After calculating the Information Gain Ratio of all attributes, the C5.0 algorithm selects the attribute with the largest value and creates a new node for each subset. For each new node, the procedure described above is repeated. In order to avoid over-fitting, the decision tree must be pruned by removing the less reliable branches to generate more accurate results. The C5.0 algorithm uses two pruning methods: prepruning [40] and postpruning [6, 39]. The prepruning method splits the tree until the stop threshold of the classification tree is reached. In contrast, the postpruning approach builds a complete tree and determines which branch is to be pruned according to the error rate. Therefore, the prepruning approach is more subjective and is often based on local information, while postpruning is grounded on global information. In this study, the postpruning method is adopted by the C5.0 technique.

3 The proposed CSFSC model

Figure 1 shows the flowchart of the presented CSFSC model. First, the CSR reports of China's listed firms in 2010 from RKS database were collected as input for this study. Then, data-preprocessing procedures, including CFSBFS, SMOTE and FCM, were performed. The CFSBFS approach was used to select important condition attributes of CSR data. After attribute selection, the data

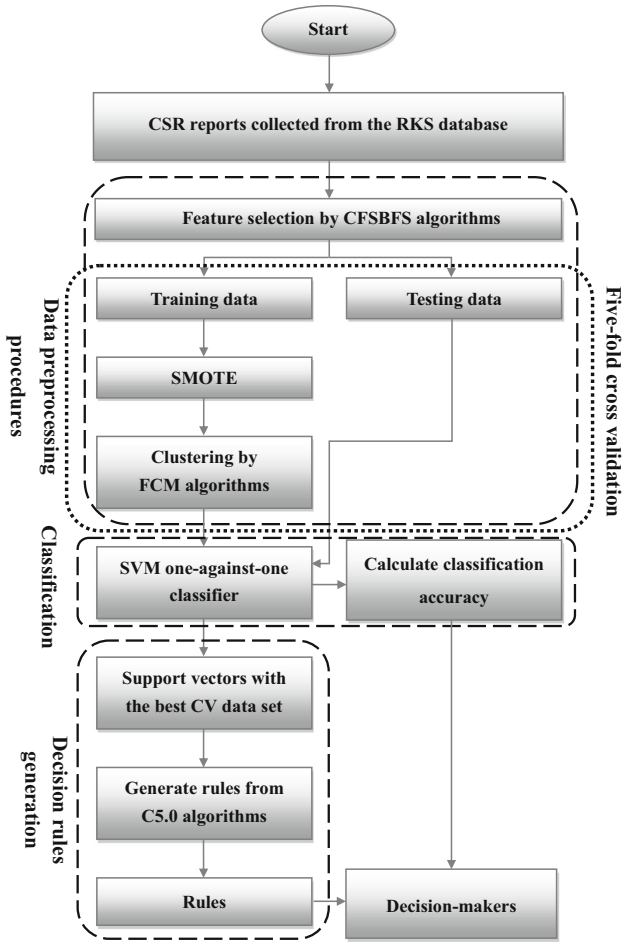

Fig. 1 Flowchart of the CSFSC model

were divided into five datasets to make a fivefold cross-validation (CV). Each dataset was split into two subsets: a training dataset (376 data), and a testing dataset (94 data). The training dataset was only for modeling the CSFSC model; the testing dataset was used to examine the performance of the proposed CSFSC model. Due to the imbalanced data for each category in the training datasets, we oversampled the training data by using the SMOTE approach in order to make the amount of training data of each class very close. Subsequently, extreme outliers were excluded by the FCM clustering algorithm. The preprocessed training data were input to the SVMOAO model to obtain training accuracy, and the testing dataset was employed to measure the performance of the developed CSFSC model. Finally, support vectors with the best CV results were used as input of the C5.0 decision tree to generate rules for decision makers. When the social environment changes, the measures of corporate social responsibility vary, too. Thus, new indices and data of corporate social responsibility can be processed by the proposed CSFSC model to yield useful information.

Table 1 Data distribution

	AAA	AA	A	BBB	BB	B	CCC	CC	C
Amount	0	17	24	27	102	186	91	21	2

4 A numerical example and experiment results

4.1 The data

The data used in this investigation were gathered from Rankins CSR Ratings in 2010. In addition to 40 data with missing values, in total, 470 original data were divided into 9 categories in descending order AAA, AA, A, BBB, BB, B, CCC, CC and C. The distributions of the data are shown in Table 1. The CSR report begins with four major indicators: macrocosm, content, technique and industry. The four major indicators were subdivided into 15 first-level indicators depicted in Table 2. The 15 indicators were further subdivided into 64 s level indicators. Appendix Table 10 lists the second-level indicators. The 64-s level indicators were employed as condition attributes represented as numerical values; and the nine categories in Table 1 are decision attributes.

4.2 Experimental results

To mitigate the influence of over-fitting, a fivefold CV was used in this study. Feature selection was generated using the CFSBFS method, and 19 selected condition attributes are listed in Table 3. The results of classification between the proposed CSFSC model with and without data-preprocessing procedures are illustrated in Table 4. According to Table 4, it can be seen that the CSFSC model with data-preprocessing procedures obtains much higher average testing accuracy than the model without data-preprocessing procedures.

In addition, Pearson's correlation coefficient was employed to verify attributes selected by the CFSBFS feature selection method. The results showed that the selected condition attributes are highly correlated with decision attributes. Nunnally [41] noted that high correlation implies that the correlation coefficient is higher than 0.7. Thus, the 18 condition attributes determined by Pearson's correlation coefficient with absolute values higher than 0.7 are T4, M3, M1, M2, T5, T6, T16, M13, M14, M15, M6, C4, M8, C8, M11, C5, T17 and C7. Of the 18 selected condition attributes, the 11 attributes selected by CFSBFS are T4, M3, M1, M2, M15, M6, C4, M8, C8, M11 and C5. That indicates the CFSBFS method, and the Pearson correlation coefficient technique has common results of feature selection to a certain degree. In order to obtain appropriate support vectors, the dataset of CV4, whose testing accuracy

Table 2 First-level indicators of RKS

1	Strategy	9	Community participation and development
2	Governance	10	Balance in report content
3	Stakeholder	11	Information comparability
4	Economic performance	12	Innovative report
5	Labor and human rights	13	Confidence and transparency
6	Environment	14	Normativeness
7	Fair operating practices	15	The effectiveness and availability of information delivery
8	Consumers		

Table 3 Selected condition attributes by CFSBFS

No.	Attributes	Description
1	M1	Information on whole responsibility strategy
2	M2	Information on sustainable development of adaptation and response
3	M3	Information on responsibility strategy and effective corporate matching,
4	M6	Basic information of company
5	M8	Information on social responsibility management agencies
6	M9	Information on decision-making process and structure
7	M11	Risk management information
8	M15	Information on the communication of stakeholders
9	C3	Basic information of major products or services
10	C4	Information on relationship between employees and employees
11	C5	Information on employee career growth
12	C8	Information on working conditions and social security
13	C9	Information on social interaction and care
14	C11	The overall environmental management information
15	C17	Quality assurance information of products or services
16	C20	Consumer service information
17	C23	Charity donation information
18	T1	Completeness
19	T4	Quantitative data

Table 4 Results of testing accuracy

The CSFSC classifier	Without data-preprocessing procedures Testing accuracy (%)	With data-preprocessing procedures Testing accuracy (%)
CV1	59.38	76.04
CV2	55.21	79.17
CV3	47.92	66.67
CV4	50.53	83.16
CV5	46.32	78.95
Average accuracy	51.87	76.80

is the highest in the CSFSC model, was used to generate rules by the C5.0 decision tree. Appendix Table 11 lists 20 selected rules derived from the CSFSC model.

Furthermore, from the statistical aspect, the analysis of variance (ANOVA) approach was employed to explore the interactions between condition attributes. For instance, two condition attributes, completeness and information on the communication of stakeholders, were investigated. The experimental results of two-way ANOVA, in Table 5, show that the interaction effect between completeness and information on the communication of stakeholders is significant. Moreover, the simple main effect of completeness under "information on the communication of stakeholders is between 2.5 and 3.472" was analyzed, as shown in Table 6. The results indicate that the completeness has significant influence on the CSR level. Furthermore, pairwise comparisons among completeness under "information on the communication of stakeholders is between 2.5 and 3.472" were conducted, and Table 7 shows the results. In Table 8, two rules can be generated from the results in Table 7. Then, Table 9 illustrates rules yielded by the CFSFS model when two condition attributes, completeness and information on the communication of stakeholders, are considered. Both the ANOVA approach and the CFSFS model can generate rules. However, the ANOVA method has to determine condition attributes in advance, and more

Table 5 Results of two-way ANOVA

Source	Sum of squares	*df*	Mean square	*F*
Completeness	32.400	3	10.800	19.278*
Information on the communication of stakeholders	30.417	3	10.139	18.098*
(Information on the communication of stakeholders) × (Completeness)	16.532	7	2.362	4.216*
Error	255.464	456	0.560	
Total	11,071.000	470		

* Significant at the $p < 0.05$ level

Table 6 Simple main effect of completeness under "information on the communication of stakeholders is between 2.5 and 3.472"

Source	Sum of squares	*df*	Mean square	*F*
Completeness	16.028	2	8.014	12.414*
Error	20.658	32	0.646	
Total	1079.000	35		

* Significant at the $p < 0.05$ level

Table 7 Pairwise comparisons among completeness under "information on the communication of stakeholders is between 2.5 and 3.472"

Completeness		Mean of CSR level		Mean difference of CSR level
2	3	3.993	4.600	−1.37*
2	4	3.993	7.000	−1.43*
3	2	4.600	3.993	1.37*
3	4	4.600	7.000	−0.06
4	2	7.000	3.993	1.43*
4	3	7.000	4.600	0.06

* Significant at the $p < 0.05$ level

Table 8 Two rules derived by ANOVA under "information on the communication of stakeholders is between 2.5 and 3.472"

Rule number	Rules
1	The CSR level under "completeness is between 0.731 and 0.952" is better than the CSR level under "completeness is between 0.511 and 0.731"
2	The CSR level under "completeness is between 0.952 and 1.172" is better than the CSR level under "completeness is between 0.511 and 0.731"

Table 9 Rule yielded by the CSFSC model when information on the communication of stakeholders and completeness are considered

If "information on the communication of stakeholders" is "less than or equal to 3.502" and "completeness" is "more than 0.882," then "the CSR level is BBB"

condition attributes may increase the computational complexity. In addition, the CFSFS model can present much clearer and informative rules, but the ANOVA technique can only compare condition attributes.

5 Conclusion

The importance of CSR is that it can help enterprises better coordinate economic, social and environmental development and competition, enhance company culture and employee relationships, maintain friendly relationships with the community, and achieve good risk management and social opportunities. Moreover, it can also increase the competitiveness of corporations. In this investigation, a CSFSC model was presented to analyze the CSR issue in China. This investigation is the first to apply a data mining-based model to analyze CSR data; this study found that data-preprocessing procedures can effectively decrease the complexity of problems and increase classification accuracy. The proposed CSFSC model is more suitable than statistic approaches because it provides clear, efficient and informative rules in analyzing the CSR issue. By adjusting values of CSR measures, the CSFSC model can provide directions for improving CSR levels. In the future, the designed CSFSC model can be used to deal with other CSR databases to examine the feasibility of the CSFSC model. Another possible direction for future study is the creation of a user-friendly interface for decision makers.

Acknowledgments The authors would like to thank Ministry of Science and Technology of the Republic of China, Taiwan, for financially supporting this research under Contract No. NSC 101-2410-H-260-005-MY2, MOST 103-2410-H-260-020 and MOST 103-2410-H-262-010.

Appendix

See Tables 10 and 11.

Table 10 Second-level indicators

Indicators	No.	Description
Macrocosm		
Strategy	M1	Information on whole responsibility strategy
	M2	Information on sustainable development of adaptability and response
	M3	Information on responsibility strategy and valid match of enterprises
	M4	Information on the CSR of corporate executives considered at the strategic level
	M5	Information on target setting and achieving of social responsibility
Governance	M6	Basic information of company
	M7	Information on values, principles and guidelines
	M8	Information on social responsibility management agencies
	M9	Information on decision-making process and structure
	M10	Information on transparency in governance
	M11	Risk management information
	M12	Information on business ethics in governance
	M13	Information on internal practice
Stakeholders	M14	Information on definition and identification of stakeholders
	M15	Information on the communication of stakeholders
	M16	Information on the opinions of stakeholders
Content		
Economic performance	C1	Earnings and returns information
	C2	Year-on-year economic information
	C3	Basic information of major products or services
Labor and human rights	C4	Information on relationship between employees and employees
	C5	Information on employee career growth
	C6	Information on occupational health and safety
	C7	Information on human rights protection
	C8	Information on working conditions and social security
	C9	Information on social interaction and care
	C10	Information on responsibilities in education
Environment	C11	The overall environmental management information
	C12	Pollution prevention information
	C13	Information on sustainable resource use
	C14	Information on mitigation and adaptation of climate change
Fair operating practices	C15	Anti-corruption management information
	C16	Information on promoting social responsibility in the sphere of influence
Consumers	C17	Information on the quality assurance of products or services
	C18	Consumer management information
	C19	Information on the protection of the safety and health of consumers
	C20	Consumer service information
	C21	Information on the protection of consumer data and privacy
	C22	Information on consumer education
Community participation and development	C23	Charity donation information
	C24	Information on volunteer service
	C25	Information on political participation
	C26	Information on job creation
	C27	Information on technological development
	C28	Information on the creation of wealth and income
	C29	Information on promotion of health
	C30	Information on social investment

Table 10 continued

Indicators	No.	Description
Technique		
Balanced in report content	T1	Completeness
	T2	Aproposity
Information comparability	T3	Consistency
	T4	Quantitative data
Innovative report	T5	Innovativeness
	T6	Innovation effectiveness
Confidence and transparency	T7	The extent of disclosure of stakeholders' opinions
	T8	the extent of testing of third parties
	T9	Authority degree of third-party certification agencies
	T10	The effective degrees of feedback mechanism of report reader's comments and suggestions
Normativeness	T11	the normativeness of reporting policies
	T12	The degrees of report standards
	T13	The degrees of report meticulousness
The effectiveness and availability of information delivery	T14	The full extent of language version of report
	T15	The obtainable channel of reports, and whether people with special needs get the report their required way
	T16	The clipart design, typesetting, etc., of report for the degree of enhancing the effect of disclosure
	T17	The degree of chartification and graphicalization of report data and information
Industry	i	Feature index of the food and beverage industry

Table 11 Selected 20 rules derived from the CSFSC model

1. If "basic information of company" is "more than 1.416" and "information on the communication of stakeholders" is "more than 3.502" and "information on working conditions and social security" is "more than 0.952," then "the CSR level is AA"
2. If "basic information of company" is "more than 1.416" and "information on social responsibility management agencies" is "more than 1.250," then "the CSR level is AA"
3. If "basic information of company" is "less than or equal to 1.416" and "information on the communication of stakeholders" is "more than 3.502," then "the CSR level is A"
4. If "information on social responsibility management agencies" is "less than or equal to 1.250" and "information on the communication of stakeholders" is "more than 3.502" and "information on working conditions and social security" is "less than or equal to 0.952," then "the CSR level is A"
5. If "information on the communication of stakeholders" is "less than or equal to 3.502" and "quantitative data" are "more than 0.893," then "the CSR level is A"
6. If "information on whole responsibility strategy" is "more than 1.420" and "basic information of company" is "more than 1.155" and "basic information of major products or services" is "more than 1.898" and "completeness" is "less than or equal to 0.882," then "the CSR level is BBB"
7. If "information on the communication of stakeholders" is "less than or equal to 3.502" and "completeness" is "more than 0.882," then "the CSR level is BBB"
8. If "basic information of company" is "less than or equal to 1.155" and "information on decision-making process and structure" is "more than 1.146" and "completeness" is "less than or equal to 0.882," then "the CSR level is BB"
9. If "information on whole responsibility strategy" is "less than or equal to 1.420" and "basic information of company" is "more than 1.155" and "basic information of major products or services" is "more than 1.898" and "information on employee career growth" is "more than 0.356" and "completeness" is "less than or equal to 0.882," then "the CSR level is BB"
10. If "information on decision-making process and structure" is "less than or equal to 0.962" and "risk management information" is "less than or equal to 0.625" and "completeness" is "more than 0.882," then "the CSR level is BB"
11. If "basic information of company" is "more than 1.155" and "basic information of major products or services" is "less than or equal to 1.898" and "information on working conditions and social security" is "more than 0.510" and "information on working conditions and social security" is "less than or equal to 0.889" and "completeness" is "less than or equal to 0.882" and "quantitative data" are "more than 0.289," then "the CSR level is BB"
12. If "risk management information" is "more than 0.521" and "information on employee career growth" is "more than 0.397" and "completeness" is "less than or equal to 0.882" and "quantitative data" are "more than 0.289," then "the CSR level is BB"

Table 11 continued

13. If "information on whole responsibility strategy" is "less than or equal to 1" and "basic information of company" is "less than or equal to 1.155" and "information on decision-making process and structure" is "less than or equal to 1.146" and "risk management information" is "more than 0.521" and "information on employee career growth" is "more than 0.397" and "information on employee career growth" is "less than or equal to 0.714" and "quality assurance information of products or service" is "less than or equal to 0.926," then "the CSR level is B"
14. If "completeness" is "less than or equal to 0.882," then "the CSR level is B"
15. If "basic information of company" is "more than 0.419" and "information on the communication of stakeholders" is "less than or equal to 2.721" and "information on relationship between employment and employment" is "less than or equal to 0.476" and "completeness" is "more than 0.828" and "quantitative data" are "less than or equal to 0.289," then "the CSR level is CCC"
16. If "information on social responsibility management agencies" is "more than 0.169" and "charity donation information" is "more than 0.143" and "completeness" is "less than or equal to 0.828" and "quantitative data" are "more than 0.146" and "quantitative data" are "less than or equal to 0.289," then "the CSR level is CCC"
17. If "information on whole responsibility strategy" is "less than or equal to 1" and "basic information of company" is "less than or equal to 1.155" and "information on employee career growth" is "less than or equal to 0.397" and "quantitative data" are "more than 0.289," then "the CSR level is CCC"
18. If "information on social responsibility management agencies" is "less than or equal to 0.169" and "basic information of major products or services" is "less than or equal to 1.481" and "completeness" is "less than or equal to 0.828" and "quantitative data" are "more than 0.146" and "quantitative data" are "less than or equal to 0.289," then "the CSR level is CC"
19. If "basic information of company" is "less than or equal to 0.419" and "quantitative data" are "more than 0.146" and "quantitative data" are "less than or equal to 0.289," then "the CSR level is CC"
20. If "quantitative data" are "less than or equal to 0.146," then "the CSR level is C"

References

1. Baden DA, Harwood IA, Woodward DG (2009) The effect of buyer pressure on suppliers in SMEs to demonstrate CSR practices: an added incentive or counter productive? Eur Manag J 27:429–441
2. Batista GEAPA, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl 6:20–29
3. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York
4. Boli J, Hartsuijker D (2001) world culture and transnational corporations: sketch of a project. International Conference on Effects of and Responses to Globalization, Istanbul
5. Bowen HR (1953) Social responsibilities of the businessman. Harper & Row, New York
6. Breiman L, Friedman JH, Olshen RA (1984) Classification and regression trees. Wadsworth International Group, Belmont
7. Carroll AB (1979) A three-dimensional conceptual model of corporate performance. Acad Manag Rev 4:497–505
8. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. Artif Intel Res 16:321–357
9. Chen T (2007) Incorporating fuzzy c-means and a back-propagation network ensemble to job completion time prediction in a semiconductor fabrication factory. Fuzzy Sets Syst 158:2153–2168
10. Chiu SC, Sharfman M (2011) Legitimacy, visibility, and the antecedents of corporate social performance: an investigation of the instrumental perspective. J Manag 37:1558–1585
11. Choi JS, Kwak YM, Choe C (2010) Corporate social responsibility and corporate financial performance: evidence from Korea. Aust J Manag 35:291–311
12. Consolandi C, Innocenti A, Vercelli A (2009) CSR, rationality and the ethical preferences of investors in a laboratory experiment. Res Econ 63:242–252
13. Davis K, Blomstrom RL (1975) Business and society: environment and responsibility, 3rd edn. McGraw-Hill Book Company, New York, p 39
14. Feng J, Jiao LC, Zhang X, Gong M, Sun T (2013) Robust non-local fuzzy c-means algorithm with edge preservation for SAR image segmentation. Sig Process 93:487–499
15. Foroughi H, Rezvanian A, Paziraee A (2008) Robust fall detection using human shape and multi-class support vector machine. In: Sixth Indian conference on computer vision, graphics & image processing, 2008. ICVGIP '08, pp 413–420
16. Freeman RE (1984) Strategic management: a stakeholder approach. Pitman Publishing Inc., Boston
17. Friedman M (1970) The social responsibility of business is to increase its profits. The New York Times Magazine
18. Galbreath J, Shum P (2012) Do customer satisfaction and reputation mediate the CSR–FP link? Evidence from Australia. Aust J Manag 37:211–229
19. Garay L, Font X (2012) Doing good to do well? Corporate social responsibility reasons, practices and impacts in small and medium accommodation enterprises. Int J Hosp Manag 31:329–337
20. Gao M, Hong X, Chen S, Harris CJ (2011) A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems. Neurocomputing 74:3456–3466
21. Ghaderi H, Kabiri P (2012) Fourier transform and correlation-based feature selection for fault detection of automobile engines. In: The 16th CSI international symposium on artificial intelligence and signal processing (AISP 2012), pp 514–519
22. Ghoul SE, Guedhami O, Kwok Chuck CY, Mishra Dev R (2011) Does corporate social responsibility affect the cost of capital? J Bank Finance 35:2388–2406
23. Hall MA (1999) Correlation-based feature subset selection for machine learning, PhD thesis, Department of Computer Science. University of Waikato. Hamilton, New Zealand
24. Hasan A, Adnan Md A (2012) High dimensional microarray data classification using correlation based feature selection. In: 2012 International conference on biomedical engineering (ICoBE), pp 319–321
25. Hua Z, Wang Y, Xu X, Zhang B, Liang L (2007) Predicting corporate financial distress based on integration of support vector machine and logistic regression. Expert Syst Appl 33:434–440
26. Hung WL, Yang MS, Chen DH (2008) Bootstrapping approach to feature-weight selection in fuzzy c-means algorithms with an

application in color image segmentation. Pattern Recogn Lett 29:1317–1325
27. Huseynov F, Klamm BK (2012) Tax avoidance, tax management and corporate social responsibility. J Corp Finance 18:804–827
28. Husted BW, Allen DB (2007) Strategic corporate social responsibility and value creation among large firms: lessons from the Spanish experience. Long Range Plan 40:594–610
29. Iliadis LS, Vangeloudh M, Spartalis S (2010) An intelligent system employing an enhanced fuzzy c-means clustering model: application in the case of forest fires. Comput Electron Agric 70:276–284
30. Inoue Y, Lee S (2011) Effects of different dimensions of corporate social responsibility on corporate financial performance in tourism-related industries. Tour Manag 32:790–804
31. Kaur P, Soni AK, Gosain A (2012) A Robust Kernelized Intuitionistic Fuzzy C-means clustering algorithm in segmentation of noisy medical images. Pattern Recogn Lett 34:163–175
32. Karegowda AG, Jayaram MA (2009) Cascading GA & CFS for feature subset selection in medical data mining. In: 2009 IEEE international advance computing conference (IACC 2009) Patiala, India, pp 6–7
33. Knerr S, Personnaz L, Dreyfus G (1990) Single-layer learning revisited: a stepwise procedure for building and training a neural network. In: Fogelman J (ed) Neurocomputing: algorithms, architectures and applications. Springer, New York
34. Kong D (2012) Does corporate social responsibility matter in the food industry? Evidence from a nature experiment in China. Food Policy 37:323–334
35. Liu B, Hao Z, Tsang ECC (2008) Nesting one-against-one algorithm based on SVMs for pattern classification. IEEE Trans Neural Networks 19:2044–2052
36. Maciejewski T, Stefanowski J (2011) Local neighbourhood extension of SMOTE for mining imbalanced data. In: IEEE symposium on computational intelligence and data mining (CIDM) 2011, pp 104–111
37. Maignan I, Ferrell OC (2003) Nature of corporate responsibilities perspectives from American, French, and German consumers. J Bus Res 56:55–67
38. McWilliams A, Siegel D (2000) Research notes and communications: corporate social responsibility and financial performance: correlation or misspecification? Strat Manag J 21:603–609
39. Muata K, Bryson O (2007) Post-pruning in decision tree induction using multiple performance measures. Comput Oper Res 34:3331–3345
40. Niblett T (1987) Constructing decision trees in noisy domains. In: Proceedings of the second European working session on learning. Sigma Press, Bled, pp 67–78
41. Nunnally JC (1978) Psychometric theory, 2nd edn. McGraw-Hill, New York
42. Park DC (2009) Classification of audio signals using Fuzzy c-Means with divergence-based Kernel. Pattern Recogn Lett 30:794–798
43. Porter ME, Kramer MR (2006) Strategy & society—the link between competitive advantage and corporate social responsibility and environmental management. Harvard Bus Rev 84:78–92
44. Quinlan JR (1993). C4.5: programs for machine learning. Morgan Kaufmann, San Mateo, CA
45. Rahim RA, Jalaludin FW, Tajuddin K (2011) The importance of corporate social responsibility on consumer behaviour in Malaysia. Asian Acad Manag J 16:119–139
46. Siegel DS, Vitaliano DF (2007) An empirical analysis of the strategic use of corporate social responsibility. J Econ Manag Strat 16:773–792
47. Smith V, Langford P (2011) Responsible or redundant? Engaging the workforce through corporate social responsibility. Aust J Manag 36:425–447
48. Taft LM, Evans RS, Shyu CR, Egger MJ, Chawla N, Mitchell JA (2009) Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery. J Biomed Inform 42:356–364
49. Tan KS, Isa NAM (2011) Color image segmentation using histogram thresholding—Fuzzy C-means hybrid approach. Pattern Recogn Lett 44:1–15
50. Tang L, Li H (2009) Corporate social responsibility communication of Chinese and global corporations in China. Public Relat Rev 35:199–212
51. Usunier JC, Furrer O (2011) A. Furrer-Perrinjaquet, The perceived trade-off between corporate social and economic responsibility: a cross-national study. Int J Cross Cult Manag 11:279–302
52. Vapnik VN (1995) The nature of statistical learning theory. Springer, London
53. Wang TY, Chiang HM (2009) One-against-one fuzzy support vector machine classifier: an approach to text categorization. Expert Syst Appl 36:10030–10034
54. Wang Y, Simon M, Bonde P, Harris BU, Teuteberg JJ, Kormos RL, Antaki JF (2012) Prognosis of right ventricular failure in patients with left ventricular assist device based on decision tree with SMOTE. IEEE Trans Inf Technol Biomed 16:383–390
55. Weber M (2008) The business case for corporate social responsibility: a company-level measurement approach for CSR. Eur Manag J 26:247–261
56. Zairi M, Peters J (2002) The impact of social responsibility on business performance. Manag Audit J 17:174–178