**Determinants of Intangible Assets:**

**The Application of Data Mining Technologies**

**Yu-Hsin Lu**

Assistant Professor of Accounting

Feng Chia University

Taiwan, R.O.C.

Email: lomahsin@gmail.com

Tel: 886-4-24517250#4214

**Determinants of Intangible Assets:**

**The Application of Data Mining Technologies**

**Abstract**

Since there is a lack of regulation and disclosure of intangible capital, it is very difficult for investors and creditors to evaluate a firm’s intangible value before making investment and loan decisions. Therefore, valuation of intangible assets has become a widespread topic of interest in the future of the economy. This paper uses data mining technologies different from traditional statistical methods to analyze and evaluate intangible assets. At first, feature selection methods are employed to find out important features (or factors) affecting intangible assets. Then, numbers of different classification techniques based on the identified important factors are developed and compared in order to find out the optimal intangible assets classification model.

In feature selection process, five feature selection methods are considered. In addition, multi-layer perceptron (MLP) neural networks are used as the baseline classification model in order to understand which features selected from these five methods can allow the classification model to perform the best.

Sequentially, the important and representative factors identified from feature selection are used to develop and compare different types of machine learning based classification techniques in order to identify the optimal classification model for intangible assets. Specifically, five classification algorithms are considered. In addition, classifier ensembles and hybrid classifiers to combine these classification techniques are developed. Consequently, thirty-one classification models are constructed for comparisons, including the six single classifiers, boosting and bagging based classifier ensembles, and the combination of *k*-means clustering, single classifiers and classifier ensembles respectively.

The experimental result shows that combining *k*-means with boosting/bagging based classifier ensembles perform much better than the others in terms of prediction accuracy, Type I and II errors. Specifically, while the best single classifier, *k*-NN provides 78.24% prediction accuracy, *k*-means + bagging based DT ensembles provide the best performance to predict intangible firm value for 91.6% prediction accuracy and 18.65% and 6.34% Type I and II errors respectively.

*Keywords*: Firm value, intangible assets, data mining, feature selection, machine learning, classification technology

**1. Introduction**

Since the knowledge-based economy era has evolved, some important factors in the success of companies are the capability and the efficiency in creation, expansion, and application of knowledge (Kessels, 2001). The primary method for creating firm value transfers from traditional physical production factors to intangible knowledge. In this situation, a large part of a firm's value may reflect its intangible assets. To evaluate the firm’s value, we not only consider tangible assets, but also respect the power of intangible assets (Chan et al., 2001; Eckstein, 2004).

Intangible assets are a firm’s dynamic capability created by core competence and knowledge resources, including organization structure, employee expert skills, employment centripetal force, R&D innovation capability, customer size, recognizable brand, and market share. Recently, with the increased importance of intangible assets value, many studies (Gleason and Klock, 2006; Fukui and Ushijima, 2007) have begun to investigate various types of important factors in intangible assets value. Gleason and Klock (2006) and Black et al. (2006) indicate that advertising and R&D expenditure are positively related to Tobin’s *Q*, a proxy for intangible firm value, but the firm size has a negative relation with Tobin’s *Q.* Fukui and Ushijima (2007) investigate the industry diversification of the largest Japanese manufacturers. Regression results show that the average relationship between diversification and intangible assets value is negative. Regression results show that the average relationship between diversification and intangible assets value is negative. However, research to date (Wiwattanakantang, 2001; Lins, 2003) provides mixed evidence on the various factors affecting intangible assets.

In a knowledge-intensive industry, knowledge and innovation are the most significant resources and are far more important than physical assets (Tzeng and Goo, 2005). Therefore, intangible assets determine a large part of a firm's value. However, financial reporting cannot reflect intangible assets value because of fewer regulations and disclosure in intangible capital. The problem with the traditional financial accounting framework is that reporting lacks the recognition of intangible capital value and creates an information gap between insiders and outsiders (Vergauwen et al., 2007). In order to provide other useful information different from financial statements for investors or creditors when evaluating investment opportunities or loans, and help them make more exact decisions effectively, it is important to find out critical affecting factors of intangible assets and build a more effective and accurate intangible assets value evaluation model.

The traditional approaches to exploring and evaluating the intangible assets or other business issues are based on some statistical methods, such as logistic regression and discriminant analysis. However, related studies in many business domains (Huang et al., 2004; Burez and Ven den Poel, 2007; Coussement and Ven den Poel, 2008) have shown that machine learning techniques or data mining techniques, such as neural networks, support vector machines, etc., are superior to statistical methods. The data mining task can be used to discover interesting patterns or relationships in the data and predict or classify the behavior of the model based on available data. In other words, it is an interdisciplinary field with a general goal of predicting outcomes and employing sophisticated algorithms to discover mostly hidden patterns, associations, anomalies, and structure from extensive data stored in data warehouses or other information repositories, then filter out unnecessary information from large datasets (Han and Kamber, 2006).

Therefore, this study first reviews related literature from diverse domains including accounting, finance, management, and marketing to collect relatively important factors affecting intangible assets. Then, we consider feature selection to select important features (or factors) from a given dataset. In data mining, feature selection is a very important step for obtaining quality mining results (Guyon and Elisseeff, 2003), as it aims to filter out redundant or irrelevant features from the original data (Yang and Olafsson, 2006). The remaining selected features are more representative and have more discriminative power over a given dataset. After feature selection process, these critical are used to construct prediction models.

The consideration of prior studies, which employ data mining techniques to build prediction models, is to identify the single best model for prediction. However, many researches have realized that there exist some limitations on using single classification techniques. This observation has motivated recent studies to utilize combination of multiple classifiers, such as classifier ensembles or hybrid classifiers for better performances of prediction (West et al., 2005; Tsai and Wu, 2008; Nanni and Lumini, 2009). In general, classifier ensembles are based on a combination of multiple classifiers in a parallel manner and hybrid classifiers are based on combining two different machine learning techniques sequentially. For example, clustering is used at first and the clustering result is then used to construct the classifier (Chauhan, et al., 2009; Chandra, et al., 2010; Verikas, et al., 2010).

In literature, each of these two kinds of approaches has shown that they can provide better prediction performances than single techniques in many domains. However, these combinations of multiple classifiers are rarely compared in order to make the final conclusion. In addition, they have not been examined in the domain of predicting intangible assets. Therefore, the second aim of this paper is to develop intangible assets prediction models using single classification, classifier ensembles, and hybrid classifiers techniques, respectively, for a large scale comparison.

The contribution of this paper is two-fold. For investors and creditors, the findings are able to help them better evaluate the investment or lending opportunities, and to make more exact decisions. In addition, from the technical point of view, we can understand whether classifier ensembles or hybrid classifiers perform the best for intangible assets prediction in terms of higher prediction accuracy and lower Type I/II errors.

The remainder of this paper is organized as follows: Section 2 reviews related studies about intangible assets and briefly describes the feature selection and classification techniques used in this paper. Section 3 describes the experimental methodology. Experimental results are present in Section 4. Finally, the conclusion is provided in Section 5.

# 2. Literature Review

## 2.1 Intangible Assets

### *2.1.1 Definition*

The terms applications of knowledge and information technology as key driving forces have mainly triggered dramatic changes in the structure of companies. These changes in conjunction with increased customer demands challenge companies to shift their perspective from tangible to intangible resources. These intangible assets have always played a certain role, and now their systematic handling is seen as being an essential competitiveness factor (Durst and Gueldenberg, 2009).

Intangible assets represent the future growth opportunities and profitability which go toward increasing market-based value of firm. Actually, they have prevailed as a measure of core competency and competitive advantage which explains the gap between the market-based value and book value of an organization at a time of decreasing usefulness of current financial reporting (Han and Han, 2004). Therefore, many researches are interested in describing the structure of intangible assets and trying to define the main component that affects the market value recently. There is no uniformity about this problem in the researchers’ environment, although a certain general understanding of intangible assets composition still exists.

Thus, Stewart (1997) defines intangible assets as knowledge, information, intellectual property, experience that can be put to use to create wealth. Sveiby (1997) determines that intangible assets of a firm consist of internal (patents, administrative system, organizational structure etc.) and external (brands, trademarks, relations with customers and suppliers etc.) organization structures as well as of the competence of its personnel. According to Edvinsson and Malone, (1997), Roos et al., (1997) and Petty and Guthrie (2000) intangible assets of a firm include organizational and human capital (internal and external). In Brooking (1996) the following constituents of intangible assets are distinguished: market assets, intellectual property assets, human centered assets and infrastructure assets.

Intangible assets and tangible assets combined create the firm market value, but the value created by intangible assets in firm is hard to tell with the value created by tangible assets (Cao, 2009) since the financial reporting cannot completely reflect the value of intangible assets because of fewer regulations and disclosure requirements for intangible capital. In a trend that among firms want provide additional information regarding intangible assets on a voluntary basis (Vandemaele et al., 2005; Burgman and Roos, 2007), it is important to find out determinants of intangible assets and then build an intangible assets prediction model for providing other information different from financial statements for investors or creditors.

### *2.1.2 Factors Affect Intangible Assets*

In literature, the factors affecting intangible assets can be classified into six categories: intangible capital, ownership structure, corporate governance, firm characteristics, industry characteristics, and reactions of analysts and customers. They are described as follows.

In intangible capital, many empirical models (Rao et al., 2004; Gleason and Klock, 2006; Fukui and Ushijima, 2007) use the intangible assets value as a forward-looking performance measure. This value represents the market’s valuation of the expected future stream of profits, based on the assessment of the return that can be generated from the firm’s tangible and intangible assets. Therefore, any intangible investment increases a firm’s value as tangible assets would. Innovation and brand loyalty are viewed as investments that can increase a firm’s intangible assets with predictably positive effects on future cash flow and intangible assets (Gleason and Klock, 2006).

Ownership structure of firms in Taiwan an emerging country, unlike the companies in many developed countries (e.g. US, UK, and Japan) are under the common administrative and financial control of a few wealthy old families whose ownership is concentrated in controlling shareholders (Morck and Yeung, 2003; Khanna and Yafeh, 2007). Recently, many studies indicate that the controlling shareholder always obtains effective control of the firm and causes the agency problem between themselves and minority shareholders (Lemmon and Lin, 2003). They extract wealth from the firm by holding high voting rights, but only bear a little cost with holding low cash flow rights. In this situation, they could make decisions for the entrenchment of minority shareholders’ interests that could result in the degradation of intangible assets value. In business groups, the situation of entrenchment is more serious (Morck and Yeung, 2003; Silva, et al. 2006).

When the agency problem arises in companies, which can affect firm intangible assets value, corporate governance may play an important role in monitoring (Lins, 2003). These monitoring mechanisms are usually based on the board of directors (Xie et al., 2003; Larcker et al., 2007) as they are charged with monitoring management to protect shareholders’ interests and avoid intangible assets being entrenched. The empirical evidence on the efficacy of the monitoring that outsiders provide (proxy for board independence) appears in many studies (Oxelheim and Randoy 2003; Xie et al., 2003). Otherwise, large non-management shareholders or institutional shareholders play a role in restraining managerial agency costs (Lins, 2003). If there exists more than one large shareholder in a firm, the large shareholders may monitor each other, hence reducing the agency costs (Wiwattanakantang, 2001).

Otherwise, a firm’s intangible assets value may be affected directly or indirectly by factors related to the nature of the firm. Sales growth is a proxy for growth opportunities that increase intangible assets, but the firm size is likely to be inversely related to expected growth opportunities (Gleason and Klock, 2006; Fukui and Ushijima, 2007). Rao et al. (2004) find that firms with higher growth opportunities have lower leverage. However, previous studies (e.g. McConnell and Servaes, 1990) show that firms with higher leverage can enjoy a tax benefit. They can deduct interest costs, which results in greater cash flow and thus has a positive relationship with intangible assets. Capital intensity also affects intangible assets value, because it is a proxy for investment opportunities.

Besides firm characteristics, difference characteristics of various industries will affect the intangible assets value of firms. The degree of industry concentration should affect the firm’s relative bargaining power. When an industry is fragmented and concentration is low, the degree of competition in the industry is likely to be more intense and the firm’s bargaining power decreased. Therefore, Anderson et al. (2004) indicate that higher concentration can provide more market power that can lead to a higher intangible assets value. On the other hand, Rao et al. (2004) argue that a higher intangible assets reflects better market efficiency rather than market power. The effect of the concentration index on intangible assets value is negative.

Finally, Lang et al., (2003) indicate that more analysts following a company means that more information is available, the firm’s information environment is better, and the cost of capital is reduced. Otherwise, an analyst is one of the outside users of financial statements and owns professional domain knowledge, while additional analyst following should bring about more scrutiny, especially when the agency cost exists. Therefore, to improve intangible assets by increasing the cash flows that accrues to shareholders (Lang et al., 2003), analyst following is important.

Table 1 lists thirty factors (among of them *INDUSTRY* includes 32 industries) affecting intangible assets, which belong to the six categories. These factors are found from diverse domains, such as accounting, finance, management, and marketing. However, they have not been considered all together in order to allow us to understand what factors are important to affect intangible assets in general.

Table 1 The factors affecting intangible assets

|  |  |  |
| --- | --- | --- |
| **Category**  | **Variables** | **Reference** |
| ***Intangible capital*** |
|  | *R&D INTENSITY* | Gleason and Klock (2006), Fukui and Ushijima (2007) , Jo and Harjoto (2011), Boujelben and Fedhila (2011). |
|  | *ADVERTISING INTENSITY* | Gleason and Klock (2006), Fukui and Ushijima (2007), Boujelben and Fedhila (2011) |
| ***Ownership structure*** |
|  | *FAMILY* | Wiwattanakantang (2001) , Jo and Harjoto (2011). |
|  | *GOVERNMENT* | Wiwattanakantang (2001). |
|  | *FOREIGN INVESTOR* | Wiwattanakantang (2001), Oxelheim and Randoy (2003). |
|  | *CASH FLOW RIGHT* | Wiwattanakantang (2001), Claessens et al. (2002). |
|  | *DIVERGENCE* | Claessens et al. (2002). |
|  | *PARTICIPATION IN MANAGEMENT* | Wiwattanakantang (2001), Lins (2003). |
|  | *NONPARTICIPATION IN MANAGEMENT* | Lins (2003). |
|  | *MANAGEMENT OWNERS* | Wiwattanakantang (2001), Lins (2003), Ellili (2011). |
|  | *PYRAMIDS* | Wiwattanakantang (2001), Lins (2003). |
|  | *BUSINESS GROUP* | Wiwattanakantang (2001). |
| ***Corporate governance*** |
|  | *BOARD SIZE* | Oxelheim and Randoy (2003), Xie et al. (2003). |
|  | *BOARD INDEPENDENCE* | Oxelheim and Randoy (2003), Xie et al. (2003), Jo and Harjoto (2011). |
|  | *BLOCKHOLDER* | Lins (2003), Yang (2011), Ellili (2011) , Jo and Harjoto (2011). |
|  | *MULTI CONTROL* | Wiwattanakantang (2001), Lins (2003). |
|  | *FOREIGN LISTING* | Oxelheim and Randoy (2003), Lang et al. (2003). |
| ***Firm characteristics*** |
|  | *SALE GROWTH* | Wiwattanakantang (2001), Fukui and Ushijima (2007). |
|  | *SIZE* | Gleason and Klock (2006), Fukui and Ushijima (2007), Bozec et al. (2010), Jo and Harjoto (2011). |
|  | *LEVERAGE* | Fukui and Ushijima (2007), Bozec et al. (2010), Ellili (2011) , Jo and Harjoto (2011). |
|  | *CAPITAL INTENSITY* | Claessens et al. (2002), Lins (2003). |
|  | *DIVIDEND* | Allayannis and Weston (2001). |
|  | *PROFITABILITY* | Allayannis and Weston (2001), Lang et al. (2003), Rao et al. (2004). |
|  | *AGE* | Wiwattanakantang (2001), Rao et al. (2004). |
|  | *DIVERSIFICATION* | Allayannis and Weston (2001), Fukui and Ushijima (2007) , Jo and Harjoto (2011). |
|  | *EXPORT* | Allayannis and Weston (2001). |
| ***Industry characteristics*** |
|  | *CONCENTRATION* | Anderson et al. (2004), Rao et al. (2004). |
|  | *INDUSTRY* | Oxelheim and Randoy (2003), Lang et al. (2003). |
| ***Reactions of analysts and customers*** |
|  | *ANALYST FOLLOWING* | Lang et al. (2003) , Jo and Harjoto (2011). |
|  | *MARKET SHARE* | Anderson et al. (2004), Morgan and Rego (2009). |

## 2.2 Feature Selection

In data mining, feature selection or dimensionality reduction is one of the most important steps to pre-process data in order to filter out unrepresentative features from a given dataset (Guyon and Elisseeff, 2003, Tsai, 2009). In particular, feature selection is used to find the minimally-sized feature subset that is necessary and sufficient to the target concept. It can also improve prediction accuracy or decrease the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features (Kira and Rendell, 1992; Koller and Sahami, 1996).

In literature, there are many well-known feature selection techniques. This paper shows five popular and commonly used feature selection methods in current literature (Questier et al., 2005; Sugumaran et al., 2007) and considers them for comparisons. They are principle component analysis, stepwise regression, decision trees, association rules, and genetic algorithms and describing as follows.

### *2.2.1 Principal component analysis*

The purpose of principal component analysis (PCA) is to find out the relationship between the large sets of variables and then to identify representative dimensions (i.e. features) that can explain the target or reduce the dimensionality of a data set in which there are a large number of interrelated variables (Canbas et al., 2005; Tsai; 2009). This reduction is achieved by creating an entirely new set of variables (i.e. principle components), much smaller in number, to partially or completely replace the original set of variables. By computing eigenvalues and eigenvectors of the principle components, the original variables are combined in linearity that makes the greatest variance. The first principle component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible (Jolliffe, 1986).

### *2.2.2 Stepwise regression*

Stepwise regression is a common traditional statistical technique used to perform feature selection (Shin and Lee, 2002; Tsai, 2009). To select important variables from a given large set of features, it starts by selecting the best predictor of the dependent variable. Sequentially, additional independent variables are selected in terms of the incremental explanatory power they can add to the regression model. Independent variables are added as long as their partial correlation coefficients are statistically significant. However, they may also be dropped if their predictive power drops to a non-significant level when other independent variables are added to the model. The result is a combination of predictor variables, all of which have significant coefficients.

### *2.2.3 Decision trees*

In previous studies (Questier et al., 2005; Sugumaran et al., 2007), decision trees are the popular method for feature selection. Decision trees are constructed by many nodes and branches on different stages and various conditions. They are multistage decision systems in which classes are sequentially rejected until an accepted class is finally reached. To this end, the critical feature space is split into unique regions, corresponding to the classes, in a sequential manner (Theodoridis and Koutroumbas, 2006).

### *2.2.4 Association rules*

The association rule (AR) is a well-known data mining technique. It is usually adopted to discover the relationship between variables in a database, and each relationship (also known as an association rule) may contain two or more variables. These relationships are found by analyzing the co-occurrences of variables in the database. Therefore, an association rule may be interpreted when the variable *A* (i.e. antecedent) occurs in a database, the variable *B* (i.e. consequent) also occurs. This is defined as an implication of the form *A* => *B*and can be interpreted as *A* and *B* are important variables in some event or situation (Tsai and Chen, 2010).

In addition, two measures are generally used to decide the usefulness of an association rule: *support* and *confidence*. The support of an association rule *A* => *B* is the percentage of *A*∪*B*. The confidence of an association rule *A* => *B* is the ratio of the number of *A*∪*B* to the number of *A*. Support measures how frequently an association rule occurs in the entire set, and confidence measures the reliability of the rule. In AR, rules are selected only if they satisfy both a minimum support and a minimum confidence threshold (Goh and Ang, 2007).

### *2.2.5 Genetic algorithms*

Genetic algorithms (GA) is a general adaptive optimization search methodology based on a direct analogy to Darwinian natural selection and genetics in biological systems. According to the Darwinian principle of ‘survival of the fittest’, GA obtains the optimal solution after a series of iterative computations (Huang and Wang, 2006). It has been investigated recently and is effective in exploring a complex space in an adaptive way, guided by the biological evolution mechanisms of reproduction, crossover, and mutation (Adeli and Hung, 1995; Kim and Han, 2000).

**2.3 Classification Techniques**

In order to construct an effective and accurate model for predicting intangible assets, supervised classification, one of the major data mining techniques, can be applied. In literature, many data mining methods are widely used for many different business domains (Buckinx and Van den Poel, 2005; Coussement and Van den Poel, 2008; Tsai and Wu, 2008).

Development of a classification model is based on creating a function from a given set of training data (Pendharkar and Rodger, 2004). The training data is composed of pairs of input objects and their corresponding outputs (i.e. class labels), respectively. The output of the function can be a continuous value, and can predict a class label of the input object.

*2.3.1 Single Classification Techniques*

There are many classification algorithms available in literature and research (Han and Kamber, 2001). Wu, et. al. (2008) indicates top 10 algorithms in data mining, among them decision trees(e.g. C4.5), *k*-Nearest Neighbor, naïve Bayes, and support vector machinesare used popularly in classification and prediction model. Otherwise, ANN has been applied to numerous classification and forecasting problems (Pendharkar and Rodger, 2004). The following briefly describes these classification techniques.

* *Decision Trees*

A decision tree is a very popular classification approach for many prediction and classification problems. It is constructed by developing many leaf nodes and branches for different stages and various conditions, and it can be used for multistage decision systems in which classes are sequentially rejected until a final accepted class is reached. To this end, the critical feature space is split into unique regions (i.e. leaf node), corresponding to the classes, in a sequential manner (Theodoridis and Koutroumbas, 2006). Specifically, each node represents some attribute of the condition, and each branch corresponds to one of the possible values for this attribute.

* *Artificial Neural Networks*

Artificial Neural Networks (ANN) which attempt to simulate biological neural systems is a class of input-output models capable of learning through a process of trial and error, and collectively constitute a particular class of nonlinear parametric models where learning corresponds to a statistical estimation of model parameters (Li and Tan, 2006). It can be regarded as a black box system, which means that it is not required to understand its internal architecture for the final output decision.

Neural Networks is usually used as a classifier. It is easily recognizable by single-layer perceptron and multilayer perceptron (MLP). Particularly, the multilayer perceptron consists of multiple layers of simple, two taste, sigmoid processing nodes or neurons that interact by using weighted connections. The first or lowest layer of MLP network is an input layer where external information is received. The last or highest is an output layer where the problem solution is obtained. It may contain several intermediary layers between input and output layers. Such intermediary layers are called hidden layers which can connect the input and output layers. Based on prior studies (Zhan et al., 1998; Hung, et al. 2006), multilayer perception is the most influential and relatively accurate neural network model.

* *Naïve Bayes*

Bayesian classification is based on Bayes’ theorem which uses all kinds of beforehand probabilities, and probabilities that are observed in the population to predict posterior probabilities. The naïve Bayesian classifier assumes that the effect of a feature value on a given class is independent of the values of the other features. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered “naïve” (Han and Kamber, 2006). Consequently, the naïve Bayesian classifier is constructed by using the training data to estimate the probability of each class given the features vectors of a new instance.

* *Support Vector Machines*

SVM produces a binary classifier which uses a linear model to implement nonlinear class boundaries through some nonlinear mapping input vectors into a high-dimensional feature space, and the so-called optimal separating hyperplane (OSH) can separate two classes in the new space. In particular, the training points that are closest to the maximum margin hyperplane are called support vectors. All other training examples are irrelevant for determining the binary class boundaries.

However, in general cases where the data is not linearly separated, SVM employs non-linear machines to find a hyperplane that minimizes the number of errors for the training set (Min and Lee, 2005; Shin et al., 2005). Although the training time of even the fastest SVM can be extremely slow, it is highly accurate, leading to the ability to model complex nonlinear decision boundaries. SVMs are much less prone to overfitting than other methods (Han and Kamber, 2006).

* *k-Nearest Neighbor*

The *k*-nearest-neighbor (*k*-NN) method was first described in the early 1950s. It has since been widely used in the area of pattern classification since it is simple and easy to implement. Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar. The training tuples are described by *n* attributes. Each tuple represents a point in an *n*-dimensional space. In this way, all of the training tuples are stored in an *n*-dimensional pattern space. For classification, given an unknown tuple, the *k*-NN classifier searches the pattern space for the *k* training tuples that are closest to the unknown tuple. These *k* training tuples are the *k* “nearest-neighbors” of the unknown tuple (Han and Kamber, 2006).

*2.3.2 Combinations of Multiple Classifiers*

* *Classifier Ensembles*

Recently, many studies have realized that there exist limitations on using single classification techniques. To improve the performance of single classifiers, the combination of multiple classifiers, such as classifier ensembles, has been proposed in the field of machine learning. Research provides the superiority of these approaches with multiple classifiers and features over single classification techniques (West et al., 2005; Tsai and Wu, 2008; Nanni and Lumini, 2009). Specifically, Hayashi and Setiono (2002) indicate increased accuracy diagnosing hepatobiliary disorders from ensembles of 30 MLP neural networks. Hu and Tsoukalas (2003) and Sohn and Lee (2003) tested both bagging and boosting ensembles, two combination methods to combine the outputs of multiple classifiers, and provide a reduction in generalization error for the bagging neural network ensemble.

The main idea of using ensembles is that the combination of multiple classifiers (e.g. neural network, naïve Bayes, and decision trees etc.) can lead to an improvement in the performance of a pattern recognition system in terms of better generalization and/or in terms of increased efficiency and clearer design (Canuto et al., 2007). The advantage of using ensembles lies in the possibility that the different results caused by the variance of input data may be reduced by combining each classifier’s output.

There are two families of multiple classifier combination: serial combination and parallel combination. The parallel combining method is based on combining classifiers in parallel. If an input is given, multiple classifiers classify it concurrently, and then the classification results are integrated by a combination method, such as majority voting, weighted voting, bagging, and boosting, etc. (Nanni and Lumini, 2009). Among them, boosting (i.e. AdaBoost) and bagging are two popular methods.

* *Hybrid Classifiers*

Besides classifier ensembles, to improve the performance of single classifier approaches, hybridization is another approach. Hybrid systems have the potential of addressing more complex tasks because of their combination of different techniques (Hsieh, 2005; Huysmans et al., 2006). Hybrid models are based on combining two or more data mining or machine learning techniques (e.g. clustering and classification techniques). For the first hybrid model, clustering, as the unsupervised learning technique, cannot predict data as accurately the supervised model (i.e. classifier). Therefore, a classifier can be trained at first, and then the data which is distinguished correctly is subsequently used as the input for the cluster to improve the clustering results (Huysmans et al., 2006; Tsai and Chen, 2010). On the other hand, in the second hybrid model, clustering can be used in the pre-processing stage to identify pattern classes for subsequent classification (Hsieh, 2005). Then, the clustering results become the new training set to train and create a prediction model based on some classification technique. In other words, the first component of the hybrid model can simply perform the task of outlier detection or data reduction in order for the second component to develop a prediction model.

**3. Research Methodology**

## 3.1 The Experimental Process in Feature Selection

In feature selection process, there are three stages to complete the experiment. Given a dataset, the first stage is to build a multi-layer perceptron (MLP) neural network as the baseline prediction model. MLP is used because it is the most widely used of the many prediction domains (Huysmans et al., 2006; Tsai and Wu, 2008). In this stage, feature selection is not considered.

For the second stage, the five feature selection techniques mentioned in literature review are used individually to select important features from the original dataset, which results in five different datasets respectively. In addition, we also examine the performance of different combinations of two or three feature selection methods based on the intersection or union of the results from them. Note that regarding our experimental results, there are six different combinations of multiple features selection methods, which provide significantly different features selected and prediction performances. Therefore, each of the different datasets with different numbers of features selected can be used to train and test the MLP model respectively. Finally, the third stage is to evaluate the twelve models’ performance (including the baseline) in terms of prediction accuracy, Type I & II errors, and the feature extraction rate.

### *3.1.1 Variables Measurement*

* *Intangible assets* – ­*Tobin’s Q*

Regarding related literature (Lins, 2003; Fukui and Ushijima, 2007), this paper uses Tobin's *Q* as the proxy for intangible assets. Tobin’s *Q* means the differences between the market value of the firm and the replacement cost of the tangible assets represents the value of intangible assets value. The construction of Tobin’s *Q* involves more complicated issues and choices. The standard definition of *Q* is the market value of all financial claims on the firm divided by the replacement cost of assets (Tobin, 1969). However, there are practical problems associated with implementing this definition because neither of these variables are observable. This study uses a modified approach adopted by Gleason and Klock (2006) as the proxy for *Q*. There is a great deal of research that indicates that this is a good approximation, such as Dadalt et al. (2003) and Gleason and Klock (2006). The modified function is shown as following:

*Q* = (Market value of common stock + Book value of preferred stock) / Book value of total assets (1)

When the Tobin’s *Q* ratio of firm is more than one, it represents that market value of firm is greater than the book value of its assets. Rao et al. (2004) indicate that this excess value reflects an unmeasured source of value attributed to the intangible assets. Therefore, the *Q* is designed as a dummy variable, taking the value of 1 if the Tobin’s *Q* ratio is more than 1 which means a firm owns higher intangible assets, otherwise it is 0. This measurement can classify the firm with (i.e. *Q* = 1) and without (i.e. *Q* = 0) intangibles and help outsiders analyze and evaluate weather invest or lend. Especially, in the investment aspect, there are incentives to invest when *Q* is equal to 1 since securities can be sold for more than the cost of the underlying assets and incentives to disinvest when securities can be purchased cheaper than the assets (Lustgarten and Thomadakis, 1987; Megna and Klock, 1993).

* *Research variables*

All the 30 variable measurements are summarized in Table 2. Among them, *INDUSTRY* includes 32 industries in Taiwan. Totally, there are 61 variables or features which have been found to be representative to affect intangible firm value.

Table 2 The measurement of variables affecting intangible assets

|  |  |  |
| --- | --- | --- |
| **Category**  | **Variables** | **Measurement**  |
| ***Intangible capital*** |
|  | *R&D INTENSITY* | Research and development expenditures to total assets. |
|  | *ADVERTISING INTENSITY* | Advertising expenditures to total assets. |
| ***Ownership structure*** |
|  | *FAMILY* | Dummy variables; indicating if the firm has a controlling shareholder who is an individual or a family. |
|  | *GOVERNMENT* | Dummy variables; indicating if the firm has a controlling shareholder who is government. |
|  | *FOREIGN INVESTOR* | Dummy variables, indicating if the firm has a controlling shareholder who is a foreign investor or a foreign company. |
|  | *CASH FLOW RIGHT* | Cash flow right of controlling shareholders. |
|  | *DIVERGENCE* | Voting rights of controlling shareholders minus cash flow rights. |
|  | *PARTICIPATION IN MANAGEMENT* | Dummy variable; indicating if the controlling shareholder and his family are present among management. |
|  | *NONPARTICIPATION IN MANAGEMENT* | If controlling shareholders are not management the variable is 1; otherwise is 0. |
|  | *MANAGEMENT OWNERS* | Cash flow rights of controlling shareholders who are also management. |
|  | *PYRAMIDS* | Dummy variable; indicating if there exists pyramids ownership structure and/or cross-shareholdings. |
|  | *BUSINESS GROUP* | Dummy variable; taking the value of 1 if the firm belongs to one of the 100 largest business groups in Taiwan. |
| ***Corporate governance*** |
|  | *BOARD SIZE* | The number of directors on the board. |
|  | *BOARD INDEPENDENCE* | The percentage of independent outsider directors. |
|  | *BLOCKHOLDER* | Dummy variable defined that if the percentage of shares of the second largest shareholder is more than 5%. |
|  | *MULTI CONTROL* | Dummy variable, if the firm has more than one controlling shareholder. |
|  | *FOREIGN LISTING* | Dummy variables; identify firms that are listed or traded on one or more foreign exchanges. |
| ***Firm characteristics*** |
|  | *SALE GROWTH* | Growth rate in sales. |
|  | *SIZE* | The log of total assets. |
|  | *LEVERAGE* | The ratio of total debt to total assets. |
|  | *CAPITAL INTENSITY* | The ratio of fixed capital (i.e. property plant and equipment) to total sales. |
|  | *DIVIDEND* | Dummy variable; which equals 1 if the firm paid a dividend in the current year. |
|  | *PROFITABILITY* | The ratio of net income to total assets. |
|  | *AGE* | The years since establishment. |
|  | *DIVERSIFICATION* | The number of subsidiary companies. |
|  | *EXPORT* | The ratio of export sales to total sales. |
| ***Industry characteristics*** |
|  | *CONCENTRATION* | The sum of the squared market shares of the firms in the industry. |
|  | *INDUSTRY* | Dummy variable for four-digit or two-digit industries traded on Taiwan stock exchange or Gretai securities market. Contain thirty two industries.  |
| ***Reactions of analysts and customers*** |
|  | *ANALYST FOLLOWING* | The number of analysts that report estimates for each company. |
|  | *MARKET SHARE* | Firm’s share of total sales by all firms in the same four-digit industries or two-digit industries. |

### *3.1.2 Feature Selection Methods*

Five feature selections for the original dataset are principle component analysis, stepwise, decision trees, association rules, and genetic algorithms. As a result, five different datasets are produced based on the five different feature selection methods respectively. After selecting five different groups of critical features, this paper uses the method of intersection or union by considering the selected features from two or three feature selection methods. This results in other critical feature sets. As a result, these new datasets with different numbers of features are used to train and test the MLP model individually.

### *3.1.3 The Performance Evaluation Model*

In order to compare the performance of the feature selection methods to obtain the highly important factors affecting intangible assets, we use the multi-layer perceptron neural network based on the back-propagation learning algorithm widely used in prior literature as the classification model (Smith and Gupta, 2000; Olafsson et al., 2008). To avoid overtraining, related work constructing MLP as the baseline model to examine different parameter settings in order to obtain the ‘best’ MLP model is necessary. This paper designs five different numbers of hidden nodes and learning epochs, respectively. The numbers of hidden nodes are 8, 12, 16, 24, and 32 and the learning epochs are 50, 100, 200, 300, and 500. As a result, there are twenty- five models constructed for each dataset. For a given dataset, the average of prediction accuracy is used to compare with other MLP models by different feature selection methods. Moreover, the cross-validation method is used, which is able to avoid the variability of samples and minimize any bias effect as shown in Tam and Kiang (1992).

### *3.1.4 Evaluation Methods*

To assess the prediction performance of MLP models, prediction accuracy and Type I/II errors are examined. The rate of prediction accuracy can be obtained by the ratio of correctly predicted data over the given set of testing data. For the error rate, Type I error means that the model classifies the firms with high level intangible assets into the group with low level intangible assets. On the other hand, the Type II error means that the model classifies the firm with low level intangible assets into the group with high level intangible assets.

In addition, the ANOVA analysis is used to analyze the significance level of the prediction performance between these methods including the baseline MLP model. In order to compare these methods and make a more reliable conclusion, this paper only considers the results that have a high level of significance. On the other hand, the time of training and testing classification models with and without feature selection is also considered for comparisons. This measurement can show the efficiency impact of the classification models using different numbers of features.

## 3.2 The Experimental Process in Prediction Models

### *3.2.1 Classification Models Development*

In classification model, there are three different types of intangible assets classification models that are developed and compared in this paper. They are single classification models, two types of classifier ensembles model (by bagging and boosting), and hybrid classifiers model (i.e. cluster + classifier). For the single techniques used to construct these prediction models, five difference techniques are applied individually, which are decision trees, multi-layer perceptron, naïve Bayesian classifier, support vector machines, and *k*-Nearest Neighbor.

* *Classifier Ensembles*

As classifier ensembles are based on combining multiple classifiers, there are two families of combining multiple classifiers: serial combination and parallel combination. In parallel combination used in the paper, system performance depends on the combination of different classification techniques (Kim et al., 2002). Therefore, to build the prediction model based on classifier ensembles, this paper combines different numbers of the same classifier techniques (e.g. decision trees, MLP, and so on), starting at combining two of the same classifiers and increasing one gradually until the prediction accuracy does not arise. In addition, the bagging and boosting combination methods are considered. As a result, there are ten different models developed in this type of method. There are bagging based five classifier ensembles such as decision trees, ANN, naïve Bayes, SVM, and *k*-NN and boosting based five classifier ensembles.

* *Hybrid Classifiers*

In general, there are two combination methods in hybrid classifiers such as a classifier combined with a cluster technique, and a cluster combined with classifier technique, respectively. In the second hybrid model, the cluster technique can simply perform the task of outlier detection or data reduction for the classifier technique to develop a prediction model and increase the performance of the model (Hsieh, 2005; Tsai and Chen, 2010a), which is the aim of the paper.

Therefore, this paper uses *k*-means, which is a popular and efficient cluster algorithm as the first component of the hybrid prediction model for reducing and detecting unrepresentative data in the original dataset (Kuo et al., 2002; Hsieh, 2005). The number of clusters (i.e. the *k* value) is set from 2 to 5 to examine and obtain the best clustering result. In particular, two out of *k* clusters can be well ‘classified’ into the high value and low value groups respectively, that is, these two clusters contain the largest proportions of the high value and low value companies respectively. Next, the clustering result (i.e. the data in the two clusters) is used to train the five different classification techniques individually.

In particular, two types of hybrid classifiers are constructed for comparisons. The first type of hybrid classifier is based on combining *k*-means with each one of the five single classifiers respectively. As a result, there are five different hybrid classifiers constructed. The second one is to combine *k*-means with each of the five classifier ensembles based on bagging and boosting methods respectively. Consequently, in total, ten hybrid classifiers by based classifier ensembles are developed respectively. Moreover, the 10-fold cross-validation is also used when construct these classification models to avoid the variability of samples and minimize any bias effect.

*3.2.2 Evaluation Methods*

To assess the performance of these classification models, prediction accuracy and Type I/II errors are examined. They can be measured by a confusion matrix shown in Table 3.

Table 3 The Confusion Matrix

|  |  |  |
| --- | --- | --- |
| **actual \ predicted**  | **High firm value** | **Low firm value** |
| **High firm value** |  (a) | II (b) |
| **Low firm value** | I (c) |  (d) |

The rate of prediction accuracy can be obtained by the ratio of correctly predicted data over a given testing data.

**Prediction accuracy** =  (2)

Otherwise, in intangible firm assets prediction, the Type I error occurs when the model classifies the firms with high level intangible assets into the group with low level intangible assets. Opposed to the Type I error, the Type II error occurs when the model classifies the firm with low level intangible assets into the group with high level intangible assets.

## 3.3 The Case Dataset

Nowadays, a knowledge economy is prevailing in developed countries and emerging markets, including Taiwan and mainland China. Taiwan and China share the same culture and celebrate the same holidays, and many private enterprises in China are invested by Taiwanese enterprises. For example, Taiwan Semiconductor Manufacturing Co., Ltd. directly invested USD$371,000 in TSMC Shanghai. At the end of 2008, the accumulated investment of Formosa Plastics Corporation in China was USD$398,770. Therefore, in this study, we use the sample firms from manifold industries in Taiwan, excluding regulated utilities and financial institutions due to the unique aspects of their regulatory environments. We hope to take the Taiwan economy as a lesson and gain some insights about their business practice and apply them to Chinese cases.

In order to increase the accessibility of the sample data, this study considers publicly listed companies with December 31 fiscal year-ends and draw from the Taiwan Economic Journal (TEJ) database. The controlling shareholder’s ownership structure data is accessed from corporate governance databases and the financial data is received from the financial database within TEJ. In the experiments, the period of the dataset is from 1996 to 2007 in order to collect large data for more accurate analysis. After excluding some data with missing values, 1,380 companies including 9,020 observations in total are used for the final analysis.

# 4. Experimental Results

## 4.1 Descriptive Statistics of Variables

Table 4 provides the descriptive statistics of variables for the overall samples. The consequent variable, Tobin’s Q, indicates that about two-thirds sample companies own intangible assets. In research variables, the average of R&D intensity is 1.974%, higher than the advertising intensity which is 0.403%. For the ownership structure variables, most of the sample companies are controlled by family members, exist pyramids construct, and most controlling shareholders participate in management since their Q1 value is 1. These results are consistent with the findings from prior literature (La Porta et al. 2002; Morck and Yeung, 2003; Silva et al., 2006).

In terms of corporate governance variables, most companies do not own these monitoring mechanisms, since the medians of board independence, blockholder, multi control, and foreign listing are 0. About 70% companies pay dividend in accordance with dividend variables. The average and median age of samples are about twenty-three and twenty-one years, respectively. The diversification variable indicates that one company has 3.5 subsidiary companies in average. Most of the sample companies do not have any analyst to report and analyze their information.

Table 4 The descriptive statistics of variables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variables\* | Average | St. | Min | Q1 | Median | Q3 | Max |
| *Tobin's Q* | 0.669  | 0.470  | 0  | 0  | 1  | 1  | 1  |
| *R&D INTENSITY* | 1.974  | 3.200  | 0  | 0  | 0.818  | 2.533  | 39.868  |
| *ADVERTISING INTENSITY* | 0.403  | 1.215  | 0  | 0  | 0.023  | 0.265  | 25.002  |
| *FAMILY* | 0.859  | 0.348  | 0  | 1  | 1  | 1  | 1  |
| *GOVERNMENT* | 0.020  | 0.141  | 0  | 0  | 0  | 0  | 1  |
| *FOREIGN INVESTOR* | 0.004  | 0.065  | 0  | 0  | 0  | 0  | 1  |
| *CASH FLOW RIGHT* | 23.769  | 16.897  | 0  | 10.290  | 20.335  | 34.165  | 97.750  |
| *DIVERGENCE* | 5.529  | 9.987  | 0  | 0  | 1.280  | 5.930  | 81.360  |
| *PARTICIPATION IN MANAGEMENT* | 0.738  | 0.439  | 0  | 1  | 1  | 1  | 1  |
| *NONPARTICIPATION IN MANAGEMENT* | 0.262  | 0.439  | 0  | 0  | 0  | 1  | 1  |
| *MANAGEMENT OWNERS* | 3.563  | 5.354  | 0  | 0  | 1.240  | 5.210  | 46.350  |
| *PYRAMIDS* | 0.963  | 0.189  | 0  | 1  | 1  | 1  | 1  |
| *BUSINESS GROUP* | 0.703  | 0.457  | 0  | 0  | 1  | 1  | 1  |
| *BOARD SIZE* | 7.047  | 2.863  | 2  | 5  | 7  | 8  | 27  |
| *BOARD INDEPENDENCE* | 9.367  | 14.756  | 0  | 0  | 0  | 22.222  | 66.667  |
| *BLOCKHOLDER* | 0.277  | 0.447  | 0  | 0  | 0  | 1  | 1  |
| *MULTI CONTROL* | 0.047  | 0.212  | 0  | 0  | 0  | 1  | 1  |
| *FOREIGN LISTING* | 0.044  | 0.205  | 0  | 0  | 0  | 0  | 1  |
| *SALE GROWTH* | 15.270  | 76.790  | -197.400  | -5.403  | 7.345  | 23.845  | 3897.660  |
| *SIZE* | 6.583  | 0.568  | 5.018  | 6.178  | 6.519  | 6.903  | 8.793  |
| *LEVERAGE* | 40.035  | 17.245  | 1.550  | 27.600  | 39.620  | 50.710  | 307.380  |
| *CAPITAL INTENSITY* | 11.203  | 317.175  | -15.377  | 0.673  | 2.349  | 7.329  | 30022.682  |
| *DIVIDEND* | 0.690  | 0.462  | 0  | 0  | 1  | 1  | 1  |
| *PROFITABILITY* | 3.771  | 11.462  | -249.945  | 0.570  | 4.452  | 9.007  | 58.359  |
| *AGE* | 22.967  | 11.758  | 1  | 14  | 21  | 31  | 62  |
| *DIVERSIFICATION* | 3.535  | 3.553  | 0  | 1  | 3  | 5  | 41  |
| *EXPORT* | 59.251  | 1026.070  | 0.000  | 3.952  | 41.623  | 78.362  | 72128.073  |
| *CONCENTRATION* | 1248.915  | 1190.441  | 310.481  | 514.859  | 787.726  | 1571.453  | 9884.513  |
| *ANALYST FOLLOWING* | 0.649  | 0.899  | 0  | 0  | 0  | 1  | 5  |
| *MARKET SHARE* | 3.241  | 7.022  | 0  | 0.277  | 0.925  | 2.955  | 99.419  |

\*The measurements of variables are shown in section Table 2.

## Experimental Results in Feature Selection Process

## *Single Feature Selection Methods*

Table 5 shows the performance of six MLP models using five single feature selection methods, which are decision trees (DT), stepwise regression (STEPWISE), genetic algorithms (GA), association rules (AR), and principal component analysis (PCA), and the baseline which uses no feature selection .

Table 5 Performances of single feature selection methods unit: %

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | No. of selected features | Extraction rate (%) | Avg. Accuracy (%) | Type I error | Type II error | Avg. Time for training & testing |
| DT | 7 | 11.5  | 74.68  | 42.91  | 16.63  | 6 min. |
| STEPWISE | 36 | 59.0  | 74.43  | 43.99  | 16.47  | 1 hr. 5 min. |
| GA | 42 | 68.9  | 74.08  | 45.39  | 16.31  | 1 hr. 25 min. |
| AR | 6 | 9.8  | 73.73  | 43.26  | 17.88  | 5 min. |
| PCA | 17 | 27.9  | 73.66  | 45.09  | 17.09  | 18 min. |
| Baseline | 61 | 100.0  | 73.91  | 45.53  | 16.48  | 2 hr. 52 min. |
| *F* value |  |  | 14.875\*\* | 3.477\* | 8.786\*\* |  |

\* Represents the level of significance is higher than 95% by ANOVA.

\*\* Represents the level of significance is higher than 99% by ANOVA.

Regarding Table 5, all of the three performance measurements contain a high level of significant difference between these six prediction models. In particular, DT performs the best for prediction accuracy and the Type I error. GA performs the best with respect to Type II error.

On the other hand, the results indicate that DT, while producing the highest avg. accuracy, extracts the second least numbers of features (i.e. the second lowest extraction rate). In other words, DT is a good feature selection method that can extract the better informative variables to increase the accuracy of prediction models and decrease the time for training & testing (i.e. about 6 min.). This method improves not only effectiveness but also efficiency successfully.

In short, although all of these six prediction models perform very similar in terms of prediction accuracy and error rates, using a smaller number of features can efficiently construct a prediction model. Therefore, considering both effectiveness and efficiency results, DT outperforms the others.

In considering the ANOVA results to have a high level of significance difference, we can rank these feature selection methods not shown in the paper. DT is the best feature selection method to provide the highest rate of prediction accuracy and the lowest rate of Type I errors. STEPWISE is another method which provides relatively better performances in prediction accuracy and Type II errors. However, since it selects more features (i.e. the extraction rate is 58.10%), the average time for training & testing is larger than DT, AR, and PCA.

## *Multiple Feature Selection Methods*

Table 6 shows the prediction performances of using six multiple feature selection methods. These include the intersection of GA and STEPWISE, the union of DT and PCA, the union of AR, DT, and PCA, the union of AR and DT, the union of AR and PCA, and the intersection of PCA and STEPWISE respectively.

Table 6 Performances of multiple feature selection methods unit: %

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | No. of selected features | Extraction rate (%) | Avg. Accuracy (%) | Type I error | Type II error | Avg. Time for training & testing |
| GA∩STEPWISE | 26 | 42.6  | 75.06  | 43.08  | 15.95  | 38 min. |
| DT∪PCA | 22 | 36.1 | 74.75  | 42.53  | 16.14  | 29 min. |
| AR∪DT∪PCA | 26 | 42.6  | 74.45  | 42.87  | 16.99  | 38 min. |
| AR∪DT | 12 | 19.7  | 74.34  | 44.13  | 16.54  | 12 min. |
| AR∪PCA | 21 | 34.4  | 73.86  | 45.55  | 16.55  | 25 min. |
| PCA∩STEPWISE | 10 | 16.4  | 73.53  | 42.68  | 18.46  | 9 min. |
| Baseline | 61 | 100.0  | 73.91  | 45.53  | 16.48  | 2 hr. 52 min. |
| *F* value |  |  | 23.843\*\* | 5.002\*\* | 20.154\*\* |  |

\*\* Represents the level of significance is higher than 99% by ANOVA.

Regarding Table 6, all three of the performance measurements contain a high level of significant difference between these seven prediction models, which are the six multiple selection methods and the baseline model. In particular, the result indicates that GA∩STEPWISE selects (or extracts) the most features and provides relatively better performances in term of prediction accuracy. For the time to train and test a prediction model, the difference between these models is not significantly different, except the baseline model. This is because the number of selected features is similar.

For average accuracy, GA∩STEPWISE and DT∪PCA are the top two methods to provide the highest rate of prediction accuracy and the lowest rate of Type II errors. Therefore, GA∩STEPWISE is the best multiple feature selection method.

## *Comparisons*

Based on the ANOVA analyses, in single feature selection methods, DT is the only method that provides better average accuracy and is significantly different from the others, including AR, PCA, AR∪PCA, PCA∩STEPWISE, and the baseline model (i.e. the level of significance is higher than 95%). For multiple feature selection methods, the average accuracy of DT∪PCA is better than AR, PCA, GA, AR∪PCA, PCA∩STEPWISE, and the baseline at a high level of significant difference, which is higher than 95% or 99% by ANOVA.

GA∩STEPWISE can produce significantly better prediction accuracy than the others, except DT, DT∪PCA, and AR∪DT∪PCA. To sum up, DT, DT∪PCA, and GA∩STEPWISE are the top three feature selection methods, which allow the prediction model to provide better prediction performances. Table 7 compares these three feature selection methods with the baseline in terms of their effectiveness and efficiency measurements. In addition, Tables 8 to 10 show the ANOVA results for prediction accuracy and the Type I/II errors of these three feature selection methods respectively.

Table 7 Performances of the top three feature selection methods

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | No. of selected features | Extraction rate (%) | Avg. Accuracy (%) | Type I error | Type II error | Avg. Time for training & testing |
| GA∩STEPWISE | 26 | 42.6  | 75.06  | 43.08  | 15.95  | 38 min. |
| DT∪PCA | 22 | 36.1  | 74.75  | 42.53  | 16.14  | 29 min. |
| DT | 7 | 11.5  | 74.68  | 42.91  | 16.63  | 6 min. |
| Baseline | 61 | 100.0  | 73.91  | 45.53  | 16.48  | 2 hr. 52 min. |

Table 8 The ANOVA analysis of average accuracy of the top feature selection methods

(*p* value)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | GA∩STEPWISE | DT∪PCA | DT | Baseline |
| GA∩STEPWISE |  | 0.962 | 0.846 | **0.000** |
| DT∪PCA |  |  | 1.000 | **0.002** |
| DT |  |  |  | **0.008** |
| Baseline |  |  |  |  |

Table 9 The ANOVA analysis of Type I errors of the top feature selection methods (*p* value)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | GA∩STEPWISE | DT∪PCA | DT | Baseline |
| GA∩STEPWISE |  | 1.000 | 1.000 | 0.266 |
| DT∪PCA |  |  | 1.000 | **0.000** |
| DT |  |  |  | **0.074** |
| Baseline |  |  |  |  |

Table 10 The ANOVA analysis of Type II errors of the top feature selection methods (*p* value)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | GA∩STEPWISE | DT∪PCA | DT | Baseline |
| GA∩STEPWISE |  | 1.000 | 0.852 | 0.974 |
| DT∪PCA |  |  | 0.986 | 1.000 |
| DT |  |  |  | 1.000 |
| Baseline |  |  |  |  |

Although the top three feature selection methods do not perform significantly different, they almost all perform better than the baseline at the high level of significant difference for both prediction accuracy and Type I error. This shows that these selected features can be regarded as the most representative features or important factors in the original 61 affecting factors. Table 11 lists all the 37 features extracted from top three feature selection methods, where 1, 2, and 3 in the bracket represent GA∩STEPWISE, DT∪PCA, and DT respectively.

Table 11 The 37 important factors that affect intangible assets

|  |  |
| --- | --- |
| *R&D INTENSITY* (1) | *Textiles Industry* (1) |
| *ADVERTISING INTENSITY* (1, 2) | *Electrical and Cable Industry* (1, 2) |
| *FAMILY* (1) | *Paper and Pulp Industry* (1) |
| *CASH FLOW RIGHT* (1, 2) | *Automobile Industry* (1) |
| *PARTICIPATION IN MANAGEMENT* (2) | *Building Material and Construction Industry* (1) |
| *NONPARTICIPATION IN MANAGEMENT* (2) | *Tourism Industry* (1) |
| *BUSINESS GROUP* (1, 2) | *Trading and Consumers' Goods Industry* (1) |
| *BOARD INDEPENDENCE* (2, 3) | *Oil, Gas and Electricity Industry* (1, 2) |
| *BLOCKHOLDER* (2) | *Electronic Parts/Components* *Industry* (1) |
| *SALE GROWTH* (1) | *Computer and Peripheral Equipment Industry* (2) |
| *SIZE* (1, 2) | *Semiconductor Industry* (1, 2, 3) |
| *LEVERAGE* (1) | *Electronic Equipment Industry* (1) |
| *CAPITAL INTENSITY* (2, 3) | *Communications and Internet Industry* (2) |
| *DIVIDEND* (1) | *Optoelectronic Industry* (1, 2) |
| *PROFITABILITY* (1, 2, 3) | *Other Electronic Industry* (2) |
| *AGE* (1, 2, 3) | *Other Industry* (2) |
| *EXPORT* (1) | *ANALYST FOLLOWING* (2, 3) |
| *CONCENTRATION* (2, 3) | *MARKET SHARE* (1, 2) |
| *Cement Industry* (1) |  |

## *Discussion of critical affecting factors extracted from the paper*

The mass of prior studies (Gleason and Klock, 2006; Fukui and Ushijima, 2007) provide that intangible capital variables, such as R&Dinvestment and advertising expense have statistically significant effects on future cash flow and market-based value. These results show that in a knowledge-based economy, enormous competitive pressure always push the firms to produce innovation products through investing more and more R&D expenditures, and then create larger market share and meet more consumer's demands. Innovative customized products not only satisfy customers’ needs, but also increase customer goodwill and brand loyalty which represents the customer retention. Shapiro and Varian (1999) argue that brand loyalty and customer base are major source of value in an information-driven economy. Therefore, besides R&D expenditures, advertising expense a proxy for customer goodwill or brand loyalty is also a critical factor. It means that innovation and brand loyalty are important factors affecting intangible assets and market-based value of firm in knowledge economy. Similar to majority related literature, the result from Taiwan data also indicates that R&Dintensity and advertising intensity are associated with Tobin’s *Q* as a proxy for intangible assets.

Unlike the companies in some developed countries that have widely dispersed ownership, most companies in developing countries are under single common administrative and financial control of few wealthy old families and their ownership is concentrated in family members. Therefore, the discussion of agency problem from family controlling shareholder always appears in emerging countries researches (e.g. Claessens et al. 2000; La Porta et al. 2002; Morck and Yeung, 2003), although some results of studies are not significant. However, Claessens et al., (2000) provide that about half of the sample firms of Taiwan exists pyramid ownership construct, and 79.8% firms indicate that the controlling shareholder and their respective families are present among management. Otherwise, the firms have a controlling shareholder who is an individual or family member with about 65.6% of sample firm. These results show that the agency problems from controlling shareholder may indeed exist in many firms in Taiwan and influence market-based value of firm. Therefore, among the ten ownership structure variables this paper finds out five variables including family, cash flow right, participation in management, nonparticipation in management, and business group which are all critical variables in ownership concentrated firms meaning that they are more important variables affecting intangible assets and market-based value of firm in Taiwan.

In order to mitigate the agency conflict, corporate governance mechanism actually plays an important role. In this study, we find out the independent outsider directors which proxy for board independence and the second largest shareholder which has none relationship with controlling shareholder are critical monitoring mechanism similar to many prior studies (Oxelheim and Randoy, 2003; Lins, 2003). However, Fan and Wong (2005) indicate that the conventional corporate control systems (e.g. boards of directors and institutions) in developing countries do not have a strong governance function, since they have weaker legal environments. Therefore, in these countries, the outside corporate control system (e.g. auditors) may play a more critical role for corporate governance, and then conventional corporate control systems may not be more important in emerging countries. Indeed, in this paper which uses Taiwan data, the percentage of critical affecting factors in corporate governance variables is less than the other five categories.

It is easy to understand that a firm’s market-based value may be affected directly or indirectly by the nature of the firm. For example, when the firm has higher sales growth ratio it means this firm owns growth opportunities in revenue which increases firm’s market value. A profitable firm triggers expectations among investors of higher cash flow potential and drives firm’s market-based value. Furthermore, there are evidences that higher intangible assets are significantly associated with higher profitability (Rao et al., 2004). Therefore, in firm characteristic variables, the results in this paper indicate that all of variables are important features affecting the firm’s value except diversification.

In industry characteristic variables, the degree of industry concentration and the industry variables are critical affecting features in this paper. These results indicate that firm’s bargaining power is stronger since it is in the high concentration industry the intangible assets are higher (Anderson et al., 2004). Besides, in the knowledge-intensive industry, knowledge and innovation are the dominating resources and are far more important than physical assets (Tseng and Goo, 2005). Therefore, intangible assets determine a large part of a firm's market-based value. Some researches show that in communications industry, the market value is about ten times higher than book value. But in traditional industries, most firms’ Tobin’s *Q* is nearly equal to one or less than one. Especially, in Taiwan, high technology industry is flourishing and intangible assets and market-based value indeed vary between electronic industry and traditional industries. After analyzing the results in the experiment, traditional industries (e.g. Cement Industry, Paper and Pulp Industry) own low-level intangible assets and high technology industries (e. g. Semiconductor Industry) exist high-level intangible assets indeed.

Similar to the prior literature, the result in this paper indicates that more analysts follow means that the firm’s information environment is better and the cost of capital is reduced. Besides, by depending on Analysts professional domain knowledge firm value will improve since the cash flows that accrue to shareholders is increased (Lang et al., 2003). In marketing theory, a firm’s market share within its industry may react to customer satisfaction, bring profitability, and thus affect intangible assets and firm market value. Morgan and Rego (2009) show that market share is positively related to Tobin’s *Q* proxy for market-based value of firms. Similar to prior literature, the market share is also an important variable affecting intangible assets in this paper

According to above findings, we hope provide other information different from financial statements to investors and creditors and help them to make the more correct decisions in investment or lending opportunities.

# 4. 3 Experimental Results in Prediction Models

After employ feature selection stage, this stage uses 26 (including 13 industries) critical affecting factors which are extracted from the feature selection method with the best average accuracy (i.e. GA∩STEPWISE) to construct intangible assets classification models. The three different types of prediction models are compared using five different classification techniques in order to determine the best one, which provides the highest rate of prediction accuracy and lowest rates of Type I and II errors.

## *4.3.1 Single Classifiers*

Table 12 shows the performance of prediction models using five single classifiers, which are multilayer perceptron (MLP), decision trees (DT), Naïve Bayesian (naïve Bayes), support vector machines (SVM), *k*-Nearest Neighbor (*k-*NN), and tow traditional statistic methods- logistic regression (LR), and linear discriminant analysis (LDA), respectively.

Table 12 Prediction performances of single classifiers

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Type I error | Type II error |
| MLP | 75.06% (3) | 43.08% (5) | 15.95% (2) |
| DT | 76.33% (2) | 35.58% (4) | 17.79% (5) |
| Naïve Bayes | 46.20% (7) |  5.13% (1) | 77.84% (7) |
| SVM | 72.22% (5) | 48.09% (6) | 17.75% (4) |
| *k*-NN | 78.24% (1) | 30.45% (2) | 17.47% (3) |
| LR | 74.24% (4) | 52.04% (7) | 12.79% (1) |
| LDA | 72.00% (6) | 31.12% (3) | 25.15% (6) |
| STDEV | 10.98 | 15.58 | 22.96 |
| *t*-value | 17.01\*\*\* | 5.96\*\*\* | 3.04\*\* |

\*\*\* Represents the level of significance is higher than 99% by *t*-test.

\*\* Represents the level of significance is higher than 95% by *t*-test.

Regarding *t*-value in Table 12, all the three performance measurements contain a high (i.e. 99% and 95%, respectively) level of significant difference between the accuracy, Type I and II errors of these five prediction models. In particular, *k*-NN performs the best in terms of prediction accuracy. In addition, it can provide relatively low rates of Type I and II errors. Note that the *k* value of *k*-NN is set by 1 at first, but we found that when the value is 2, minimum error rate is reached. On the other hand, it is interesting that naïve Bayes and LR perform the best in terms of Type I and II errors. However, naïve Bayes performs the worst for the Type II error, which results in the worst prediction model.

In short, the best single classifier is *k*-NN, which provides the highest rate of prediction accuracy and the second lowest rate of Type I and Type II errors. The data mining technologies outperform traditional statistic methods in accuracy and Type I error.

## *4.3.2 Classifier Ensembles*

Table 13 shows the performance of each prediction model by boosting and bagging to construct classifier ensembles, in which the number underlined represents the best performance.

Table 13 Prediction performances of classifier ensembles

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Type I error | Type II error |
| *Boosting* |  |  |  |
| MLP | 77.34% (2) | 35.75% (4) | 16.20% (3) |
| DT | 78.89% (1) | 33.23% (3) | 15.12% (2) |
| Naïve Bayes | 50.93% (5) | 7.95% (1) | 69.38% (5) |
| SVM | 73.40% (4) | 65.63% (5) | 12.29% (1) |
| *k*-NN | 76.96% (3) | 30.85% (2) | 19.18% (4) |
| STDEV | 11.69 | 20.53 | 24.14 |
| *t*-value | 13.69\*\*\* | 3.77\*\* | 2.45\* |
| *Bagging* |  |  |  |
| MLP | 77.16% (3) | 41.15% (4) | 13.80% (2) |
| DT | 79.55% (1) | 32.46% (3) | 14.52% (1) |
| Naïve Bayes | 48.17% (5) | 5.77% (1) | 74.58% (5) |
| SVM | 72.24% (4) | 47.99% (5) | 17.77% (4) |
| *k*-NN | 78.48% (2) | 30.48% (2) | 17.09% (3) |
| STDEV | 13.13 | 16.04 | 26.35 |
| *t*-value | 12.11\*\*\* | 4.40\*\* | 2.34\* |

\*\*\* Represents the level of significance is higher than 99% by *t*-test.

\*\* Represents the level of significance is higher than 95% by *t*-test.

\* Represents the level of significance is higher than 90% by *t*-test.

For classifier ensembles using boosting, all the three performance measurements contain a high level of significant difference between these five prediction models. Particularly, DT ensembles perform the best in terms of prediction accuracy. Similar to single classifiers, naïve Bayes ensembles can provide the lowest rate of Type I error. For the Type II error, SVM ensembles outperform the others. On the other hand, for classifier ensembles using bagging, all three performance measurements also contain a high level of significant difference between these five prediction models. Among them, DT ensembles can provide the highest rate of prediction accuracy and second lowest rate of the Type II error. Again, naïve Bayes ensembles perform the best in terms of the Type I error.

To compare with single classifiers, prediction accuracy of each model using bagging and/or boosting slightly increases ranging from 0.24% to 3.22%. In addition, most of Type I and II errors of these ensembles decreases. This is consistent with the literature that the superiority of these approaches with multiple classifiers is over single classification techniques (Tsai and Wu, 2008; Nanni and Lumini, 2009). Regarding Table 13, DT ensembles using bagging can provide the highest rate of prediction accuracy. For the Type I error, naïve Bayes ensembles using bagging outperform the other ensembles. On the other hand, SVM ensembles using boosting can provide the lowest rate of the Type II error.

## *4.3.3 Hybrid Classifiers*

To construct the hybrid classifiers, *k*-means is used for the first component. We found that *k* = 3 (i.e. 3 clusters) can produce the best clustering result. That is, two clusters can mainly represent high and low intangible assets groups. As a result, these two clusters contain 7,250 observations to train the single classifiers as the second component. Furthermore, regarding the above results, the performance of classifier ensembles is almost better than the performance of single classifiers. Therefore, this paper constructs two types of hybrid prediction models for comparisons, which are hybrid classifiers by single classifiers (i.e. *k*-means + single classifiers) and hybrid classifiers by classifier ensembles (i.e. *k*-means + boosting/bagging based classifier ensembles).

* *Hybrid Classifiers: K-Means + Single Classifiers*

Table 14 shows the performances of combining *k*-means with the five single classifiers respectively. All of the three performance measurements contain a high level of significant difference between these five prediction models.

Table 14 Prediction performances of combining *k*-means with the five single classifiers

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Type I error | Type II error |
| MLP | 89.42% (3) | 16.25% (3) | 9.44% (3) |
| DT | 90.33% (2) | 20.38% (4) | 7.52% (2) |
| Naïve Bayes | 86.94% (5) | 4.79% (1) | 14.72% (5) |
| SVM | 88.58% (4) | 11.88% (2) | 11.33% (4) |
| *k*-NN | 91.34% (1) | 22.77% (5) | 5.83% (1) |
| STDEV | 1.68 | 7.16 | 3.45 |
| *t*-value | 118.71\*\*\* | 4.75\*\*\* | 6.33\*\*\* |

\*\*\* Represents the level of significance is higher than 99% by *t*-test.

The result shows that this type of hybrid classifier can largely improve the performance of single classifiers, including prediction accuracy and Type I and II errors. In particular, these hybrid classifiers can provide above 86% accuracy and below 20% Type I and II errors. It is interesting to note that naïve Bayes improved the most from 46.2% to 86.94% accuracy.

For prediction accuracy, *k*-means + *k*-NN performs the best, and *k*-means + DT stands in second place. These two hybrid classifiers can provide over 90% accuracy. For prediction errors, *k*-means + naïve Bayes and *k*-means + *k*-NN can provide the lowest rates of Type I and II errors respectively. This result also indicates that the first component of the hybrid model (i.e. clustering) can filter out some unrepresentative data, which may affect prediction performance.

In short, combining *k*-means with single classifiers, *k*-means + *k*-NN provides the highest rate of prediction accuracy and the lowest rate of the Type II error. *K*-means + DT provide the second highest and lowest rates of prediction accuracy and Type II error, respectively.

* *Hybrid Classifiers: K-Means + Classifier Ensembles*

Table 15 shows the performance of combining *k*-means with boosting/bagging based classifier ensembles, in which the number underlined represents the best performance.

Table 15 Prediction performances of combining *k*-means with boosting/bagging based classifier ensembles

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Type I error | Type II error |
| *Boosting* |  |  |  |
| MLP | 90.48% (3) | 26.82% (5) | 6.05% (3) |
| DT | 91.43% (1) | 22.19% (3) | 5.83% (1) |
| Naïve Bayes | 87.21% (5) | 4.95% (1) | 14.36% (5) |
| SVM | 89.21% (4) | 15.51% (2) | 9.84% (4) |
| *k*-NN | 91.09% (2) | 23.93% (4) | 5.90% (2) |
| STDEV | 1.72 | 8.70 | 3.75 |
| *t*-value | 117.00\*\*\* | 4.77\*\*\* | 5.02\*\*\* |
| *Bagging* |  |  |  |
| MLP | 90.73% (3) | 20.13% (4) | 7.09% (3) |
| DT | 91.60% (1) | 18.65% (3) | 6.34% (2) |
| Naïve Bayes | 87.27% (5) | 5.12% (1) | 14.26% (5) |
| SVM | 88.54% (4) | 12.54% (2) | 11.25% (4) |
| *k*-NN | 91.46% (2) | 22.36% (5) | 5.76% (1) |
| STDEV | 1.92 | 7.00 | 3.69 |
| *t*-value | 104.61\*\*\* | 5.05\*\*\* | 5.44\*\*\* |

\*\*\* Represents the level of significance is higher than 99% by *t*-test.

For *k*-means + classifier ensembles using boosting, all of the three performance measurements contain a high level of significant difference between these five prediction models. Particularly, *k*-means + DT ensembles perform the best in terms of prediction accuracy and Type II error. Similar to *k*-means + naïve Bayes, *k*-means + naïve Bayes ensembles can provide the lowest rate of Type I error. On the other hand, *k*-means + classifier ensembles using boosting, all the three performance measurements also contain a high level of significant difference between these five prediction models. Among them, *k*-means + DT ensembles can provide the highest rate of prediction accuracy and second lowest rate of the Type II error. Again, *k*-means + naïve Bayes ensembles perform the best in terms of the Type I error.

To compare these two types of hybrid classifiers, prediction accuracy of *k*-means + classifier ensembles using bagging and/or boosting is slightly higher except *k*-means + *k*-NN ensembles using boosting and *k*-means + SVM ensembles using bagging. In addition, most of these *k*-means + classifier ensembles provide lower Type I and II errors. Similar to the comparative result between single classifiers and classifier ensembles, *k*-means + classifier ensembles outperform *k*-means + single classifiers. Regarding Table 15, *k*-means + DT ensembles using bagging can provide the highest rate of prediction accuracy. For the Type I error, *k*-means + naïve Bayes ensembles using boosting outperform the others. On the other hand, *k*-means + *k*-NN ensembles using bagging can provide the lowest rate of the Type II error.

## *4.3.4 Comparisons and Discussions*

Regarding the above results, DT is the only single classifier that provides relatively better prediction performances than other single classifiers. However, in classifier ensembles and hybrid classifiers, DT and *k*-NN are the best two classifiers. Therefore, we further compare them in terms of prediction accuracy and Type I/II errors in order to find out the best model to predict intangible assets. Table 16 shows the comparative result.

Table 16 Comparisons of the best prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Type I error | Type II error |
| Single classifier |  |  |  |
| *k*-NN | 78.24% | 30.45% | 17.47% |
| Classifier ensembles |  |  |  |
|  *Boosting* |  |  |  |
|  DT | 78.89% | 33.23% | 15.12% |
|  *Bagging* |  |  |  |
|  DT | 79.55% | 32.46% | 14.52% |
|  *k*-NN | 78.48% | 30.48% | 17.09% |
| Hybrid classifiers (single classifiers) |  |  |  |
|  *k*-NN | 91.34% | 22.77% | 5.83% |
|  DT | 90.33% | 20.38% | 7.52% |
| Hybrid classifiers (classifier ensembles) |  |  |  |
|  *Boosting* |  |  |  |
|  DT | 91.43% | 22.19%  | 5.83%  |
|  *Bagging* |  |  |  |
| DT | 91.60%  | 18.65%  | 6.34%  |
| *k*-NN | 91.46%  | 22.36%  | 5.76%  |

This shows that classifier ensembles and two types of hybrid classifiers perform better than single classifiers. In other words, classifier ensembles and hybrid classifiers can improve the prediction performances of single classifiers by combining multiple classifiers and reducing unrepresentative data at the first stage, respectively.

Specifically, the performances of combining k-means with boosting/bagging based classifier ensembles are much better than others in terms of prediction accuracy and Type I and II errors. This finding shows that using a clustering technique to perform the data reduction task can allow the later classifiers to perform better without considering this data pre-processing stage including single classifiers and classifier ensembles. Thus, the hybrid classifiers that combine *k*-means with boosting/bagging based classifier ensembles, can provide the top three prediction accuracies, especially *k*-means + bagging based DT ensembles, which provide the best performance intangible assets prediction.

## 5. Conclusion

Intangible assets have become key drivers of economic performance, which has prompted a growing number of firms to emphasize intangible assets investment. However, the lack of recognition of intangible assets in financial statements causes the information gap between insider and outsider. As a result, it is very important to accurately evaluate intangible assets and market-based value of firms, especially when investors assess an initial public offerings (IPO) firm. Therefore, this dissertation uses data mining technologies different from traditional statistical methods used in prior literature to evaluate intangible assets and market-based of firms and provide a new viewpoint to analyze intangibles.

In feature selection process, this paper first of all reviews related literature from diverse domains including accounting, finance, management, business and marketing to collect relatively important factors affecting intangible assets. Then, five feature selection methods, which are PCA, STEPWISE, DT, AR and GA are used to select important factors and to evaluate the intangible assets of firms in Taiwan. In addition, combining different feature selection results by the union and intersection strategies can result in other selected features for further comparisons. To assess the effectiveness of the identified features of these methods, MLP neural networks are constructed as the evaluation model to examine the prediction performances.

Regarding the experimental results over the chosen dataset containing 61 variables, for single feature selection methods, DT is the best method to provide the highest rate of prediction accuracy and lowest rate of Type I errors. In addition, it only selects 7 features, which constructs the prediction model in a very efficient manner. On the other hand, for the multiple feature selection methods created by combining the results from different single methods, GA∩STEPWISE and DT∪PCA are the top two methods. They select 26 and 22 features respectively and provide the best prediction accuracy and outperform single method-DT.

Therefore, these selected features from the best method (i.e. GA∩STEPWISE) can then be regarded as the important factors affecting intangible assets and market-based value of firms in Taiwan. They are *R&D INTENSITY* and *ADVERTISING INTENSITY* in intangible capital category, *FAMILY*, *CASH FLOW RIGHT*, and *BUSINESS GROUP* in ownership structure category, *SALE GROWTH*, *SIZE*, *LEVERAGE*, *DIVIDEND*, *PROFITABILITY*, *AGE*, and *EXPORT* in firm characteristics, 13 industries in industry characteristics category, and *MARKET SHARE* in reactions of analysts and customer category.

Sequentially, this paper compares various machine learning techniques to figure out a more accurate evaluation and prediction model after employing feature selection provided in the feature selection process. Particularly, in addition to seven single classification algorithms, which are DT, ANN, naïve Bayes, SVM, *k*-NN, LR, and LDA, classifier ensembles and hybrid classifiers are considered for the comparative study. In single classifier technologies, this paper compares traditional approach (i.e. LR and LDA) employed in prior related literature with machine learning techniques or data mining techniques, such as neural networks, support vector machines, etc., which are superior to statistical methods in prior studies. Sequentially, classifier ensembles and hybrid classifiers have shown that they can provide better prediction performances than single techniques in many domains are constructed.

Regarding the experimental results, classifier ensembles and two types of hybrid classifiers can improve the prediction performances over single classifiers. In particular, the hybrid classifiers by combining *k*-means with boosting/bagging based classifier ensembles perform much better than the others in terms of prediction accuracy and Type I and II errors. In particular, *k*-means + bagging based DT ensembles provide the best performance to evaluate intangible firm value. This prediction model as the best one can help investors and creditors to make more accurate decisions of investments and loans.

This paper focus on a new viewpoint- explore affecting factor of intangible assets and expect the selected factors and empirical results allow us not only to understand which category of factor and classifier model can be used to evaluate intangible assets effectively but also to provide other information which are not disclosed in financial statements for outside users. This will help investors or creditors to better evaluate the new investment or lending opportunities, and help them make decisions more accurately.

Although this paper has tried to collect as many related important factors as possible in all kinds of business disciplines in literature, some affecting variables of intangible assets might be missing (e.g. employee of R&D, goodwill and brand loyal in intangible capital category, and so on). Therefore, for future work, newer related literature or minor factors found in related studies could be incorporated to conduct additional analyses. Otherwise, it should be noted that although this study considers several widely used techniques to develop the evaluation models, there are other algorithms available in literature, which can be applied, for example, self-organizing map (SOM) as the clustering technique, and genetic algorithms for the feature selection method or classification techniques. However, it is difficult to conduct a comprehensive study on all existing clustering and classification techniques. Thus, for future work, other clustering and classifier methods can also be applied to compare with the prediction models provided by this study in order to make a more reliable conclusion.

Although most prior literature and this study use Tobin’s *Q* as the proxy for intangible assets, the excess value (i.e. market value of firms is greater than the book value of its assets) may not all reflect an unmeasured source of value attributed to the intangible assets. Excess value may reflect an unmeasured value of tangible assets. Therefore, for future work some more accurate measurement method can be used as the proxy for intangible assets to compare with Tobin’s *Q*. In addition, this study mainly focuses on evaluating the intangible assets and market-based value problems. Other domain problems can be applied, such as bankruptcy prediction, stock price prediction and financial distress forecasting in finance domain; audit opinion prediction, auditor’s going concern uncertainty decision prediction and litigation prediction in account/auditing domain, etc. in the future work.

# References

Adeli, H. and Hung, S. 1995. *Machine learning: neural networks, genetic algorithms, and fuzzy systems*, New York: Wiley.

Anderson, E. W., Fornell, C., and Mazvancheryl, S. K. 2004. “Customer Satisfaction and Shareholder Value.” *Journal of Marketing* 68(4): 172-185.

Black, B. S., Jang, H., and Kim, W. 2006. “Does Corporate Governance Predict Firms’ Market Values? Evidence from Korea.” *The Journal of Law, Economics, & Organization* 22(2): 366-413.

Bozec, R., Dia, M., and Bozec, Y. 2010. “Governance-performance relationship: a re-examination using technical efficiency measures.” *British Journal of Management* 21(3): 684-700.

Brooking, A. 1996. *Intellectual Capital*, London: International Thomson Business Press.

Buckinx, W. and Van den Poel, D. 2005. “Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting.” *European Journal of Operational Research* 164(1): 252-268.

Burez, J. and Van den Poel, D. 2007. “CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services.” *Expert Systems with Applications* 32(2): 277-288.

Burgman, R. and Roos, G. 2007. “[The importance of intellectual capital reporting: evidence and implications](http://assets.emeraldinsight.com/Insight/viewContentItem.do;jsessionid=276F7ACE7CD23060012C905EF4369B3F?contentType=Article&contentId=1585614).” *Journal of Intellectual Capital* 8(1): 7-51.

Canbas, S., Cabuk, A., and Kilic, S. B. 2005. “Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The Turkish case.” *European Journal of Operational Research* 166(2): 528-546.

Canuto, A. M. P., Abreu, M. C. C., De Melo Oliverira, L., Xavier, Jr., J. C. and Santos, A. D. M. 2007. ”Investigating the influence of the choice of the ensemble members in accuracy and diversity of selection-based and fusion-based methods for ensembles.” *Pattern Recognition Letters* 28(4): 472-486.

Cao, Y. H. 2009. “The research on the recognition and measurement of intangible assets for high-tech enterprise.” *Management Science and Engineering* 3(2): 55-60.

Chan, L. K. C., Lakonishok, J., and Sougiannis, T. 2001. “The Stock Market Valuation of Research and Development Expenditures.” *The Journal of Finance* 56(6): 2431-2456.

Chandra, D. K., V. Ravi and P. Ravisankar (2010) “Support Vector Machine and Wavelet Neural Network hybrid: Application to Bankruptcy Prediction in Banks.” *International Journal of Data Mining, Modeling and Management* 1(2): 1-21.

Chauhan, N. J., V. Ravi and D. K. Chandra (2009) “Differential Evolution trained Wavelet Neural Network: Application to bankruptcy prediction in banks.” *Expert Systems with Applications* 36(4): 7659-7665.

Claessens, S., Djankov, S., Fan, J. P. H., and Lang, L. H. P. 2002. “Disentangling the incentive and entrenchment effects of large shareholdings.” *The Journal of Finance* 57(6): 2741-2771.

Coussement, K. and Van den Poel, D. 2008. “Churn prediction in subscription services: An application of support vector machines with comparing two parameter-selection techniques.” *Expert Systems with Applications* 34(1): 313-327.

Dadalt, P. J., Donaldson, J. R., and Garner, J. L. 2003. “Will any Q do?” *Journal of Financial Research* 26(4): 535-551.

Durst, S. and Gueldenberg, S. 2009. “The Meaning of Intangible Assets: New Insights into External Company Succession in SMEs.” *Electronic Journal of Knowledge Management* 7(4): 437-446.

Eckstein, C. 2004. “The Measurement and Recognition of Intangible Assets: Then and Now.” *Accounting Forum* 28(2): 139-158.

Edvinsson, L. and Malone, M. 1997. *Intellectual Capital: Realizing Your Company’s True Value by Finding its Hidden Brainpower*, HarperBusiness, New York.

Ellili, N. O. D., 2011. “Ownership Structure, Financial Policy and Performance of the Firm: UK Evidence.” *International Journal of Business and Management* 6(10): 80-93.

Fan, J. P. H. and Wong, T. J. 2005. “Do external auditors perform a corporate governance role in emerging markets? Evidence form East Asia.” *Journal of Accounting Research* 43(1): 35-72.

Fukui, Y. and Ushijima, T. 2007. “Corporate diversification, performance, and restructuring in the largest Japanese manufacturers.” *Journal of the Japanese and International Economies* 21(3): 303-323.

Gleason, K. I. and Klock, M. 2006. “Intangible capital in the pharmaceutical and chemical industry.” *The Quarterly Review of Economics and Finance* 46(2): 300-314.

Goh, D. H. and Ang, R. P. 2007. “An introduction to association rule mining: An application in counseling and help-seeking behavior of adolescents.” *Behavior Research Methods* 39(2): 259-266.

Guyon, I. and Elisseeff, A. 2003. “An introduction to variable and feature selection.” *Journal of Machine Learning Research* 3(7/8): 1157-1182.

Han, J. and Kamber, M. 2001. *Data Mining: Concepts and Techniques*, Morgan Kaufmann.

Han, J. and Kamber, M. 2006. *Data Mining: Concepts and Techniques*, 2nd Edition. Morgan Kaufmann.

Han, D. and Han, I., 2004. “Prioritization and Selection of Intellectual Capital Measurement Indicators Using Analytic Hierarchy Process for the Mobile Telecommunications Industry.” *Expert Systems with Applications* 26: 519-527.

Hayashi, Y. and Setiono, R. 2002. “Combining neural network predictions for medical diagnosis.” *Computers in Biology and Medicine* 32(4): 237-246.

Hsieh, N. C. 2005. “Hybrid mining approach in the design of credit scoring models.” *Expert Systems with Applications* 28(4): 655-665.

Hu, M. Y. and Tsoukalas, C. 2003. “Explaining consumer choice through neural networks: the stacked generalization approach.” *European Journal of Operational Research* 146(3): 650-660.

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H. and Wu, S. 2004. “Credit rating analysis with support vector machines and neural networks: a market comparative study.” *Decision Support Systems* 37(4): 543-558.

Huang, C. L. and Wang, C. J. 2006. “A GA-based feature selection and parameters optimization for support vector machines.” *Expert Systems with Applications* 31(2): 231-240.

Hung, S. Y., Yen, D. C. and Wang, H. Y. 2006 “Applying data mining to telecom churn management.” *Expert Systems with Applications* 31(3): 515–524.

Huysmans, J., Baesens, B., Vanthienen, J, and Van Gestel, T. 2006. “Failure prediction with self organizing maps.” *Expert Systems with Applications* 30(3): 479-487.

Jo, H. and Harjoto, M. A., 2011. “Corporate Governance and Firm Value: The Impact of Corporate Social Responsibility.” *Journal of Business Ethics* 103(3): 351-383.

Jolliffe, I. T. 1986. *Principal Component Analysis*, Springer, New York.

Kessels, J. 2001. *Verleiden tot kennisproductiviteit*. Enschede, Universiteit Twente.

Khanna, T. and Yafeh, Y. 2007. “Business Groups in Emerging Markets: Paragons or Parasites?” *Journal of Economic Literature* 45(2): 331-372.

Kim, K. J. and Han, I. 2000. “Genetic algorithm approach to feature discretization in artificial neural network for the prediction of stock price index.” *Expert Systems with Applications* 19(2): 125-132.

Kira, K. and Rendell, L. 1992. “A practical approach to feature selection.” *Proceedings of the Ninth International Conference on Machine Learning*,Aberdeen, Scotland: Morgan Kaufmann. 249-256.

Koller, D. and Sahami, M. 1996. “Toward Optimal Feature Selection.” *Proceedings of International Conference on Machine Learning*.

Kuo, R. J., Ho, L. M. and Hu, C. M, 2002. “Integration of self-organizing feature map and K-means algorithm for market segmentation.” *Computers and Operations Research* 29(11): 1475–1493.

Lang, M. H., Lins, K. V., and Miller, D. P. 2003. “ADRs, Analysts, and Accuracy: Does Cross Listing in the United States Improve a Firm's Information Environment and Increase Market Value?” *Journal of Accounting Research* 41(2): 317-345.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., and Vishny, R. 2002. “Investor protection and corporate valuation.” *Journal of Finance* 57(3): 1147-1170.

Larcker, D. F., Richardson, S. A., and Tuna, I. 2007. “Corporate Governance, Accounting Outcomes, and Organizational Performance.” *The Accounting Review* 82(4): 963-1008.

Lemmon, M. L. and K. V. Lins. 2003. “Ownership structure, corporate governance, and firm value: Evidence from the East Asian financial crisis.” *Journal of Finance* 58(4): 1445-1468.

Li, C. T. and Tan, Y. H. 2006. “Adaptive control of system with hysteresis using neural networks.” *Journal of Systems Engineering and Electronics* 17(1): 163-167.

Lins, K. V. 2003. “Equity Ownership and Firm Value in Emerging Markets.” *Journal of Financial and Quantitative Analysis* 38(1): 159-184.

Lustgarten, S. and Thomadakis, S. 1987. “Mobility Barriers and Tobin’s *q*.” *Journal of Business* 60(4): 519-537.

Megna, P. and Klock, M. 1993. “The Impact of Intangible Capital on Tobin’s *q* in the Semiconductor Industry.” *The Value of Intangible Assets* 83(2): 265-269.

McConnell, J. J. and Servaes, H. 1990. “Additional Evidence on Equity Ownership and Corporate Value.” *Journal of Financial Economics* 27(2): 595-612.

Min, J. H. and Lee, Y. C. 2005. “Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters.” *Expert Systems with Applications* 28(4): 603-614.

Morgan, N. A. and Rego, L. L. 2009. “Brand Portfolio Strategy and Firm Performance.” *Journal of Marketing* 73(1): 59-74.

Morck, R. and Yeung, B. 2003. “Agency problems in large family business group.” *Entrepreneurship Theory and Practice* 27(4): 367-382.

Nanni, L., and Lumini, A. 2009. “An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring.” *Expert Systems with Applications* 36(2): 3028-3033.

Olafsson, S., Li, X., and Wu, S. 2008. “Operations research and data mining.” *European Journal of Operational Research* 187(3): 1429-1448.

Oxelheim, L. and Randoy, T. 2003. “The impact of foreign board membership on firm value.” *Journal of Banking and Finance* 27(12): 2369-2392.

Questier, F., Put, R., Coomans, D., Walxzak, B., and Heyden, Y.V. 2005. “The use of CART and multivariate regression tree for supervised and unsupervised feature selection.” *Chemometrics and Intelligent Laboratory System* 76: 45-54.

Pendharkar, P. C. and J. A. Rodger (2004) “An empirical study of impact of crossover operators on the performance of non-binary genetic algorithm based neural approaches for classification.” *Computers & Operations Research*, 31(4): pp. 481-498.

Petty, R. and Guthrie, J. 2000. “Intellectual capital literature overview: measurement, reporting and management.” *Journal of Intellectual Capital* 1(2): 155-176.

Rao, V. R., Agarwal, M. K., and Dahlhoff, D. 2004. “How Is Manifest Branding Strategy Related to the Intangible Value of a Corporation?” *Journal of Marketing* 68(4): 126-141.

Roos, G. and Roos, J. 1997. “Measuring your company’s intellectual performance.” *Long Range Planning* 30(3): 413–426.

Shapiro, C. and Varian, H. 1999. *Information Rules*, Boston: Harvard Business School Press.

Shin, K. S. and Lee, Y. J. 2002. “A genetic algorithm application in bankruptcy prediction modeling.” *Expert Systems with Applications* 23(3): 321-328.

Silva, F., Majluf, N., and Paredes, R. D. 2006. “Family ties, Interlocking Directors and Performance of Business Groups in Emerging Countries: The Case of Chile.” *Journal of Business Research* 59(3): 315-321.

Smith, K. A. and Gupta, J. N. D. 2000. “Neural networks in business: techniques and applications for the operations researcher.” *Computers & Operations Research* 27(11-12): 1023-1044.

Sohn, S. Y. and Lee, S. H. 2003. “Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea.” *Safety Science* 41(1): 1-14.

Stewart, T. A. 1997. *Intellectual Capital*, London: Nicholas Brealey Publishing.

Sugumaran, V., Muralidharan, V., and Ramachandran, K. I. 2007. “Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing.” *Mechanical Systems and Signal Processing* 21(2): 930-942.

Sveiby K. 1997. *The New Organizational Wealth: Managing and Measuring Knowledge-based Assets*, Barrett-Kohler Publishers, San Francisco.

Tam, K. Y. and Kiang, M. Y. 1992. “Managerial applications of neural networks: the case of bank failure prediction.” *Management Science* 38(7): 926-947.

Theodoridis, S. and Koutroumbas, K. 2006. *Pattern Recognition*, 3 ed. Academic Press.

Tobin, J. 1969. “A general equilibrium approach to monetary theory.” *Journal of Money, Credit, and Banking* 1(1): 15-29.

Tsai, C. F. 2009. “Feature selection in bankruptcy prediction.” *Knowledge-Based Systems* 22(2): 120-127.

Tsai, C. F. and Chen, M. L. 2010a. “Credit rating by hybrid machine learning techniques.” *Applied Soft Computing* 10(2): 374-380.

Tsai, C. F. and Wu, J. W. 2008. “Using neural network ensembles for bankruptcy prediction and credit scoring.” *Expert Systems with Applications* 34(4): 2639-2649.

Tzeng, M. and Goo, Y. J. J. 2005. “Intellectual Capital and Corporate Value in an Emerging Economy: Empirical Study of Taiwanese Manufacturers.” *R&D Management* 35(2): 187-201.

Vandemaele, S. N., Vergauwen, P. G. M. C. and Smits, A. J. 2005. “[Intellectual capital disclosure in The Netherlands, Sweden and the UK: A longitudinal and comparative study](http://www.emeraldinsight.com/Insight/viewContentItem.do;jsessionid=07E40536EF2CCBF6FB988B87790224F4?contentType=Article&contentId=1513194).” *Journal of Intellectual Capital* 6(3): 417-426.

Vergauwen, P., Bollen, L., and Oirbans, E. 2007. “Intellectual capital disclosure and intangible value drivers: an empirical study.” *Management Decision* 45(7): 1163-1180.

Verikas, A., Z. Kalsyte, M. Bacauskiene, and A. Gelzinis 2010. “Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: A survey.” *Soft Computing* 14(9): 995-1010.

West, D., Dellana, S. and Qian, J. 2005. ”Neural network ensemble strategies for financial decision applications.” *Computers & Operations Research* 32(10): 2543-2559.

Wiwattanakantang, Y. 2001. “Controlling shareholders and corporate value: Evidence from Thailand.” *Pacific-Basin Finance Journal* 9(4): 323-362.

Wu, X., V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, D. Steinberg 2008. “Top 10 algorithms in data mining.” *Knowledge Information System* 14: 1-37.

Xie, B., Davidson III, W. N., and Dadalt, P. J. 2003. “Earnings management and corporate governance: the role of the board and the audit committee.” *Journal of Corporate Finance* 9(3): 295-316.

Yang, J. and Olafsson, S. 2006. “Optimization-based feature selection with adaptive instance sampling.” *Computers & Operations Research* 33(11): 3088-3106.

Zhan, G., Patuwo, B. E. and Hu, M. Y. 1998. “Forecasting with artificial neural network: The state of the art.” *International Journal of Forecasting* 14(1): 35-62.