Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# CLASSIFICATION MODELS BASED FORWARD SELECTION FOR BUSINESS PERFORMANCE PREDICTION

[1]EDI NOERSASONGKO, [2]PURWANTO, [3]GURUH FAJAR SHIDIK
[1, 2, 3] Dian Nuswantoro University, Semarang, Indonesia
E-mail: [1]edi-nur@dosen.dinus.ac.id , [2]purwanto@dsn.dinus.ac.id , [3]guruh.fajar@research.dinus.ac.id

**ABSTRACT**

This paper proposes a classification model to improve the accuracy of prediction for business performance. The proposed model uses a combination of forward selection method to select the optimum attributes and classification models. Business performance data set is used to evaluate the accuracy of the proposed model. From results of experiments show that the combination of forward selection and Naïve Bayes model can improve the prediction accuracy of business performance compared to the other classification models, namely Logistic Regression, k-NN, Naïve Bayes, C4.5 and Support Vector Machine models significantly. The proposed model also yields better result compared to the other attribute selection using backward elimination method.

**Keywords:** *Forward Selection, Naïve Bayes, Entrepreneur, Business performance, Classification Models*

## 1. INTRODUCTION

The economic crisis that occurred in 2008 was a very heavy blow for entrepreneurs. The impact of the crisis is felt directly by the entrepreneur is a large increase in the rate of inflation, weakening currency values, so that the costs of various increases and many companies close their business. However, in terms of production, the crisis can also encourage positively the growth of small business output. These positive effects are obtained through the labor market because of the increasing number of business units and the increasing number of workers and entrepreneurs in small business due to the large number of workers in the formal sector of the medium and large business that are exposed to termination. Due to the pressure to earn a living, many former employees make any economic activities, including opening small business, with their available capital and resources [1].

Small business is strategic business if it is managed properly. Small business can survive because they can produce various unique products at affordable prices. Small business development program in Indonesia has not been carried out as well as that in developed countries such as America and Australia [2]. Many studies on the factors that influence the success of small business in Indonesia have been conducted. They can be divided into three groups, namely studies on small business development strategies, micro-economic studies, and studies on detailed cases [1].

Business performance is influenced the internal and external factors. The factors are the problems faced by entrepreneurs of small business who want to grow and tied face business competition. Each small business will compete for the existing markets so that consumers will be selective and critical in determining their choice. This situation requires ability improvement of small businesses [3].

In this study, the entrepreneur characteristic, entrepreneurship, leadership and business ability influence the business performance. This study presents business performance prediction using classification models such as C4.5, Logistic Regression (LR), *k*-NN, Naïve Bayes (NB) and Support Vector Machine (SVM) models. These models are employed to predict business performance based on 19 attributes, namely (1). education, (2). gender, (3). training, (4). experience, (5). age, (6). vision, (7). planning, (8). motivation, (9). innovation, (10). opportunity, (11). self-recognition, (12). risk, (13). ethics, (14). adaptation, (15). dictator, (16). participation, (17). delegation, (18). consideration, (19). business ability, and (20). business performance.

Improving classification accuracy using multivariate data still remains a challenging task for researchers. Classification models using

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

multivariate data have been applied for prediction in various applications, such as logistic regression, naïve Bayes, decision tree and support vector machine. Logistic regression model has been used for classification [4, 5]; Naïve Bayes has been used in economics and health [6, 7]; Decision tree model has been applied for classifying blood donors [9]; and support vector machine model has been employed in medicine and clinical studies and engineering [4, 10, 11]. The results show that the accuracies of classification models vary depending on the type of data. There is no particular classification model that yields the best for all the data sets [12].

In classification, attributes selection is an important aspect in analysis of data [13]. It is possible that only several attributes contribute to the final output. The selection of appropriate attributes will perform better prediction accuracy results in applying classification models. In this study, we propose method using a combination of forward selection to select the attributes and classification models such as logistic regression, C4.5, *k*-NN, Naïve Bayes, and support vector machine to obtain the best accuracy for business performance prediction.

## 2. CLASSIFICATION MODELS

In this research, some models are used for business performance prediction. This section describes C4.5, logistic regression (LR), naïve Bayes (NB), *k*-NN and support vector machine (SVM) models.

### 2.1 C4.5 Model

Decision tree (DT) model, especially C4.5 model has been implemented in various fields such as modeling, text mining, data mining and many other areas. C 4.5 model is a classification method to predict label that is constructed using training data. Decision tree can be arranged in a structure that can easily be understood by humans [14, 15].

C4.5 algorithm steps are as follows:

1. Preparing training data
2. Determining the root of the tree. The root of the tree is determined by calculating the highest Information Gain of each attribute.

$$Gain(D,A) \equiv Entropy(D) - \sum_{i=1..n} \frac{|D_i|}{|D|} Entropy(D_i) \quad (1)$$

$$Entropy(D) \equiv \sum_{i=1}^{n} -p_i \log_2 p_i \quad (2)$$

where $p_i$ is the probability of class in data set D and A is attributes.

3. Repeat step 2 until all tuples is partitioned.

### 2.2 Logistic Regression (LR) Model

Logistic regression model is one of the models used in machine learning for classification. This model can accept the mixture of numerical and categorical data attributes. Equation (3) represents the logistic function.

$$f(z) = \frac{1}{1+e^{-z}} \quad (3)$$

The logistic function *f(z)* ranges between 0 and 1 and the logistic model is designed to describe a probability, which is always some number between 0 and 1. The logistic model is constructed from the logistic function. Using Equation (3), *z* is written as [16]:

$$z = \alpha + \beta_1 X_1 + ... + \beta_k X_k \quad (4)$$

where $X_1$, $X_2$, ..., $X_k$ are attributes and α is a constant, $i$=1...$k$

The logistic regression model is calculated as [16]:

$$P(Y=1|X_1,X_2,...,X_k) = P(X)$$

$$= \frac{e^{\sum_{i=0}^{k} \beta_i X_i}}{1+e^{\sum_{i=0}^{k} \beta_i X_i}} \quad (5)$$

### 2.3 Naïve Bayes (NB) Model

Naïve Bayes method is one of the methods used for classification. This method employs the conditional probability as the basis. Assume that the label (dependent variable) $Y$ is a Boolean-valued random variable. The $X_i$ are also Boolean-valued random variables, where $i$= *1,...,n*. Using Bayes' theorem can be represented as [14]:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$P(Y=y_k|X_1...X_n)=\frac{P(X_1...X_n|Y=y_k)P(Y=y_k)}{\sum_j P(X_1...X_n|Y=y_j)P(Y=y_j)} \quad (6)$$

where $y_k$ is the $k$-th possible value of $Y$ and $X_1$, $X_2$, ... $X_n$ are discrete-valued or real attributes.

Naïve Bayes classifier estimates the class conditional probability by assuming that the attributes are conditionally independent given the class label $Y$. Conditional independent assumption can be expressed as [14]:

$$P(X_1...X_n|Y)=\prod_{i=1}^{n} P(X_i|Y) \quad (7)$$

Assume that $X_i$, where $i$=1,2,…, $n$ are conditional independent variables for a given $Y$ and using Bayes' theorem, Equation (7) can be rewritten as:

$$P(Y=y_k|X_1...X_n)=\frac{P(Y=y_k)\prod_{i=1}^{n}P(X_i|Y=y_k)}{\sum_j P(Y=y_j)\prod_{i=1}^{n}P(X_i|Y=y_j)} \quad (8)$$

### 2.4 K-Nearest Neighbor (k-NN) Algorithm

K-Nearest Neighbor (k-NN) algorithm is a classification method in which a new object is labeled by $k$-nearest neighbor objects [15]. KNN including supervised learning algorithm where the result of new instance query is classified based on the majority of categories on KNN.

The steps of the $k$-Nearest Neighbor algorithm are as follows:

1. Determining parameter, namely $k$ (number of nearest neighbors)
2. Calculating the Euclidean distance of each object to the given sample data.
   The Euclidean distance between two points can be calculated as [15]:
   $$dist(X_1,X_2)=\sqrt{\sum_{i=1}^{n}(x_{1i}-x_{2i})^2} \quad (9)$$
   where $X_1$= ($x_{11}$,$x_{12}$,…,$x_{1n}$), $X_2$ = ($x_{21}$,$x_{22}$,…,$x_{2n}$) and $i$=1,2, …, $n$.
3. Sorting these objects into groups that have the smallest Euclidean distance
4. Collecting nearest neighbor classification categories.
5. Majority of categories on the nearest neighbor will be used for prediction.

### 2.5 Support Vector Machine Model

The SVM is a useful model that is used for classification and regression [17, 18]. The support vector machine model can be described as follows:

Suppose that ($x_i$, $y_i$), is a sample point of attribute-label pair data, where, $x_i \in \Re^D$, $y_i \in \{+1,-1\}$ and $i$ = 1, . . . , $n$. Let the positive class be denoted as +1 and negative class be denoted as -1. The SVM method requires the formulation of the optimization problem as follows:

Minimize: $\frac{1}{2}\|w\|^2$. (10)

Subject to: $y_j(\omega x_j+\beta)\geq 1, \forall j,$ (11)

where,
$\omega$ : weight vector,
$\beta$ : a bias.

This optimization can be solved using Lagrange Multipliers as follows:

$$L(\omega,\beta,\gamma)=\frac{1}{2}\|\omega\|^2-\sum_{j=1}^{n}\gamma_j[y_j(\omega,x_j+\beta)-1]. \quad (12)$$

with $\gamma_j \geq 0,\ j=1,2,...,n.$

The solution can be obtained as follows:

$$\text{Max: } L(\gamma)=\sum_{j=1}^{n}\gamma_j-\frac{1}{2}\sum_{j,k=1}^{n}\gamma_j\gamma_k y_j y_k x_j x_k. \quad (13)$$

Subject to: $\sum_{j=1}^{n}\gamma_j y_j=0$ and $\gamma_j \geq 0,\ j=1,...,n.$ (14)

The solutions of this problem may result in many values of $\gamma_j$ which are equal to 0. The pattern vectors $x_j$ for which $0<\gamma_j$ are support vectors that have the shortest distance to the hyper plane.

## 3. BUSINESS PERFORMANCE DATA

The classification models such as C4.5, $k$-NN, Naïve Bayes, Logistic Regression and support vector machine models are used to predict business performance data set. The combination using forward selection and classification models are also employed to predict business performance. The business performance data set is collected from small business in Indonesia.

The data pertain to entrepreneur characteristic, entrepreneurship, leadership, business ability, and business performance. The attributes are as follows:

Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

1. Entrepreneur characteristics comprise five attributes, those are: education, gender, training, experience, and age.
2. Entrepreneurship consists of nine attributes, those are: vision, planning, motivation, innovation, opportunity, self-recognition, risk, ethics, and adaptation.
3. Leadership style covers dictator, participation, delegation, and consideration
4. Business ability is the mean of production ability, marketing ability, and finance ability.
5. Business performance that has a binary value representing whether the person success or not success.

Business performance data set comprises 160 data points. The attributes of business performance data set are shown in Table 1. The minimum, maximum, mean, and descriptions of the data are also given in Table 1.

*Table 1: Descriptive Statistics of Business Performance Data Set*

| Name | Min. | Max. | Mean | Descriptions |
|---|---|---|---|---|
| x1 | 1 | 5 | 3.538 | Education |
| x2 | 1 | 2 | 1.163 | Gender |
| x3 | 1 | 5 | 3.219 | Training |
| x4 | 1 | 5 | 3.113 | Experience |
| x5 | 1 | 5 | 2.269 | Age |
| x6 | 2 | 5 | 3.719 | Vision |
| x7 | 2 | 5 | 3.713 | Planning |
| x8 | 2 | 5 | 3.900 | Motivation |
| x9 | 2 | 5 | 4.088 | Innovation |
| x10 | 2 | 5 | 3.394 | Opportunity |
| x11 | 2 | 5 | 3.875 | Self-recognition |
| x12 | 2 | 5 | 3.269 | Risk |
| x13 | 2 | 5 | 3.806 | Ethics |
| x14 | 2 | 5 | 3.963 | Adaptation |
| x15 | 2 | 5 | 3.088 | Dictator |
| x16 | 2 | 5 | 3.756 | Participation |
| x17 | 2 | 5 | 3.275 | Delegation |
| x18 | 3 | 5 | 4.019 | Consideration |
| x19 | 2 | 4 | 3.400 | Business ability |
| BP | 0 | 1 | 0.750 | Business performance |

## 4. PROPOSED METHOD

In this study, the proposed method is used for business performance prediction. The steps of proposed method are shown in Figure 1.

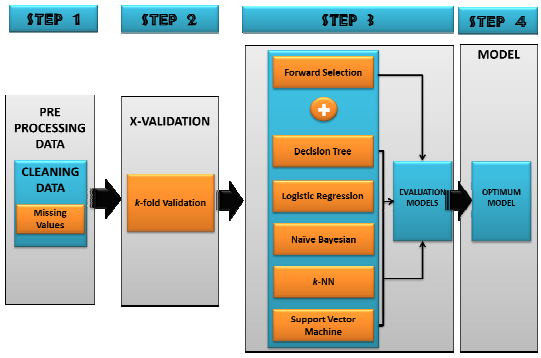


*Figure 1: Proposed Method For Business Performance Prediction*

The proposed method is divided into four steps. Step by step of proposed method can be explained as follows:

STEP 1 : In this step, pre-processing is performed to clean up the business performance data. The business performance data set is preprocessed by replacing all missing values by mean values before any further analysis.

STEP 2 : In step 2, the experiments setup using k-fold validation. We use 10-fold validation to construct and evaluate the classification models.

STEP 3 : The classification models, namely C4.5, naïve Bayes, k-NN, Logistic Regression and SVM models are employed for business performance prediction. Combination forward selection and the classification models are employed for business performance prediction. The evaluation models to obtain the best model are performed based on a comparison of accuracies of the different classification models.

STEP 4 : The optimum model is selected based on the values of accuracies of prediction. The optimum classification model is the model that has the highest accuracy value.

Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

## 5. RESULTS AND DISCUSSION

### 5.1 Performance Measure

In this work, the classification models, namely C4.5, naïve Bayes, *k*-NN, Logistic Regression and SVM models are used for business performance prediction, individually. The combination forward selection and the classification models are also employed for business performance prediction. The performances of these classification models are calculated based on business performance data set. A confusion matrix is used to calculate the performances of prediction [19].

*Table 2: Confusion matrix for a 2-class model*

| | | PREDICTED CLASS | |
|---|---|---|---|
| | | Class= Yes | Class=No |
| ACTUAL CLASS | Class=Yes | ***a*** | ***b*** |
| | | (True Positive) | (False Negative) |
| | Class=No | ***c*** | ***d*** |
| | | (False Positive) | (True Negative) |

*Accuracy* is calculated as follows:

$$Accuracy = \left( \frac{a+d}{a+b+c+d} \right) x100\% \qquad (15)$$

### 5.2 Results Experiment

**Pre-processing:** In this work, the preprocessing data is conducted to clean the business performance data. To obtain good business performance data, missing values are replaced by mean value. The total business performance data is 160 records (data points).

**X-Validation**: To obtain the optimum classification model, x-validation technique is used. In this study, 10-fold validation is used.

**Classification Models**: We construct classification models to predict business performance. The five classification models, namely C4.5, logistic regression, naïve Bayes, k-NN and support vector machine models are used to predict business performance, individually. First experiment, all of the attributes in business performance data are applied as inputs for the five models. The accuracies of classification models are calculated to select the best model. The second experiment, the combination forward selection and classification models are also performed.

For C4.5 model, different criterion, namely information gain, gain ratio and gini index are applied to determine the optimum configuration. For k-NN model, we used different values of *k*, where *k*=1, 3, 5, 7, 9, 11, and 13. SVM model is employed using different kernels, namely dot, radial and polynomial.

The results of accuracy using C4.5, logistic regression, naïve Bayes, k-NN and support vector machine models to predict business performance are shown in Table 3.

*Table 3: The accuracy of the classification models (without attributes selection)*

| No | Models | Performance Accuracy (%) |
|---|---|---|
| 1 | C4.5 (Information gain) | 76.25 |
| 2 | C4.5 (Gain ratio) | 73.75 |
| 3 | C4.5 (Gini Index) | 75.00 |
| 4 | Logistic Regression | 79.38 |
| 5 | Naïve Bayes | 84.38 |
| 6 | k- Nearest Neighbor (k=1) | 80.00 |
| 7 | k- Nearest Neighbor (k=3) | 79.38 |
| 8 | k- Nearest Neighbor (k=5) | 81.25 |
| 9 | k- Nearest Neighbor (k=7) | 80.00 |
| 10 | k- Nearest Neighbor (k=9) | 81.25 |
| 11 | k- Nearest Neighbor (k=11) | 81.88 |
| 12 | k- Nearest Neighbor (k=13) | 79.38 |
| 13 | Support Vector Machine (dot) | 78.75 |
| 14 | Support Vector Machine (radial) | 75.00 |
| 15 | Support Vector Machine (polynomial) | 79.38 |

From Table 3, it is found that C4.5 model using Information gain criterion yields the better results compare to C4.5 model using gain ratio or gini index criterion. K-Nearest Neighbor algorithm using *k*=11 yields the better results compare to *k*-NN using *k* = 1, 3, 5, 7, 9, or 13. SVM model using polynomial kernel yields the better results compare to SVM using radial and dot kernels.

Figure 2 represents a comparison of accuracy values obtained using optimum C4.5, logistic regression, naïve Bayes, optimum *k*-Nearest Neighbor and optimum support vector machine

Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

models. It is seen that the Naïve Bayes gives the best result compared to the other models.

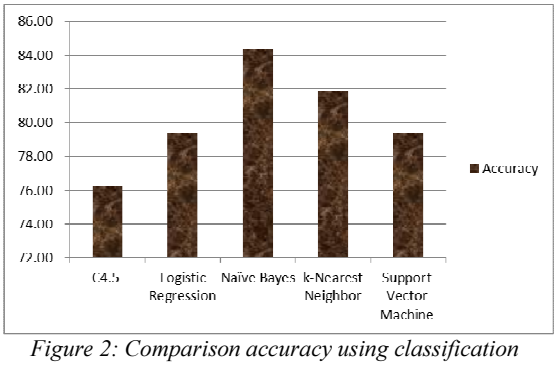


*Figure 2: Comparison accuracy using classification models for business performance*

In the proposed method, to improve the accuracy of classification is performed using combination forward selection and classification models for business performance prediction. If we carried out the appropriate selection of the independent variables, the accuracy of prediction can be improved. From the experiment results, a combination forward selection and Naïve Bayes method gives the best result compared to combination forward selection and the other classification models.

This study also shows that out of the total 19 attributes 14 attributes, namely *education, gender, experience, planning, motivation, opportunity, self-recognition, risk, ethics, adaptation, dictator, participation, delegation* and *consideration* are eliminated. Five attributes of business performance, namely *training, age, vision, innovation* and *business ability*, are used as inputs for the models.

Table 4 shows a comparison of accuracy values obtained using combination forward selection and classification models namely C4.5 using information gain criterion, logistic regression, naïve Bayes, k-Nearest Neighbor using $k$=11 and support vector machine using polynomial kernel.

*Table 4: Comparison accuracy of the classification models based forward selection*

| Models | Accuracy (%) |
|---|---|
| C4.5 + Forward Selection | 79.38 |
| Logistic Regression + Forward Selection | 83.75 |
| Naïve Bayes + Forward Selection | **88.75** |
| k-Nearest Neighbour + Forward Selection | 86.25 |
| Support Vector Machine + Forward Selection | 81.25 |

Figure 3 shows a comparison of accuracy values obtained using the proposed method (combination forward selection and classification models) and individual classification models (without forward selection). A combination forward selection and Naïve Bayes method provides an improvement of accuracy for business performance compared to the individual model. Improvement of accuracy for business performance prediction is 5.18 %.

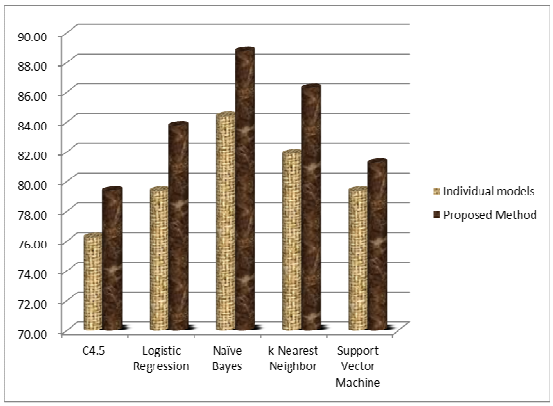


*Figure 3: Comparison accuracy of proposed method and individual classification models*

The experiments using the other attributes selection method, namely backward elimination method is also performed. The combination using backward elimination and Naïve Bayes method has been applied to business performance data set. Using this method, 2 attributes of business performance, namely *gender*, and *opportunity*, are eliminated. Seventeen attributes of business performance, namely *education, training, experience, age, vision, planning, motivation, innovation, self-recognition, risk, ethics, adaptation, dictator, participation, delegation, consideration,* and *business ability*, are used as inputs for the models. The accuracy results obtained by using 10-fold validation of business performance data set are 86.25%. It is seen that the proposed method using combination forward selection and Naïve Bayes (see Table 4) yields better accuracy result.

## 6. CONCLUSION

The proposed method to improve the accuracy of business performance prediction has been done. A data pre-processing, forward selection method for attribute selection to be combined with classification models for obtaining the optimum result is performed for business performance

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

prediction. From the experimental results show that combines forward selection and Naïve Bayes can significantly improve the accuracy of prediction compared to individual models. The accuracy improvements achieved business performance data is 5.18%. The combination forward selection and Naïve Bayes also yields better results compared to using other attribute selection method i.e. backward elimination.

**REFERENCES:**

[1] T. Tambunan, Small and Medium Enterprises in Indonesia: Some important Issues, Jakarta: Salemba Publisher, 2002.

[2] G. G. Meredith, Entrepreneurship: Theory & Practice, Jakarta: PPM Publisher, 2002

[3] R. Dalimunthe, Influence of Individual Characteristics, Entrepreneurship, Leadership Styles and Success of Business Capability Small Industrial Business Weaving and Embroidery in North Sumatra, West Sumatra and Riau, University of Airlangga, Surabaya: Dissertation, 2002.

[4] J. Raza, J. P. Liyanage, H. A. Atat and J. Lee, "A comparative study of maintenance data classification based on neural networks, logistic regression and support vector machine", *Journal of Quality in Maintenance Engineering*, Vol. 16 No. 3, pp. 303-318, 2010

[5] B.Eftekhar, K. Mohammad, H.E.Ardebili, M.Ghodsi and E.Ket bchi, 1Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical", *BMC Medical Informatics and Decision Making*, Vol. 5 No.3, 2005

[6] M.Karim, and R. M. Rahman, "Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing", Journal of Software Engineering and Applications, No. 6, pp. 196-206, 2013

[7] G. Subbalakshmi, K. Ramesh , and M. ChinnaRao, "Decision support in heart disease prediction system using naive Bayes". *Indian Journal of Computer Science and Engineering,* Vol 2 No. 2, pp. 170-176, 2011.

[8] V. Bhuyar, "Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District", *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, Vol. 3, No. 2, 2014, pp. 200-203.

[9] P. Ramachandran, N. Girija, and T. Bhuvaneswari, "Classifying blood donors using data mining techniques", *International Journal of Computer Science & Emerging Technologies*, Vol. 1 No. 1, pp. 10-13, 2011

[10] R. M. Balabin, and E. I. Lomakina, "Support vector machine regression (LS-SVM) - an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data?.", *Phys. Chem. Chem. Phys,* No. 13, pp. 11710-11718, 2011

[11] A. Widodo, and B. Yang B, "Support vector machine in machine condition monitoring and fault diagnosis", *Mechanical Systems and Signal Processing*, No. 21, pp. 2560-2574, 2007

[12] V. U. B. Challagulla, F. B. Bastani, I. Yen, and R. A. Paul, "Empirical assessment of machine learning based software defect prediction techniques*", International Journal on Artificial Intelligence Tools*, Vol. 17 No. 2, pp. 389-400, 2010

[13] W. Gauvin, C. Chen, X. Fu, and B. Liu "Classification of commercial and personal profiles on MySpace", *In Proceedings of the 3rd International Workshops on Security and Social Networking*, pp 276-28, 2011

[14] T. M. Mitchell, Machine learning, New York: McGraw-Hill, 1997

[15] J. Han and M. Kamber, Data Mining Concepts and Techiques, Second Edition, San Francisco: Elsevier, 2007

[16] D. G. Kleinbaum, M. Klein, Logistic regression a self-learning text, New York: Springer, 2002

[17] F. Markowet, L. Edler, and M. Vingron, "Support vector machines for protein fold class prediction", *Biometrical Journal* , No. 45, pp. 377-389, 2003

[18] Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou, "Support vector machines for predicting HIV protease cleavage sites in protein", *Journal of Computational Chemistry*, Vol. 23 No. 2, pp. 267-274, 2002

[19] F. Gorunescu, Data Mining: Concepts, models and techniques, *Intelligent Systems Reference Library*, Vol. 12, pp.319–330, 2011