INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue2)

Automated Diagnosis of Heart Disease using Data Mining Techniques

Prof. Priya R. Patil
Marathwada Institute of Technology

Prof. S. A. Kinariwala
Marathwada Institute of Technology

Abstract: ***The accurate diagnosis of a heart diseases, is one of the most important biomedical problems whose administration is imperative. In biomedical field, the classification of disease using data mining is the critical task. The prediction accuracy plays a vital role in disease data set. More data mining classification algorithms like decision trees, neural networks, Bayesian classifiers, Support vector machines, Association Rule, Ensemble techniques are used to diagnosis the heart diseases.***

Keywords: Data Mining, Heart Disease.

1. INTRODUCTION

An important task of any diagnostic system is the process of attempting to determine and/or identify a possible disease or disorder and the decision reached by this process. For this purpose, machine learning algorithms are widely employed. As people are generating more data everyday so there is a need for such a classifier which can classify those newly generated data accurately and efficiently. This System mainly focuses on the supervised learning technique called the Random forests for classification of data by changing the values of different hyper parameters in Random Forests Classifier to get accurate classification results. In the proposed system, the improvement of the random forests classification algorithm, which meets the aforementioned characteristics, is addressed. This is achieved by determining automatically the only tuning parameter of the algorithm, which is the number of base classifiers that compose the ensemble and affects its performance. The proposed method has some advantages over the aforementioned methods since it does not include any tuning parameter, which can be related to the number of base classifiers, such as the pre selection methods, and it does not contain an overproduction phase, such as the post selection methods; thus, it does not construct base classifiers in advance that may not be needed. The proposed method determines the members of the ensemble dynamically taking into account the combination performance of the base classifiers, in contrast to the ranking methods. It does not differentiate the members of the ensemble depending on the instance being classified and on how the neighbors of this instance were classified by the initial pool, like weighted voting methods, but it creates an ensemble that works well for all the instances. the proposed system aims to construct an ensemble with optimal accuracy and correlation. The proposed system incorporates into the termination criterion both the features that an ensemble classifier should fulfill: high accuracy and low correlation. More specifically, the construction of the forest is initiated by adding a tree. As a new tree is added each time, the new accuracy and the new correlation of the forest are computed, and an online fitting procedure is applied on the curves expressing the variation of accuracy and correlation, respectively. The procedure is terminated when the differences 1) between the curve of the accuracy and the fitted curve and 2) between the curve of the correlation and the fitted one meet a specific criterion. The aforementioned characteristics permit the proposed method to be fully integrated into any diagnostic or therapeutic

system since it improves random forests algorithm, thus, providing a classification algorithm of high performance, time, and computational effective that works independently of the medical problem and the nature of data, it can handle noisy or missing data, a common characteristic of medical datasets, and it does not require any human intervention since the only tuning parameter of the algorithm is determined automatically [1][2][3].

2. LITERATURE SURVEY

Data mining is an important and interesting field in Computer Science and has received a lot of attention from the research community particularly over the past decade. Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis [4]. Data mining uses sophisticated mathematical algorithm to segment the data and evaluate the probability of future events. Data mining is the explosion of large datasets to extract hidden and previously unknown patterns, relationships and Knowledge that are difficult to detect in traditional statistical methods. Data mining is rapidly growing successful in wide range of applications such as analysis of organic components, financial forecasting, healthcare and weather forecasting. The key properties of data mining are:

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large data sets and databases

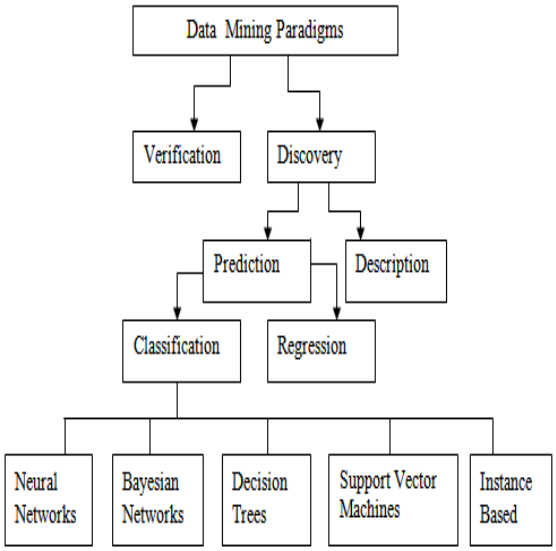

Figure 2.1: Taxonomy of Data Mining Methods

It is useful to distinguish between two main types of data mining: verification-oriented discovery-oriented. Fig 2.1 illustrates this taxonomy. Each type has its own methodology [5]. Discovery methods, which automatically identify patterns in the data, involve both prediction and description methods. Description methods focus on understanding the way the underlying data operates while prediction-oriented methods aim to build a behavioral model for obtaining new and unseen samples and for predicting values of one or more variables related to the sample. Verification methods, on the other hand, evaluate a hypothesis proposed by an external source. These methods include the most common methods of traditional statistics, like the goodness-of-fit test, the t-test of means, and analysis of variance.

2.1 Data mining Algorithm and Techniques used for Heart Disease

In Data Mining two techniques are available for the data analysis: Data Classification and Data Prediction. The Classification techniques are mainly used to predict the discrete class labels for the new observation or new data on the basis of training data set provided to the classifier algorithm, and prediction techniques generally works with the continuous valued functions [6].

2.1.1 Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified example to develop a model that can classify the population of records at large. The data classification process involves learning and classification. In Learning the training data is analyzed by classification algorithm. In classification test data is used to estimate the accuracy of the classification rules. Classification is a classical data mining technique based on machine

learning. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics.

A two step process is involved in classification.

• Model construction

• Model usage

Model construction describes a set of predetermined classes. Each sample is assumed to belong to a predefined class as determined by the class label attribute. The set of samples used for model construction: training set. The model is represented as classification rules, decision trees or mathematical formula. Model usage is used for classifying future and unknown objects. Estimate accuracy of the model. Test set should be an independent of training set, otherwise over-fitting will occur. Test set should be an independent of training set, otherwise over-fitting will occur [7]. Types of Classification models are: Classification by decision tree induction, Bayesian classification,. Neural networks, Support vector machines, Classification based on Association Rule, Classification based on ensemble techniques

2.1.2 Clustering

Clustering can be said identification of similar classes of objects. Clustering technique can further identify dense and sparse regions in object space and discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as `preprocessing approach for attribute subset
selection and classification.

2.1.3 Prediction

Regression technique can be adapted for prediction. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are to be predicted. Unfortunately many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex technique (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the classification, regression and decision tree algorithm can be used to build both classification trees (to classify categorical response variables) regression trees (to forecast continuous response variables).Neural networks too can create both classification and regression models.

2.2 Types of Classification model:

- Classification by decision tree induction
- Bayesian classification
- Neural networks
- Support vector machines
- Classification based on Association Rule
- Classification based on ensemble techniques

2.2.1 Classification by Decision Tree Induction

It is a tree in which internal nodes are represented by features, edges represent tests to be done at feature weights and leaf nodes represent categories which results from above tests. It categorizes a document by starting at the tree root and moving successfully downward via the branches (whose conditions are satisfied by the document) until a leaf node is reached. The document is then classified in the category that labels the leaf node. Decision Trees have been used in many applications in speech and language processing [8].

2.2.2 Bayesian classification

Naïve Bays Classifier

This classifier is based on the probability statement that was given by Bayes. This theorem pro-vides conditional probability of occurrence of event E_1 when E_2 has already occurred, the vice versa can also be calculated by following mathematical statement. $P(E1/E2) = \frac{P(E2/E1)P(E1)}{P(E2)}$

This basically helps in deciding the polarity of data in which opinions / reviews / arguments can be classified as positive or negative which is facilitated by collection of positive or negative examples already fed.

Naïve Bayes algorithm is implemented to estimate the probability of a data to be negative or positive. The aforesaid

probability is calculated by studying positive and negative examples & then calculating the frequency of each pole which is technically termed as learning. This learning is actually supervised as there is an existence of examples[1]. Thus, the conditional probability of a word with positive or negative meaning is calculated in view of a plethora of positive and negative examples & calculating the frequency of each of class.

2.2.3 Neural networks

Neural Network is a set of connected input/output units and each connection has a weight present with it .During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples .Neural network have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are to complex to be noticed by either humans or other computer techniques .These are well suited for continuous valued inputs and outputs. for example handwritten character recognition, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries.

2.2.4 Support Vector Machine (SVM)

The basic goal of support vector machine is to search a decision boundary between two classes that is excellently far away from any point in the training data. SVM develops a hyper planes or a set of hyper planes in infinite dimension space. This distance from decision surface to closest data point determines the margin of classifier. So the hyper planes act as decision surface which act as criteria to decide the distance of any data point from it. The margin of classifier is calculated by the distance from the closest data point. This success-fully creates a classification but a slight error will not cause a misclassification. Fig 2.2 Shows Principles of SVM [9].

Separating hyper planes

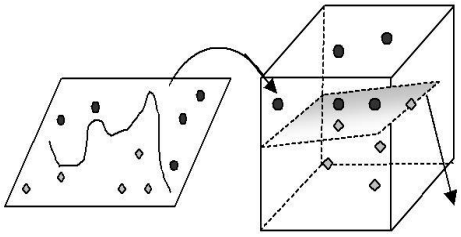

Figure 2.2: Principle of SVM

2.2.5. Classification based on Association Rule

Association and Correlation is used to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible association rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any)value. Types of association rules are:

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule
-

2.2.6. Classification based on ensemble techniques

Ensemble classification is a data mining approach that utilizes a number of classifiers that work together in order to identify the class label for unlabeled instances. Several methods of constructing and combining an ensemble of classifiers have been proposed to improve the accuracy of learning algorithm. Ensemble classification is an application of ensemble learning to boost the accuracy of classification. Ensemble learning is a machine learning paradigm where multiple models are used to solve the same problem. In ensemble classification, multiple classifiers are used and are more accurate than the individual classifiers in the ensemble [10].

2.3 Ensemble methodology

The main purpose of an ensemble methodology is to combine a set of models, each of which solves the same original problem, in order to obtain a better composite global model with more accurate and reliable estimates or decisions than can be obtained from using a single model. The main discovery is that the ensemble classifier is constructed by ensemble machine learning algorithms, such as bagging and boosting approaches, often performs much better than the single classifiers that make them up. The idea of ensemble methodology is to build a predictive model by integrating

multiple models. It is well known that ensemble methods can be used for improving prediction performance. The learning procedure for ensemble algorithms can be divided into the following two parts: Constructing base classifiers/base models [11]. The main tasks of this division are:

(i) Data processing: prepare the input training data for building base classifiers and attributes selection to reduce the dimensionality of the attributes.

(ii) Base classifier constructions: build base classifiers on the data set with a learning algorithm.

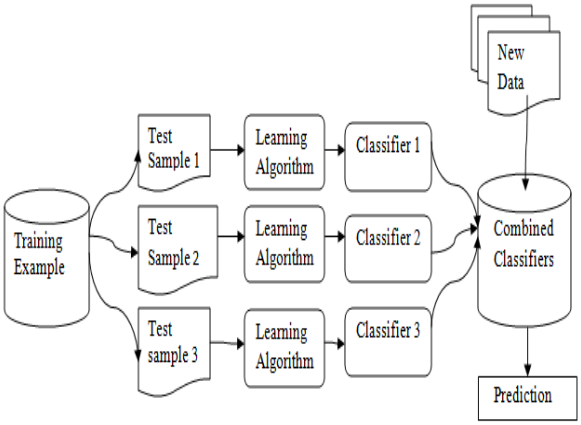

Figure 2.3: Ensemble Classifier

Fig 2.3 shows Ensemble of classifiers. It has been proved to be very effective way to improve classification accuracy because uncorrelated errors made by a single classifier can be removed by voting. A classifier which utilizes a single minimal set of classification rules to classify future examples may lead to mistakes. An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way to classify new example. Many research results illustrated that such multiple classifiers, if appropriately combined during classification, can improve the classification accuracy. Research in ensemble methods has largely revolved around designing ensembles consisting of competent yet complementary models.

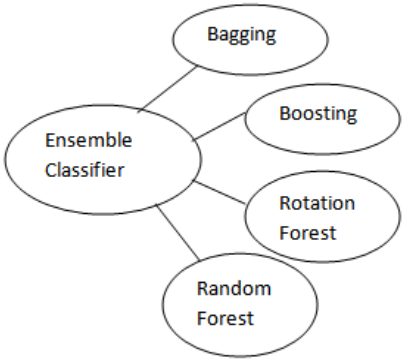

Figure 2.4 : Variants of Ensemble Classifiers

Fig 2.4 shows Three widely used ensemble approaches, namely, boosting, bagging, and Random Forest.

2.3.1 Boosting:

Boosting is an incremental process of building a sequence of classifiers, where each classifier works on the incorrectly classified instances of the previous one in the sequence.

Bagging is based on allowing each base classifier to be trained with different random subset of training set with the goal of bringing diversity in the base classifier. AdaBoost (Freund & Schapire, 1997) is the representative of this class of techniques. However, AdaBoost is prone to overfitting. Fig 2.5 shows Boosting technique for combining multiple base

classifiers whose combined performance is significantly better than that of any of the base classifiers.

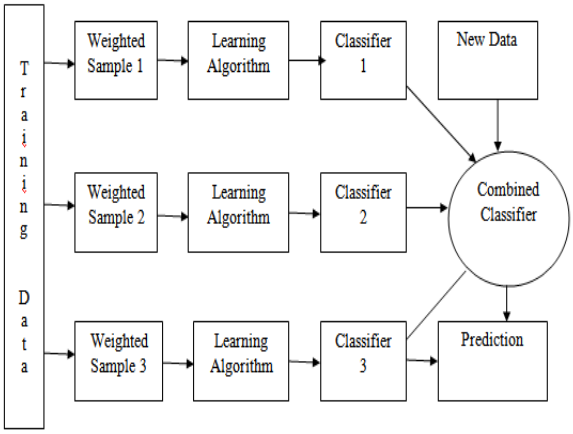

Figure 2.5 : Boosting

2.3.2 Bagging

The other class of ensemble approaches is the Bootstrap Aggregating (Bagging) (Breiman, 1996a). Bagging involves building each classifier in the ensemble using a randomly drawn sample of the data, having each classifier giving an equal vote when labeling unlabeled instances. Bagging is known to be more robust than boosting against model overfitting. RF is the main representative of bagging (Breiman, 2001). Fig 2.6 shows Bagging technique which combines predictions of independent base classifiers for arriving at final prediction. Bagging works because if a learning algorithm (i.e. decision tress) is unstable a small change in training set causes large changes in the learned classifier and Bagging always.

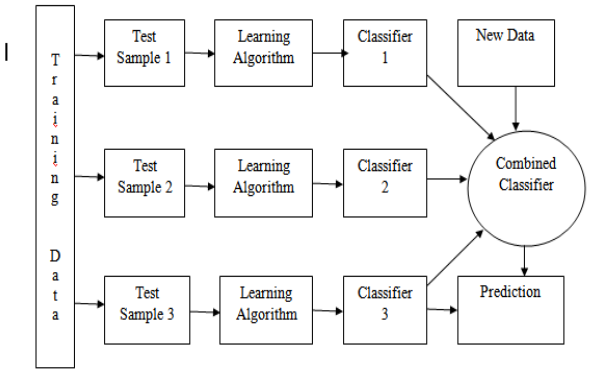

Figure 2.6: Bagging

2.3.3 Random forest:

Random Forest is essentially an ensemble of unpruned classification trees. It gives excellent performance on a number of practical problems, largely because it is not sensitive to noise in the data set, and it is not subject to overfitting. It works fast, and generally exhibits a substantial performance improvement over many other tree-based algorithms.

Random forests are built by combining the predictions of several trees, each of which is trained in isolation. Unlike in boosting (Schapire & Freund, 2012) where the base models are trained and combined using a sophisticated weighting scheme, typically the trees are trained independently and the predictions of the trees are combined through averaging. There are three main choices to be made when constructing a random tree. These are

- The method for splitting the leafs.
- The type of predictor to use in each leaf.
- The method for injecting randomness into the trees.

. In Brieman's early work each individual tree is given an equal vote and later version of Random Forest allows weighted and unweighted voting [12]. The technique on which Random Forest ensemble is formed can be considered over following parameters:

i) Base Classifier: It describes the base classifier used in the Random Forest ensemble. Base classifier can be decision tree, Random tree, or extremely randomized tree.

ii) Split Measure: If base classifier of Random Forest is decision tree, then which split measure is found at each node of the tree to perform the splitting. To perform splitting Gini index, Info gain etc are used.

iii) Number of Passes: For building Random Forest classifier, if single pass is sufficient or multiple passes through data are needed.

iv) Combine Strategy: In Random Forest ensemble, all the base classifiers generated are used for classification. At the time of classification, how the results of individual base classifiers are combined is decided by the combine strategy.
v) Number of attributes used for base classifier generation: This parameter gives the number of how many attributes are to be used which are randomly selected from the original set of attributes at each node of the base decision tree. Filter and Wrapper these are main techniques used for feature selection and extraction [13].

- Each tree of Random Forest is grown, are described as follows:

Suppose training data size containing N number of records, then N records are sampled at random but with replacement, from the original data, this is known as bootstrap sample along with M number of attributes. This sample will be used for the training set for growing the tree. If there are N input variables, a number n << N is selected such that at each node, n variables are selected at random out of N and the best split on these m attributes is used to split the node. The value of m is held constant during forest growing. The decision tree is grown to the largest extent possible. A tree forms "inbag" dataset by sampling with replacement member from the training set. It is checked whether sample data is correctly classified or not using out of bag error with the help of out of bag data which is normally one third of the "inbag" data. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble.

2.4 Summary of developments to RFs

Ho (1995) proposed a method to overcome a fundamental limitation on the complexity of decision tree classifiers derived with traditional methods.The proposed method uses oblique decision trees which are convenient for optimizing training set accuracy.

Amit and Geman (1997) proposed a shape recognition approach based on the joint induction of shape features and tree classifiers.

Ho (1998) proposed a method to solve the dilemma between overfitting and achieving maximum accuracy.

Breiman (2001) Proposed a RF ensemble learning method used for classification and regression.

Latinne et al. (2001) Used McNemar non-parametric test of significance to a priori limit the number of trees that will participate in majority voting and without loss in accuracy.

Robnik-Šikonja (2004) Decreased correlation between trees by using several attribute evaluation measures. Used weighted voting instead of majority voting. Tsymbal et al. (2006) Replaced majority voting with more sophisticated dynamic integration techniques: DS, DV, and DVS

Amaratunga et al. (2008) Improved the declining performance when the number of features is large and the number of truly informative features is small by using weighted random sampling instead of simple random sampling when picking features to split each node.Saffari et al. (2009) Introduced a novel online RF algorithm to remedy the limitations of the off-line algorithm.Bader-El-Den and Gaber (2012) Used genetic algorithms to boost the performance of RF.

Xu et al. (n.d.) Proposed a hybrid RF approach for classifying very high-dimensional data that outperformed the traditional RF [14][15].

2.5 Conclusion from Literature Survey

Supervised learning algorithms are commonly described as performing the task of searching through a hypothesis space to fine a suitable hypothesis that will make good predictions with a particular problem. an ensemble is a technique for combining many weak learners in an attempt to produce a strong learner. Evaluating the prediction of an ensemble typically requires more computation than evaluating the prediction of a single model, so ensemble may be thought of as a way to compensate for poor learning algorithms by performing a lot of extra computation. For example fast algorithms such as decision tree sometime have relatively poor accuracy compared to other knowledge models like neural networks. In order to overcome this problem, a large number of decision trees are generated for the same data set, and used simultaneously for prediction. Random forest is one such ensemble based method which is commonly used with decision trees.

There are often two main criticism of ensemble based classification research; the dearth of publicly available real data to perform the experiments on; and the lack of published well researched methods and techniques. To counter both of them, this paper garners all related literature for categorization ad comparison, selects some innovative methods and techniques for discussion; and point towards other data sources as possible alternatives.

There are basically two motivations behind building a ensemble of classifier.

i. Reduced variance: Results are less dependent on the peculiarities of a single training set.

ii. Reduced bias: A combination of multiple classifiers may learn a more expressive concept class than a single classifier.

REFERENCES

[1] Evanthia E. Tripoliti, "Automated Diagnosis of Diseases Based on Classification: Dynamic Determination of the Number of Trees in Random Forests Algorithm", IEEE, 2012

[2] Jehad Ali1, Rehanullah Khan2, Nasir Ahmad3, Imran Maqsood4 "Random Forests and Decision Trees" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, 2012

[3] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective", Artif. Intell. Med., vol. 23, no. 1, pp. 89–109, 2001.

[4] G. D. Magoulas and A. Prentza, "Machine learning in medical applications", Mach. Learning Appl. Berlin/Heidelberg, Germany: Springer, vol. 2049, pp. 300–307, 2001.

[5] Jiawai Han, Micheline Kamber ,"Data Mining:Concepts and Techniques",2nd Edition, Elsevier, 2006

[6] Mr. Hitesh H. Parmar, Prof GloryH. Shah, "Experimental and Comparative Analysis of Machine Learning Classifiers",ISSN:22 77 128X,volume 3.Issue 10, 2013

[7] P. Deepika, P. Vinothini, "Heart Disease Analysis And Prediction Using Various Classification Models-A survey",ISSN-2250-1991,Volume:4,Issue:3,Mar2015 [8] Simon Bernard, Laurent Heutte, Sehastien Adam, "Forest-RK: A New Random Forest Induction Method", ICIC (2), Springer, pp.430-437, Lecture Notes in Computer Science, vol. 5227, 2009

[9] Shashikant Ghumbre, Chetan Patil, Ashok Ghatol, "Heart Disease Diagnosis using Support Vector Machine", ICCSIT, 2011

[10] Robert E. Banfield, Lawrence O. Hall, Kevin W. Bowyer, Divya Bhadoria, "A Comparison of Ensemble Creation Techniques", Fifth International Conference on Multiple Classifier System, 2004

[11] Abdelhmid Salih Mohmed Salih1 and Ajith Abraham, "Novel Ensemble Decision Support and Health Care Monitoring System", Journal of Network and Innovative Computing ISSN 2160-2174, Volume 2, 2014

[12] Sarika Pachange, Bela Joglekar, "Random Forest approach for characterizing Ensemble Classifiers", ISSN 2348-4853, 2014

[13] A Sheik Abdullah, R.R. Rajalaxmi, "A Data Mining Model for Predicting The Coronary Heart Disease using Random Forest Classifier", ICON3C, IJCA, 2012

[14] Khaled Fawagreh, Mohamed Medhat Gaber & Eyad Elyan, " Random forests: from early developments to recent advancements", ISSN, 2014

[15] Gayathri P. N. Jaisankar, "Comprehensive Study of Heart Disease Diagnosis using Data Mining and Soft Computing Techniques", IJET, vol 5 No 3, 2013

