Volume 5, Issue 4, April 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Conceptual Study of Data Mining with Special Reference to Breast Cancer

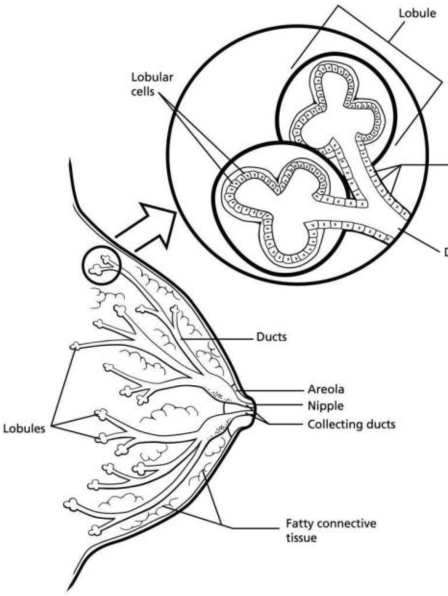
Dr. Shivaji D. Mundhe*, **Prof. Sunil Joshi, Prof. Krishna Priya S**
Sinhgad Institute Of Management & Computer Application (MCA)
STES Technical Campus, Building No. 5, Narhe Top, Narhe, Pune, India

Abstract—Breast cancer is considered a major health problem in women. India is a growing breast cancer epidemic with an increasing number of younger women becoming susceptible to the disease. A new global study estimates that by 2030, the number of new cases of breast cancer in India will increase from the current 115,000 to around 200,000 per year. According to Globocan data (International Agency for Research on Cancer), India is on top of the table with 1.85 million years of healthy life lost due to breast cancer. The study confirmed conclusions from earlier research: that breast cancer is now the second most common cancer diagnosed in Indian women after cervical cancer. Studies have also shown that Indian women develop breast cancer roughly a decade earlier than women in western countries. Poor survival may be largely explained by lack of or limited access to early detection services and treatment. It is one of leading causes of death in the world (World Health Organization, 2010) and the second cause of death of women in United States (National Cancer Institute, 2010 According to the extensive database of International Agency for Research on Cancer (IARC), in 2000, more than one million people around the world were diagnosed with breast cancer and about one-third of women died from the disease, despite that it can be cured at early stages. In this paper the researcher has theoretically explained DMT in detection of Breast Cancer.

Keywords—Data Mining Techniques (DMT), Breast Cancer, R-Programming techniques, SVM, Ada Boost Model, Random Forest Model

I. INTRODUCTION

The body is made up of trillions of living cells. Normal body cells grow, divide into new cells, and die in an orderly way. During the early years of a person's life, normal cells divide faster to allow the person to grow. After the person becomes an adult, most cells divide only to replace worn-out, damaged, or dying cells. Cancer begins when cells in a part of the body start to grow out of control. There are many kinds of cancer, but they all start because of this out-of-control growth of abnormal cells.



Source: www.cancer.org Fig 1**:** Normal parts of the breasts

Breast cancer is a malignant (cancer) tumour that starts in the cells of the breast. It is found mostly in women, but men can get breast cancer, too. A woman's breast is made up of glands that can make breast milk (lobules), small tubes that carry milk from the lobules to the nipple(ducts), fatty and connective tissue, blood vessels, and lymph vessels. Most breast cancers begin in the cells that line the ducts. Fewer breast cancers start in the cells lining the lobules Cancers can also start in cells of the other tissues in the breast. To understand breast cancer, it helps to know something about the normal parts of the breasts, as shown in the figure 1.Most women have more than one known risk factor for breast cancer, yet will never get the disease. The most common risk factors for breast cancer is not only being female and growing older. There may be more than one cause of breast cancer.

These may be:

- Being a woman
- Getting older
- Having an inherited mutation in the BRCA1 or BRCA2 breast cancer gene
- Lobular carcinoma in situ (LCIS)
- A personal history of breast or ovarian cancer
- A family history of breast, ovarian or prostate cancer
- having high breast density on a mammogram
- having a previous biopsy showing atypical hyperplasia
- Starting menopause after age 55
- Never having children
- having your first child after age 35
- Radiation exposure, frequent X-rays in youth
- High bone density
- being overweight after menopause or gaining weight as an adult
- Postmenopausal hormone use (current or recent use) of estrogen or estrogen plus progestin

II. OBJECTIVES OF THE STUDY

1. To analyse the usage of Data Mining Techniques in Medicine.
2. Introduction of Data Mining Techniques in detection of Breast Cancer.

III. SIGNIFICANCE OF DATA MINING IN MEDICINE

Hospitals are gathering enormous amount data daily. This data is valuable source of medical information that has potential to be very useful in diagnosing and treatment. This data may comprise thousands of records which may contain valuable patterns and dependencies hidden deep among them. The volume of the dataset and complexity of the medical domain make it very difficult for a human to analyse the data manually to extract hidden information. Various data mining algorithms have been developed which analyse the data in order to extract underlying knowledge.

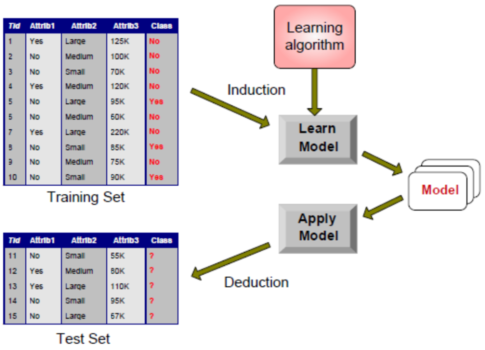
IV. DATA MINING TECHNIQUES IN DETECTION OF BREAST CANCER

The fast -growing, tremendous amount of data, collected and stored in large and numerous repositories, has far exceeded the human ability for comprehension without powerful tools, resulted in a 'data rich but information poor' situation. Data mining also called Knowledge Discovery in Databases (KDD) has attracted a great deal of attention in the IT industry and in society as whole in recent years.

Data mining uses a combination of an explicit knowledge base, sophisticated analytical skills, and domain knowledge to uncover hidden trends and patterns. These trends and patterns form the basis of predictive models that enable analysts to produce new observations from existing data. Gartner Inc.'s definition of data mining is the most comprehensive: "…the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories, and by using pattern recognition technologies, as well as statistical and mathematical techniques." Data mining should be performed on very large or raw datasets using either supervised or unsupervised data mining algorithms. There are four major tasks in data mining like Data Classification, Clustering, Association and Prediction.

V. CLASSIFICATION

Classification is a two-step process consisting of learning, or model construction (where a model is constructed based on class-labelled tuples from a training data set), and classification, or model usage (where the model is used to predict the class labels of tuples from new data sets).A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it. The class label must be a discrete attribute. This is a key characteristic that distinguishes classification from regression.

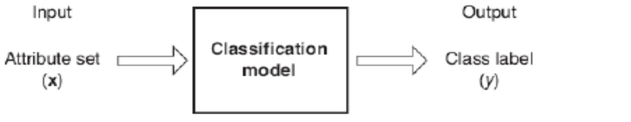


Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
5	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Classification as the task of mapping an input attribute set $\mathbf{x}$ into its class label y.

Fig2. Illustration of classification task

The research utilizes 4 data mining algorithms for classification.

- Support Vector Machine
- Decision Tree Induction
- Ada Boost algorithm
- Random Forest algorithm

They have been all implemented in the R programming packages and that is the reason why R programming Language is chosen for the research. The following subsections describe the data mining methods in detail.

i) Support Vector Machines (SVM)

SVM is a useful technique for data classification. The Support Vector Machine (SVM) is a state-of-the-art classification method introduced in 1992 by Boser, Guyon, and Vapnik. The SVM classier is widely used in bioinformatics (and other disciplines) due to its high accuracy, ability to deal with high-dimensional data such as gene expression and flexibility in modeling diverse sources of data [46]. SVMs belong to the general category of kernel methods. The SVM technique is illustrated in the following figure.

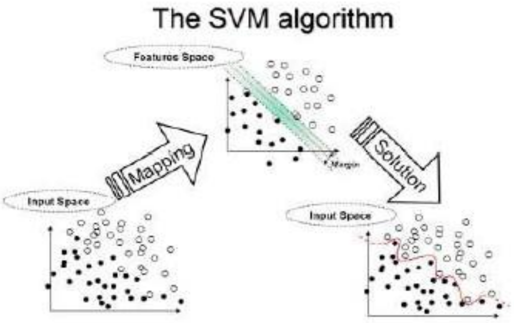


Fig. 3 Illustration of SVM Algorithm

A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one target value" (i.e. the class labels) and several attributes" (i.e. the features or observed variables). The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes.

Given a training set of instance-label pairs $(\mathbf{x}_i, y_i), i = 1, \ldots, l$ where $\mathbf{x}_i \in R^n$ and $\mathbf{y} \in \{1, -1\}^l$, the support vector machines (SVM) [1,6] require the solution of the following optimization prooiem:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

Here training vectors $\mathbf{x}_i$ are mapped into a higher (maybe infinite) dimensional space by the function ϕ. SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. $C > 0$ is the penalty parameter of the error term. Furthermore, $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ is called the kernel function.

The four basic kernels are

- linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$.
- polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma\mathbf{x}_i{}^T\mathbf{x}_j + r)^d$, $\gamma > 0$.
- radial basis function (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, $\gamma > 0$.
- sigmoid: $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma\mathbf{x}_i{}^T\mathbf{x}_j + r)$.

Here, γ, r, and d are kernel parameters.

ii) Decision Tree Induction

A decision tree model is one of the most common data mining models. It is popular because the resulting model is easy to understand. Decision tree inducers are algorithms that automatically construct a decision tree from a given dataset. The algorithms use a recursive partitioning approach.

The decision tree is a structure that includes root node, branch and leaf node. Each internal node denotes a test on attribute, each branch denotes the outcome of test and each leaf node holds the class label. The topmost node in the tree is the root node.

In this algorithm there is no backtracking, the trees are constructed in a top down recursive divide-and-conquer manner. There are various top–down decision trees inducers such as ID3, C4.5and CART. Some consist oftwo conceptual phases: growing and pruning (C4.5 and CART). Other inducers perform only the growing phase.

In most of the cases, the discrete splitting functions are univariate. Uni-variate means that an internal node is split according to the value of a single attribute. Consequently, the inducer searches for the best attribute upon which to split. There are various univariate criteria. These criteria can be character- ized in different ways, such as:

- According to the origin of the measure: information theory, dependence, and distance.
- According to the measure structure: impurity based criteria, normalized impurity based criteria and Binary criteria.

Information gain is an impurity-based criterion that uses the entropy measure (origin from information theory) as the impurity measure.

$$InformationGain(a_i, S) =$$
$$Entropy(y, S) - \sum_{v_{i,j} \in dom(a_i)} \frac{|\sigma_{a_i=v_{i,j}} S|}{|S|} \cdot Entropy(y, \sigma_{a_i=v_{i,j}} S)$$

where:

$$Entropy(y, S) = \sum_{c_j \in dom(y)} -\frac{\left|\sigma_{y=c_j} S\right|}{|S|} \cdot \log_2 \frac{\left|\sigma_{y=c_j} S\right|}{|S|}$$

Gini index is an impurity-based criterion that measures the divergences between the probability distributions of the target attribute's values and it is defined as

$$Gini(y, S) = 1 - \sum_{c_j \in dom(y)} \left(\frac{\left|\sigma_{y=c_j} S\right|}{|S|}\right)^2$$

Consequently the evaluation criterion for selecting the attribute ai is defined as:

$$GiniGain(a_i, S) = Gini(y, S) - \sum_{v_{i,j} \in dom(a_i)} \frac{\left|\sigma_{a_i=v_{i,j}} S\right|}{|S|} \cdot Gini(y, \sigma_{a_i=v_{i,j}} S)$$

Decision tree induction is a nonparametric approach for building classification models. It doesn't require any aprior assumptions regarding the type of probability distributions satisfied by the class and other attributes. Finding an optimal decision tree is an NP-complete problem. Many decision tree algorithms employ a heuristic-based approach to guide their search in the vast hypothesis space. Constructing decision tree techniques are computationally inexpensive, making it possible to quickly construct models even when the training set size is very large. Once a decision tree has been built, classifying a test record is extremely fast, with a worst-case complexity of O(w), where w is the maximum of the tree depth.

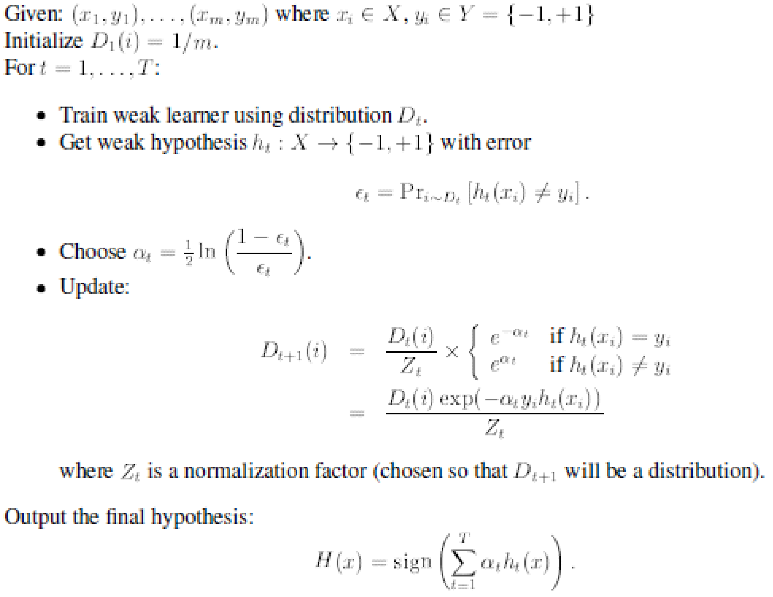
iii) Ada Boost Model

Boosting is a general method for improving the accuracy of any given learning algorithm. This is a widely used and powerful prediction technique that sequentially constructs an ensemble of weak classifiers. A weak classifier is a very simple model that has just slightly better accuracy than a random classifier, which has 50% accuracy on the training data set. The set of weak classifiers is built iteratively from the training data over hundreds or thousands of iterations. At each iteration or round, the examples in the training data are reweighted according to how well they are classified (larger weights given to is classified examples). Weights are computed for the weak classifiers based on their classification accuracy. The weighted predictions from the weak classifiers are combined using voting to compute a final prediction of the outcome

AdaBoost is an algorithm for constructing a "strong" classifier as linear combination of "simple" "weak" classifiers ht(x).

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

AdaBoost (Adaptive Boosting) is a powerful classifier that works well on both basic and more complex recognition problems. AdaBoost works by creating a highly accurate classifier by combining many relatively weak and inaccurate classifiers. AdaBoost therefore acts as a meta algorithm, which allows you to use it as a wrapper for other classifiers. AdaBoost is adaptive in the sense that subsequent classifiers added at each round of boosting are tweaked in favor of those instances misclassified by previous classifiers. The default number of boosting rounds for AdaBoost is 20, however this can easily be set using the



Given: $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \in X$, $y_i \in Y = \{-1, +1\}$

Initialize $D_1(i) = 1/m$.

For $t = 1, \ldots, T$:

- Train weak learner using distribution D_t.
- Get weak hypothesis $h_t : X \to \{-1, +1\}$ with error

$$\epsilon_t = \Pr_{i \sim D_t}\left[h_t(x_i) \neq y_i\right].$$

- Choose $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$.
- Update:

$$\begin{aligned} D_{t+1}(i) &= \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases} \\ &= \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t} \end{aligned}$$

where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis:

$$H(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$

AdaBoost is a powerful classification algorithm that has enjoyed practical success with applications in a wide variety of fields, such as biology, computer vision, and speech processing. Unlike other powerful classifiers, such as SVM, AdaBoost can achieve similar classification results with much less tweaking of parameters or settings (unless of course you choose to use SVM with AdaBoost). The user only needs to choose: (1) which weak classifier might work best to solve their given classification problem; (2) the number of boosting rounds that should be used during the training phase. The AdaBoost algorithm will select the weak classifier that works best at that round of boosting.

iv) Random Forest Model

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark. The term came from random decision forests that was first proposed by Tin Kam Ho of Bell Labs in 1995. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho and Amit and Geman in order to construct a collection of decision trees with controlled variance.

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, \ldots, x_n$ with responses $Y = y_1, \ldots, y_n$, bagging repeatedly selects a bootstrap sample of the training set and fits trees to these samples:

For $b = 1, \ldots, B$:

1. Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.
2. Train a decision or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x')$$

or by taking the majority vote in the case of decision trees.

In the above algorithm, B is a free parameter. Typically, a few hundred to several thousand trees are used, depending on the size and nature of the training set. Increasing the number of trees tends to decrease the variance of the model, without increasing the bias. As a result, the training and test error tend to level off after some numbers of trees have been fit. An optimal number of trees B can be found using cross-validation, or by observing the out-of-bag error: the mean prediction error on each training sample x_i, using only the trees that did not have x_i in their bootstrap sample.

The above procedure describes the original bagging algorithm for trees. Random forests differ in only one way from this general scheme: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. This process is sometimes called "feature bagging". The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the B trees, causing them to become correlated. Typically, for a dataset with p features, $\sqrt{p}$ features are used in each split.

VI. CONCLUSION

The main goal medical data mining algorithm is to get best algorithms that describe given data from multiple aspects. In this research paper researcher has analyse the usage of Data Mining Techniques in Medicine and Introduction of Data Mining Techniques in detection of Breast Cancer. In series of this research researcher is going to apply DMT on Datasets using R-Programming to get the result.

ACKNOWLEDGMENT

I would like to thank my colleagues Prof. Krishna Priya S and Prof. Sunil Joshi, who helped me in making this research paper.

REFERENCES

[1] G. Ravi Kumar et al , "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques", International Journal of Innovations in Engineering and Technology (IJIET)

[2] K.R.Lakshmi et al , " PERFORMANCE COMPARISON OF DATA MINING TECHNIQUES FOR PREDICTION AND DIAGNOSIS OF BREAST CANCER DISEASE SURVIVABILITY", Asian Journal Of Computer Science And Information Technology 3 : 5 (2013) 81 - 87. Contents lists available at www.innovativejournal.in Asian Journal of Computer Science And Information Technology Journal Homepage: http://www.innovativejournal.in/index.php/ajcsit ISSN 2249 - 5126

[3] Mohammad Taha Khan et al , "A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining", International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11 (2012) © Research India Publications; http://www.ripublication.com/ijaer.htm

[4] VikasChaurasia&Saurabh Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering *(An ISO 3297: 2007 Certified Organization)* Vol. 2, Issue 1, January 2014 Copyright to IJIRCCE www.ijircce.com ISSN (ONLINE) : 2320 – 9801, ISSN (Print) : 2320 -9798

[5] G.SUJATHA &K. USHA RANI, "A SURVEY ON EFFECTIVENESS OF DATA MINING TECHNIQUES ON CANCER DATA SETS", Vol 04, Special Issue 01, 2013 International Journal of Engineering Sciences Research-IJESR http://ijesr.in/ ACICE-2013 ISSN: 2230-8504; e-ISSN-2230-8512

[6] Reda Al-Bahrani et al , "Colon cancer survival prediction using ensemble data mining on SEER data"

[7] K.Rajesh&Dr. Sheila Anand, "Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm", *International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 2, April 2012* Copyright to IJARCCE www.ijarcce.com ISSN : 2278 -1021

[8] AbdelghaniBellaachia&ErhanGuven, "Predicting Breast Cancer Survivability Using Data Mining Techniques"

[9] Mariammal.D et al ,"Major Disease Diagnosis and Treatment Suggestion System Using Data Maining Techniques", International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014) © 2014, IJARCST All Rights Reserved 338 Vol. 2 Issue Special 1 Jan-March 2014, ISSN (Online) : 2347 – 8446, ISSN(Print) : 2347 -9817

[10] SHELLY G.et al, "DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS", Indian Journal of Computer Science and Engineering (IJCSE)

[11] S.Kharya&D. Dubey , S. Soni, "Predictive Machine Learning Techniques for Breast Cancer Detection", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6) , 2013, 1023-1028 ISSN : 0975 - 9646

[12] S. S.Shajahaan et al , "Application of Data Mining Techniques to Model Breast Cancer Data", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal,Volume 3, Issue 11, November 2013)