Journal of Computer Science

Original Research Paper

Performance Analysis of Classification Algorithms on Medical Diagnoses-a Survey

Vanaja, S. and K. Rameshkumar

Research Scholar and Research guide, Research and Development, Bharathiar University, Coimbatore, Tamil Nadu, India

Article history
Received: 02-12-2013
Revised: 28-12-2013
Accepted: 25-07-2014

Corresponding Author:
Vanaja, S.,
Research Scholar and Research guide, Research and Development, Bharathiar University, Coimbatore, Tamil Nadu, India
Email: vanajasha@yahoo.com

Abstract: The aim of this research paper is to study and discuss the various classification algorithms applied on different kinds of medical datasets and compares its performance. The classification algorithms with maximum accuracies on various kinds of medical datasets are taken for performance analysis. The result of the performance analysis shows the most frequently used algorithms on particular medical dataset and best classification algorithm to analyse the specific disease. This study gives the details of different classification algorithms and feature selection methodologies. The study also discusses about the data constraints such as volume and dimensionality problems. This research paper also discusses the new features of C5.0 classification algorithm over C4.5 and performance of classification algorithm on high dimensional datasets. This research paper summarizes various reviews and technical articles which focus on the current research on Medical diagnosis.

Keywords: Classification, Medical Diagnosis, C5.0, High Dimensional Dataset

Introduction

Data mining is the process of discovering actionable information from large sets of data. Data mining uses mathematical analysis to derive patterns and trends that exist in data. Typically, these patterns cannot be discovered by traditional data exploration because the relationships are too complex or because there is too much data. These patterns and trends can be collected and defined as a data mining model. Building a mining model is part of a larger process that includes everything from asking questions about the data and creating a model to answer those questions, to deploying the model into a working environment. This process can be defined by using the following six basic steps:

- Defining the problem
- Preparing data
- Exploring data
- Building models
- Exploring and validating models
- Deploying and updating models

The Fig. 1 describes the relationships between each step in the process.

The steps illustrated in the diagram is circular, each step does not necessarily lead directly to the next step. Creating a data mining model is a dynamic and iterative process. After explore the data, if the data is insufficient to create the appropriate mining models, more data is to be added. The redefining the problem and updating of the models is carried out after they have been deployed because more data has become available. Each step in the process might need to be repeated many times in order to create a good model.

Classification is one of the data mining methodologies used to predict and classify the predetermined data for the specific class. There are different classifications methods propose by researchers. The basic methods are given by (Han and Kamber, 2001):

- Bayesian classification (Statistical classifier)
- Decision tree induction
- Rule based classification (IF THEN Rule)

- Classification using back propagation (Neural network algorithm)
- Support Vector Machine (for linear and non-linear data)
- Classification using Association Rule
- Lazy learners
- K-nearest neighbor classifier
- Case-based reasoning classifier
- Other classification algorithms
- Rough set approach
- Genetic algorithm
- Fuzzy set approach

Any one of the above mentioned classification techniques can be applied to classify the application oriented data. The appropriate classification method is to be chosen according to the type of application and the dimensionality of the data.

It is a very big challenge to the researcher to apply the appropriate data mining classification algorithm to diagnose medical related problems. Choosing the correct method is a challenging task. The exact method can be chosen only after analysing all the available classification methods and checking its performance in term of accuracy. Various researches have been carried out in the area of medical diagnoses by using classification methodology. The most important fact in medical diagnosis system is the accuracy of the classifier. This research paper analyses the different classification methods applied in medical diagnoses and compares the performance of classification accuracy.

This study is organized in such a way that the section 2 and 3 discussed about the preprocessing of data and introduction of the classification. Section 4 and 5 discuss about the different classification algorithms and the problem statement. The section 6 gives the details about the binary and multiclass. Section 7 states the impact of feature selection methods on classification algorithms, section 8 and 9 discuss about the various kinds of medical dataset, related works and its performance analysis. The future enhancements on classification algorithm on medical dataset, conclusion is given in section 10 and 11.

Pre-Processing

The data collected from the real world is incomplete, inconsistent, inadequate and it consisting of noise, redundant groups. The Knowledge discovery using the training data with such an irrelevant, inconsistent and redundant data will reduce the mining quality.

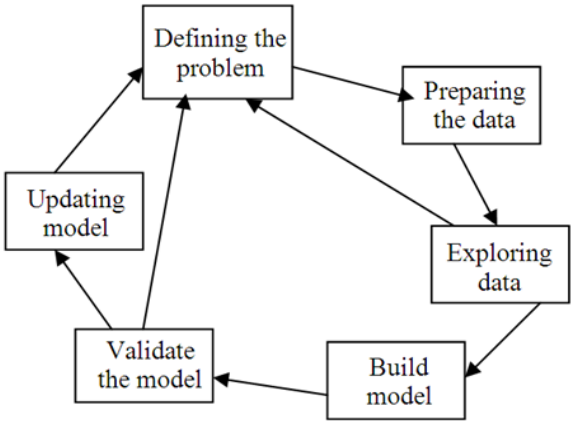


Fig. 1. Relationship between data mining steps

To improve the quality of mining, the data preprocessing techniques are applied. The main objective of data preprocessing is cleaning of noise, filling up of missing values, reduce the redundancy and normalize the data. The preprocessing steps are data cleaning, data integration, data transformation, data reduction and data discretization. After preprocessing has been done the data will be complete, noise free and ready for classification. Any classification algorithm can be applied for classifying the data.

Classification

Classification is one of the data analysis used to predict the categorical data. Classification is a two phase process. The Training phase and the Testing phase. In training phase the pre determine data and the associated class label are used for classification. The tuples used in training phase is called training tuples. This is also known as supervised learning. Figure 2 shows the Training and Testing Phases of the classification.

In Testing step the Test data tuple are used to estimate the accuracy of the classification rule. If the accuracy of the classifier rule on test data is acceptable the rule can be applied for further classification of unseen data.

Classification Steps

The Real world data is classified using number of steps. Initially the dataset is preprocessed to remove the noise and fill the missing values. The preprocessing step make the dataset be ready for classification. Figure 3 depicts the processing steps for classification.

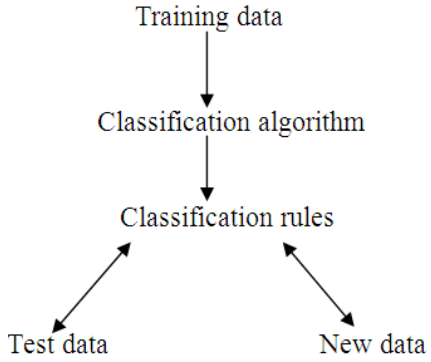


Fig. 2. Training and testing phases

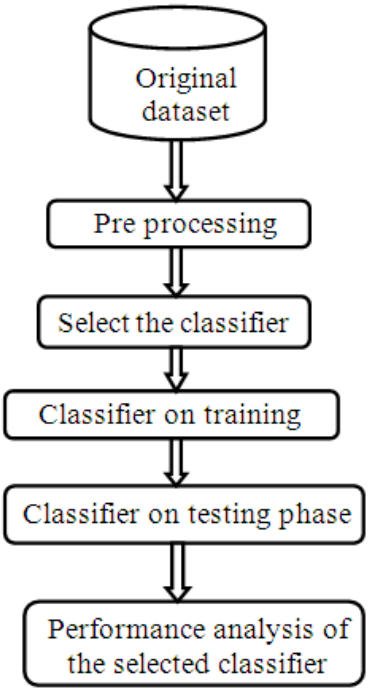


Fig. 3. Classification steps

Existing Algorithms

Decision Tree Algorithm

The decision tree is one of the classification algorithm. It is frequently used by the researchers to classify the data. The decision tree is very popular because it is easy to build and require less domain knowledge. Also the decision tree method is scalable for large database. The first decision tree algorithm is developed in early 1980s is Iterative Dichotomiser (ID3). Quinlan and Kaufmann (1993) presents the C4.5 which is the successor of ID3. In 1984 Classification And Regression Tree (CART) is introduced. It is mainly support for the binary tree classification. All the three algorithms adopt the greedy approach and construct the decision tree in top down, recursive, divide and conquer manner.

The following are the basic steps used for decision tree algorithm:

Input: Data partition, D, which is a set of training tuples and their associated class labels;

Attribute list: The set of candidate attributes;

Attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset.

Output: A decision tree:

(1) Create a node N
(2) If tuples in 'D' are belongs to same class C, then return N as a leaf node labeled with the class C
(3) If attribute list is empty then return N as a leaf node labeled with the majority class in D;
(4) Apply Attribute selection method (D, attribute list) to find the "best" splitting criterion
(5) label node N with splitting criterion
(6) If splitting attribute is discrete-valued and multiday splits allowed then//not restricted to binary trees
(7) Attribute list ← attribute list-splitting attribute; //remove splitting attribute
(8) For each outcome j of splitting criterion//partition the tuples and grow subtrees for each partition
(9) Let Dj be the set of data tuples in D satisfying outcome j;//a partition
(10) If Dj is empty then attach a leaf labeled with the majority class in D to node N
(11) Else attach the node returned by Generate decision tree (Dj, attribute list) to node N; endfor
(12) Return N

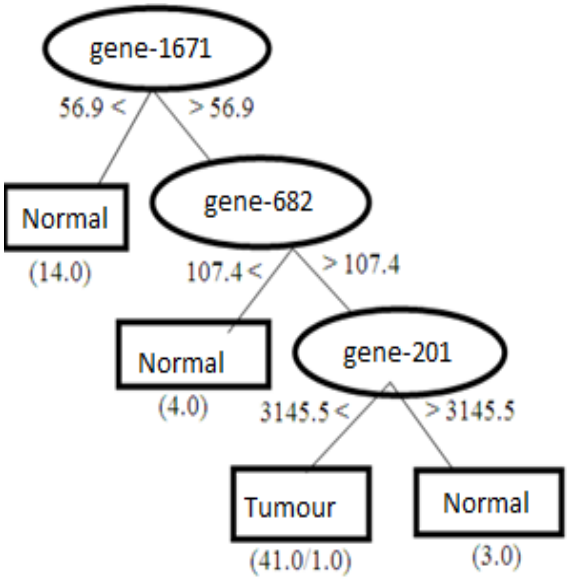
The decision tree algorithm is very robust and learning efficiency with its learning time complexity of $O(n\log_2 n)$. The outcome of a decision tree that can be easily represented as a set of symbolic rules (IF….THEN). This rule can be directly interpreted and compared with available knowledge and provide useful information.

According to Quinlan, the learning algorithm applies a divide and-conquer strategy to construct the tree. The sets of instances are associated by a set of attributes. A decision tree comprises of node and leaves, where nodes represent a test on the values of an attribute and leaves represent the class of an instance that satisfies the conditions. The outcome is

'true' or 'false'. Rules can be derived from the path starting from the root node to the leaf and utilizing the nodes along the way as preconditions for the rule, to predict the class at the leaf. The tree Pruning has to be carried out to remove unnecessary preconditions and duplications.

Figure 4 show the decision tree induced using colon tumor data and its corresponding decision rule (Srinivasa and Sujatha, 2011).

The nodes represent genes and branches represent the expression conditions. The leaves of the tree represent the decision outcome (in this case either 'is a tumor tissue' or 'is a normal tissue'). The brace under a leaf denotes the number of instances correctly and incorrectly classified by the leaf (TP/FP) and the equivalent decision rules are derived from the decision trees (Srinivasa and Sujatha, 2011).



Decision Rule:
If gene-1671 < 56.9 Then Normal

If gene_1671 > 56.9 and
gene_682 < 107.4 Then Normal

If gene_1671 > 56.9 and
gene_682 > 107.4 and
gene_201 < 3145.5 Then Tumour

If gene_1671 > 56.9 and
Gene_682 > 107.4 and
Gene 201 > 3145.5 Then Normal

Fig. 4. A decision tree induced from the colon tumor dataset

C4.5 Algorithm

The C4.5 classification algorithm is the extraction of ID3 algorithm. The C4.5 algorithm not support the over fitting of data which is the disadvantage in ID3. It also reduce the errors considerably. It support continuous attributes and shows the best accuracy on attribute with missing values. The entropy and the information gain are used for pruning the tree. The entropy is a measure of similar kind of sets. The information gain is an attribute measure which indicates how much percentage the given attribute separate the training dataset according to their final classification. The Entropy for a set S is calculate as:

$$Entropy(s) = \sum_{i=1}^{n} p_i \ log_2 \ p_i$$

Where:

'n' = The number of classes and the
P_i = The probability of S belongs to class i

The gain of A and S is calculated as:

$$Gain(A) = Entropy(s) - \sum_{k=1}^{m} \frac{|s_k|}{|s|} Entropy(s_k)$$

C5.0 Algorithm

The C5.0 classification algorithm is the successor of C4.5. Both C4.5 and C5.0 can produce classifiers expressed either as decision trees or rulesets. In many applications, rulesets are preferred because they are simpler and easier to understand than decision trees, but C4.5's ruleset methods are slow and need more memory. C5.0 embodies new algorithms for generating rulesets and the improvement is substantial.

Accuracy: The C5.0 rulesets have noticeably lower error rates on unseen cases.

Speed: C5.0 is much faster; it uses different algorithms and is highly optimized. If C4.5 required nine hours to find the ruleset, C5.0 completed the task within 73 sec.

Memory: C5.0 commonly uses less memory than C4.5 during ruleset construction.C4.5 needs more than 3GB, but C5.0 requires less than 200MB.

Based on the research of Freund and Schapire, this is an exciting new development that has no counterpart in C4.5. Boosting is a technique for generating and combining multiple classifiers to improve predictive accuracy.

The boosting doesn't always help when the training cases are noisy, boosting can actually reduce classification accuracy. C5.0 uses a novel variant of

boosting that is less affected by noise, thereby partly overcoming this limitation. C5.0 supports boosting with any number of trials. Naturally, it takes longer to produce boosted classifiers, but the results can justify the additional computation! Boosting should always be tried when peak predictive accuracy is required, especially when unboosted classifiers are already quite accurate.

C5.0 incorporates several new facilities such as variable misclassification costs. In C4.5, all errors are treated as equal, but in practical applications some classification errors are more serious than others. C5.0 allows a separate cost to be defined for each predicted/actual class pair; by using this option the C5.0 constructs classifiers to minimize expected misclassification costs rather than error rates.

The C5.0 has provision for a case weight attribute that quantifies the importance of each case; C5.0 attempts to minimize the weighted predictive error rate.

C5.0 has several new data types in addition to those available in C4.5, including dates, times, timestamps, ordered discrete attributes and case labels. In addition to missing values, C5.0 allows values to be noted as not applicable. Further, C5.0 provides facilities for defining new attributes as functions of other attributes.

Some recent data mining applications are characterized by very high dimensionality, with hundreds or even thousands of attributes. C5.0 can automatically winnow the attributes before a classifier is constructed, discarding those that appear to be only marginally relevant. For high-dimensional applications, winnowing can lead to smaller classifiers and higher predictive accuracy and can often reduce the time required to generate rulesets.

C5.0 is also easier to use. Options have been simplified and extended to support sampling and cross-validation, for instance and C4.5's programs for generating decision trees and rulesets have been merged into a single program.

Some Special Features of C5.0

C5.0 has been designed to analyse substantial databases containing thousands to millions of records and tens to hundreds of numeric, time, date, or nominal fields. C5.0 also takes advantage of computers with up to eight cores in one or more CPUs to speed up the analysis. To maximize interpretability, C5.0 classifiers are expressed as decision trees or sets of if-then rules, forms that are generally easier to understand. C5.0 is easy to use and does not presume any special knowledge of Statistics or Machine Learning.

Bayesian Algorithm

Bayesian algorithm is a statistical classification method. It is also known as naïve Bayesian classification. The base for Bayesian classifier is Bayes theorem. The Bayesian classifier has produced high accuracy and is speed when applied in the large database. It uses the assumption of class conditional independence. The naïve Bayesian classifier is a simple classification algorithm and it can be comparable with decision tree induction classifier and neural network classifier. The posterior probability is calculated for estimation. When H is a hypothesis on attribute V and V is an 'n' dimensional attribute vector the posterior probability is calculated as:

$$P(H/V) = P\frac{(V/H)P(H)}{P(V)}$$

The attribute vector belongs to the class C_1 only when $P(C_1/V)$ has maximum posterior hypothesis. The (C_1/H) is calculated using the above equation as:

$$P(C_i/\mathrm{V}) = \frac{P(\mathrm{V}/\mathrm{C}_i)\mathrm{P}(\mathrm{C}_i)}{p(\mathrm{V})}$$

The $P(V/C_i)$ is to be calculated for each class C_i where i takes the value $i = 1, 2, 3 \ldots k$.

The $P(V/C_i)$ is the product of $P(V_1/C_i),P(V_2/C_i),..,P(V_n/s)$. When probability value of some $P(V_x/C_i)$ is zero since it involves the product, it cancels all other values which are involved with C_i. The Laplacian correction or the Laplace estimation is used to correct this kind of error.

Bayesian Networks (BNs)

The naïve Bayes assumes the class independence. But there exists dependencies between the variables. It specifies joint conditional probability distribution. BNs are probability distribution. It is also known as belief network or probabilistic networks. The Bayes network is a graphical model representing a set of random variables and their conditional dependencies via a Directed Acyclic Graph (DAG). Each node in the graph represents a random variable. A random variable denotes a feature about which we may be uncertain. Each random variable has a set of mutually exclusive and collectively exhaustive possible values. That is, exactly one of the possible values is or will be the actual value and we are uncertain about which one it is. The graph represents direct qualitative dependence relationships; the local distributions represent quantitative information about the strength of those dependencies. A belief network has one Conditional Probability Table (CPT) for each variable which specifies the conditional distribution of the

variable to indicates each possible combination of values of its parents.

The graph and the local distributions together represent a joint distribution over the random variables denoted by the nodes of the graph (Neapolitan, 2003). One of the most important features of BNs is the fact that they provide an elegant mathematical structure for modeling complicated relationships among random variables while keeping a relatively simple visualization of these relationships (Han amd Kamber, 2001). Any network topology use the following algorithm steps for Bayesian network

Step1 = The gradient value is calculated using the predefine formulae
Step2 = The weights are update using the gradient value calculated in step1
Step3 = The calculated weight in step 2 is in the form of probability values, those values are renormalized for simplification purpose

The algorithm which follows the above learning steps is known as adaptive probabilistic network.

To select the best BN for grading raisins, different search algorithms were evaluated.

Rule Based Classifiers

The IF…THEN structure is used to represent a learning model.

The rule can be extracted from the decision tree. The decision tree can be converted to classification IF…THEN rule by tracing the tree from the root. The training tuple and its associated class labels are used to estimate the accuracy. The rule is prune if it is not satisfy the estimated accuracy.

The sequential covering algorithm is use to extract the IF…THEN rules directly from the training tuple. It is widely used to mine disjunctive set of classification rule. The AQ, CN2 and RIPPER are some of the sequential covering algorithms. It is contrast to decision tree induction. The rule is learned on the selected tuple and is continue on the remaining tuples by removing the previously selected tuple.

Algorithm:

Sequential covering. Learn a set of IF-THEN rules for classification.

Input: D, a data set class-labeled tuples;Att vals, the set of all attributes and their possible values.
Output: A set of IF-THEN rules.
Method:
Rule set = {}; // initial set of rules learned is empty
for each class c do
repeat
Rule = Learn One Rule(D, Att vals, c);
remove tuples covered by Rule from D; until terminating condition;
Rule set = Rule set + Rule; // add new rule to rule set
endfor
return Rule Set;

Back Propagation Classifier

The back propagation is used for both classification and prediction. The back propagation is a neural network algorithm. The multilayer feed forward network is a type of neural network which the neural network algorithm is performed. The multilayer feed forward consisting of input layer, output layer and the hidden layers. The number of hidden layer is arbitrary. The initial network is designed by assigning the initial weights in the training process. The network topology is designed by specifying the number of units in the input, output and hidden layers. The network design is a trial an error process and it affects the accuracy of the classifier. The training process is repeated for getting better accuracy in different network topologies.

For each training tuple the weight is modified in the backward direction to minimize the mean square error between the network prediction and the actual target value.

Algorithm:

Neural network learning for classification or prediction using the back propagation algorithm:

Input: D, a data set consisting of the training tuples and their associated target values; l, the learning rate;
Network: A multilayer feed-forward network.
Output: A trained neural network.
Method:
Initialize all weights and biases in network;
while terminating condition is not satisfied {
for each training tuple X in D
{
//Propagate the inputs forward:
for each input layer unit j
{
$O_j = I_j$;//output of an input unit is its actual input value
for each hidden or output layer unit j
{
$I_j = \sum_i w_{ij} O_i + q_j$//compute the net input of unit j with respect to the previous layer, i
$O_j = 1/1+e^{-1}j$;
}
//compute the output of each unit j
//Backpropagate the errors:

for each unit j in the output layer
$Err_j = O_j(1-O_j)(T_j - O_j)$;//compute the error
for each unit j in the hidden layers, from the last to the first hidden layer
$Err_j = O_j(1-O_j)\sum_k Err_k w_{jk}$;//compute the error with respect to the next higher layer, k
for each weight wi j in network
{
$\Delta w_{ij} = (l)Err_j O_i$;//weight increment
$w_{ij} = w_{ij} + \Delta w_{ij}$; g//weight update
for each bias Θ_j in network
{
$\Delta\Theta_j = (l)Err_j$;//bias increment
$\Theta_j = \Theta_j + \Delta\Theta_j$;}//bias update
}}

There are many variations of the back propagation algorithm, several of which are described here (Paulin and Santhakumaran, 2011).

Batch Training

In batch mode the weights and biases of the network are updated only after the entire training set has been applied to the network. The gradients calculated at each training example are added together to determine the change in the weights and biases.

Batch Gradient Descent

This is the batch steepest descent training algorithm. The weights and biases are updated in the direction of the negative gradient of the performance function.

Batch Gradient Descent with Momentum

In provides faster convergence, steepest descent with momentum. Momentum allows a network to respond not only to the local gradient, but also to recent trends in the error surface. Acting like a low pass filter, momentum allows the network to ignore small features in the error surface. Without momentum a network can get stuck in a shallow local minimum.

Conjugate Gradient Algorithms

The basic back propagation algorithm adjusts the weights in the steepest descent direction (negative of the gradient), the direction in which the performance function is decreasing most rapidly. It turns out that, although the function decreases most rapidly along the negative of the gradient, this does not necessarily produce the fastest convergence. In the conjugate gradient algorithms a search is performed along conjugate directions, which produces generally faster convergence than steepest descent directions.

Quasi-Newton Algorithms

Newton's method is an alternative to the conjugate gradient methods for fast optimization. There is a class of algorithms that is based on Newton's method, but which doesn't require calculation of second derivatives. These are called quasi-Newton (or secant) methods. They update an approximate Hessian matrix at each iteration of the algorithm. The update is computed as a function of the gradient.

Levenberg-Marquardt

Like the quasi-Newton methods, the Levenberg-Marquardt algorithm was designed to approach second-order training speed without having to compute the Hessian matrix.

Resilient Back propagation

Multilayer networks typically use sigmoid transfer functions in the hidden layers. These functions are often called "squashing" functions, because they compress an infinite input range into a finite output range. Sigmoid functions are characterized by the fact that their slopes must approach zero as the input gets large. This causes a problem when you use steepest descent to train a multilayer network with sigmoid functions, because the gradient can have a very small magnitude and, therefore, cause small changes in the weights and biases, even though the weights and biases are far from their optimal values. The purpose of the resilient back propagation training algorithm is to eliminate these harmful effects of the magnitudes of the partial derivatives. Only the sign of the derivative is used to determine the direction of the weight update; the magnitude of the derivative has no effect on the weight update. The size of the weight change is determined by a separate update value.

SVM

SVM algorithms are based on the learning system which uses the statistical learning methodology and they are popularly used for classification. In SVM technique the optimal boundary, known as hyperplane, of two sets in a vector space is obtained independently on the probabilistic distribution of training vectors in the set. This hyperplane locates the boundary that is most distant from the vectors nearest to the boundary in both sets. The vectors that are placed near the hyperplane are called supporting vectors. If the space is not linearly separable there may be no separating hyperplane. The kernel function is used to solve the problem. The Kernel function analyses the relationship among the data and it creates a complex divisions in the space (Isabelle *et al*., 2002).

KNN

KNN classification classifies instances based on their similarity. It is one of the most popular algorithms for pattern recognition. It is a type of Lazy learning where the function is only approximated locally and all computation is deferred until classification. An object is classified by a majority of its neighbors. K is always a positive integer. The neighbors are selected from a set of objects for which the correct classification is known.

The KNN algorithm steps are:

- The number of nearest neighbors 'k' is found out
- Using the distance measure, calculate the distance between the query instance and all the training samples
- The distance of all the training samples are sorted and nearest neighbor based on the k minimum distance is determined
- Since the KNN is supervised learning, get all the categories of the training data for the sorted value which fall under k
- The prediction value is measured by using the majority of nearest neighbors

The Code of KNN Algorithm

```
Function KNN(train_patterns, train_targets, test_patterns )
end
Uc-a set of unique labels of train-targets;
N-size of test-patterns
for i = 1…N,
dist: = EQ-Dis(train-patterns, test-patterns(i))
idxs: = sort(dist)
topk Classes: = train_targets(idxs(1: Knn))
cls: = DominatingClass (topkClasses)
test-targets(i): = cls
```

Genetic Algorithms (GAs)

Genetic algorithm optimization strategies that are inspired by the principles observed in natural evolution of a collection of potential problem solutions that compete with each other, the best solutions are selected and combined with each other. Genetic algorithms and evolutionary programming are used in data mining to formulate hypotheses about addictions between variables, in the form of association rules or some other internal formalism.

Fuzzy Sets Approach

Fuzzy sets form a key methodology for representing and processing uncertainty. Uncertainty arises in many forms in today's databases: Imprecision, non-specificity, inconsistency, vagueness. Fuzzy sets exploit uncertainty in an attempt to make system complexity manageable. As such, fuzzy sets constitute a powerful approach to deal not only with incomplete, noisy or imprecise data, but may also be helpful in developing uncertain models of the data that provide smarter and smoother performance than traditional systems.

Neural Networks

Neural Networks (NN) are those systems modeled based on the human brain working. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it. The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input.

It required long training time and poor interpretability.

Support for highly noisy data and classify the untrained data. It is well suitable for classifying the continuous valued data.

The feature of neural networks is an iterative learning process in which data cases are presented to the network one at a time and the weights associated with the input values are adjusted each time. After all cases are presented, the process often starts over again. During this learning phase, the weights of the network nodes are used to predict the correct class label of input samples. After a network has been structured, the network is ready to be trained. The initial weights are chosen randomly. Then the training or learning, begins. The most popular neural network algorithm is back-propagation algorithm. There are many types of neural networks can be used for classification purposes. One of them is feedforward multilayer networks or multilayer perceptions which is widely studied and used as neural network classifiers. The feedforward, back-propagation architecture were developed in 1970's. This back-propagation architecture is the most popular, effective and easy-to-learn model for complex, multi-layered networks. The typical back-propagation network has an input layer, an output layer and at least one hidden layer. There is no limit on the number of hidden layers. Each layer is fully connected to the succeeding layer.

Training inputs are applied to the input layer of the network and desired outputs are compared at the output layer. During the learning process, a forward sweep is made through the network and the output of each element is computed layer by layer. The difference between the output of the final layer and

the desired output is back-propagated to the previous layers, usually modified by the derivative of the transfer function and the connection weights are normally adjusted. This process proceeds for the previous layers until the input layer is reached.

The advantages of Neural Networks for classification are:

- Neural Networks are more robust because of the weights and is more robust in noise environment
- The Neural Networks improves its performance by learning. This may continue even after the training set has been applied
- The use of Neural Networks can be parallelized as specified above for better performance
- There is a low error rate and thus a high degree of accuracy once the appropriate training has been performed

Rough Sets

A rough set is determined by a lower and upper bound of a set. Every member of the lower bound is a certain member of the set. Every nonmember of the upper bound is a certain non-member of the set. The upper bound of a rough set is the union between the lower bound and the so-called border line region. A member of the boundary region is possibly (but not certainly) a member of the set. Therefore, rough sets may be viewed as with a three valued membership function (yes, no, perhaps). Rough sets are a mathematical concept dealing with uncertainty in data. They are usually combined with other methods such as rule induction, classification, or clustering methods.

Classifier Performance

The performance of the classifier is depends on the characteristics of the data to be classified. The performance of the classifiers indicates the accuracy and the error rate produce by the classifier. Various kind of test can be conducted on classifiers to evaluate the performance. Some of the performance evaluator methods are holdout, random sub-sampling, k-fold cross validation and bootstrap.

In k-fold cross validation the dataset D is divided into k equal size subsets or k folds randomly such as d_1, d_2, …, dk. The training and testing is performed k times. In the first iteration, the subsets d_2, …, d_k serve as the training set in order to obtain a first model, which is tested on first subset d_1; the second iteration is trained in subsets d_1,d_3,…, d_k and tested on second subset d_2; and so on. The accuracy of the classifier refers to the ability of a given classifier to correctly predict the class label for new or unseen data.

The Accuracy and Error rate can be defined as:

$$Accuracy = (TP + TN) / (TP + FP + TN + FN)$$
$$Error\ rate = (FP + \mathrm{FN}) / (TP + FP + TN + FN)$$

Where:
TP = The number of True Positives
TN = The number of True Negatives
FP = The number of False Positives
FN = The number of False Negatives

The TP, TN, FP and FN are the values from the confusion matrix

Problem Statement

Complex data analysis and mining on huge amount of data can take a long time and is infeasible. So the data reduction techniques such as data cube aggregation, attribute subset selection, dimensionality reduction, numerosity reduction, data discretization and concept hierarchy generations are used. When the user applies any one of the data reduction methodology it must reduce the dataset without missing of any valuable data. Another problem is the accuracy of the classifier. The accuracy of the classifier is depends not only with the classification algorithm but also on the feature selection method. The feature selection method places the major role in classification accuracy. Even there are different kind of methods are available, for selecting the appropriate features the best algorithm should be chosen to maximize the accuracy.

Binaryversus Multiclass Datasets

Originally we have two categories in datasets Binary and multiclass datasets. When the values are split into two categories, it is consider as a binary class dataset. For the Binary class the Dataset D is split into two data tuples say D1, D2. Tuples in D1 represents the attributes related to the Class1 and tuples on D2 represent the attributes related to the Class2. The mid-point or split-point is used to split the given attribute a into binary class values. The confusion matrix is generated to analyse the accuracy of the classifier. The confusion matrix list out the instances assigned to each class. For the binary class dataset there are only two possible instances so that the 2×2 confusion matrix is generated.

Example case for binary class may get the possible values of Yes or No instances. When we take the cases:

Class 1 = Yes
Class 2 = No

The possible values for confusion matrix are:

- TP-True Positive-Classified as true, actually true
- TN-True Negative-Classified as False, actually true
- FP-False Positive-Classified as True, actually false
- FN-False Negative-Classified as False, actually false

The confusion matrix is used to analyse the accuracy of the classifier. The sensitivity or true positive and specificity or true negative can also be used for analysing the classifier.

The sensitivity and the specificity measure can be calculate using:

$$Sensitivity = No.of\ TP\ /\ No.of\ PT$$
$$Specificity = No.of\ TN\ /\ No.of\ NT$$

where, PT, NT represent Positive tuple and Negative tuple

Accuracy and the Error Rate can be calculated as:

$$Accuracy = (TP + TN) / (TP + FP + TN + FN)$$
$$Error\ Rate = (FP + FN) / (TP + FP + TN + FN)$$

The datasets like microarray consists of not only the two classes of data but also has number of categories which comes under multiple classes. The multiclass, the data set is split into multiple classes C_i where i = 1, 2….n. The class C1 corresponds to the tuples which satisfies the constraint1, C2 corresponds to the constraint 2 and so on.

Feature Selection Models

The data set D can consists of N number of attributes, each related with any one of M number of classes. The related attributes to the specific classes must be selected and is to be placed under the related class. To compute the relevance between the attributes we must use any one of the feature selection algorithm. Figure 5 shows the general attribute selection method.

Different kinds of feature selection algorithms have been proposed. Every feature selection algorithm uses any one of the three feature selection techniques. The feature selection technique can be categorized into three. Wrapper method, Filter method and embedded method.

The filter method filters the undesirable feature. The Fig. 6 shows steps in filter method. The filter method is faster and is appropriate method for analysing the large datasets. The filter method has the high rank in supporting of feature selection for medical data analysis.

The FSDD, (Liang *et al*., 2008), RFS, (Robnik-Sikonja and Kononenko, 2003), CFS, (Michalak and Kwasnicka, 2006), are some of the feature selection algorithm which uses the Filter methodology. The relevance score is calculated for the features to check the correlation between the features. The calculated score is high with some threshold value then the particular feature is selected for further classification. When the ranking is low those feature are removed. This method is independent of classification algorithm.

The CFS Algorithm

CFS is a Correlation-based Feature Selection algorithm which uses the filter method of selection of attributes. It is described by Hall Correlation. The CFS algorithm uses a heuristic which measures the usefulness of individual features for predicting the class label along with the level of inter-correlation among them. The highly correlated and irrelevant features are avoided. The equation used to filter out the irrelevant, redundant feature which leads the poor prediction of the class is calculate using the equation:

$$Fs = \frac{N\overline{r_{ci}}}{N + N(N-1)\overline{r_{ii}}}$$

N = The number of features in the subset,
r_{ci} = The mean feature correlation with the class and
r_{ii} = The average feature inter-correlation

For computing the correlations necessary for equation a number of information based measures of association were proposed such as: The uncertainty coefficient, the gain ratio or the minimum description length principle. The best results however were achieved with the gain ratio used for feature-class correlations and symmetrical uncertainty coefficients used for feature inter correlations.

Chi-Square Test

The correlation between the attributes can also be measured by using the Chi-square test. The Chi-square (χ^2) is a non-parametric test of statistical significance for bivariate tabular analysis. The χ^2 test is defined by the equation.

$$\chi^2 = \sum \frac{(f_0 - f)^2}{f}$$

Where:
f_0 = An observed frequency and
f = An expected frequency

The χ^2 test is commonly used for testing independence and/or correlation between two vectors. The test compares observed frequencies with the corresponding expected frequencies. The value of χ^2 equal to 0 means the corresponding two vectors are statistically independent with each other. The higher value of χ^2 indicates that there is a high correlation between the vectors.

The SVM-RFE (Isabelle *et al*., 2002), GLGS (Liu *et al*., 2009), are some of the feature selection algorithm which uses the Wrapper method. The Fig. 7 dipicts the steps in wrapper method. The Wrapper method need some predefine learning algorithm to identify the relevant feature. It has interaction with classification algorithm. The wrapper method uses learning algorithm with a statistical resampling method (cross validation) for accuracy estimation of feature subset. The over fitting of feature is avoided using the cross validation. But it takes much time comparing with the filter method.

The embedded method consumes less time than the wrapper method. The embedded method can use the support vector mechanism to select the feature.

When we consider the microarray datasets it has thousands of features and is treated as high dimensional dataset. The feature selection algorithm places the important role to maximize the accuracy of such high dimensional datasets. To maximize the accuracy of such a high dimensional dataset we must follow the steps:

Step 1 = Select the appropriate feature selection method
Step 2 = Select the suitable classification algorithm

The specified feature selection method and the chosen classification algorithm must support for both classes of datasets and maximize it accuracy. For simplification purpose the multiclass dataset can be treated into number of binary class using subsets and can be merged to treat as multiclass. Not all the classification algorithms produce better accuracy on binary as well as multiclass datasets. Some of the algorithm performs well on binary dataset and the other performs well on multiclass datasets.

The algorithm is a better one, when it performs well on both binary as well as multiclass datasets. For better performance of an algorithm it must incorporate the best feature selection algorithm.

Classification with Medical Dataset

People get affected by different diseases like cancer, heart related problems, liver disease, diabetes and kidney related problems. The data classification of medical data is important to early prediction of disease and improvement in cure rates. Also using classification various kinds of treatments can be suggested for those diseases.

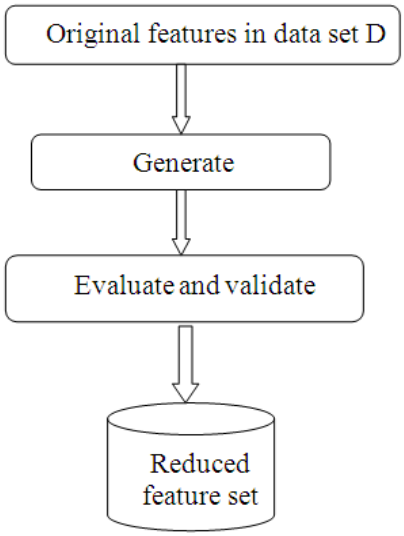


Fig. 5. General attribute selection model

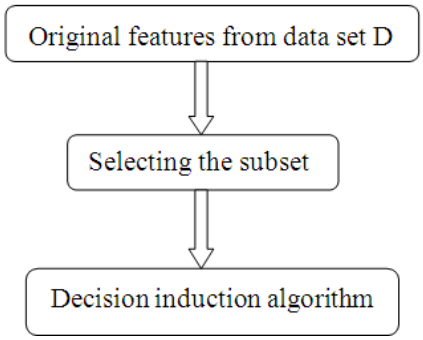


Fig. 6. Filter feature selection model

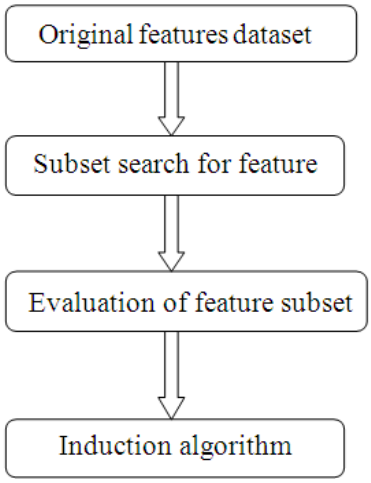


Fig. 7. Wrapper feature selection model

Cancer Dataset

According to the statistical reports of WHO, the incidence of breast cancer is the number one form of cancer among women. In the United States (US), approximately one in eight women have a risk of developing breast cancer. An analysis of the most recent data has shown that the survival rate is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis. Hence, it can be seen from the study that an early diagnosis improves the survival rate. In 2007, it was reported that 202,964 women in the United States were diagnosed with breast cancer and 40,598 women in the United States died because of breast cancer. A comparison of breast cancer in India with US obtained from Globocon data, shows that the incidence of cancer is 1 in 30. However, the actual number of cases reported in 2008 was comparable; about 1, 82,000 breast cancer cases in the US and 1, 15,000 in India. A study at the Cancer Institute, Chennai shows that breast cancer is the second most common cancer among women in Madras and southern India after cervix cancer (Rajesh and Anand, 2012). Agrawal *et al*. (2011) use the SEER cancer data sets to analyze the lung cancer. The SEER database is also used for analyzing the Breast cancer. SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 28% of the US population. SEER database is a premier source for cancer statistics in the United States, which has information on incidence, prevalence and survival from specific geographic areas of the US population as also cancer mortality for the entire country. The dataset used contained data that pertained to all types of cancer cases for the period 1973-2008. Of this, 1403 record samples, each having 124 attributes pertained to breast cancer (Rajesh and Anand, 2012). But the SEER data base require the preprocessing to make the raw data into normalized format for analyzing.

The SEER data files were requested through the Surveillance, Epidemiology and End Results (SEER) web site (http://www.seer.cancer.-gov).

Dursun *et al*. (2004). Uses the SEER Program which is a part of the Surveillance Research Program (SRP) at the National Cancer Institute (NCI) and is responsible for collecting incidence and survival data from the participating nine registries and disseminating these datasets (along with descriptive information of the data itself) to institutions and laboratories for the purpose of conducting analytical research.

The SEER public use data consists of nine text files, each containing data related to cancer for specific anatomical sites (i.e., breast, colon and rectum, other digestive, female genital, lymphoma and leukemia, male genital, respiratory, urinary and all other sites).

There are 72 variables in each file and each record in the file relates to a specific incidence of cancer. The data in the file is collected from nine different registries (i.e., geographic areas). These registries contain a population that is representative of the different racial/ethnic groups residing in the United States. The cancer incidence trends and mortality rates in SEER are assumed to be representative of the cancer incidence trends and mortality rates for the total United States. The SEER database is considered to be the ‘‘most comprehensive source of information on cancer incidence and survival in the USA’’.

The SEER program emphasizes quality and completeness of the data and it has been estimated that databases provided by SEER program are 98% complete (or better) for each of these nine registries. The SEER database is used often for analytical research purposes in a variety of projects. “A search of the national library of Medicine’s Database (PUBMED) using only the term SEER identified in excess of 570 publications for the time period of 1978-1999, many of which either include and analysis of SEER data or refer to cancer statistics based on SEER data’’.

Wisconsin Breast Cancer Data (WBCD) is analyzed by various researchers on medical diagnosis of breast cancer in neural network literature. The Wisconsin breast cancer database was originally provided by MacMillan *et al*. (1978; Paulin and Santhakumaran, 2011) and used by a number of researchers in pattern recognition and machine learning. The database available in the UCI database repository contains 699 cases. The original dataset contains 11 attributes including both sample id number and class label, which are removed in the actual dataset that are used in this application. The class of each instance is either benign or malignant. The remaining 9 attributes represent 9 cytological characteristics of breast fine-needle aspirates (Paulin and Santhakumaran, 2011).

Pitchumani and Kamal (2011) proposed association rule based decision tree induction model to deduct the breast cancer in mammogram images. This methodology gives the 99% accuracy. Paulin and Santhakumaran (2011) use the Back propagation algorithm as the training algorithm to classify the

breast cancer. The algorithm is used to train the Feed Forward Artificial Neural Networks. The performance of the network is evaluated using Wisconsin breast cancer data set for various training algorithms. Six training algorithms are proposed in this research paper. The highest accuracy of 99.28% is achieved in Levenberg Marquardt algorithm.

Rajesh and Anand (2012) applied the C4.5 to analyse the SEER data set for breast cancer and classify the patients either in the beginning stage or the pre-cancer stage. Five hundred records are used and 93% accuracy is achieved in testing datasets.

Micro calcification Clusters (MCs) in the mammogram image is classified by using the multidimensional genetic association rule to predict the breast cancer (Thangavel and Mohideen, 2009). The rule based approach of classification reaches the accuracy of 85%. Gupta *et al*. (2011) analyzes the various classification techniques applied to diagnosis and prognosis of breast cancer. The author analyse the papers (Sarvestan *et al*., 2010; Orlando *et al*., 2010; Abdelaal *et al*., 2010; Chang and Liou, 2005; Gandhi *et al*., 2010; Padmavati, 2011; Chul *et al*., 2001; Hassanien and Jafar, 2004; Sudhir *et al*., 2006; Jamarani *et al*., 2005; Abdelghani and Guven, 2006; Choi *et al*., 2009; Lundin *et al*., 1999; Street, 1998; Chi *et al*., 2007; Dursun *et al*., 2004; Khan *et al*., 2008) and concludes that any classification method is acceptable for diagnosis. But for the prognosis ANN classification method gives higher accuracy than any other classification methods.

Chang and Liouc (2002) compare three datamining techniques such as logistic regression, Decision tree and ANN along with Genetic algorithm to analyse breast cancer data. The accuracy of each method is estimated with 10 fold cross validation. The logistic regression gains the accuracy of 0.96; the Decision tree yields the accuracy of 0.94, while the ANN method yields the accuracy of 0.95.

Othman and Yau (2007) Compares the different kinds of classification techniques such as Bayes network, Decision tree and the Nearest Neighbours' method using WEKA, software in breast cancer datasets. The bayes yield 89.71% accuracy; the Decision tree gives the 85.71% and finally the Nearest Neighbour produces 84.57% accuracy.

For diagnoses and prognoses the breast cancer, the Naïve bayes, ANN and the C4.5, decision tree algorithms are used (Shweta, 2012). The accuracy of each algorithm is stated. The first method shows 84.5% and the second method gains 86.5% and the third method yields 86.7%. Above all, it state that the Decision tree is the best method with 93.62% accuracy for the bench mark database. Two datamining algorithms and logistic regression are used to diagnose the breast cancer survivability (Dursun *et al*., 2004) for the large dataset. It also uses 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. The results indicate that the decision tree is the best predictor with 93.6% accuracy. The ANN and logistic regression model gives 91.2%, 89.2% of accuracy.

Heart Disease Dataset

Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 sec in the United States. The different types of heart disease widely in the world are Coronary heart disease, Heart failure, Coronary artery disease, Ischemic heart disease, Cardiovascular disease, Hypoplastic left heart syndrome, Atherosclerosis, Chronic obstructive pulmonary disease, Congenital heart disease, Volvuli heart disease. The term "cardiovascular disease" includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardiovascular Disease (CVD) results in several illness, disability and death.

Generally, artificial intelligence techniques were used in prediction of heart disease. Machine learning algorithm was used in medical diagnostic problem for heart disease. Case-Based Reasoning (CBR) was considered as a suitable technique for diagnosis, prognosis and prescription in the medical domain puts more stress on real cases than other domains. Coronary Artery Disease was diagnosed using two techniques called Binary Particle Swarm Optimization (BPSO) and Genetic Algorithm (GA). For the diagnosis of heart disease various classification and regression processes was used. It provides medical knowledge for diagnosis purpose (Gayathri and Jaisankar, 2013).

Most of the heart disease datasets are available in heart-c.arff at UCI repository.

Cleveland Heart Disease database which consists of 303 records and Statlog Heart Disease database consists of 270 records are some of the public heart disease database.

Soni *et al*. (2011) different datamining methods are applied to predict the heart disease. The accuracy of

each algorithm is verified and is listed as Naive Bayes has got 86.53% accuracy, Decision Tree shows 89% accuracy while the ANN 85.53%. It concludes that the Decision Tree method outperforms others and it is suitable for heart disease prediction.

Chaitrali *et al*. (2012) the Decision Trees, Naïve Bayes and Neural Networks algorithms are used to analyze the Heart disease. The accuracy of these techniques are compared and the result shows the Neural Networks, Decision Trees and Naive Bayes have 100, 99.62 and 90.74% accuracy respectively.

The ECG signals are interpreted to analyze the heart related problem using the data mining classification algorithms ANN, C4.5 and the Decision Tree (Ramalingam *et al*., 2012). Each algorithm uses the different features. Among the three algorithms the decision tree algorithm outperform best and gives the accuracy of 97.5%.

Liver Disease Dataset

Liver tissue is composed of thousands of lobules and each lobule is made up of hepatic cells, the basic metabolic cells of the liver. The Excessive consumption of alcohol can cause an acute or chronic inflammation of the liver and may even harm other organs in the body, alcohol induced liver disease remains a major problem. When the liver becomes diseased, it may have many serious consequences. Liver disease (also called hepatic disease) is a broad term describing any single number of diseases affecting the liver. Many are accompanied by jaundice caused by increased levels of bilirubin in the system. The bilirubin results from the breakup of the hemoglobin of dead red blood cells; normally, the liver removes bilirubin from the blood and excretes it through bile.

Several diseases states can affect the liver. Some of the diseases are Wilson's disease, hepatitis (an inflammation of the liver), liver cancer and cirrhosis (a chronic inflammation that progresses ultimately to organ failure).

Alcohol alters the metabolism of the liver, which can have overall detrimental effects if alcohol is taken over long periods of time. Hemochromatosis can cause liver problems.

Fatty liver

It also known as steatorrhoeic hepatosis or steatosis hepatitis is a reversible condition where large vacuoles of triglyceride fat accumulate in liver cells via the process of steatosis. It can occur in people with a high level of alcohol consumption as well as in people who never had alcohol.

Hepatitis

Usually caused by a virus spread by sewage contamination or direct contact with infected body fluids.

Cirrhosis

of the liver is one of the most serious liver diseases. It is a condition used to denote all forms of diseases of the liver characterized by the significant loss of cells. The liver gradually contracts in size and becomes leathery and hard. The regenerative activity continues under liver cirrhosis but the progressive loss of liver cells exceeds cell replacement.

Liver Cancer

The risk of liver cancer is higher in those who have cirrhosis or who have had certain types of viral hepatitis; but more often, the liver is the site of secondary (metastatic) cancers spread from other organs.

Symptoms of Liver Disease

The external signs include a coated tongue, itchy skin, excessive sweating, offensive body odor, dark circles under the eyes, red swollen and itchy eyes, acne rosacea, brownish spots and blemishes on the skin, flushed facial appearance or excessive facial blood vessels. Other symptoms include jaundice, dark urine, pale stool, bone loss, easy bleeding and itching, small, spider-like blood vessels visible in the skin, enlarged spleen and fluid in the abdominal cavity, chills and gallbladder. The symptoms related to liver dysfunction include both physical signs and a variety of symptoms related to digestive problems, blood sugar problems, immune disorders, abnormal absorption of fats and metabolism problems. Nervous system disorders include depression, mood changes, especially anger and irritability, poor concentration and "foggy brain", overheating of the body, especially the face and torso and recurrent headaches (including migraine) associated with nausea. The blood sugar problems include a craving for sugar, hypoglycaemia and unstable blood sugar levels and the onset of type 2 diabetes. Abnormalities in the level of fats in the bloodstream, whether too high or too low levels of lipids in the organism. Hypercholesterolemia: Elevated LDL cholesterol, reduced HDL cholesterol, elevated triglycerides, clogged arteries leading to high blood pressure heart attacks and strokes, buildup of fat in other body

organs (fatty degeneration of organs), lumps of fat in the skin (lipomas and other fatty tumors), excessive weight gain (which may lead to obesity), inability to lose weight even while dieting, sluggish metabolism, protuberant abdomen (pot belly), cellulite, fatty liver and a roll of fat around the upper abdomen (liver roll). Pitchumani and Kamal (2011) Or too low levels of lipids: Hypocholesterolemia: Low total cholesterol, low LDL, VLDL cholesterol and low triglyderides.

Symptoms may include:

- Jaundice
- Tendency to bruise or bleed easily
- Ascites
- Impaired brain function
- General failing health

Risk Factors

Hepatitis is an inflammation of the liver that can be caused by a virus, inherited disorders and sometimes by certain medications or toxins such as alcohol and drugs. Scientists have identified four main types of viral hepatitis: Hepatitis A, hepatitis B and hepatitis C and hepatitis D.A fifth type, hepatitis E, is generally not found in North America.

Hepatitis A is waterborne and spread mainly via sewage and contaminated food and water. Hepatitis B is transmitted by contact with infected semen, blood, vaginal secretions and from mother to newborn. Hepatitis B is most commonly spread by unprotected sex and by sharing of infected needles (including those used for tattooing, acupuncture and ear piercing). Hepatitis C spreads via direct blood-to-blood contact. Hepatitis D is spread by infected needles and blood transfusions.

Improved screening of donated blood has greatly reduced the risk of catching hepatitis B or C from blood transfusions.

Both hepatitis B and C can be spread through sharing of razors, toothbrushes and nail clippers. The main cause of cirrhosis is chronic infection with the hepatitis C virus. Other causes include:

- Long-term, excessive alcohol consumption
- Chronic infection with hepatitis B virus
- Inherited disorders of iron and copper metabolism
- Severe reactions to certain medications
- Fatty liver caused by obesity
- Infections from bacteria and parasites usually found in the tropics
- Repeated episodes of heart failure with liver congestion and bile-duct obstruction

With cirrhosis, the liver tissue is irreversibly and progressively destroyed as a result of infection, poison or some other disease. Normal liver tissue is replaced by scars and areas of regenerating liver cells (Rajeswari and Reen, 2010).

The liver disease dataset can be collected from the UCI repository. The liver disorder data warehouse contains the screening the data of liver disorder patients. Initially, the data warehouse is pre-processed to make the mining process more efficient.

Aneeshkumar and Venkateswaran (2012) Naive Bayesian, C4.5 Decision Tree algorithms are used to estimate the surveillance of liver disorder. The results of both the algorithms are verified in terms of its accuracy. Among the algorithm the C4.5 gives the better performance than the Naïve Bayes method. The C4.5 yields the 99.20% accuracy while the Bayes algorithm gives the accuracy of 89.60%.

Rajeswari and Reen (2010) uses the three machine learning algorithms naïve Bayes, KStar an FT tree derived from WEKA software tool for classifying the liver disease dataset collected from UCI repository. The performance were compared and is stated as Bayes 96.52%, FT Tree 97.10% and KStar 83.47% which concludes the naïve Bayes gives the better accuracy with less time for liver disease dataset. Even the FT tree algorithm gives more accuracy than the Naïve Bayes the time take for FT Tree algorithm is considerably more than the Naïve Bayes, the Naive Bayes is consider being the best algorithm.

Diabetes Dataset

Diabetes is the most common endocrine disease in all populations and all age groups. According to the World Health Organization, it affects around 194 million people worldwide and that number is expected to increase to at least 300 million by 2025. Diabetes has become the fourth leading cause of death in developed countries and there is substantial evidence that it is reaching epidemic proportions in many developing and newly industrialized nations with evidence pointing to avoidable factors such as sedentary lifestyle and poor diet.

Diabetes describes a metabolic disorder characterized by chronic hyper glycaemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both. The long-term complications

include progressive development of the specific complications of retinopathy with potential blindness, nephropathy that may lead to renal failure and/or neuropathy with risk of foot ulcers, amputation, Charcot joints and features of autonomic dysfunction, including sexual dysfunction, known as micro-vascular complications.

People with diabetes are also at a greatly increased risk of cardiovascular, peripheral vascular and cerebrovascular disease, known as macro vascular complications.

There are two main classes of diabetes, which are diagnosed ultimately by the severity of the insulin deficiency. Insulin-dependent diabetes mellitus or type1 diabetes is an insulin openic state, usually seen in young people, but it can occur at any age. Non-insulin-dependent diabetes mellitus or type2 diabetes is the more common metabolic disorder that usually develops in overweight, older adults, but an increasing number of cases occur in younger age groups. Pinhas-Hamiel and Zeitler predict serious public health challenges given the rise in pediatric cases and the poor medication adherence in teens. The prevalence of type 2 diabetes has risen from 3 to 45% of adolescent diabetes in the last 15 years. In this age group for girls, complications in pregnancy add to the social cost. In Northern Ireland in 2004/5, 25% girls and 20% boys in the age range 4.5-5.5 years were classified as overweight or obese and research indicates that 85% of obese children will become obese adults.

In type 2 diabetes, the pancreas may produce adequate amounts of insulin to metabolize glucose (sugar), but the body is unable to utilize it efficiently. Over time, insulin production decreases and blood glucose levels rise. Patients with type 2 diabetes do not require insulin treatment to remain alive, although up to 20% are treated with insulin to control blood glucose levels. Type 2 diabetes accounts for up to 85% of the diabetic population in most countries; it probably affects 5-7% of western populations, 10% of people over 65 years of age and up to 50% of the cases may be currently undiagnosed. The peak age of onset of type 2 diabetes is 60 years old; most subjects are diagnosed after 40 years of age.

The burden of complications can be considerable both for the individual concerned and the health service in general. In the United Kingdom, economic costs are estimated at £2.5 billion annually. Many aspects of these complications can be limited, even prevented in some instances, with good early management of the condition, in particular the effective control of blood glucose levels. Computer-based tools can assist healthcare professionals to better manage people with type 2 diabetes, in order to reduce complications and improve quality of life. Better control of blood glucose reduces the risk of diabetic-related complications significantly. So understanding the major factors that determine overall control is important for clinicians in the prevention of diabetic complications and will assist the patient with self manageme.

UCHT collected diabetic patients' information from 2000 to 2004 in a clinical information system (Diamond, Hicom Technology). The data contained physiological and laboratory information for 3857 patients, described by 410 features. The patients included not only type 2 diabetic patients, but also type 1 and other types of diabetes such as gestational diabetes. It is very important to examine the data thoroughly.

In many research the pima diabetes datasets are taken for analysing performance of classification algorithms. The Pima are the group of indial peoples live in Southern Arizona. The possitive and negative record of Pima data set complicate the classification task.

Jayalakshmi and Santhakumaran (2010) uses the ANN method for diagnosing diabetes, using the Pima Indian diabetes dataset without missing data and produces 68.56% classification accuracy. In study (Pradhan and Sahul, 2011) suggested an ANN based classification model for classifying diabetic patients. It shows the average accuracy of 72.2%.

Huang *et al*. (2007) use the three classification algorithms naïve Bayes, IB1 and the c4.5 to predict the diabetes control. The c4.5 has got the highest accuracy of 95% and proved that the c4.5 is the stable classifier.

Biswadip (2012) applies Fuzzy Composite Programming (FCP) to build a diabetes classifier using the PIMA diabetes dataset. He has evaluated the performance of the fuzzy classifier using Receiver Operating Characteristic (ROC) Curves and compared the performance of the Fuzzy classifier against a Logistic Regression classifier. The results show that FCP classifier is better when the calculated AUC values are compared. The classifier based on logistic regression has an AUC of 64.8%, while the classifier based on FCP has an AUC of 72.2%. He proved that the performance of the fuzzy classifier was found to be better than a logistic regression classifier. Quinlan applied C4.5 and it was 71.1% accurate.

Christobel and Sivaprakasam (2012) applied KNN method to the Pima Indian Diabetes dataset. With 10-fold cross validation it gives the average accuracy of

71.94%.The algorithm is improvised and it gains the accuracy for the same dataset as much as 73.38%.

Skin Cancer Dataset

Skin cancers are divided into two major forms: Non-melanomas and Melanomas. These cancer subtypes are largely differentiated based on where in the skin layers they form.

Melanoma Melanoma is a form of skin cancer that affects the melanocytes in the epidermis. Melanocytes are special skin pigment cells that give our skin color and which allow our skin to "tan" when exposed to ultraviolet light from the sun. The darkening of the skin we call tanning provides the deeper body tissues extra protection from ultraviolet radiation.

Melanoma skin cancer will effect roughly 70,000 Americans in 2010 and roughly 12,000 Americans will die from melanoma in 2010 (ACS, 2012). The danger posed by melanoma is largely due to the risk of metastases' Melanoma is much more likely to spread to other parts of the body and to do so faster, than are non-melanoma skin cancers. As is the general case, metastasized cancers are harder to successfully treat than are localized cancers. Melanoma skin cancer is quite treatable provided it is caught early on before significant metastasis has taken place.

Non-melanoma As the non-creative name suggests, non-melanoma skin cancer is a sort of "blanket" term used to group together the types of skin cancer that aren't melanoma. There are two primary forms of non-melanoma skin cancer and a handful of other rare non-melanoma types.

Fatima *et al*. (2013) the detailed study about the diagnoses of early symptoms of skin cancer melanoma arose is tested by applying the algorithms such as k-nearest neighbourhood, Classification tree, SVM, MPECS adopted the MFNN. The Multi Parameter Extraction and Classification (MPECS) with machine learning technique by extracting the parameters is used to classify skin cancer which gives the accuracy of 81% which is more than 75% gives by SVM Classifier.

Experimental and Comparative Study

The performance analysis of each algorithm on the specific medical datasets from the existing work is carried out and the effectiveness of an algorithm is represented figuratively.

The analysis of most frequently used classification algorithms on Cancer dataset is shown in Fig. 8. The maximum accuracy of an algorithm gained in each research work for diagnoses and prognoses of breast cancer from (Pitchumani and Kamal, 2011; Paulin and Santhakumaran, 2011; Rajesh and Anand, 2012; Thangavel and Mohideen, 2009; Gupta *et al*., 2011; Sarvestan *et al*., 2010; Orlando *et al*., 2010; Abdelaal *et al*., 2010; Chang and Liou, 2005; Gandhi *et al*., 2010; Padmavati, 2011; Chul *et al*., 2001; Hassanien and Jafar, 2004; Sudhir *et al*., 2006; Jamarani *et al*., 2005; Choi *et al*., 2009; Lundin *et al*., 1999; Street, 1998; Chi *et al*., 2007; Dursun *et al*., 2004; Khan *et al*., 2008; Burke *et al*., 1997; Chang and Liouc, 2002; Othman and Yau, 2007; Shweta, 2012) are taken as the input data values and is used for the comparative analysis which is shown in Fig. 8.

The Decision tree, C4.5 and the Back propagation perform well on any kind of cancer datasets and shows maximum accuracy. The Naïve Bayes, ANN and Logistic regression has given the less performance on the same kind of datasets. Among the algorithms the Back propagation gives higher accuracy than the other two. Comparatively there is no much differece in all the three algorithms. It indicates that the decision tree, C4.5 and the Backe probagation are suitable algorithm for any kind of cancer analysis.

The performance of specifically used classification algorithms on diabetes datasets (Pradhan and Sahul, 2011; Biswadip, 2012; Huang *et al*., 2007) are shown in Fig. 9. Three algorithms are more frequently used are Logistic regression and KNN which gives the less performance on diabetes datasets and the C4.5 algorithm performs well with the higher accuracy and it can be used as the best algorithm for analysing the diabetes dataset.

The Decision tree, Bayes, ANN, C4.5 and Neural networks are more frequently used algorithms to analyse the heart disease datasets (Soni *et al*., 2011; Aneeshkumar and Venkateswaran, 2012; Chaitrali *et al*., 2012) and is shown on Fig. 10. It shows the maximum accuracy achieved in different researches on heart disease dataset. The decision tree and the c4.5 algorithms perform equally well. The Neural network algorithm performs better and gives 100% accuracy. The difference in accuracy of DT, C4.5 and Neural network is considerably less. All the three algorithms are best and suitable for analysing heart disease datasets.

The Performance of c4.5, Naïve Bayes, FT tree and KStar classification algorithm for surveillance of liver disorder (Jayalakshmi and Santhakumaran, 2010), (Rajeswari and Reen, 2010) on Liver data set is shown in the Fig. 11. (Jayalakshmi and Santhakumaran, 2010)

the naïve bayes and c4.5 classification algorithms were used and the maximum accuracy is gained from C4.5 as 92.2%. Rajeswari and Reen (2010) the Naïve Bayes, FT tree and the KStar algorithms Shows the accuracies as 96.52, 97.10 and 83.47%. The comparative study of the classification algorithms on liver disease data set is shown in the Fig. 11.

Figure 12 indicates the performance of most frequently used classification algorithms Multi Parameter Extraction for Classification (MPECS), SVM on skin cancer datasets (Fatima *et al*., 2013). The MPECS shows the maximum accuracy on skin cancer data set and is considered to be the best algorithm for skin cancer data analysis is shown in Fig. 12.

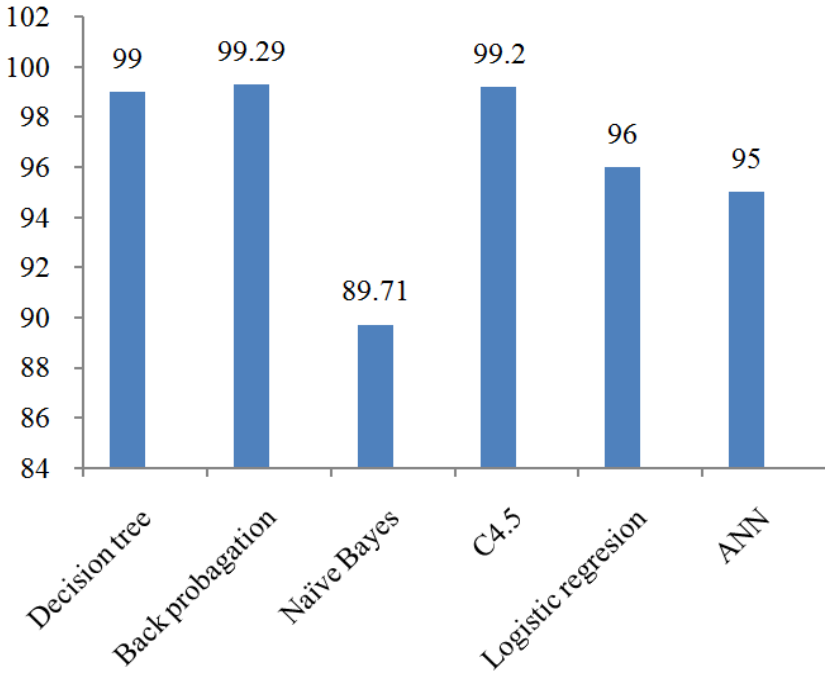


Fig. 8. Maximum accuracy of different classification algorithm on Cancer data set

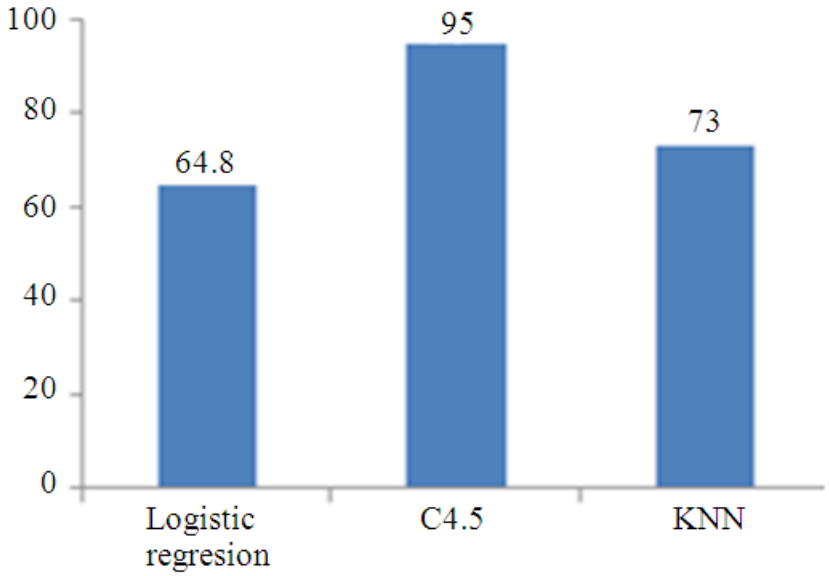


Fig. 9. Maximum accuracy of different classification algorithms on diabetes data set

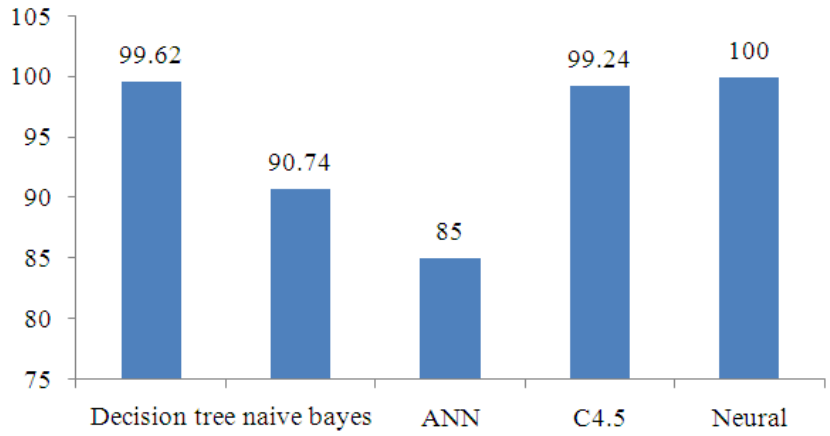


Fig. 10. Maximum accuracy of different classification algorithms on Heart disease data set

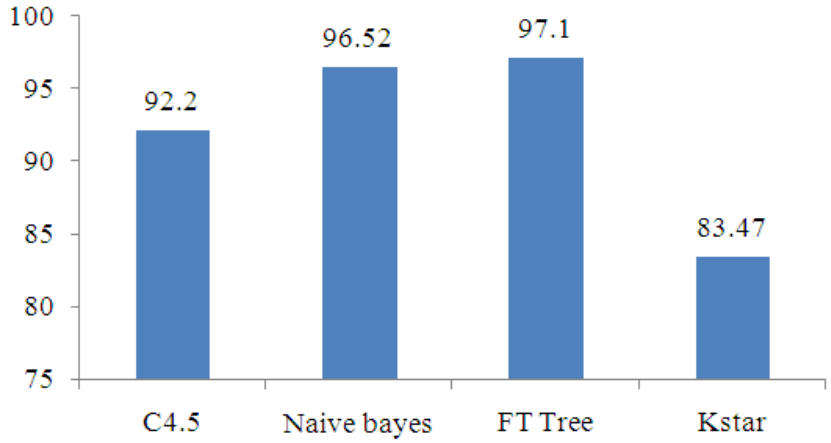


Fig. 11. Maximum accuracy of different classification algorithms on liver disease data set

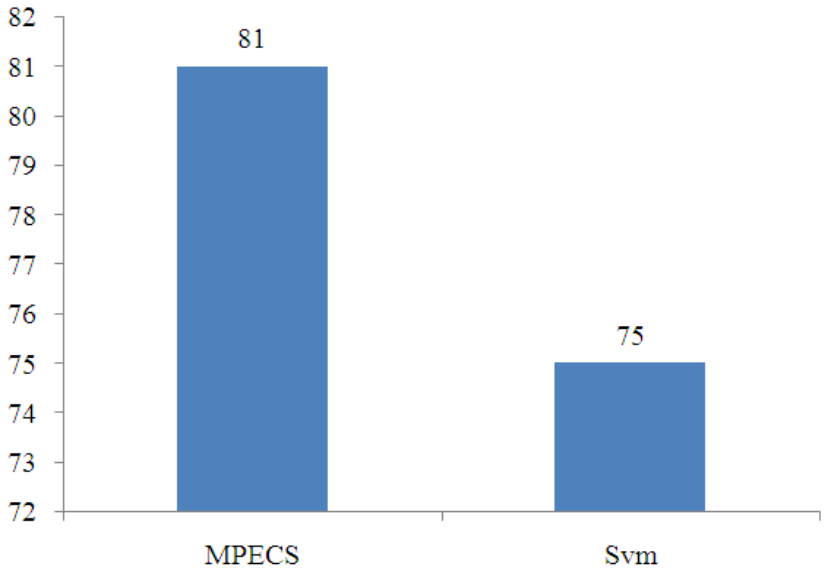


Fig. 12. Maximum accuracy of MPECS SVM classifier on skin cancer data set

Future Direction

The classification algorithms we have taken from the existing work deals with maximum of thousands of data, which uses minimum number of features and dimensionality. When the volume of data is increased from thousands to lakhs or crores, the features are in thousands and the dimensionality is more, what happens to the accuracy of an algorithm? By dimensionality reduction the accuracy of an algorithm can be improved. When the dimensionality is reduced it may lead to missing of some valuable data. Reducing the dimensionality of the dataset without missing any valuable data will degrade the performance of an algorithm. The algorithm which produces the maximum accuracy with maximum dimensionality must incorporate better feature selection algorithm along with the classification algorithm to classify the high dimensional datasets. When the volume of data is increased the existing algorithms requires more memory space for execution and the error rate is also increased.

The performance analysis of classification algorithm on different kind of medical dataset shows the best algorithms suitable for the specific kind of medical data and maximum accuracy gained on each algorithm. The accuracy of classification algorithm is a primary constraint for any kind of datasets. The classification algorithm must show higher accuracy for high dimensional dataset. When the dataset is high dimensional, by application of dimensionality reduction some of the valuable data are unconsidered and the accuracy of an algorithm is increased with missing of those data. The feature selection algorithm for high dimensional dataset plays an important role. With enhanced feature selection algorithm the classifications algorithm can improve its performance in terms of its accuracy. Also the enhanced feature selection algorithm must support for both the binary and multi class datasets. Any classification algorithm is considered to be the best algorithm only when it supports high dimensional datasets, handle large volume of data with less space requirement and produce less error rate with maximum accuracy. This research paper states the best algorithm in terms of accuracy constraint. With suitable feature selection algorithm, the specified classification algorithm can be implemented on high dimensional dataset to maximize it accuracy with less space requirement and less error rate. We have chosen only the medical dataset for our analysis but when the algorithm is implemented it can be extracted to other kind of high dimensional dataset.

Conclusion

Most of the medical diagnoses use different classification algorithms. Only few algorithms give better performance. Comparatively the C4.5 algorithm gives the better performance than any other classification algorithms. Still the improvisation of C4.5 algorithm is required to maximise accuracy, handle large amount of data, reduce the space requirement for large amount of datasets, support new data types and reduce the error rate. The improved version of C4.5 is C5.0 algorithm which produces more accuracy; requires less space when volume of data is increased from thousands to lakes or crores it performs better than C4.5. It is easier to understand and takes less time for execution. It also has lower error rate and supports new data types like timestamp, ordered discrete attributes, case labels which are not supported by C4.5. It also minimizes the weighted predictive errors and is less affected by noise. The study analysis shows the C5.0 algorithm is the potentially suitable algorithm for any kind of medical diagnoses. When the volume of data is increased and the numbers of features are in thousands the C5.0 algorithm work faster and give the better accuracy with less memory consumption. The C5.0 algorithm also works faster for high dimensional datasets. When C5.0 algorithm is applied on high dimensional dataset it must incorporate any one of the best feature selection algorithm for better performance which is our future work.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abdelaal, A.M.M., H.A. Sena, M.W. Farouq and A.M. Salem, 2010. Using data mining for assessing diagnosis of breast cancer. Proceedings of the International Multiconference on Computer Science and Information Technology, Oct. 18-20, IEEE Xplore Press, Wisla, pp: 11-17.
DOI: 10.1109/IMCSIT.2010.5679647

Abdelghani, B. and E. Guven, 2006. Predicting breast cancer survivability using data mining techniques. Proceedings of the International Conference on Data Mining, Ninth Workshop on Mining Scientific and Engineering Datasets in Conjunction, (SIAM' 06), pp:1-4.

ACS, 2012. Cancer Facts and Figures. Am. Cancer Society.

Agrawal, A., S. Misra, R. Narayanan, L. Polepeddi and A. Choudhary, 2011. Poster: A lung cancer mortality risk calculator based on SEER data. IEEE proceedings of the 1st International Conference on Computational Advances in Bio and Medical Sciences, Feb. 3-5, IEEE Xplore Press, Orlando, FL., pp: 233-233. DOI: 10.1109/ICCABS.2011.5729887

Aneeshkumar, A.S. and C.J. Venkateswaran, 2012. Estimating the surveillance of liver disorder using classification algorithms. Int. J. Comput. Applic., 57: 39-42. DOI: 10.5120/9121-3281

Biswadip, G., 2012. Using fuzzy classification for chronic disease. Ind. J. Econom. Bus.

Burke, H.B., P.H. Goodman, D.B. Rosen, D.E. Henson and J.N. Weinstein *et al*., 1997. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer, 79: 857-862. PMID: 9024725

Chaitrali, S., D. Sulabha and S. Apte, 2012. Improved study of heart disease prediction system using data mining classification techniques. Int. J. Comput. Applic., 47: 44-48. DOI: 10.5120/7228-0076

Chang, P.W. and M.D. Liou, 2005. Comparison of three Data Mining techniques with Genetic Algorithm in analysis of Breast Cancer data.

Chang, W.P. and B.D.M. Liouc, 2002. Comparison of three data mining techniques with genetic algorithm in the analysis of breast cancer data. National Yang-Ming University.

Chi, C.L., W.H. Street and W.H. Wolberg, 2007. Application of artificial neural network-based survival analysis on two breast cancer datasets. Annu. Symp. Proc.

Choi, J.P., T.H. Han and R.W. Park, 2009. A hybrid bayesian network model for predicting breast cancer prognosis. J. Korean Society Med. Inform., 15: 49-57. DOI: 10.4258/jksmi.2009.15.1.49

Christobel, Y.A. and P. Sivaprakasam, 2012. Improving the performance of k-nearest neighbor algorithm for the classification of diabetes dataset with missing values. Int. J. Comput. Eng. Tech., 3: 155-167.

Chul, L.H., S.H. Seon and C.C. Sang, 2001. Rule discovery using hierarchical classification structure with rough sets. Proceedings of the International Conference on IFSA World Congress and NAFIPS, Jul. 25-28, IEEE Xplore Press, Vancouver, pp: 447-452. DOI: 10.1109/NAFIPS.2001.944294

Dursun, D., W. Glenn and A. Kadam, 2004. Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelli. Med., 32: 113-127. DOI: 10.1016/j.artmed.2004.07.002

Fatima, R., M.Z.A. Khan, Dr.A. Govardhan and K. Dhruve, 2013. Detecting in-situ melanoma using multi parameters. Int. J. Comput. Eng., 4: 16-33.

Gandhi, R.K., K. Marcus and S. Kannan, 2010. Classification rule construction using particle swarm optimization algorithm for breast cancer data sets. Proceedings of the International Conference on Signal Acquisition and Processing, Feb. 9-10, IEEE Xplore Press, Bangalore, pp: 233-237. DOI: 10.1109/ICSAP.2010.58

Gayathri, P. and N. Jaisankar, 2013. Comprehensive study of heart disease diagnosis using data mining and soft computing techniques. Int. J. Eng. Technol.

Gupta, S., K. Dharminder and S. Anand, 2011. Data mining classification techniques applied for breast cancer diagnosis and prognosis. Ind. J. Comput. Sci. Eng., 2: 188-195.

Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. 7th Edn., Morgan Kaufmann, San Francisco, ISBN-10: 1558604898. pp: 550.

Hassanien, E.A. and A.H.M. Jafar, 2004. Rough set approach for generation of classification rules of breast cancer data. J. Inform., 15: 23-38.

Huang, Y., P. McCullagh, N. Black and R. Harper, 2007. Feature selection and classification model construction on type 2 diabetic patients data. Artificial Intell. Med. Elsevier 41: 251-262.

Isabelle, G., W. Jason, B. Stephen and V. Vapnik, 2002. Gene selection for cancer classification using support vector machines. Mach. Learn., 46: 389-422. DOI: 10.1023/A:1012487302797

Jamarani, S.M.H., H. Behnam and G.A.R. Rezairad, 2005. Multiwavelet based neural network for breast cancer diagnosis. IEEE Trans. Image Process., 9: 792-800.

Jayalakshmi, T. and A. Santhakumaran, 2010. A novel classification method for diagnosis of diabetes mellitus using artificial neural networks. Proceedings of the International Conference on Data Storage and Data Engineering, Feb. 9-10, IEEE Xplore Press, Bangalore, pp: 159-163. DOI: 10.1109/DSDE.2010.58

Khan, M.U., J.P. Choi, H. Shin and M. Kim, 2008. Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Aug. 20-25, IEEE Xplore Press, Vancouver, BC., pp: 5148-5151. DOI: 10.1109/IEMBS.2008.4650373

Liang, H.L., D.H. Yi, Q.J. Zheng, J.F. Du and Y.X. Cao *et al*., 2008. Improvement of heart allograft acceptability associated with recruitment of CD4+CD25+ T cells in peripheral blood by recipient treatment with granulocyte colony-stimulating factor. Transplant Proc., 40: 1604-11. DOI: 10.1016/j.transproceed.2008.02.078

Liu, Q.X., M. Hiramoto, H. Ueda, T. Gojobori and Y. Hiromi *et al*., 2009. Midline governs axon pathfinding by coordinating expression of two major guidance systems. Genes Dev., 23: 1165-1170. PMID: 19451216

Lundin, M., J. Lundin, B.H. Burke, S. Toikkanen and L. Pylkkänen *et al*., 1999. Artificial neural networks applied to survival prediction in breast cancer. Oncology, 57: 281-286. PMID: 10575312

MacMillan, W.E., H. William, G. Wolberg and P. Welling, 1978. Pharmacokinetics of fluorouracil in humans. Cancer Res., 38: 3479-3482. PMID: 688233

Michalak, K. and H. Kwasnicka, 2006. Correlation-based feature selection strategy in neural classification. Proceedings of the 6th International Conference on Intelligent Systems Design and Applications, Oct. 16-18, IEEE Xplore Press, Jinan, pp: 741-746. DOI: 10.1109/ISDA.2006.128

Neapolitan, J.L., 2003. Explaining variation in crime victimization across nations and within nations. Int. Crim. Just. Rev., 13: 76-89. DOI: 10.1177/105756770301300104

Orlando, A., G.C. Bruno, V. Susana, G. Jorge and O.L. Arlindo *et al*., 2010. A data mining approach for the detection of high-risk breast cancer groups. Adv. Soft Comput., 74: 43-51. DOI: 10.1007/978-3-642-13214-8_6

Othman, M.F.B. and T.M.S. Yau, 2007. Comparison of different classification techniques using weka for breast cancer. Proceedings of the 3rd Kuala Lumpur International Conference on Biomedical Engineering, Dec. 11-14, Springer Berlin Heidelberg, Kuala Lumpur, Malaysia, pp: 520-523. DOI: 10.1007/978-3-540-68017-8_131

Padmavati, J., 2011. A comparative study on breast cancer prediction using RBF and MLP. Int. J. Sci. Eng. Res., 2: 5-4.

Paulin, F. and A. Santhakumaran, 2011. Classification of breast cancer by comparing back propagation training algorithms. Int. J. Comput. Sci. Eng., 3: 327-332.

Pitchumani, AS. and N.B. Kamal, 2011. Association rule mining based decision tree induction for efficient detection of cancerous masses in mammogram. Int. J. Comput. Applic., 31: 5-5. DOI: 10.5120/3825-5309

Pradhan, M. and R.K. Sahu, 2011. Predict the onset of diabetes disease using Artificial Neural Network (ANN). Int. J. Comput. Sci. Emerg. Technol., 2: 303-311.

Quinlan, R.J. and C.A.M. Kaufmann, 1993. C4.5: Programs for Machine Learning. 1st Edn., Revised, Morgan Kaufmann, ISBN-10: 1558602380, pp: 302.

Rajesh, K. and S. Anand, 2012. Analysis of SEER dataset for breast cancer diagnosis using C4.5 classification algorithm. Int. J. Adv. Res. Comput. Commun. Eng., 1: 72-77.

Rajeswari, P. and G.S. Reen, 2010. Analysis of liver disorder using data mining algorithm. Global J. Comput. Sci. Technol.

Ramalingam, V.V., S.G. Kumar and V. Sugumaran, 2012. Analysis of EEG signals using data mining approach. Int. J. Comput. Eng. Technol., 3: 206-212.

Robnik-Sikonja, M. and I. Kononenko, 2003. Theoretical and Empirical Analysis of ReliefF and RReliefF. Mach. Learn., 53: 23-69. DOI: 10.1023/A:1025667309714

Sarvestan, S.A., A.A. Safavi, M.N. Parandeh and M. Salehi, 2010. Predicting breast cancer survivability using data mining techniques. Procedings of the 2nd International Conference on Software Technology and Engineering, Oct. 3-5, IEEE Xplore Press, San Juan, pp: 227-231. DOI: 10.1109/ICSTE.2010.5608818

Shweta, K., 2012. Using data mining techniques for diagnosis and prognosis of cancer disease. Int. J. Comput. Sci. Eng. Inf. Technol. 37: 52-52.

Soni, J., U. Ansari, D. Sharma and S. Soni, 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. Int. J. Comput. Applic., 17: 43-48. DOI: 10.5120/2237-2860

Srinivasa, R. and P.R. Sujatha, 2011. Analysis of colon cancer dataset using k-means based algorithms and See5 algorithms. School Inform. Tech. Eng., 2: 482-484.

Street, W.N., 1998. A neural network model for prognostic prediction. Proceedings of the 15th International Conference on Machine Learning, (ICML '98), Morgan Kaufmann, San Francisco, pp: 540-546.

Sudhir, D.S., A.A. Ghatol and A.P. Pande, 2006. Neural network aided breast cancer detection and diagnosis. Proceedings of the 7th WSEAS International Conference on Neural Networks, (NN'06), Wisconsin, USA, pp: 158-163.

Thangavel, K. and A.K. Mohideen, 2009. Classification of microcalcifications using multi-dimensional genetic association rule miner. Int. J. Recent Trends Eng., 2: 233-235.