International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN : 2456-172X | Vol. 2, No. 4, December 2017- February, 2018 | Pages 227-235

Analysis of Classification Performance Using Hybrid Algorithm

Shiny R.M[1] , Jetlin C P[2]

Assistant Professor-Department of CSE,
Agni College of Technology, Chennai

ABSTRACT

This study focuses on feature subset selection from high dimensionality databases and presents modification to the existing Random Subset Feature Selection(RSFS) algorithm and Recursive Feature Elimination(RFE) for the random selection of feature subsets and for improving stability. A standard k nearest-neighbour (kNN) classifier is used for classification. The RSFS and RFE algorithms are used for reducing the dimensionality of a data set by selecting useful novel features. It is based on the random forest algorithm. The current implementation suffers from poor dimensionality reduction and low stability when the database is very large. In this study, an attempt is made to improve the existing algorithm's performance for dimensionality reduction and increase its stability. The proposed algorithm was applied to scientific data to test its performance. With 10 fold cross-validation and modifying the algorithm classification accuracy is improved. From the results it is concluded that the improved algorithm is superior in reducing the dimensionality and improving the classification accuracy when used with a simple kNN classifier. The data sets are selected from public repository. The datasets are scientific in nature and mostly used in cancer detection. From the results it is concluded that the algorithm is highly recommended for dimensionality reduction while extracting relevant data from scientific datasets.

1.0 Introduction

Due to the advancement in technology there are large amount of unprocessed information. It is time consuming to view or extract the needed information. In such a situation we are in need to develop a strategy which is useful to obtain the necessary information. Since there are large amount of data decision making process is tedious. To overcome these pitfalls the concept of Data Mining is used. The techniques of data mining will help the users to acquire the essential information .

Data mining is an area of computer science with a huge prospective, which is the process of discovering or extracting information from large database or datasets. Data mining is the process to extract information from a data set and transform it into an understandable structure. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. It is

a technique which is used primarily for discovering unknown patterns and that converts raw data into user understandable information. Nowadays it is being increasingly used in science and technology to extract the vast amount of data.

Data Mining has three major components Clustering or Classification, Association Rules and Sequence Analysis. By simple definition, in classification/clustering analyze a set of data and generate a set of grouping rules which can be used to classify future data. Classification also can be implemented through a number of different approaches or algorithms**.** Data mining techniques broadly classified into two categories. They are predictive and descriptive. Both of these methods are used to extract the hidden patterns from huge amount of data.

Classification:
Classification is the process of converting the data records into set of classes. It is divided into Supervised classification and unsupervised classification. In supervised classification, the data that are to be classified is previously known based on few assumptions. In Unsupervised classification, the set of cases were not predicted by the users. By some assumption it is the job of the user to classify the given data and try to assign the name for those cases. This type of classification is known as clustering. Classification involves predicting a certain outcome based on a given input. In order to predict the results, it needs to fetch the data already available. Based on this data the records are classified. The data sources can be categorized into training set and test set. The training set contains the data which are classified before and it used as a reference for classification purpose. With the help of the attributes the results are predicted. Next the test data is supplied to the algorithm. These data are checked against the attribute which are stored previously and based on these assumptions the data are classified. The algorithm analyses the data given and predicts the results.

Classification is a major technique in data mining and widely used in various fields. Classification is a data mining (machine learning) technique used to predict group membership for data instances. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. The data analysis task classification is where a model or classifier is constructed to predict categorical labels (the class label attributes). Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating. Classification has many applications in customer segmentation, business modelling, marketing, credit analysis, and biomedical and drug response modelling.

The paper is arranged as follows: Section 2 delve into the details regarding the literature survey that has been undertaken and the inference from the same. Section 3 brings out the broad view of the existing system, its limitations and the system design of the proposed methods along with its module specifications. Section 4 renders about the experimental analysis and the results of the experiments. Section 5 concludes the report with the future work that can be carried out.

2.0 Literature Review

Feature Subset Selection(FSS) is an important step in the data mining process to select the relevant feature subset from a large dataset before classification. The purpose of the feature subset selection is to improve performance and select effective predictors by addressing the dimensionality. Selection of the most important and relevant features from high dimensional scientific data ,for the classification task currently faced by many data mining professionals.

Lei Yu et.al(2004) illustrated about feature selection to reduce the number of features in many applications where data has hundreds or thousands of features. Existing feature selection methods mainly focus on finding relevant features. In this paper, it shows the feature relevance alone is insufficient for efficient feature selection of high-dimensional data. It defines feature redundancy to perform explicit redundancy analysis in feature selection. A new framework is introduced that decouples relevance analysis and redundancy analysis. Correlation-based method for relevance and redundancy analysis are developed and an empirical study is conducted for measuring the efficiency and effectiveness comparing with representative methods.

In this paper, it is identified the need for explicit redundancy analysis in feature selection, provided a formal definition of feature redundancy, and investigated the relationship between feature relevance and redundancy. New framework is proposed for efficient feature selection via relevance and redundancy analysis, and a correlation-based method which uses C-correlation for relevance analysis and both C- and F-correlations for redundancy analysis.
A new feature selection algorithm FCBF is implemented and evaluated through extensive experiments comparing with three representative feature selection algorithms. The feature selection results are further verified by two different learning algorithms. The method demonstrates its efficiency and effectiveness for feature selection in supervised learning in domains where data contains many irrelevant and/or redundant features.

Thomas Abeel et.al (2008) demonstrated the Robust Feature Selection Techniques. Robustness or stability of feature selection techniques is a topic of recent interest , and is an important issue when selected feature subsets are subsequently analysed by domain experts to gain more insight into the problem modelled. In this work, the use of ensemble feature selection techniques are investigated , where multiple feature selection methods are combined to yield more robust results. It shows these techniques show great promise for high-dimensional domains with small sample sizes, and provide more robust feature subsets than a single feature selection technique. In addition, it also investigate the effect of ensemble feature selection techniques on classification performance, giving rise to a new model selection strategy. The robustness of feature selection techniques will gain importance in the future, and the topic of ensemble feature selection techniques might open many new avenues for further research.

G.Kesavaraj et al (2013) demonstrated on the Classification techniques in Data mining. Data mining is a process of inferring knowledge from such huge data. Data Mining has three major components Clustering or Classification, Association Rules and Sequence Analysis. It is one of the kind of computational process of discovering patterns in large datasets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The

actual data mining task is the automatic or semiautomatic analysis of large quantities of data to extract previously unknown interesting patterns. Classification is a major technique in data mining and widely used in various fields. Classification is a data mining(machine learning) technique used to predict group membership for data instances. In this paper, the basic classification techniques are presented. Several major kinds of classification method including decision tree induction, Bayesian networks, k nearest neighbour classifier. Decision tree commonly used data mining algorithm. A Bayesian network, Bayesian networks are Directed Acyclic Graphs(DAG) whose nodes represent random variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters or hypotheses. Edges represent conditional dependencies; nodes which are not connected represent variables which are conditionally independent of each other. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node. The goal of this study is to provide a comprehensive review of different classification techniques in data mining. The goal is to generate more certain, precise and accurate system results.

Lakshmi Padmaja Dhyaram(2018) illustrated about the Random Subset Feature Selection Method for Classification. Feature Subset Selection(FSS) is a process to select subset of relevant features, whereas Random Feature Selection (RSFS) process , randomly selects the subset of relevant features from the dataset to avoid the bias and over fitting selected subsets, for classifying the features from high dimensional datasets. Feature selection is a supervised inducting the learning algorithm to find the relevant features and to eliminate the irrelevant features or redundant features for better accuracy. Feature selection to improve performance, to visualize the data for model selection and to reduce dimensionality and remove noise. This paper describes the overview of Random subset feature selection, techniques used in and its application areas in different fields. The main goal of data mining techniques is to discover the knowledge from active data. Out of all technique’s classification is one of the prominent areas to find relevant features from large amounts of data. In future work various learning algorithms for random subset feature selection method can be developed.

To reduce the number of features in large datasets a new framework is introduced that performs relevance analysis and redundancy analysis. Random Subset Feature Selection Algorithm is used to find the relevant features and to eliminate irrelevant features for better accuracy. The important problem is to enhance the accuracy and performance. Feature Selection Algorithms is used to increase performance and stability. Different algorithms with classifier provide the improvement in classification performance.

3.0 System Architecture

Random subset selection is a useful approach to find the variability of static data such as location and scaling estimates. When no outliers or data anomalies the random subset results provides a useful measure of the inherent variability of the data characterization of interest. The main goal of the random subset selection is to generate q subsets to generate from the data set D each of the same size M. The simple version of random subset selection is probably random selection with replacement in which each of the M elements of the subset Si is randomly drawn from the size N data set independently with probability 1/N.

The existing RSFS algorithm has three tasks, namely (1)preprocessing,(2)random subset feature selection and (3)classification

Challenges:

Several challenges are encountered with the existing methods of feature subset selection methods resulting a needed for improved one. The Challenges are:

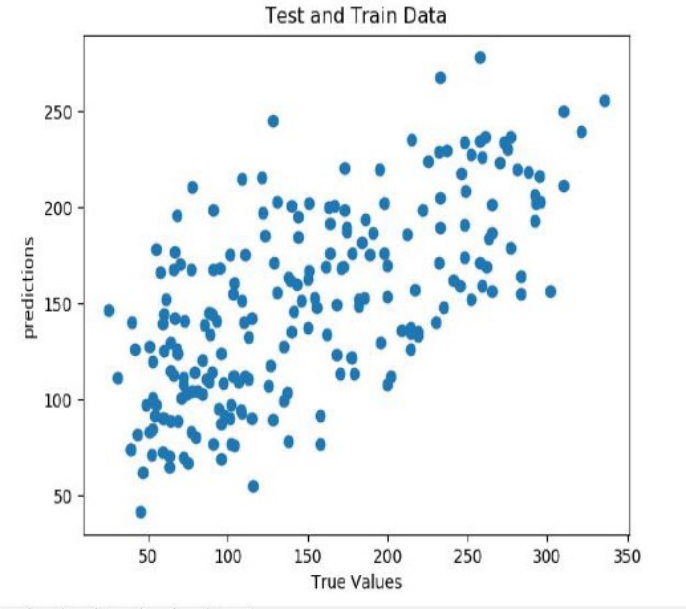
In filter based methods, the feature utility is computed as the correlation between feature and class label. This is carried out by ranking the features in descending order and selection the top scores. This method cannot capture the interaction between the features. This results in suboptimal results and it is very difficult to decide the optimal subset. Using Wrapper methods, the feature subset selection relatively more accurate when compared to the filter method but the feature subsets that are overly specific to the used classifier.

Both Filter and Wrapper methods uses search strategy for identifying the subsets but the challenge is based on scalability. The existing RSFS algorithm works well as long as the data set has a moderate to high number of features. However the number of features in the dataset exceeds ,the classification accuracy starts to decline, and reliability in the dataset becomes very poor. As a result there is a need for modification of the existing algorithm to enhance the performance and improve reliability.

After considering the challenges in the existing RSFS algorithm ,the modified algorithm is proposed, In order to achieve improved accuracy and reliability for high dimensional datasets, the modified RSFS algorithm is presented as,

- Read the true feature dataset and randomly generate the dummy feature set.
- Split the data sets into 2/3 training datasets and 1/3 testing datasets.

Figure shows the data set split into train and test data



- Randomly selects the subsets for the true feature and dummy feature dataset.
- Classify the subsets using the KNN classifier and calculate the relevancies using performance and expected criteria.
- Convert all the relevancies into a standard normal distribution and select the subsets based on the threshold values.
- Sort the features in descending order and select the top most features and display the results.

Advantages:
The main advantages to observe in proposed(improved) RSFS algorithm are, which are not there in existing algorithm.

- Separate training and testing datasets are given as input for realistic accuracy in improved RSFS algorithm.
- Consistency of the final condensed subset is high.
- Relatively lesser number of iterations will be taken for achieving the same result.
- The benefits can be observed better for large data sets.
- The number of features is reduced in proposed algorithm compared to existing algorithm.

4.0 System Implementation

Step 1-Data Pre-processing : Data pre-processing is the datamining technique that involves transforming raw data into understandable format. Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. The dataset used is a binary classification problem where all of the attributes are numeric and have different scales. When our data is comprised of attributes with varying scales, rescaling the attributes is done to all have the same scale**.** Transformation of data using a binary threshold. is needed to transform all values above the threshold to 1 and all equal and below are marked as 0. This is called binarizing the data. Data Pre-processing is used as database driven applications such as customer relationship management and rule based applications. There are several transformation techniques such as min-max, Z –score and decimal scaling of these methods Z score is the most powerful. It converts all indicators to a common scale with an average of zero and standard deviation. The average of zero means that it avoids introducing aggregation distortions stemming from differences in indicators means. Data goes through a series of steps during pre-processing:
Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
Data Integration: Data with different representations are put together and conflicts within the data are resolved.
Data Transformation: Data is normalized, aggregated and generalized.
Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.
Data Discretization: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.
Steps in Data Preprocessing:
Step 1 : Import the libraries
Step 2 : Import the data-set

Step 3 : Check out the missing values
Step 4 : See the Categorical Values
Step 5 : Splitting the data-set into Training and Test Set
Step 6 : Feature Scaling

Step 2-Random subset feature selection: Random Subset Feature Selection(RSFS), is an algorithm that is used for selecting relevant features from a large data set which is based on the Random Forest algorithm .A random forest is an ensemble of random decision tree classifiers, that makes predictions by combining the predictions of the individual trees. Different random forests differ in how randomness is introduced in the tree building process. The relevant feature set is a reduced data set, which helps to improve the performance of classification task .

In RSFS, the features are selected by choosing a random subset from a feature set and then classifying with a KNN classifier. In each iteration, the relevance of each feature is computed and updated based on its performance. With more iterations, the quality of relevant feature selection improves gradually. In this algorithm, a set of dummy features are selected to create a random walk process and shape parameter. In each iteration, the relevancies are computed using the same process.

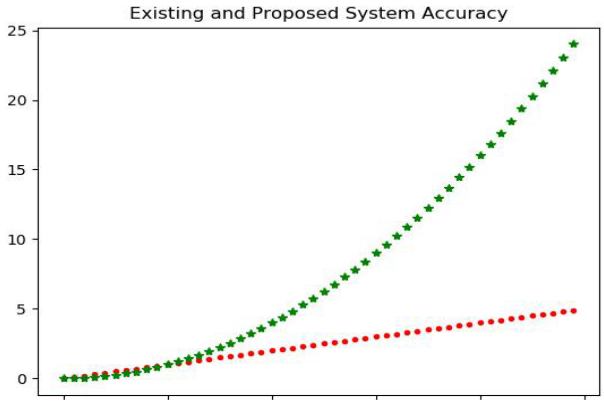
Step 3-Classification: It is a Data analysis task. The process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (sub populations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known. During the classification, kNN classifiers are used to classify the random subsets generated by the random forest algorithm. Since kNN is stable, re-sampling is not necessary for kNN. Each kNN classifier classifies a test point by the majority, or weighted majority class, of its k nearest neighbors. The final classification in each case is determined by a majority vote of random kNN classifications.

A KNN model can be implemented by following the below steps:
1. Load the data
2. Initialize the value of k
3. For getting the predicted class, iterate from 1 to total number of training data points
i)Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.
ii)Sort the calculated distances in ascending order based on distance values
iii)Get top k rows from the sorted array
iv)Get the most frequent class of these rows
v)Return the predicted class

In a high dimensional dataset, there remain some entirely irrelevant, insignificant and unimportant features. These features has to be removed for getting the good accuracy during classification .The Recursive Feature Elimination (or RFE) works by recursively removing

attributes and builds a model on those attributes that remain and the RSFS algorithm identifies the best possible feature subset from the large data set.

KNN classifiers are used to classify the data sets and the results of each prediction are compared to the actual class value in the test set. At the end of the run the accuracy of the model is found as 99%.which is more higher than the existing system and the graph is visualized comparing the existing and proposed system which is shown in the **Figure**



5.0Conclusion and Future Work

Based on the results, it is observed that the various feature selection algorithm is used for reducing the dimensionality of the scientific datasets in the existing algorithm and it does not compromise the accuracy. In the current study, the improved algorithm was iteratively applied on the training data set to enhance the classification accuracy. In the final result, it is evident that the informative features are better chosen by the classifier even for the high dimensional datasets. In every iteration the selected features are compared with final subset array to discard the duplicate features. In every iteration, the training data set is modified by omitting the strong two features which are selected in the previous iteration. This enables to converge the solution faster and more relevant features are selected.

Although the modification was successfully implemented on high dimensional scientific data, to fully address the curse of dimensionality, more study is required to understand the behaviour of the solution when applied to sparse datasets. It is also required to undertake future study, on multi-class data with a combination of classifiers. In future the detailed study is required using the different type of classifiers such as Support Vector Machines(SVM) , Decision Tree Artificial Neural Network(ANN) to understand the performance of the algorithm.

References:

a) *Priyank Pandey, Manoj Kumar and Prakhar Srivastava ,"Classification Techniques for Big Data :A Survey ", IEEE International Conference on Computing for Sustainable Global Development , pp 3625 - 3629, Mar. 2016.*
b) *D. L. Padmaja, B. Vishnuvardhan, "Comparative study of feature subset selection methods for dimensionality reduction on scientific data", in: Advanced Computing (IACC), 2016 IEEE 6th International Conference on, IEEE, pp. 31–34.*

c) *Arguello, et al., "A survey of feature selection methods: algorithms and software", Ph.D. thesis, 2015.Z. M. Hira, D. F. Gillies, A review of feature selection and feature extraction methods applied on microarray data, Advances in bioinformatics 2015 (2015).*
d) *Girish Chandrashekar and Ferat Sahin," A Survey On Feature Selection Methods", Journal of Computers and Electrical Engineering, vol.40 , pp 16 - 28, Dec. 2013.*
e) *A. Sisto, C. Kamath, "Ensemble Feature Selection in Scientific Data Analysis", Technical Report, Lawrence Livermore National Laboratory (LLNL), Livermore, CA, 2013*
f) *Vens, F. Costa, "Random forest based feature induction", in: 2011 IEEE 11th International Conference on Data Mining, IEEE, pp. 744–753.*
g) *Li, S, Harner, J., Adjeroh, D, "Random kNN feature selection – a fast and stable alternative to random forests". BMC Bioinformatics (2011).*
h) *Y.Saeys, T. Abeel, Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques", in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer 2008, pp. 313–325.*
i) *H. Peng, F. Long, C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", IEEE Transactions on pattern analysis and machine intelligence 27 (2005) 1226–1238*
j) *F. Livingston, "Implementation of breiman's random forest machine learning algorithm" ece591q machine learning journal paper, 2005.*
k) *L.Yu, H. Liu, "Efficient feature selection via analysis of relevance and redundancy", Journal of machine learning research 5 (2004) 1205–1224.*
l) *L. C. Molina, L. Belanche, À. Nebot, "Feature selection algorithms: a survey and experimental evaluation", in: Data Mining, 2002. ICDM 2003.Proceedings. 2002 IEEE International Conference on, IEEE, pp. 306–313.*
m) *A. L. Blum, P. Langley, "Selection of relevant features and examples in machine learning", Artificial intelligence (1997) 245–271.*
n) *G. V. Trunk, "A problem of dimensionality":A simple example, IEEE Transactions on pattern analysis and machine intelligence (1979) 306–307.*