International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S3, September 2019

PCA based Regression Decision Tree Classification for Somatic Mutations

Anuradha Chokka, K Sandhya Rani

Abstract.The analization of cancer data and normal data for the predication of somatic mu-tation occurrences in the data set plays an important role and several challenges persist in de-tectingsomatic mutations which leads to complexity of handling large volumes of data in classifi-cation with good accuracy. In many situations the dataset may consist of redundant and less significant features and there is a need to remove insignificant features in order to improve the performance of classification. Feature selection techniques are useful for dimensionality reduction purpose. PCA is one type of feature selection technique to identify significant attributes and is adopted in this paper. A novel technique, PCA based regression decision tree is proposed for classification of somatic mutations data in this paper.The performance analysis of this clas-sification process for the detection of somatic mutation is compared with existing algorithms and satisfactory results are obtained with the proposed model.

Keywords: Somatic mutations, Feature Selection, Regression Based Decision Trees (RDT).

I. INTRODUCTION

Present day's machine learning techniques are useful for programming domain to classify the programming modules as adequate or not adequate with the goal that early detection of damaged modules can be amended and tested before finalizing the module. This may prompt the quality result of the module and furthermore there might be cost advantage. Classification is a mainstream approach for prediction and classifies the attributes into deficient attributes or not deficient, which can be done by different classification methods [1]. One of the classification techniques is Decision Tree learning which is the most frequently utilized and useful techniques over supervised learning information. Based on different properties decision tree signifies a strategy that classifies the categorical as well as numerical data. This decision tree is also used for handling huge amount of information and thus has the use in machine learning applications. Hence decision trees are proper for experimental learning detection with their portrayal of enhanced information in hierarchical tree construction which is instinctive and makes understandable to all. The overhead of utilizing the decision tree is logarithmic for the preparation of a tree in view of the consideration of dimensions of data points.Decision trees can deal with multi-yield issues. Decision trees can perform well regardless of whether the presumptions are somewhat damaged by the considered dataset [2]. When decision tree learning approach continues to improve hypotheses, causes to decrease training data error at the rate of an improved test data error which causes to make a huge size of decision tree procedure called an over-fitting. Because of over-fitting, the decision tree may lose some generalization capability. Over-fitting is formed by using noisy information and insignificant attributes and makes misclassification and data imbalance [3]. There by over fitting decreases the performance of decision tree with higher amount of dimensions in classification model. So as to decrease large data dimensions a typical methodology used for attributes is a feature reduction to acquire lesser dimensional information that depends upon features which is considered for the problem scenario. Feature reduction is automatically followed by feature selection which is used by correlation and gain ratio techniques [4]. Feature Selection has been generally implemented in various parts of software engineering and machine learning. Feature Selection is a significant method to decrease the dimensions of huge dimensional samples. The benefits of feature selection are that, it enables the entire strategy to be executed computationally high effective and well-organized and feature selection accomplishes the proliferation of its accuracy. In this paper, a Feature Selection technique, called Principal component analysis (PCA) used for the selection of features [5]. PCA has several applications in various fields like Pattern Recognition, Machine Learning and information compression. The strategy projected in this paper isn't just totally novel but also easy and instinctive. In this paper, applying PCA as a feature selection technique is yet achievable and feasible process to choose significant feature components from all the feature components of actual samples. The main purpose of using PCA is to remove insignificant features by choosing applicable and non-correlated features without changing the data enclosed in the actual information and after that classification is established using regression decision tree algorithm to analyze the data [6]. The covariance matrix is calculated for finding the Eigen values, Eigen vectors and after that principal components are selected by taking the Eigen vector with highest Eigen values of the cancer data collection.

Manuscript published on 30 September 2019.
***** Correspondence Author (s)
Anuradha Chokka, Research Scholar, Dept. of Computer Science, Sri PadmavathiMahilaVisvavidyalayam, Tirupati, AndhraPradesh, India.(Email: akshayagokul2009@gmail.com)
Dr. K Sandhya Rani, Professor, Dept. Of Computer Science, Sri PadmavathiMahilaVisvavidyalayam, Tirupati, A.P, India.(Email: sandhyaranikasireddy@yahoo.co.in)

In order to select the significant features and to eliminate the insignificant features, the Eigen values are arranged in ascending order. By this process, higher dimensional information is decreased to lesser dimensions. The main objective of this paper is to obtain pertinent features by PCA technique and after that an anticipation of classification model is developed by using regression decision tree to enhance the accuracy of decision tree. The earlier work related to the proposed is presented in the next section.

II. RELATED WORKS

In [5], the authors implemented the principal components analysis (PCA) for the selection of features and the anticipated technique inclined well for the consideration of feature selection concern. The study demonstrated that PCA chosen various significant entities from the entire feature components. The contrived algorithm is exposed to the idea of PCA and furthermore computationally productive and effective. The computed results on face recognition demonstrated that the proposed technique can incredibly decrease the dimensionality of the actual data without decreasing the recognition accuracy.

Andreas G.K. Janecek and WilfriedN.Gansterer [7], investigated the relationship between a few attribute space reduction methods and the subsequent classi?cation accuracy for two various application regions. Subsets of the actual attributes are constructed by ?lter and wrapper procedures, developed by the principal component analysis (PCA) are compared in the form of classi?cation performances accomplished with different machine learning algorithms along with runtime performance. They progressively decreased the size of the feature sets and investigated the relationship between the variance attained in the linear combinations within PCA and also improved the classi?cation accuracy.

The Author [8] discussed the performance of various classification techniques in analyzing Breast Cancer data through the analization of mammogram images. Execution analization of classification algorithms like J48, CART and Decision Tree are indicated with its precision. The performance of taken algorithms is estimated with the parameters like specificity, sensitivity and kappa statistics.

For dealing with high dimensions of dataset, Decision tree is one of the complex and profound procedures. The objective of this study [9] is to analyze the kidney transplant patient's report by considering the group of predicted factors using collective strategies. This paper compared the classification accuracy of various decision tree algorithms such as ID3, C4.5 and CART, and also ensemble methods like Random forest, Boosting and Bagging techniques with the approaches C4.5 and CART. The result showed that the technique Classification and Regression Tree (CART) along with Boosting approach yielded the better results than remaining methods.

III. CANCER DATASETS

The six cancer data sets which are considered in this paper are obtained from https://github.com/ikalatskaya/ISOWN [10] and datasets are prepared from COSMIC repository. The various attributes in cancer data sets are shown in Table 1.

Table 1: Attributes Information in Cancer Data Set

```
@relation Somatic Vs Germ line
@attribute ExAc {true, false}
@attribute dbSNP {true, false}
@attribute CNT numeric
@attribute fre numeric
@attribute VAF numeric
@attribute mutAss {'neutral','low','medium','high','stopgain','stoploss'}
@attribute pattern {'CG', 'CA', 'CT', 'TA', 'TC', 'TG'}
@attribute SeqContext
{'ATT','CTT','GTT','TAT','AAA','CAA','AAC','CAC','GAA','AAG','CAG','GAC','GAG','TGA','TGC','TCA','AAT'
,'TCC','TGG','CAT','TCG','GAT','TGT','TTA','TTC','TCT','TTG','TTT','AGA','CGA','AGC','CGC','ACA','CCA','
GGA','ACC','AGG','CCC','CGG','GGC','GCA','ACG','CCG','GCC','GGG','GCG','AGT','ATA','ATC','CGT','CTA',
'ACT','CTC','ATG','CCT','GGT','GTA','TAA','CTG','GTC','TAC','GCT','GTG','TAG',}
 @attribute isFlanking numeric
 @attribute polyphen {'benign', 'probably', 'possibly'}
 @attribute  {true, false}
```

IV. PROPOSED MODEL

The main aim of proposed model is to develop a Regression based Decision Tree classifier for the detection of somatic mutations and to improve the performance of the classifier, PCA-Feature Selection technique [11] is considered to find the significant attributes, which plays a vital role for identification of somatic mutations. The cancer data set discussed in the previous section is considered for expermental purpose. The given data set consists of numerical as well as categorical values andfor performance purpose all categorical values are converted into numerical values. For example the attribute mutAsswhichiscategorical value and can take six different values such as neutral, low, medium, high, stopgain, stoploss. This variable information can be represented in numerical form by splitting into six numerical values for computational purpose. This attribute is having any one of the six values. The value presented for the corresponding column is 1 and other columns are 0.

This procedure is adopted for all categorical attributes such as Pattern, Seq Context, Polyphen etc. for converting into numarical values for computational purposes. Now we have appplied PCA algorithm in order to find out most significant attributes.

To perform feature selection with PCA the following algorithm is devised. The first step is to find the significant attributes by using PCA Feature Selection method [4]. The various steps in PCA algorithm is explained in the following section.

4.1 PCA Algorithm

Input: Cancer Dataset with numeric values.

Output: Reduced set of attributes.

Step 1: Standardize each attribute and Calculate variance, standard deviation by

using the following formula.

$$\text{Var (P)} = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - y_j)^2$$

Step 2: Compute correlation of every attribute that constructed on covariance

formula.

Covar (p, q) =

Step 3: Compute Eigenvalues of every attribute or feature to generate

Principal Components (PC).

Step 4: Calculate Principal Components by taking the proportion of variance.

These components collectively augmented for 95% of principal

component from computed Eigen vectors.

Step 5: Find Eigenvectors by transposing and multiplying the values of matrices.

Eigenvector's every element signifies the impact of given attributes to

Principal Component (PC).

Based on the correlations among the actual attributes and the principal components compute the new attributes. The considered threshold for this computation of attributes is 0.5.The attributes which are having the threshold above 0.5 are considered as significant attributes. Only these attributes are considered for further development of classifiers.

4.2 Regression Decision Tree Classifier

After obtaining the significant attributes by applying PCA on Dataset, the next step is to develop a classifier based on these significant attributes. In this paper, Regression Decision Tree Classifier is considered for classification of somatic mutations. A brief description of Regression based Decision Tree is presented in the next section.The main approach for developing decision trees is to split the input dataset into hyper cubes. This is done in a hierarchical manner using a sequence of binary splitting rules for the given inputs. Decision tree techniques are for classification and regression.The principle used for enhancing decision trees is the consideration of prediction error as mean-square error for regression trees. Regression is a very important machine learning problem. To split the attributes of a tree the impurity measure is considered and the split point to be taken is the standard deviation of the predictor for the dataset at the tree node [12]. The mean square error of the tree node is a more appropriate impurity measure for the regression decision trees. Regression based decision Tree classifies the dataset for numeric and continuous target variables. It uses standard deviation method for splitting dataset and calculates lowest standard square error for splitting the dataset. The most common method for building a regression decision tree model based on a sample is to obtain the model parameters that maximize the decrease in the error of the tree [13], which means the best split is the split that maximizes least square error reduction.

4.2.1.Regression Decision Tree Algorithm

Input: Dataset with subset of attributes or features.

Output: A regression Decision Tree with leaf nodes as class labels.

Step 1: Start at the root node for Decision Tree

Step 2: For every attribute P, compute standard deviation using the formula

$$\text{SD (P)} = \sqrt{\frac{\sum(p-\bar{p})^2}{n}}$$

Step 3: Calculate the standard deviation for two attributes using the formula

$$\text{SD (T, P)} = \sum_{a \in p} Q(a)SD(a)$$

Step 4: Calculate the standard deviation reduction by using the formula

SDR (T, P) = SD (T) - SD (T, P)

Step 5: Find the attribute with the highest standard deviation reduction, which

is selected for the decision node.

Step 6: The data set is divided based on the values of the selected attributes.

This process is to be performed iteratively on the non-leaf branches

until the entire data is processed and reaches the predicted leaf class labels.

V.EXPERIMENTAL RESULTS

The proposed model is implemented on cancer Dataset by using developed PCA based Regression Decision Tree. The obtained results at different stages are given below. For performance purpose we have converted the categorical values of dataset's attributes into numerical values. After conversion we got 86 attributes. The correlation matrix for the set of features is computed and some portion of the sample results is shown in below Table 2.

1	0.68	0.04	-0.28	-0.41	-0.22	0.01	0.14	0.07	0.14	0.05	0.11	0.14	0.03	-0.17	-0.06	-0.09	0.05	
0.68	1	0.05	-0.38	-0.27	-0.19	-0.01	0.17	0.03	0.1	0.03	0.09	0.14	0.02	-0.19	-0	-0.07	0.02	0
0.04	0.05	1	0.03	-0.03	-0.02	-0.03	0.05	-0.01	0.01	-0	0.03	0.01	-0.01	-0.02	-0.01	-0.02	-0.01	
-0.28	-0.38	0.03	1	0.12	0.14	0	-0.12	-0.05	-0.04	-0.01	0.01	-0.1	0.03	0.03	0.04	-0.04	-0.03	0
-0.41	-0.27	-0.03	0.12	1	0.06	-0.01	-0.04	0.02	-0.07	-0.03	-0.15	-0.07	-0.01	0.2	-0.04	0.14	0.04	
-0.22	-0.19	-0.02	0.14	0.06	1	-0.53	-0.4	-0.13	-0.1	-0.03	-0.05	-0.04	-0.03	0.06	0.04	0.03	0.01	
0.01	-0.01	-0.03	0	-0.01	-0.53	1	-0.43	-0.14	-0.11	-0.03	0.04	-0.05	-0.02	0.03	-0.03	-0.02	-0.06	
0.14	0.17	0.05	-0.12	-0.04	-0.4	-0.43	1	-0.11	-0.08	-0.03	0.02	0.01	-0	-0.03	0	-0.01	0.07	-0
0.07	0.03	-0.01	-0.05	0.02	-0.13	-0.14	-0.11	1	-0.03	-0.01	-0.02	0.05	0.07	-0.06	0.01	0.04	-0.02	
0.14	0.1	0.01	-0.04	-0.07	-0.1	-0.11	-0.08	-0.03	1	-0.01	-0.04	0.22	0.06	-0.14	-0.04	-0.02	-0.02	
0.05	0.03	-0	-0.01	-0.03	-0.03	-0.03	-0.03	-0.01	-0.01	1	-0.02	-0.02	-0.01	0.05	-0.01	-0.01	-0	
0.11	0.09	0.03	0.01	-0.15	-0.05	0.04	0.02	-0.02	-0.04	-0.02	1	-0.21	-0.15	-0.49	-0.15	-0.08	-0.05	
0.14	0.14	0.01	-0.1	-0.07	-0.04	-0.05	0.01	0.05	0.22	-0.02	-0.21	1	-0.13	-0.41	-0.13	-0.07	-0.04	
0.03	0.02	-0.01	0.03	-0.01	-0.03	-0.02	-0	0.07	0.06	-0.01	-0.15	-0.13	1	-0.29	-0.09	-0.05	-0.03	
-0.17	-0.19	-0.02	0.03	0.2	0.06	0.03	-0.03	-0.06	-0.14	0.05	-0.49	-0.41	-0.29	1	-0.29	0.16	0.11	
-0.06	-0	-0.01	0.04	-0.04	0.04	-0.03	0	0.01	-0.04	-0.01	-0.15	-0.13	-0.09	-0.29	1	-0.05	-0.03	
-0.09	-0.07	-0.02	-0.04	0.14	0.03	-0.02	-0.01	0.04	-0.02	-0.01	-0.08	-0.07	-0.05	0.16	-0.05	1	-0.02	
0.05	0.02	-0.01	-0.03	0.04	0.01	-0.06	0.07	-0.02	-0.02	-0	-0.05	-0.04	-0.03	0.11	-0.03	-0.02	1	
-0.02	0	-0.01	0.05	-0	0.08	-0.05	-0.04	0.02	-0.02	-0.01	-0.07	-0.06	-0.04	0.14	-0.04	-0.02	-0.01	

Table 2: The Resultant Correlation Matrix

The obtained Eigenvectors v1, v2, etc. for the sample data are as shown in Table 3.

Table 3: Obtained Eigen Vectors

V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22
V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38 V39 V40 V41
V42 V43 V44 V45 V46 V47 V48 V49 V50 V51 V52 V53 V54 V55 V56 V57
-0.3675 -0.0783 -0.271 -0.2165 0.0468 -0.0144 -0.0064 -0.1125 -0.0626 0.0217 0.0074 0.0048 -0.0394 0.0968 -0.0605 -0.0619 0.1176 0.0342 0.0341 0.0032 0.005 0.0274 0.0017 -0.0057 0.0048 -0.0086 -0.0006 0.0138 0.0029 -0.0134 -0.0215 0.0038 -0.0061 0.0051 -0.0039 -0.0119 -0.0134 -0.0107 0.0003 -0.006 0.0018 -0.0116 0.0009 -0.0025 0.0016 0.002 0.0032 0.0498 -0.069 0.0043 -0.0585 -0.0082 -0.1231 -0.0259 0.0946 0.0871 0.0704 inExAct=false
-0.3569 -0.0945 -0.2442 -0.1591 0.075 -0.0117 -0.0354 -0.1564 -0.1617 0.01 0.043 -0.008 -0.0931 0.0803 0.039 -0.1578 -0.0818 -0.0273 -0.0189 -0.0006 -0.0112 -0.0352 -0.0028 0.0159 0.001 0.0209 0.0005 -0.0044 0.0023 0.0047 0.0124 0.0009 0.0122 0.0024 -0.0018 0.0083 0.0025 -0.0082 -0.0016 0.0059 -0.0081 0.0038 0.0005 -0.0084 0.0046 -0.0039 0.0025 0.004 -0.0741 0.0926 -0.1356 0.0088 -0.0398 0.0199 0.2339 0.1073 0.0716 dbSNP=false
-0.0429 0.0076 -0.0211 -0.0365 0.0065 0.0906 0.1748 0.1562 0.0285 -0.1405 -0.0147 0.0013 0.3576 0.2974 0.3559 -0.0287 0.0024 0.0301 0.0099 0.0171 0.0041 -0.0323 0.0257 -0.0049 0.0145 -0.016 -0.0052 -0 -0.0103 0.0211 0.0155 -0.0102 0.0052 -0.0018 -0.0138 0.0053 0.0035 -0.0143 0.0049 0.0226 -0.0015 -0.0098 -0.0191 -0.0076 0.0074 -0.0021 0.012 -0.3228 0.1335 0.1371 -0.3362 0.4311 -0.2678 -0.0357 0.0147 -0.0066 0.1453 CNT
0.2091 0.1556 0.1614 0.0606 -0.0048 0.1237 0.0243 0.1437 0.2934 -0.1074 -0.1182 0.0871 -0.0248 0.2397 -0.211 -0.1079 0.0029 0.0168 0.0061 -0.0012 -0.0056 0.0025 -0.0083 0.0078 0.0185 0.0137 0.0032 0.0063 -0.0117 -0.0103 0.0041 -0.0007 0.0033 0.0046 0.0074 -0.006 -0.004 -0.0135 0.0053 0.0111 -0.0051 -0.0033 0.0059 -0.0028 0.0036 -0.0003 0.0065 0.0083 0.1963 -0.0049 0.145 0.1237 -0.0993 -0.1154 -0.1942 -0.0717 -0.0249 fre
0.256 -0.0641 0.2041 0.2304 -0.0591 -0.1196 -0.0621 -0.0875 -0.1786 -0.1178 -0.0969 -0.1774 0.0539 -0.0693 0.0548 -0.2107 0.0445 0.0154 -0.0008 -0.0034 -0.0018 0.0028 0.0163 0 0.0074 -0.0043 0.0057 0.0133 -0.0015 -0.0105 0.0029 0.0015 -0.0047 -0.0021 -0.0056 0.0141 -0.0083 -0.0132 0.0021 0.0043 -0.007 -0.0073 0.004 -0.0034 0.0057 0.003 0.0109 0.0003 0.0735 0.0411 0.0257 0.1703 0.0126 -0.0431 0.2321 0.0674 -0.0837 VAF
0.234 0.25 -0.2568 0.2154 0.0567 0.2275 0.1873 -0.2069 0.0042 0.1596 0.1261 0.1264 -0.1113 -0.014 0.1012 -0.0282 0.0203 0.0127 0.0165 0.0034 0.0037 0.0055 -0.0012 0.008 -0.008 0.0013 -0.0016 0.0055 0.0035 0 -0.0044 0.0003 -0.0013 0.0025 -0.0108 -0.0023 -0.0007 0.0011 -0.0045 -0.008 -0.0018 -0.0064 -0.0044 -0.0005 0.0014 -0.0029 0.0057 0.0045 -0.1237 -0.1893 -0.0804 0.0196 -0.0808 0.1455 0.0203 -0.188 0.2442 mutAss=neutral
0.0065 0.0204 0.2673 -0.4351 -0.1735 -0.381 -0.2336 0.0484 0.0894 -0.0236 0.0876 0.0374 -0.007 0.0518 0.0432 -0.0109 -0.002 0.0134 -0.0044 0.0074 -0.0027 -0.0083 0.0085 -0.0074 0.0067 -0.0001 -0.0142 -0.0026 -0.0036 0.0016 0.0024 0.0019 0.0043 -0.0019 0.0082 0.0018 -0.0045 0.0002 0.0025 0.0076 -0.0053 -0.0006 0.0008 -0.0047 0.0027 0.0023 -0.0066 -0.0342 0.0245 0.2895 0.0402 -0.1345 -0.0872 -0.0887 -0.0384 -0.0237 0.0988 mutAss=low

In this methodology, PCA Feature Selection technique is applied on 86 attributes to find the significant attributes by considering the threshold value as 0.5. The attributes which are having more than threshold value 0.5 are being selected as significant attributes and other attributes are not considered for further analysis. In this way 86 attributes in cancer dataset are reduced to 24 significant attributes whose threshold value is greater than 0.5 by using PCA feature selection. The obtained significant ranked attributes with ranked values are shown in Table 4.

Rank Value	Ranked Attribute
0.9515	polyphen=possibly
0.9181	SeqContent=GAA
0.8911	SeqContent=AAC
0.8643	SeqContent=TGA
0.839	SeqContent=GAG
0.8147	SeqContent=GAC
0.7942	SeqContent=CTT
0.7742	SeqContent=ATT
0.7548	pattern=TG
0.7361	pattern=TC
0.7177	mutAss=neutral
0.6998	VAF
0.6826	Fre
0.6662	CNT
0.6499	dbSNP=false
0.6342	mutAss=low
0.6187	mutAss=medium
0.6034	mutAss=high
0.5881	pattern=CA
0.5729	pattern=TA
0.5578	pattern=CG
0.5426	mutAss=stopgain
0.5275	mutAss=stoploss
0.5124	SeqContent=TGG

Table 4: Obtained Significant Ranked Attributes with ranked values.

The plot representation of selected features using PCA is shown in Figure 1. The x-axis represents the number of attributes. The y-axis represents the variance for every attribute and the threshold value above 0.5 is considered to select the significant attributes.

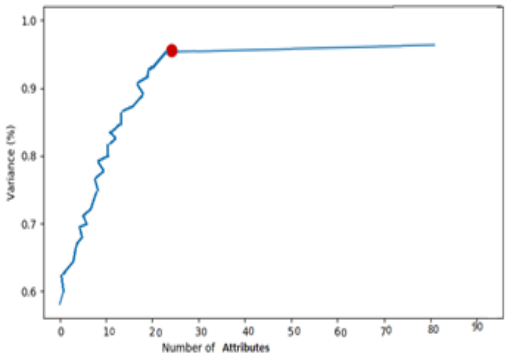

Figure 1:Selecetion of Significant Attributes

These ranked attributes are considered for the development of Regression based Decision Tree and other attributes are ignored. The technique Regression based Decision Tree classifier is implemented on attributes without using PCA and its sample result is shown in below Table 5.

```
| fre< 0.01
| | mutAss=low < 0.5
| | | SeqContent=GAG <0.5 : FALSE (132.86/0)
| | | SeqContent=GAG >= 0.5
| | | | CNT <0.5 : FALSE (3/0)
| | | | CNT >= 0.5 : TRUE (1/0)
| | mutAss=low >= 0.5
| | | pattern=TG < 0.5
| | | | VAF < 44.58
| | | | | SeqContent=ATT < 0.5
| | | | | | CNT < 0.5
| | | | | | | polyphen=possibly < 0.5
| | | | | | | | pattern=CA < 0.5
| | | | | | | | | pattern=TC <0.5 : FALSE (6.54/0)
| | | | | | | | | pattern=TC >= 0.5
| | | | | | | | | | VAF < 39.72
| | | | | | | | | | | VAF <38.95 : FALSE (8.07/0)
```

```
| | | | | | | | | | | VAF >= 38.95 : TRUE (2/0)
| | | | | | | | | | VAF >= 39.72 : FALSE (9.54/0)
| | | | | | | | pattern=CA >= 0.5 : FALSE (3.76/0)
| | | | | | | polyphen=possibly >= 0.5 : FALSE (5.48/0)
| | | | | | CNT >= 0.5 : TRUE (2/0)
| | | | | SeqContent=ATT >= 0.5 : TRUE (1/0)
| | | | VAF >= 44.58 : FALSE (41.76/0)
| | | pattern=TG >= 0.5 : FALSE (4/0)
| fre>= 0.01 : FALSE (140/0)
dbSNP=false >= 0.5
| pattern=CG < 0.5
| | inExAct=false < 0.5
```

```
| | | | | VAF < 33.88
| | | | | | polyphen=possibly < 0.5
| | | | | | | pattern=TG < 0.5
| | | | | | | | SeqContent=GAG < 0.5
| | | | | | | | | pattern=TA < 0.5
| | | | | | | | | | fre<0.04 : TRUE (4.86/0)
| | | | | | | | | | fre>= 0.04 : FALSE (0.93/0)
| | | | | | | | | pattern=TA >= 0.5 : FALSE (2/0)
| | | | | | | | SeqContent=GAG >= 0.5
| | | | | | | | | mutAss=low < 0.5
| | | | | | | | | | VAF <16.46 : FALSE (0.93/0)
| | | | | | | | | | VAF >= 16.46 : TRUE (2.93/0)
| | | | | | | | | mutAss=low >= 0.5 : TRUE (2/0)
| | | | | | | pattern=TG >= 0.5 : FALSE (7.71/0)
| | | | | | polyphen=possibly >= 0.5 : TRUE (1.64/0.43)
| | | | | VAF >= 33.88
| | | | | | SeqContent=GAG < 0.5
| | | | | | | pattern=TA < 0.5
| | | | | | | | CNT < 0.5
```

Table 5: Regression Decision Tree without PCA

After performing Regression Decision Tree classifier on the cancer dataset without using the technique PCA, the obtained classifier accuracy is 90.04%. By using PCA as a Feature Selection Technique on the attributes and the selected subset of features is given to the classification of regression decision tree technique.In this experiment the selected subset of features is 24 attributes out of 86 attributes, on which the Regression Decision Tree classification is performed. The sample results are shown in below Table 6.

```
inExAct=false < 0.5
|  VAF < 40.8
|  |  dbSNP=false < 0.5 : false (99/4) [11/0]
|  |  dbSNP=false >= 0.5
|  |  |  CNT < 0.5
|  |  |  |  pattern=TC < 0.5 : false (15/0) [9/0]
|  |  |  |  pattern=TC >= 0.5
|  |  |  |  |  mutAss=neutral < 0.5 : false (6/0) [1/0]
|  |  |  |  |  mutAss=neutral >= 0.5 : true (8/4) [1/0]
|  |  |  CNT >= 0.5 : true (14/3) [4/0]
|  VAF >= 40.8 : false (273/4) [75/3]
inExAct=false >= 0.5
|  dbSNP=false < 0.5 : false (5/0) [5/1]
|  dbSNP=false >= 0.5 : true (372/7) [93/4]
```

Table 6: Sample Results for Regression Decision Tree with PCA

The accuracy obtained by PCA based Regression Decision Tree (RDT) classifier is 97.94%. The accuracy and precision which are obtained from the classification is shown in Table 7.

	Accuracy	Precision
RDT without PCA	90.04%	70.3%
PCA+RDT	97.94%	99%

Table 7: Performance of Regression Decision Tree (RDT) without PCA and with PCA

VI. CONCLUSION

The proposed model PCA based Regression Decision Tree (RDT) is applied on cancer Dataset in order to classify the somatic mutational Dataset. The dataset is obtained from COSMIC repository. The application of PCA reduced the dataset by considering only significant attributes. Experiments are carried out with PCA and without PCA for Regression Decision Trees classifier to classify somatic mutations Dataset. The experimental results proved that the proposed Regression decision tree classifier with PCA yielded 97.94% accuracy and 99% precision. Hence it is proved that Regression Decision Tree classifier with PCA is better classifier than RDT without PCA.

REFERENCES

1. N.Gayatri, S.Nickolas, A.V.Reddy, "Feature Selection Using Decision Tree Induction in Class level Metrics Dataset for Software Defect Predictions", Proceedings of the World Congress on Engineering and Computer Science 2010 Vol I WCECS 2010, October 20-22, 2010.
2. Bhumika Gupta, Akshay Jain, Aditya Rawat, "Analysis of Various Decision Tree Algorithms for Classification in Data Mining", International Journal of Computer Applications (0975 - 8887) Volume 163 - No 8, April 2017.
3. Autsuo Higa, "Diagnosis of Breast Cancer using Decision Tree and Artificial Neural Network Algorithms", International Journal of Computer Applications Technology and Research Volume 7-Issue 01, 23-27, 2018, ISSN:-2319-8656.
4. M Z F Nasution, O S Sitompul and M Ramli, "PCA based feature reduction to improve the accuracy of decision tree c4.5 classification", 2 nd International Conference on Computing and Applied Informatics 2017 , Journal of Physics: Conf. Series 978 (2018).
5. Fengxi Song, Zhongwei Guo, Dayong Mei, "Feature selection using principal component analysis", 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization.
6. N.Gayatri, S.Nickolas, A.V.Reddy, "Feature Selection Using Decision Tree Induction in Class level Metrics Dataset for Software Defect Predictions", Proceedings of the World Congress on Engineering and Computer Science 2010 Vol 1 WCECS 2010, October 20-22, 2010, San Francisco, USA.
7. Andreas G.K.Janecek ,Wilfried N.Gansterer, "On the Relationship Between Feature Selection and Classi?cation Accuracy", JMLR: Workshop and Conference Proceedings 4: 90-105.
8. B.Padmapriya, T.Velmurugan, "Classification Algorithm Based Analysis of Breast Cancer Data", International Journal of Data Mining Techniques and Applications Volume 5, Issue 1, June 2016, Page No.43-49.
9. Yamuna N R, Venkatesan P , "A Comparative Analysis of Decision Tree Methods to Predict Kidney Transplant Survival", International Journal of Advanced Research in Computer Science, Volume 5, No. 3, March-April 2014.
10. Quang M. Trinh, Melanie Spears, John D. Mc Pherson, "ISOWN: accurate somatic mutation identification in the absence of normal tissue controls Irina Kalatskaya", Genome Medicine, (2017) 9:59.
11. Noor T. Mahmood, Salah T. Allawi, "Modified PCA Based on JK Method for Ranking to Select Features in Statistical DataSets", International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 8, August 2016.
12. Pooja Gulati, Amitasharma, Manish Gupta, "Theoretical Study of Decision Tree Algorithms to Identify Pivotal Factors for Performance Improvement: A Review", International Journal of Computer Applications (0975 - 8887) Volume 141 - No.14, May 2016.
13. Alin Dobra, Johannes Gehrke, "SECRET: A Scalable Linear Regression Tree Algorithm", in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2002.