Improved Classification of Somatic Mutations Using AdaBoost With Feture Se-lection

Anuradha Chokka, K Sandhya Rani

Abstract - The normal cells in human are transformed to cancer cells due to sequence of abnormal genetic events and cancer can be considered genetic changes of somatic mutations. To find the somatic mutations in accurate manner is the major challenge in cancer research. The main difficulty in cancer prediction analysis lies on tumor samples with the contamination and normal data samples. Identifying somatic mutations in cancer genes is a complex process. Feature extraction techniques retrieve significant features from the data and the classifiers which are developed based on these features improve the performance of the classifier. In this paper, to maximize the precision AdaBoost technique with feature selection is applied to detect the gene changes among the normal and tumor cells which are the causes of somatic mutations. The experimental results proved that AdaBoost with the feature selection method improves the performance of classifier in terms of precision, accuracy, and recall.

Keywords - Cancer Prediction, Somatic Mutations, AdaBoost T echnique, Feature Selection.

1. INTRODUCTION

Cancer is a heterogeneous disease that can develop in different tissues and cell types. Even within one cancer type, the disease may manifest itself in multiple subtypes, which are usually distinguished based on different specific mutations, and which may lead to different clinical outcomes. Somatic mutations are a primary contributor to malignancy in human cells. Accurate detection of mutations is needed to define the clonal composition of tumors whereby clones may have distinct phenotypic properties. Although analysis of mutations over multiple tumor samples from the same patient has the potential to enhance identification of clones. Genomic accumulation of somatic point mutations, can disrupt the regular activity of cells and may result in cancer initiation and progression. Identifying new cancer subtypes can help classification of patients into groups with similar clinical phenotypes or response to treatment. Machine learning is the study of various algorithms which are capable of inevitably mining of data from observed data. An ensemble [1] is a type of complex technique, which merges a sequence of week classifiers in order to build a strong classifier. Distinct classifier selects and performs the number of iterations until last prediction label returned which performs standard voting. Ensembles perform with more accuracy. Boosting is a type of ensemble algorithms which performs with a set of week classifiers to form into a strong high accurate classifier. Boosting algorithm tracks the technique where actually failed the accuracy. These Boosting algorithms [2] are not much effected by over fitting problem. Adaptive boosting [3] technique is a type of Boosting classifier in ensemble methods which was proposed by Robert Schapire and freund (1996). It merges the multiple numbers of classifiers (week classifiers) to form a strong classifier in order to maximize the classifier's accuracy. It performs number of iterations to maintain the accuracy. It sets the weights to each and every classifier and train the data for each iteration in such a way that it guarantees that the predictions of rare observations are very accurate. In this paper AdaBoost classifier with Feature Selection is proposed to classify the somatic mutational patterns by considering the six types of cancer datasets related to Breast Cancer, Colon Cancer, Pancreatic Cancer, Esophageal Cancer, Uterine Cancer and Kidney Cancer.

2. RELATED WORKS

By using the Shannon Entropy measurement for analyzing C4.5 classification [4] using Information Gain Ratio which helps to construct Decision Tree. The main aim is to classify the dataset by considering various entropies, the accuracy was evaluated efficiently.

The research of C4.5 with particle swarm optimization algorithm [5] is to measure the accuracy based on particle swarm optimization. Particle Swarm Optimization as a Feature Selection, optimizes the accuracy of C4.5 algorithm. The integration of C4.5 and Particle Swarm Optimization algorithms proved that the classification accuracy of breast cancer diagnosis is improved.

B.Padmapriya, T.Velmurugan discussed about The CART algorithm [6] which is chosen to classify the breast cancer data because of its best accuracy for medical data sets. CART and Gini Index are used for attribute selection measure to construct a decision tree. By using the training dataset CART performs the Pruning. These CART classifiers experiments are conducted on breast cancer data for better accuracy and execution time to construct the tree.

In [7], CART algorithm is used as a knowledge–discovery tool because of its interpretability. The CART model provides critical variables threshold and their directional influence on the outcomes. Their results provided a rich basis for hypotheses regarding weight loss prevention in the irradiated HNC patient.

Revised Manuscript Received on December 22, 2018.

Anuradha Chokka, Research Scholar, Dept. of Computer Science, Sri Padmavathi Mahila Visvavidyalayam, Tirupati, India.

Dr.K Sandhya Rani, Professor, Dept. Of Computer Science, Sri Padmavathi Mahila Visvavidyalayam, Tirupati,A.P,India

HNC patients from database Oncospace were identified and among 391 patients identified, Weight Loss predictors during Regression Tree planning were International Classification of Diseases diagnosis;

To obtain the feature subsets to reach the objectives of classification and clustering feature selection methods are proposed. The ideas of feature significance, evaluation criteria, and the behavior of feature selection are discussed [8]. The comparisons of Feature Selection techniques are done. This paper reviewed the classification and various characteristics of feature selection and the benefits and drawbacks of Feature Selection techniques to handle the real world applications are progressed.

The main process of Feature Selection is to eliminate the features which are unrelated to the dataset. Supervised and unsupervised learning techniques can be used to improve the efficiency of different machine learning algorithms [9]. Feature selection speed up the run time of learning, improves data quality and data understanding. This paper provides a comprehensive overview of various characteristic of feature selection.

3. 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 numbers of instances of each cancer type data set which are considered in this paper are shown in Table 1.

Table 1: Instance Count of Cancer Dataset.

SI.NO	Label	Count
1	BRCA	2478
2	COAD	47134
3	ESCA	35524
4	KIRC	6542
5	PAAD	6832
6	UCEC	18142

The various attributes in cancer data set are presented in Table 2.

Table 2: 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 isSomatic {true, false}

4. METHODOLOGY

In this study six types of cancer dataset BRCA, ESCA, KIRC, PAAD, UCEC , COAD are merged to form one cancer dataset which consists of normal data and also somatic mutational data related to six types of cancers. This merged cancer dataset is considered for further analysis. The main aim of proposed model is to classify the somatic mutations patterns with subset of features (attributes) to increase the accuracy of the classification and to decrease the execution time.

As Feature Selection technique improves the performance of a classifier [11], AdaBoost classifier with feature selection using with PCA is proposed in this paper in order to classify the somatic mutational patterns of cancer datasets which are considered to develop classifiers to identify somatic mutations patterns. Principal Components Analysis (PCA) is considered in this study to select the attribute subset combinations [12] and AdaBoost Classifier is adopted in order to classify the somatic mutational patterns. A brief description of Feature Selection-PCA and AdaBoost methods are presented in the following section.

4.1 Feature Selection

In this paper, feature selection is incorporated in the proposed model in order to reduce the dimensionality of the dataset by considering subset of important features and thus improves the efficiency of the classifiers.

If the dataset consists of large number of attributes then it is difficult to apply machine learning techniques to perform classification and clustering operations. In machine learning methods, preprocessing of data is essential in order to improve the performance. Feature selection is one of the preprocessing techniques which is widely used to find significant attributes which improves the performance machine learning techniques in various aspects

It is also known as variable selection, attribute selection, or variable subset selection in machine learning and statistics. It is the process of detecting relevant features and removing irrelevant, redundant, or noisy data. This process speeds up data mining algorithms, improves predictive accuracy, and increases comprehensibility. Irrelevant features are those that provide no useful information, and redundant features provide no more information than the currently selected feature

The two techniques namely Dimensionality reduction and feature subset selection plays vital role in classification and Regression problems.

PCA is widely used technique to find out linear dependencies among attributes of a given dataset. By identifying the strongest patterns in the data, PCA reduces the number of attributes. In other words the attribute space is reduced to subset of attributes which represents more than 95 percent of the original data. PCA generates a set of variables called principal components which are orthogonal to each other which indicates no redundant information. Significant original features for principal component can be identified based on the analysis of eigenvectors. The variance coverage factor plays important role in the identification of important features and based on this parameter best results for classifier can be obtained. Features are ranked and based on the ranks of features; subsets of significant features can be selected to build a predictor or classifier. The PCA with feature selection method which is considered in this paper is presented in the form of algorithm as follows in Table 3.

Table 3: PCA Algorithm.

Algorithm: PCA with Feature Selection
Input: Dataset with attributes Output: Highly Significant attributes •Data Preprocessing • Standardize each attribute based on the formulas of variance and standard devia tion. • Each attribute correlation is calculated by using the covariance formula. • To generate principal components (PC), each attributes Eigen values are comput ed. • Based on the computed Eigen values find principal component by proportion of variance by considering 95% of threshold. • Eigen vectors are calculated by transposing and multiplying matrices. The contri bution of attributes to principal component (PC) is represented by each element in the Eigen vectors. • Identify new significant attributes keeping in view of the correlations between the original attributes and the principal component.

The outcome of this algorithm is highly significant attributes whose contribution is performed to the prediction of class label. Instead of the attributes in the original data these prominent features are considered for the development of classifiers to identify somatic mutational patterns. As the dimensionality of the attributes is reduced compared to original dataset, the performance of the classifier is improved in various aspects.

4.2 AdaBoost Classifier

For each data set, a concatenation of features is mined at various measures from various sub regions. For this composite feature selection method, AdaBoost [8] performs the classification by choosing only those discrete features which can be best distinguish among the classes. We achieve this by designing our weak learner classifier trained on an individual advent feature, and then AdaBoost selects a hypothesis which carries less error. It indicates that every performance of AdaBoost iteration selects the hypothesis [13], along with the individual feature vector that includes the most discriminating data class allowing a correction of classification errors caused from previous steps.

The algorithm representation for AdaBoost classification is shown below.

4.2 AdaBoost Algorithm

Every instance of training data is weighted and initially it is weight(wk)=1/m. where w is the kth instance and 'm' is the total quantity of instances.

The computation of misclassification (error) rate is specified as

err= (D-m)/m. Here D is Correctly predicted instances.

The error rate for weak classifier which gives a weighting for any sort of pre

diction can be calculated as WCW = ln((1-err)/err).

Weights are updated by weight=weight*exp(err* xerr). Here xerr is the error after the weak classifier performed a prediction. xerr can be given as either 0 or 1.

5. EXPERIMENT RESULTS

As the second experiment Feature Selection-PCA is applied on merged cancer dataset in order to find the significant attribute subset from the original attribute subset. In this study, we mainly explored the somatic mutations and their relative information for cancer primary site classification. Patients with mutations are more likely have common features [14]. Feature Selection helps with the problems by reducing the insignificant features without much loss of information. Thus it induces better training to find a more reliable pattern in the model for machine learning experiments and the number of features gene rated for each patient based on the threshold of the modeling process. Now we take the total combination of attributes for the somatic mutational data. The AdaBoost algorithm classified the data with all attributes combination and gave the result with Correctly Classified Instances are 9637 with percentage 91.7984 %. The Incorrectly Classified Instances are 861 with the percentage 8.2016 %. The Kappa statistic shown for the Adaboost classification is 0.836. Kappa (Cohen) [15] statistic is the effectiveness of the classifier's performance. The change in the accuracy and the null error rate is having a high kappa score. The Mean absolute error is 0.1251. Root mean squared error after computation is 0.251. The Relative absolute error is 25.0151 %. Root relative squared error is 50.2055 %. We have taken the total number of Instances are 10498. The AdaBoost classifier's accuracy with all features representation is shown in Table 4.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table 4: Result Shown For Adaboost Classification with All Set Of Features

Total Correctly Classified Instances	9637	91.7984 %
Incorrectly Classified Instances	861	8.2016 %
Kappa statistic	0.836	
Mean absolute error	0.1251	
Root mean squared error	0.251	
Relative absolute error	25.0151 %	
Root relative squared error	50.2055 %	
Total Number of Instances	10498	

AdaBoost Classifier performance is also measured in various aspects such as TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC Area, PRC Area, Class. Recall calculates the percentage of the known somatic mutations patterns that are properly predicted. Precision is the ratio of the properly predicted somatic mutations to the total number of somatic mutations. ROC curve is a graph representation which reviews the classifier's performance, and it takes TP Rate as Y-axis and FP Rate as X-axis. F measure is a measure which is the mean of the recall measurement and the precision. Mean Absolute Error (MAE) [16] is the measure of the error's mediocre magnitude without a particular direction. The accuracy can be measured with MAE for continuous data values. Root Mean Squared Error (RMSE) is measurement which averages the error's magnitude and it is a quadratic scoring rule. RMSE provides more weight to more errors. The MAE and the RMSE can also be used combinly to find out the changes in the errors in a data set. The detailed classification accuracy with each and every class measurement and also with the weighted average is shown in Table 5.

Table 5: Adaboost Classification Accuracy For Every Class Measurement With Wa (Weighted Average)

TPRate	FPRate	Precision	Recall	F-Measure	MCC	ROCArea	PRCArea	Class
0.924	0.088	0.914	0.924	0.919	0.836	0.970	0.966	TRUE
0.912	0.076	0.922	0.912	0.917	0.836	0.970	0.971	FALSE
WA0.918	0.082	0.918	0.918	0.918	0.836	0.970	0.968	

In order to select the significant attributes for dimensionality reduction purpose, Feature Selection method PCA is applied on the original dataset. In this process principal component by proportion of Variance with a threshold 95% are computed based on the Eigen values and finally Eigen vectors are calculated. Finally six attributes are found more significant and the original dataset is reduced with these six attributes. This new dataset is used for further analysis. The experimental results when AdaBoost applied on the new dataset are shown in Table 6.

Table 6: Result Shown For Adaboost Classification With Subset Of Features

AdaBoost Correctly Classified Instances	9771	93.079 %
Incorrectly Classified Instances	727	6.921 %
Kappa statistic	0.896	
Mean absolute error	0.1051	
Root mean squared error	0.201	
Relative absolute error	0.0151 %	
Root relative squared error	40.2055 %	
Total Number of Instances	10498	

The detailed accuracy classification for each and every class measurement can be shown in below table. For the measurements TP-Rate, FP-Rate, Precision, Recall, -Measure, MCC, ROC Area, PRC Area of the data which has given the TRUE class label is shown in one row, and for FALSE class label is shown in another row. The weighted average for the AdaBoost classification of all the considered measurements is shown in below Table 7.

Table 7: Adaboost Classification Accuracy For Every Class Measurement

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.934	0.068	0.934	0.936	0.929	0.876	0.980	0.976	TRUE
0.922	0.066	0.932	0.933	0.937	0.876	0.980	0.971	FALSE
WA0.938	0.082	0.938	0.938	0.938	0.856	0.980	0.988	

CONCLUSION

When the dimensions of the data are complex then it is different to apply machine learning techniques. In order to reduce the dimensionality in terms of number of attributes, in this paper a Classification model with Feature Selection is proposed. The main focus in this paper is to find the feature selection incorporated in classification process which improves the performance of classifier. In this paper Feature Selection-PCA and AdaBoost Algorithms are considered to identify the significant attributes and based on these features Classifier is developed. The experiment results proved that AdaBoost with PCA Feature Selection yielded good results.

REFERENCES

1 LiTaiFang, Pegah Tootoonchi Afshar, Aparna Chhibber, Marghoob Mohiyuddin,"An ensemble approach to accurately detect somatic mutations using SomaticSeq", genomic Biology,16:197,December 2015.

2 R.Senkamalavalli and Dr.T.Bhuvaneswari, "Improved Classification Of Breast Cancer Data Using Hybrid Techniques" , International Journal of Advanced Engineering Research and Science (IJERS)Volume 8, No. 8, September-October 2017.

3 Jaree Thongkam, Guandong Xu, Yanchun Zhang and Fuchun Huang, "Breast Cancer Survivability via AdaBoost Algorithms", in Workshop on Health Data and Knowledge Management Volume 80 pages 55-64, Australia, January 2008..

4 Seema Sharma, Jitendra Agrawal , Sanjeev Sharma, "Classification Through Machine Learning Technique: C4.5 Algorithm based on Various Entropies", International Journal of Computer Applications (IJCA), Volume 82 – No 16, November 2013.

5 M A Muslim, S H Rukmana , E Sugiharti , B Prasetiyo and S Alimah, "Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis", Journal of Physics Conference series, IOP Publishing Ltd, 2018.

6 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.

7 Zhi Cheng, Minoru Nakatsugawa, Chen Hu," Evaluation of classification and regression tree (CART) model in weight loss prediction following head and neck cancer radiation therapy ", Advances in Radiation and Ancology, Elsevier(2018) 3, 346–355.
8 Vipin Kumar and Sonajharia Minz," Feature Selection: A literature Review", Smart Computing,vol. 4, no. 3 , June 2014.
9 Ms. Shweta Srivastava, Ms. Nikita Joshi, Ms. Madhvi Gaur," A Review Paper on Feature Selection Methodologies and Their Applications", Volume 7, Issue 6 (June 2013), PP. 57-61.
10 Quang M. Trinh, Melanie Spears, John D. McPherson, "ISOWN: accurate somatic mutation identification in the absence of normal tissue controls Irina Kalatskaya". Genome Medicine, (2017) 9:59.
11 Jiarui Ding1, Ali Bashashati, Andrew Roth, Arusha Oloumi, Kane Tse, Thomas Zeng, "Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data" , Bioinformatics,Vol. 28 no. 2 2012, pages 167–175.
12 Christos Boutsidis, Michael W. Mahoney,"Unsupervised Feature Selection for Principal Components Analysis", International Conference on Knowledge Discovery and Data Mining, USA, August 24–27, Pages 61-69,2008
13 Schapire RE, Singer Y (1999) "Improved boosting algorithms using confidence rated predictions", Journal Machine learning, 37 issue 3,1999, 297-336.
14 Liu H, Sun J, Liu L, Zhang H (2009) "Feature selection with dynamic mutual information. Pattern Recognition", Pattern Recognition, 42, issue 7,2009,
1330-1339.
15 Anthony J. Viera, MD; Joanne M. Garrett, "Understanding Interobserver Agreement: The Kappa Statistic", NCBI, May:37(5):360-3,2005.
16 Vivek Kumar, Brojo Kishore Mishra, Manuel Mazzara , Dang N. H. Thanh, Abhishek Verma ,"Prediction of Malignant & Benign Breast Cancer: A Data Mining Approach in Healthcare Applications", http://arxiv.org/pdf/1902.03825.