Rotation Forest Ensemble Algorithm for the Classification of Phytochemicals from the Medicinal Plants

Ani.R[1], and O.S.Deepa[2]

[1]Department of Computer Science and applications, Amrita School of Engineering, Amritapuri

[2]Department of Mathematics, Amrita School of Engineering, Coimbatore

Amrita Vishwa Vidyapeetham, Amrita University, India.

***Corresponding author: E-Mail: os_deepa@cb.amrita.edu**

ABSTRACT

Drug Discovery from medicinal plants is an important area in current research and has been providing important source of new drug leads. Plant extracts are proved as main source for many drugs. A major part of traditional therapy uses plant extracts or the associated active principles. Many of the traditional medicines are made as a result of applying some small synthetic modifications of naturally obtained substances. But most of the modern medicines are using synthetic substances instead of natural substances obtained from medicinal plants. Few machine learning predictive algorithms are applied to classify the compounds to the defined classes and the accuracies of different classification algorithms are analyzed. The present study shows the significance of Rotation Forest ensemble algorithms in the classification of medicinal plant compounds. The other algorithms analyzed are Decision Tree, Random Forest and Naive Bayes. The Random Forest tree based ensemble outperformed the other algorithms in this study.

KEY WORDS: Medicinal Plants, Chemical compounds, Drugs, Machine Learning, Classification, Rotation Forest, Accuracy measures.

1. INTRODUCTION

Medicinal plants were only used to cure any types of diseases in ancient days. Now there are many synthetically available compounds which are used as ingredients to the modern drugs. The use of synthetically available compounds and the combination of these compounds in modern medicine may lead to lot of side effects. Even though they cause side effects these synthetic drugs are used for the treatments at least as a last option to cure some diseases. There are many medicines which are available today in market that uses the compounds which are extracted from medicinal plants. This study aims in finding out the medicinal plants compounds which may be used to design a drug. The chemical compounds extracted from medicinal plants exhibits certain drug relevant features. There are 81 attributes selected for this preliminary study from a large set of attributes which may be used to classify the nature of acceptability to consider the natural compound extracted from medicinal plants. There are many plant extracts that could be used to replace some pharmaceutical preparations. Nearly 30 % of the medicines available in the market still uses chemical compounds extracted from medicinal plants. A machine learning approach may be useful to classify the compounds based on the attribute values. A study on the classification of chemical compounds extracted from medicinal plants are used to decide on the acceptability of the compound in drug discovery.

Machine learning algorithms are found to be very useful in drug discovery process. Classification algorithms in machine learning are used for classifying the compounds found in medicinal plants based on the properties of the compounds. It is used to create models which uses a set of instances with predictor variables and class labels. Based on the analysis of attributes in the training set and corresponding class labels, it is used to predict the outcome in unseen data to any of the defined classes. Classification algorithms in machine learning such as Decision Tree, Naïve Bayes and Random Forest are used to analyse the set of attributes of the chemical compound found in medicinal plants. There are two class labels considered in the training data sample. The class labels used in the data set are Medium and High which describes the acceptance level of the compound for drug design. Decision tree is an example of supervised predictive algorithm which uses a set of attributes and class labels and builds a tree like structure based on the splitting criteria used. The information gain and gain ratio and Gini Index are the different important splitting criteria used in decision tress. Random forest is an ensemble classifier that consists of so many decision trees and it predicts the class based on majority voting of class predictions from individual trees. Ensemble algorithms exhibits more accuracy in predictions. Different subsets of data are taken and decision tress are being built from each subset of data. The splitting criteria used in building individual tress is the gain ratio. If there are N numbers of trees, the Random forest predict the class of a new test data based on the predictions from all N tress. Some of the advantages of Random Forest includes less sensitiveness towards outlier data and pruning of trees are not needed. This ensemble algorithm can easily take care of the missing data and can handle data types like continuous, binary and categorical data very easily. It overcomes the problem of over fitting. Random sampling with replacement, ensemble strategies and bagging results in better accuracy in prediction of Random Forest.

The chemical compounds identified for the present study are Curcumin, Cyanidin, Lawsone, Rosmarininc acid, Nimbin, Demethoxy Curcumin, Tumerone, Allicin, Emodin, Ursolic acid, Luteotin, Quercetin, Epigallacatechin gallate, Aloin, Rutin, Gingerol and epigallocatechin. These chemical compounds are extracts taken

from medicinal plants. The Ursolic acid, Rosmarininc acid and Luteotin are taken from *Tulsi*. Cyanidin is from *Hibiscus sabdariffa*. Tumerone, Demethoxy Curcumin and Curcumin are from *Curcuma domestica*. Nimbin, Quercetin from *Azadirachta indica*. Epigallacatechin gallate from *Camellia sinesis*. Rutin is considered from *Calendula officinalis*. Lawsone is taken from *Lawsonia inermis*. Rutin taken from *Calendula officinalis*. Gingerol is taken from *Zingiber*. The epigallocatechin is identified from *Camellia Sinesis*. The Aloin and Emodin compounds are identified from *Aloe vera*. The 81 properties identified for the compounds are like ADMET, drug-likeness, Molecular weight, Molecular surface area, alogP, tPSA, Rotable hydrogen bond, Plat index, Randii index, Balaban index, Wiener index, bioavailablility, Lipinskies Rule of Five, Lead rule, CMC-50 Like Rule, MDDR, Ghose filter, Verbers filter, BBD likeness, UWQED, WQED, Solubility in water, etc.

Juan J. Rodrıguez (2006), proposed a new ensemble classifier based on the concept of feature extraction. The feature transformation technique called Principal Component Analysis (PCA) is applied to the subset of the feature sets before creating the training sample subset and proved the efficiency of Rotation Forest Algorithm in terms of accuracy and diversity. Juan J. Rodrıguez (2006) also compared the suggested algorithm on 33 benchmarked datasets with Bagging, Boosting and Random Forest classifiers and proved a better accuracy for their proposed algorithm. Rico Blaser and Piotr Fryzlewicz (2015), had studied on random rotation ensembles. Hasan Koyuncu and Rahime Ceylan (2013), had also studied on Artificial Neural Network based on Rotation Forest for biomedical pattern classification. Ludmila and Juan (2015), carried out the study to analyse the parameters and randomization heuristics responsible for the better accuracy in Rotation Forest Algorithm. Feature extraction was carried out using the linear transformation method and PCA proved better result when compared to non-parametric discriminant analysis - NDA and random projections. Marcy J. Balunas (2005), conducted studies related to drug discovery from medicinal plants and concluded that natural products isolated from medicinal plants can be predicted as an essential component in the search for new medicines. Salim (2008), made analysis and study of plant derived compounds in the drug discovery process. Akin Ozcif (2011), studied on random forests ensemble classifier which is trained with data resampling strategy to improve cardiac arrhythmia diagnosis.Kun-Hong Liu and De-Shuang Huang (2008), had carried out the Cancer classification using Rotation Forest .Random forest ensemble classifier to predict the coronary heart disease using risk factors was studied by Ani (2015).

2. METHODOLOGY

Random Forest: Random Forest is an ensemble method of decision tree. It is an effective tool for classification. It is also a collection of different simple decision tree. Each of these trees are capable of giving an output when shown with a set of output values. Every tree in the random forest is used to determine the final output. When an input is given to random forest algorithm, it will push to each tree and every tree produce an output. The classification of newly inputted data is done on the basis of majority voting. The number of tree that is to be created can be decided by the individuals.

Decision Tree: Decision Tree is a decision making tool that uses a tree like graph or structure of decision. Decision tree is one of the main supervised learning method for classification and regression. The main idea is to create a structure that predicts the value of a variable by learning simple decision rules from the data features. Decision tree consist of internal node, branches, leaf node. Internal node represents the test conducted on the attributes, each branches shows the output of the test and the leaf node. The algorithm for decision tree is given below.

Training Phase

Given

- X: the objects in the training data set (an $N \times n$ matrix)
- Y: the labels of the training set (an $N \times 1$ matrix)
- L: the number of classifiers in the ensemble
- K: the number of subsets
- $\{\omega_1, \ldots, \omega_c\}$: the set of class labels

For $i = 1 \ldots L$

- Prepare the rotation matrix R_i^a:
  - Split $\mathbf{F}$ (the feature set) into K subsets: $\mathbf{F}_{i,j}$ (for $j = 1 \ldots K$)
  - For $j = 1 \ldots K$
    - Let $X_{i,j}$ be the data set X for the features in $\mathbf{F}_{i,j}$
    - Eliminate from $X_{i,j}$ a random subset of classes
    - Select a bootstrap sample from $X_{i,j}$ of size 75% of the number of objects in $X_{i,j}$. Denote the new set by $X'_{i,j}$
    - Apply PCA on $X'_{i,j}$ to obtain the coefficients in a matrix $C_{i,j}$
  - Arrange the $C_{i,j}$, for $j = 1 \ldots K$ in a rotation matrix R_i as in equation (1)
  - Construct R_i^a by rearranging the the columns of R_i so as to match the order of features in $\mathbf{F}$.
- Build classifier D_i using (XR_i^a, Y) as the training set

Classification Phase

- For a given $\mathbf{x}$, let $d_{i,j}(\mathbf{x}R_i^a)$ be the probability assigned by the classifier D_i to the hypothesis that $\mathbf{x}$ comes from class ω_j. Calculate the confidence for each class, ω_j, by the average combination method:

$$\mu_j(\mathbf{x}) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(\mathbf{x}R_i^a), \quad j = 1, \ldots, c.$$

- Assign $\mathbf{x}$ to the class with the largest confidence.

Naive Bayes: Naive Bayes classification is one of the proved efficient algorithms in classification of data. It depends on the nature of features of the given data set. It predicts the probability of a given sample belonging to particular classes. Naive Bayesian classifiers assume that the effect of an attribute value on a given class which is independent of the values of the other attributes. This classification algorithms shows better result in predictive and diagnostic problems. This is ideal for problems with more feature dimensions. The method of maximum likelihood is used to do the parameter estimation in this model.

Rotation Forest: The Rotation Forest ensemble classifier algorithm proved as more accurate classifier compared to bagging, AdaBoost and Random Forest ensembles across a collection of benchmark data sets. Classifier generation is based on feature extraction and feature projection. In this paper the feature sets are randomly divided into subsets and a feature transformation method, PCA is applied to each subset. A feature rotation is applied to each subset and a new classifier is generated based on the subset. The present algorithm gave a better accuracy and diversity within the ensemble. Decision tress are very sensitive to rotation of the feature axes. So the tree based classifiers gave better a performance in the present algorithm. All principal components are retained in the proposed algorithm to maintain the variability of information stored in the data.

3. RESULTS

Table.1. Accuracy of Different algorithms on Different Dataset

	Random Forest %	Rotation Forest %	J48 %	Naive Bayes %
Medicinal compounds$_{FR1}$ (Dataset 1)	76.47	82.35	70.5	82.35
Medicinal compounds$_{FR1}$ (Dataset 2)	82.35	82.35	64.70	76.47
Medicinal compounds$_{FR1}$ (Dataset 3)	82.35	82.35	82.35	88.23
Medicinal compounds$_{FR2}$ (Dataset 4)	58.82	47.05	47.04	76.47
Medicinal compounds$_{FR2}$ (Dataset 5)	82.35	76.47	76.47	64.70
Medicinal compounds$_{FR2}$ (Dataset 6)	88.23	88.23	94.11	82.35

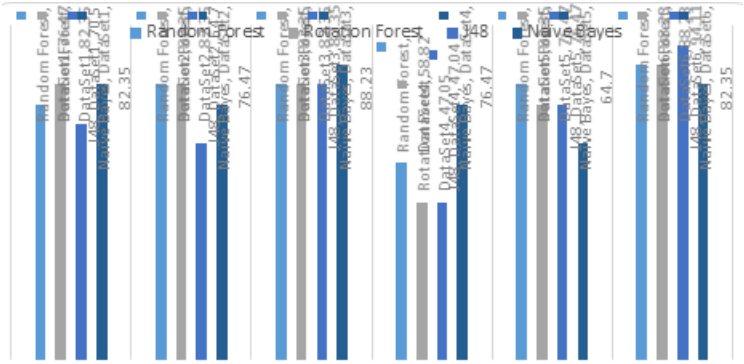

Figure.1. Accuracy Chart for different Algorithms

Table.2. Different Performance measures analyzed on each data set Precision, Recall, F-Measure and ROC Area

Dataset	Class	Precision	Recall	F-Measure	ROC Area
Medicinal compounds$_{FR1}$ (Dataset 1)	Medium	0.714	0.833	0.769	0.821
	High	0.9	0.818	0.857	0.67
Medicinal compounds$_{FR1}$ (Dataset 2)	Medium	0.833	1	0.909	0.956
	High	1	0.714	0.833	0.905
Medicinal compounds$_{FR1}$ (Dataset 3)	Medium	0.813	1	0.897	0.987
	High	1	0.25	0.4	0.917
Medicinal compounds$_{FR2}$ (Dataset 4)	Medium	0	0	0	0.442
	High	0.667	0.615	0.64	0.442
Medicinal compounds$_{FR2}$ (Dataset 5)	Medium	0.75	0.9	0.818	0.771
	High	0.8	0.571	0.667	0.771
Medicinal compounds$_{FR2}$ (Dataset 6)	Medium	0.846	1	0.917	0.848
	High	1	0.667	0.8	0.848

Two different ranges were considered:

Medicinal compounds$_{FR1}$: Data Set 1, 2 and 3 are pertaining to a fixed range in set – I of 81 descriptors for 17 compounds from medicinal plants.

Medicinal compounds$_{FR2}$: Data Set 4, 5 and 6 are pertaining to a fixed range in set – II of 81 descriptors for 17 compounds from medicinal plants.

Based on the fixed range the entire data sets are normalized.

Highly acceptable level: Compounds from the medicinal plants are classified as highly acceptable level if the compounds satisfies the fixed ranges for more than 72 descriptors, 69 descriptors and 66 descriptors.

Medium acceptable level: Compounds from the medicinal plants are classified as medium acceptable level if the compounds satisfies the fixed ranges for less than 72 descriptors, 69 descriptors and 66 descriptors.

2. CONCLUSION

There are many approaches for Drug discovery process from medicinal plants. Machine learning based prediction of class labels which shows the level of of acceptability of the compounds for the preparation of drugs which are carried out in this study. Four widely used Machine Learning Algorithms are used for the analysis. The two classes shows the level of acceptance of the compound to form a drug. Rotation Forest ensemble algorithm's performace is analysed in this study. Transformation of attributes is being done by projecting the attributes using principal component analysis before classifying using the base classifier. Figure 1 gives the accuracy Chart for the different Algorithms. Table 1 gives the accuracy of different algorithms on different Dataset and Table 2 gives the different performance measures analyzed on each data set along with Precision, Recall, F-Measure and ROC Area. It is found that data set 3 - Medicinal compounds$_{FR1}$ gave a better performance compared to other dataset for all the four different algorithms. The linear transformation method used in the Rotation Forest may be modified in future study to improve the accuracy.

REFERENCES

Akin Ozcift, Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis, Computers in Biology and Medicine, Computers in Biology and Medicine, 41 (5), 2011, 265-271.

Ani R, Aneesh Augustine, N.C. Akhil, O.S.Deepa, Random forest ensemble classifier to predict the coronary heart disease using risk factors,proceeding of the International conference on soft computing systems, Springer, Vol 397, 2015.

Hasan Koyuncu, Rahime Ceylan, Artificial Neural Network based on Rotation Forest for biomedical pattern classification, IEEE conference, 2013.

Juan J. Rodrıguez, Ludmila I. Kuncheva, Carlos J. Alonso, Rotation forest: A new ensemble classifier method, IEEE transaction on pattern analysis and machine intelligence, 28 (10), 2006.

Kun-Hong Liu, De-Shuang Huang, Cancer classification using Rotation Forest, Computers in Biology and Medicine, 38 (5), 2008, 601-610.

Ludmila I. Kuncheva, and Juan J. Rodrigue, An Experimental Study on Rotation Forest Ensembles, Springer, International Workshop on Multiple Classifier Systems, 2015, 459-468.

Marcy J. Balunas, Douglas Kinghorn A, Drug discovery from medicinal plants, Elsevier, Life Science, 78, 2005, 431-441.

Rico Blaser, Piotr Fryzlewicz, Random Rotation Ensembles, Journal of Machine Learning Research, 2, 2015, 1-15.

Salim A.A, Chin Y.W, and A.D, Drug Discovery from Plants, In:Bioactive Molecules and Medicinal Plants, Spinger (book), 2008.