Application of Feature Selection Techniques on Bioinformatics Data

[1]Divisha Khaturia, [2]Charu Khamesra, [3]Mayank Patel

[1]Associate Software Developer Innoplexus, Pune
[2]Associaate Professor, Department of Chemistry, GITS, Udaipur
[3]Associaate Professor, Department of Computer Science and Engineering, GITS, Udaipur

Abstract : Bioinformatics so far have become an integral part of the research and various other applications in big data. Determining the clinical implications through varied experiments in biomedical research is increasing rapidly. Selection of certain attributes on which the research will rely is yet another crucial task, as the outlook of the algorithm depicts how the information extracted is useful in the given field. Feature Selection so far has proved to be efficacious in this meadow of research, as it selects a subset of features for avoiding the issue of 'curse of dimensionality'. Multi-attribute feature selection is chosen which involves the selection of relevant attributes from multi-attributed datasets, results of which improve accuracy. Our aim is to imitate the survival probability of the diabetic patients and identify the associated risk factors along with it. The analysis depicts the factors which contribute to a shorter survival span of diabetes mellitus patients. The factors include glucose levels, BMI, age, insulin, blood pressure, skin thickness and pregnancies.

***IndexTerms* - Bioinformatics, Survival Probability, Diabetes Mellitus Patients, Feature Selection.**

1. INTRODUCTION

A dataset may contain thousands of attributes; the task is to find the most relevant subset of attributes for the target component. Feature selection is a technique that can be used in such situations. [1]Bioinformatics dataset usually contains a lot of attributes, from this to extract the most relevant features becomes difficult as almost all the features are essential for some problems. Dimensionality of the feature space can be reduced in order to deal with such problems. [4]By selecting or extracting a subset of features from a collection of the whole set of features and ranking these features and thereby choosing the most relevant features reduces the dimensionality. Feature ranking gives a ranking value for the most important attributes. Feature selection trains the machine learning algorithms much faster and reduces the model complexity thereby the interpretation becomes much easier. [2]Over fitting can be reduced and if the correct subset is chosen, it improves the accuracy too.

The field of research has made a wide progress in technology and around.[4] Applying data mining algorithms on medical data is yet another concern as there is plenty of information which is stored in facts and figures.[3] Whether it comes to recommending a particular medicine or treatment to the patients based on its attributes or features or any other fact that is related to bioinformatics data, converting it to facts or a language that the machine understands is no less than a hectic task to do.[2] There are many recommendations or decisions that are dependent on the responses one gets from the agonized patients using that one particular application. Thus in order to carry on a further study and provide a conclusion that is more relatable and easy to understand, a comparative study is made on the most commonly used methods and evaluating the as outcomes of the comparisons is presented the paper.

Feature selection algorithms have been applied so far in various grounds of research, each single algorithm individually. The comparisons would thus enable one to choose a better method related to their arena in which the research is to be carried on. Novelty is achieved by applying survival analysis as it helps one to predict the overall survival probability of the patients wretched with the specified illness. Accuracy measures suggest the method appropriate for the study and an approximation of survival therefore acts an advantage over the conveyed dissertation.

2. LITERATURE SURVEY

GuojunGan and Michael Kwok-Po Ng [11] proposed a KMOR algorithm for outlier detection where they are not able to control the number of outliers. The study also lacks the value of k. SumaiyaThaseenIkram a, Aswani Kumar Cherukur [12] designed a SVM model for network attacks for minimizing the test and training time thereby improving the accuracy of classification. Melanie Kershaw,Ruth Krone'[14] reviewed the survival of children with diabetes which in turn helps the healthcare experts.The review also proposed a clinic. Monica Parrya,KyleDanielsona, Sarah Brennenstuhl a, Ian R. Drennanb,c, Laurie J. Morrisond [15] have done a study to compare the relation between status of diabetes and other outcomes. It also predicts the survival. Khan, A (2004) [1] gives an idea how rough set theory can be used for analyzing

diabetes data. Nahla H. Barakat, Andrew P. Bradley, and Mohamed Nabil H. Barakat (2010) [3] proposed a SVM model for analyzing the diabetes patient data.
Tarek Helmy and Zeehasham Rasheed [4] have used a Bioinformatics dataset for classification using machine learning. Xiang Zhang, Huaixiang Zhang and Ertao Li [5] proposed an optimization technique for fuzzy systems using genetic approaches. Taoying Li, Yan Chen, Xiangwei Mu and Ming Yang [6] proposed a clustering algorithm for diabetes dataset to accure more accuracy. Ping-Hung Tang and Ming-Hseng Tseng [7] compares different techniques to find which one provides the best accuracy and performance. JayalakshmiT. and Dr. Santhakumaran A. [8] proposed ANN approach for classification of diabetes data. Ian H. Witten and Eibe Frank [10] reviewed about the machine learning techniques that can be used for various datasets.

3. BACKGROUND WORK

3.1 Feature Selection

It can either be of filter or wrapper methods. Pre-processing step is carried out in the filter methods where the essential features are selected based upon the rank/score attained. The technique is used for four reasons:
1. Simplifying the models to make an easy interpretation for researchers or users.
2. Make the training times shorter.
3. To avoid the curse of dimensionality.
4. Enhancement of generalization by overfitting reduction.

3.2 Chi-Squared

It is a method used in statistics for testing the independence of two events. In a given dataset, we get the observed (O) and expected count (E), numbers of which are derived from each other. Chi- Squared test was carried out here in to obtain the score values. [6]Chi- square test is a statistical method used for finding the goodness of fit among the observed and expected. Based on the null hypothesis that the two events are independent, we can calculate the expected value E_A using the following formula-

$$\frac{E_A}{A+C} = \frac{A+B}{N}$$

So,

$$E_A = (A+C)\frac{A+B}{N}$$

3.3 Random Forest

Wrapper methods are used to build a model using the subset features which can be further used for adding or dropping the features. Random forest is a collection of decision trees which is then presented with some modifications independently. It is applied on the entire set of attributes initially. Random forest is an ensemble learning method that can be used for tasks such as regression classification etc. It also gives the accuracy of the most relevant features.
Bootstrap aggregation or bagging to tree learners techniques are applied as training algorithms for random forests.
Given a training set A = a_1, ..., a_n along with the responses B = b_1, ..., b_n, bagging continuously (for P times) when selected a random sample and the training set is replaced and trees are fit into these samples:
For p = 1,..., P:
After replacements of the sample, B training examples from A, B; A_p, and $B_{p\ are}$ called.
Train a decision or regression tree f_p on A_p, B_p.
After training, unseen samples predictions of x' can be made by taking average of the predictions from all the regression trees on x' individually:

$$f = \frac{1}{P}\sum_{p-1}^{P} f_p(x')$$

When majority vote is taken in the case of decision trees.

3.4 One-R algorithm

One R is another algorithm used in this study that comes with a principle of one rule per predictor and finally finds the rule with minimum total error as the one rule. [7]According to the algorithm the data is split into training and test sets, where we to calculate a classified accuracy for each rule and feature.

4. PROPOSED WORK

4.1 Survival analysis

Survival analysis is applied so as to find the survival period for the selected features. It outputs a time value for the occurrence of an event in general. [9]For the bioinformatics dataset it outcomes the period until each test result attribute exists in a relevant state. In addition to it, K- means clustering is also applied to the dataset in order to group the data into various clusters depending on their target contents.

The survival function is the primary objective and denoted as S, is defined as

$$S(t) = P_r(T > t)$$

Where-

- t refers to some time.
- T stands for a random variable, denotes the time of death
- Pr stands for probability

The methodology used for carrying out this research work is as follows-

4.1.1 Data Collectio : The data used for this work is gathered from www.kaggle.com which covers the attributes like age glucose levels and BMI values (which are more focused). The procedures adopted at this stage of the research are: Data Cleaning, Data Selection, Data Transformation and Data Mining.

4.1.2 Data Cleaning

Here, a consistent format for the data model is made which takes care of missing data, finds the presence of duplicated data if any, and weeding out of noisy data. At last, the cleaned data is transformed to a format that is suitable for data mining.

4.1.3 Data Selection

At this stage, data relevant to the analysis was decided on and retrieved from the dataset. The bioinformatic dataset had four main attributes, their type and description is presented in Table 1.

4.1.4 Data Transformation

Also called data consolidation, here the selected data is transformed into forms that are suitable for data mining. The data file is saved in Commas Separated Value (CSV) file format and to lessen the effect of scaling on the data, the datasets are then normalized.

4.1.5 Data Mining Stage

The data mining stage is further subdivided into three phases where all the algorithms are used to analyze the bioinformatics dataset separately on each phase.

5. Figures and Tables

5.1 Table

ATTRIBUTE	TYPE	DESCRIPTION
Age	Numerical	age considered
Glucose	Numerical	Glucose considered
BMI	Numerical	BMI values
Outcome	Categorical	Presence and absence of diabetes is shown

Table 1 : Simulation Parameter Consideration

5.2 Graphical Representation

The chi-square method depicts the maximum contribution of an attribute in predicting the diabetes.

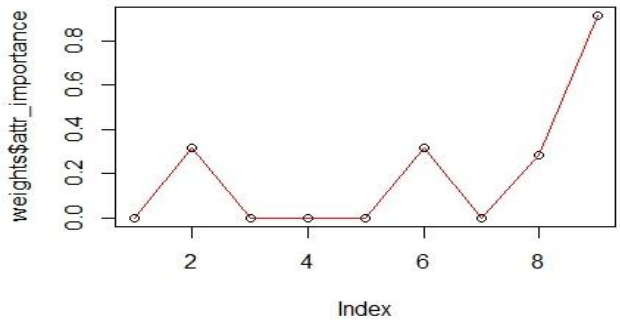


Fig 1 : Follows up the One-R algorithm.

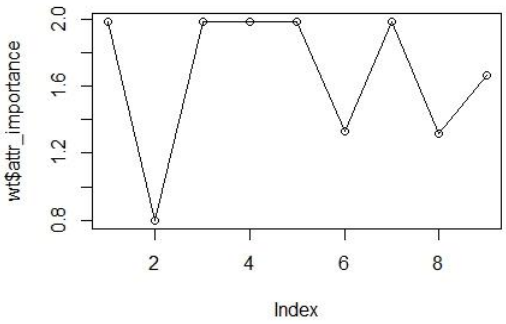
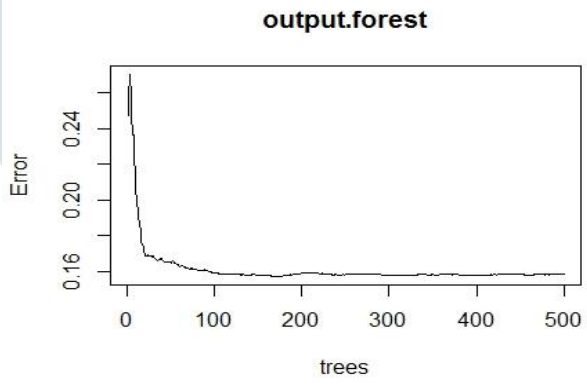


Fig 2 : And then the random forest algorithm.



The clustering graphs help one to relate two features and thus depict how the value of BMI and age varies in accordance with glucose. The clusters show how the BMI and age values vary in accordance with the glucose levels in the patients. Survival analysis when done gives the following output graph as-

Fig 3 : The values are divided into different clusters is shown with the following graphs.

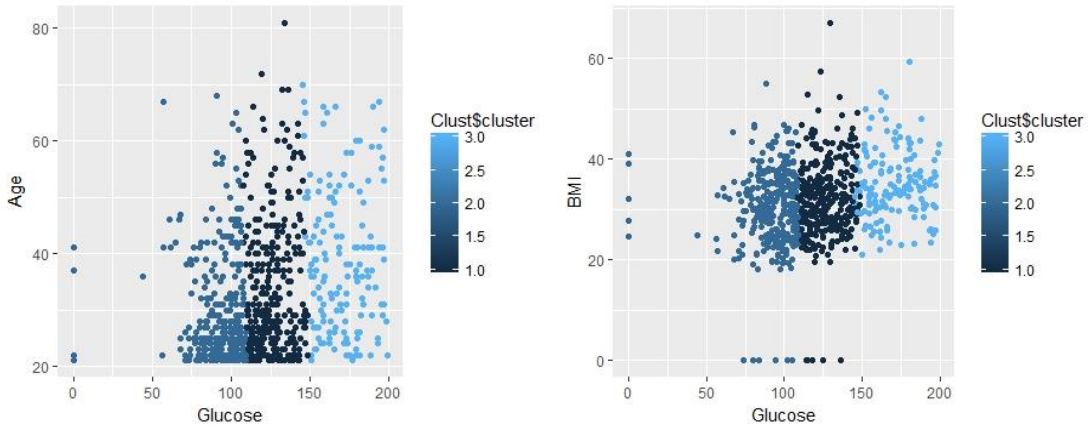


Fig 4: Analysis of Glucose based on Age and BMI

The graph shows the range and predicted values of the diabetes mellitus patients in the form of clusters. The crests represent the higher survival probability i.e. those who are less likely to be prone to diabetes while the lows represent patients who have less chances of survival. The intermediate values thus show the average probability for survival.

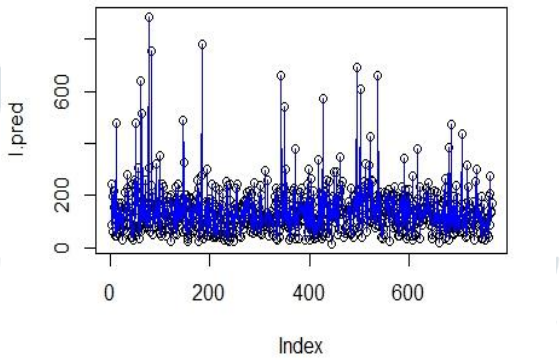


Fig 5: Diabetes mellitus patients in the form of clusters

6. RESULT

The outcomes of the applied feature selection and ranking algorithm are displayed as-

```
> print(weights)
                         attr_importance
Pregnancies                    0.0000000
Glucose                        0.3168560
BloodPressure                  0.0000000
SkinThickness                  0.0000000
Insulin                        0.0000000
BMI                            0.3185487
DiabetesPedigreeFunction       0.0000000
Age                            0.2855096
Outcome                        0.9129370
```

Fig 6. Chi-square method

```
> print(wt)
                         attr_importance
Pregnancies                    1.9835459
Glucose                        0.7994792
BloodPressure                  1.9835459
SkinThickness                  1.9835459
Insulin                        1.9835459
BMI                            1.3330696
DiabetesPedigreeFunction       1.9835459
Age                            1.3182266
Outcome                        1.6656519
>
```

Fig 7. One-R method

```
> print(importance(output.forest))
                              %IncMSE IncNodePurity
set$Pregnancies             13.663574     12.67744
set$Glucose                 50.466072     41.24933
set$BloodPressure            2.688425     12.95451
set$SkinThickness            4.677022     10.71162
set$Insulin                  8.010712     11.59945
set$BMI                     24.501649     26.45239
set$DiabetesPedigreeFunction 9.293722     18.81072
set$Age                     20.886826     21.49700
```

Fig 8. Random Forest

The survival analysis done using K-means clustering in turn makes it easy to identify the areas where diabetes can be a major concern to an individual, depending upon which how can one regulate the use of medicines and maintain a proper health and hygiene in this regard. [11]The accuracy level of K-means clustering is evaluated as 68% making it quite reliable to put into use.

7. CONCLUSIONS AND FUTURE DIRECTIONS

The conclusions drawn out of this comparative study is how various feature selection and ranking methods can be applied to the field of biomedical sciences so far to help make a better study of health of an individual or a group on the whole. Better results can be drawn using various data mining techniques and reliable maintainence, updation and retrieval of any information is achieved.

In future research works various other data mining techniques and models shall be used for the prediction of bioinformatics information. [12]This work is important to study bioinformatics as the variation in medical conditions in terms of different health related concerns can be studied with the help of these data mining techniques.

REFERENCES

[1]. Khan, A (2004) 'Data Mining the PIMA Dataset Using Rough Set Theory with a Special Emphasis on Rule Reduction', Proceedings of INMIC, Luton, UK, 2004.

[2]. Balakrishnan, S., Narayanaswamy, R., Savarimuthu, N., Samikannu, R.,(2008) 'SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases', IEEE international Conference on Systems, Man and Cybernetics.

[3]. Nahla H. Barakat, Andrew P. Bradley, and Mohamed Nabil H. Barakat (2010) 'Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus', IEEE Transactions on Information Technology in Biomedicine, Volume 14, Issue 4

[4]. Tarek Helmy and Zeehasham Rasheed (2009), 'Multi-Category Bioinformatics Dataset Classification using Extreme Learning Machine' Proceeding CEC'09 Proceedings of the Eleventh conference on Congress on Evolutionary Computation

[5]. Xiang Zhang, Huaixiang Zhang and Ertao Li (2010), 'Optimization of Fuzzy Classification System by Genetic Strategies', Sixth International Conference on Natural Computation, pp 2424-2428.

[6]. Taoying Li, Yan Chen, Xiangwei Mu and Ming Yang, (2010) 'An Improved Fuzzy K-Means Clustering with K-Center Initialization', Third International Workshop on Advanced Computational Intelligence IWACI 2010.

[7]. Ping-Hung Tang and Ming-Hseng Tseng,(2009) 'Medical Data Mining Using BGA AND RGA for Weighting of Features in Fuzzy K-NN Classification',

International Conference on Machine Learning and Cybernetics, pp 3070-3075

[8]. Jayalakshmi T. and Dr Santhakumaran A. (2010), 'A Novel Classification Method for Diagnosis of Diabetes Mellitus Using

Artificial Neural Networks', International Conference on Data Storage and Data Engineering (DSDE) Bangalore, India, pp 159-163

[9].TopTenfactsaboutDiabeteshttp://www.medindia.net/health_statistics/health_facts/diabetesfacts.htm, on 06/01/2010

[10]. Ian H. Witten and Eibe Frank, 'Data Mining: Practical Machine Learning Tools and Techniques, Second Edition

[11] Guojun Gan and Michael Kwok-Po Ng k -means clustering with outlier removal

[12] Sumaiya Thaseen Ikram a, *, Aswani Kumar Cherukur Intrusion detection model using fusion of chi-square feature selection and multi class SVM

[13] Monica Parrya,Kyle Danielsona, Sarah Brennenstuhl a, Ian R. Drennanb,c, Laurie J. Morrisond 'The association between diabetes status and survival following an out-of-hospital cardiac arrest: A retrospective cohort study '

[14] Melanie Kershaw ,Ruth Krone' A survival guide to the children's diabetes clinic'.