IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

A Review of Feature Selection Algorithms to Identify Risk Factors for Liver Disease

Suri Yaddanapudi[1], Madhanan Balaram[2]

Student, Computer Science, R.V.R &J.C, Guntur, India [1]

Student, Computer Science, VIT University, Vellore, India [2]

Abstract: Huge volumes of datasets with relatively higher number of dimensions are being collected by medical practitioners to identify the relevant features that cause a disease, which gives rise to an important technique, called feature selection, as the pre-processing strategy in obtaining knowledge and information from datasets. Feature selection is important when machine learning algorithms are applied on medical datasets to make the model easy to understand. Feature section techniques in medical domain should be model independent and at the same time should come with less number of features. Filter feature selection is independent of any model and helps in solving the curse of dimensionality. In this paper different types of filter feature selection algorithms are applied to A.P Liver dataset and performance is evaluated using sensitivity and specificity analysis.

Keywords: Feature Selection, Liver Diagnosis, Data Mining, A.P. Liver Dataset, Wrapper, Filter.

I. INTRODUCTION

Benefiting from the progress in information technology, data is being collected in huge volumes in every field. A health care organization is one such field where large amounts of data are collected to distinguish the patients from the people with illness to the people who aren't. Large volumes of data is good for building a model that easily distinguishes the patients with and without illness, but the data also comes with number of features which are irrelevant and redundant thereby not only increasing the computational complexity but also decrease the model performance.

Even though the feature selection algorithms can be applied on both supervised and unsupervised learning, this review is focused on application of different feature selection techniques on supervised learning problem. Hence, A.P liver dataset obtained from UCI Machine Learning Repository [1] is used. As opposed to other dimensionality reduction that are focused on compression or projection, these feature selection techniques doesn't alter the originality of dataset, but selects a subset of attributes.

This document is a template. An electronic copy can be downloaded from the conference website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website.

II. FEATURE SELECTION TECHNIQUES

A. Feature Selection

In the problem of classification (supervised learning), given a set of data points S, where every point associates with a number of attributes and a target variable, the learning procedure is a two-fold summary: (1) formulating a probability distribution function *F* over all the data points and (2) finding the response function *R* from data instances to the response variable [2]. To formulate the response function *R* as close to real function as possible, theoretically, we should make use of as more features as possible to generate a model that easily distinguishes all the classes from one another. But, in real-world problems, using more features doesn't necessarily helps in capturing the response function which is explained by *curse of dimensionality*.

Occam's Razor [3] bias helps us to generate a model which is as simple as possible that avoids overfitting on the datasets, thus we need only relevant features to simply the learning procedure and to infer the response function more accurately. By only considering the features that are relevant, learning models will reduce the number of rules that results in requirement of small data samples to generate a good response function. To generate the true probability distribution function that infers response function more accurately, we need to eliminate irrelevant and redundant features. In addition, the computational time and the space complexity can be reduced which helps in identifying a model that is as simple as possible.

B. Different types of feature selection techniques

Large varieties of feature selection techniques have been proposed and majority of the feature selection techniques fall under the three categories *filter, wrapper, embedded* [4]. Wrapper feature selection techniques are wrapped by a model and select the features based on the performance of the model. Filter feature selection techniques selects the features based on a certain criteria as a result modelling is excluded from the subset feature selection and only the relationship between the target variable and attributes are considered. Embedded feature selection techniques incorporate the feature selection technique as a part of modelling (Decision trees).

There are a large number of feature selection techniques that are available and few of them are discussed. In this paper we mainly focus on *filter* feature selection algorithms as they are independent of any modelling algorithm.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Chi-squared feature selection algorithm

Chi-squared feature section algorithms select the subset of features by conducting the chi-squared test on discrete attributes.

Entropy based feature selection algorithms:

Entropy based feature selection algorithm selects the subset of features based on the correlation with continuous variables and can be subdivided into

Information gain

Information in bits is measured by information gain about class prediction and this can be used in selecting the subset of features. Information gain is given by formula

$$H\,(Class) + H\,(Feature) - H\,(Class, Feature)$$

Gain ratio feature selection technique

Subset of features that are selected by this algorithm used gain ratio as a criteria to select the features. Gain ratio is given by formula

$$\frac{H\,(Class) + H\,(Feature) - H\,(Class, Feature)}{H\,(Feature)}$$

Symmetrical uncertainty feature selection

Bias of the mutual information is balanced by symmetrical uncertainty and gives a measurement for feature correlation that can be used to select the subset of features [5].

OneR feature selection algorithm

OneR selects the subset of features based on simple association rules involving one feature in condition part.

Random forest feature selection algorithm

Random Forest feature selects the subset of features based on random forest algorithm.

RReliefF feature selection algorithm

RReliefR samples the data points and finds their nearest misses and hits which can be used in selecting feature subset.

III. DATA AND METHODS USED

A.P Liver dataset is obtained from UCI machine learning repository. Liver dataset is used for research because the problems with liver diseases are not easily discovered in early stages [6], hence to facilitate medical practitioners to identify the relevant features in early stages to distinguish the patients with or without liver diseases. Andhra Pradesh liver dataset contains information of 416 liver patients and 167 non liver patients.

Table 1 gives the description of the dataset used.

Once the relevant features are selected, a classification algorithm is needed to compare the sensitivity and specificity analysis. KNN classification for designing a supervised model that easily distinguishes the liver patients from the non-liver patients. A KNN classification technique is also used for wrapper feature selection technique as wrapper feature selection techniques need a model to select the feature subset. Gender in the dataset is a categorical variable consisting either male or female.

To facilitate for KNN classification model to find the nearest neighbors, male and female are replaced with '1' and '0'.

Duda et al.,[7] suggested that *K* should be square root of number of features in a dataset to obtain optimal results. Leave one out cross validation is used in analysis on classification modelling.

TABLE I: A.P LIVER DATASET

Attribute	**Data Type**
Age	Numeric
Gender	Factor
Total bilirubin	Numeric
Direct bilirubin	Numeric
Alkphos alkaline phosphotase	Numeric
Sgpt alamine aminotransferase	Numeric
Sgot Aspartate Aminotransferase	Numeric
Total proteins	Numeric
Albumin	Numeric
Albumin and Globulin Ratio	Numeric
Class	Factor

IV. RESULTS AND DISCUSSION

Performance of different feature selection techniques and the weightage factors obtained using those feature selection techniques are presented in the below tables.

A. Chi-squared feature selection algorithm

Age, gender, total proteins and albumin are irrelevant features as per *chi-squared* feature selection algorithm. Table 2 shows the attributes and the weightage factor associated with the feature.

TABLE II: FEATURE SUBSET USING CHI-SQUARED FEATURE SELECTION

Chi-squared feature selection algorithm	
Attribute	**Attribute importance**
Total bilirubin	0.32292
Direct bilirubin	0.3085428
Alkphos alkaline phosphotase	0.2936586
Sgot Aspartate Aminotransferase	0.2894212
Sgpt alamine aminotransferase	0.2639557
Albumin and Globulin Ratio	0.1934897
Age	0
Gender	0
Total proteins	0
Albumin	0

B. Information gain feature selection algorithm

Age, gender, total proteins and albumin are irrelevant features as per *information gain* feature selection algorithm. Table 3 shows the attributes and the weightage factor associated with the feature.

TABLE III: FEATURE SUBSET USING INFORMATION GAIN FEATURE SELECTION

Information gain feature selection algorithm	
Attribute	**Attribute importance**
Direct bilirubin	0.05975293
Total bilirubin	0.0585834
Sgot Aspartate Aminotransferase	0.04658529
Alkphos alkaline phosphatase	0.0446

DOI 10.17148/IJARCCE.2015.41147

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Sgpt alamine aminotransferase	0.0415907
Albumin and Globulin Ratio	0.01954776
Age	0
Gender	0
Total Proteins	0
Albumin	0

C. *Gain ratio feature selection algorithm*

Age, gender, total proteins and albumin are irrelevant features as per *gain ratio* feature selection algorithm. Table 4 shows the attributes and the weightage factor associated with the feature.

TABLE IV: FEATURE SUBSET USING GAIN RATIO FEATURE SELECTION

Gain ratio feature selection algorithm	
Attribute	**Attribute importance**
Direct bilirubin	0.10512142
Total bilirubin	0.0890751
Sgpt alamine aminotransferase	0.074402327
Sgot Aspartate Aminotransferase	0.0717936
Alkphos alkaline phosphatase	0.06446198
Albumin and Globulin Ratio	0.02919063
Age	0
Gender	0
Total Proteins	0
Albumin	0

D. *Symmetrical uncertainty feature selection*

Age, gender, total proteins and albumin are irrelevant features as per *Symmetrical uncertainty* feature selection algorithm. Table 5 shows the attributes and the weightage factor associated with the feature.

TABLE V: FEATURE SUBSET USING SYMMETRICAL UNCERTAINTY *FEATURE* SELECTION

Symmetrical uncertainty feature selection algorithm	
Attribute	**Attribute importance**
Direct bilirubin	0.10249144
Total bilirubin	0.09332985
Sgot Aspartate Aminotransferase	0.07474767
Sgpt alamine aminotransferase	0.07192013
Alkphos alkaline phosphatase	0.06921877
Albumin and Globulin Ratio	0.03085071
Age	0
Gender	0
Total Proteins	0
Albumin	0

E. *OneR feature selection*

Age, gender, total proteins and albumin turned out to be relevant features as per *oneR* feature selection algorithm. Table 6 shows the attributes and the weightage factor associated with the feature

TABLE VI: FEATURE SUBSET USING ONER FEATURE SELECTION

OneR feature selection algorithm	
Attribute	**Attribute importance**
Age	0.5630776
Gender	0.5630776
Total Proteins	0.5630776
Albumin	0.5630776
Direct bilirubin	0.4341651
Sgpt alamine aminotransferase	0.4179176
Sgot Aspartate Aminotransferase	0.3588752
Total bilirubin	0.3579801
Albuminumin and Globulin Ratio	0.3167266
Alkphos alkaline phosphatase	0.2849741

F. *Random forest feature selection*

Table 7 shows the attributes and their importance after *random forest* feature selection technique is applied.

TABLE VII: FEATURE SUBSET USING RANDOM FOREST FEATURE SELECTION

Random forest feature selection algorithm	
Attribute	**Attribute importance**
Direct bilirubin	22.626576
Total bilirubin	19.331616
Sgot Aspartate Aminotransferase	18.579674
Sgpt alamine aminotransferase	17.966112
Age	12.789043
Gender	9.011435
Alkphos alkaline phosphatase	8.664225
Albumin	7.885841
Albuminumin and Globulin Ratio	6.333309
Total Proteins	4.29676

G. *Relief feature selection*

Table 8 shows the attributes and their importance after *relief* feature selection technique is applied.

TABLE VIII: FEATURE SUBSET USING RELIEF FEATURE SELECTION

Relief feature selection algorithm	
Attribute	**Attribute importance**
Direct bilirubin	0.023941799
Total bilirubin	0.012784232
Albuminumin and Globulin Ratio	0.007659259
Alkphos alkaline phosphatase	0.007604625
Age	0.00667528
Albumin	0.005797101
Total Proteins	0.005367687
Sgot Aspartate Aminotransferase	0.002492979
Sgpt alamine aminotransferase	0.002097525
Gender	0

Sensitivity and 1–specificity analysis are done on the A.P liver data set once with including all the features from the dataset and can be seen in Fig 1.

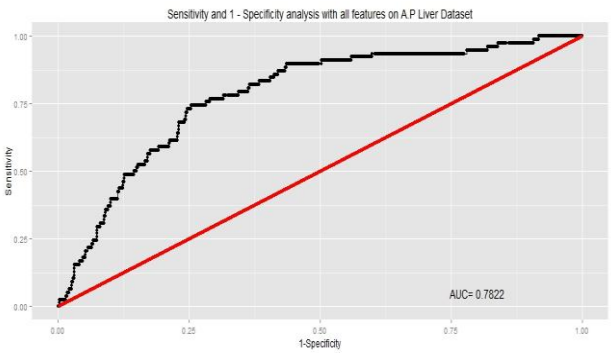

Fig. 1 Sensitivity and 1-specificity analysis with all features

DOI 10.17148/IJARCCE.2015.41147

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Fig 2 shows the sensitivity and 1–specificity analysis when 6 most dominants features direct bilirubin, total bilirubin, albuminumin and globulin ratio, alkphos alkaline phosphatase, sgot aspartate aminotransferas and sgpt alamine aminotransferase are used. All the feature selection analysis shows gender, age, albumin and total proteins are irrelevant variables and can be ignored. Furthermore sensitivity and 1-specificity analysis shown in Fig 2 shows that performance of the model can be improved by ignoring the irrelevant features in the dataset.

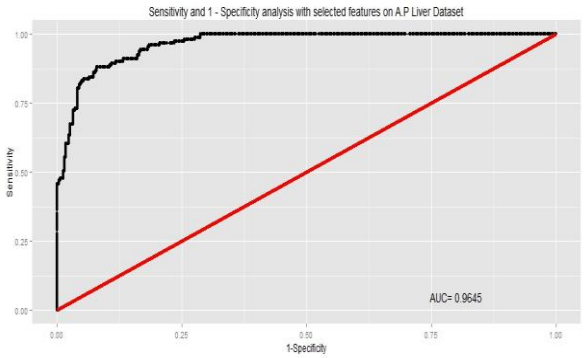

Fig. 2 Sensitivity and 1-specificity analysis with selected features

V. CONCLUSION

In this study, we implemented popular feature selection algorithms on A.P Liver Dataset and measured their performance based on sensitivity and specificity analysis. We also showed that feature selection algorithms can boost the model discriminating power by ignoring the redundant and irrelevant features. Selected dataset considers age, gender, total proteins and albumin as irrelevant features.

REFERENCES

[1] Bendi,Venkata Ramana (2012). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Blum AL, Langley P (1997) "Selection of relevant features and examples in machine learning," Artificial Intelligence, 97(1/2): 245-271.

[3] Bulmer Anselm, Ehrenfeucht Andrzej, Haussler David, Warmuth Manfred K (1987) "Occam's Razor," Information Processing Letter, 24(6): 377-380.

[4] Y. Saeys, I. Inza, and P. Larraaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507-2517, 2007.

[5] Wu, Yimin, and Aidong Zhang. "Feature selection for classifying high-dimensional numerical data." Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. Vol. 2. IEEE, 2004.

[6] Ramana, Bendi Venkata, M. Surendra Prasad Babu, and N. B. Venkateswarlu. "A critical study of selected classification algorithms for liver disease diagnosis." International Journal of Database Management Systems 3.2 (2011): 101-114.

[7] Duda, Richard O., Peter E. Hart, and David G. Stork. Pattern classification. John Wiley & Sons, 2012.

[8] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.