International Journal of Research in Engineering, Science and Management
Volume-2, Issue-3, March-2019
www.ijresm.com | ISSN (Online): 2581-5792

Effective Feature Selection Strategy for Healthcare Analysis using SVM and MLP

S. Ilangovan[1], V. Vidhiya Bharathy[2], M. Saranya[3]
[1]*Professor, Department of Information Technology, KLNCE, Madurai, India*
[2,3]*Student, Department of Information Technology, KLNCE, Madurai, India*

***Abstract*: In data mining, feature selection is the most important technique which deals with the all the applications which consists of large number of variables. Pre-processing is done on the dataset using lambda function. After that feature selection is done in two ways to measure its influence in the dataset. Recursive Feature Elimination is the first model which eliminates all the weak attributes. It constructs a model from such selected attributes. Best features algorithm termed as Extra tree classifier used a second technique to rank the attributes based on the feature score. The attributes with high score are selected for further processing. The features selected by both the techniques are taken as separate datasets. For calculating the efficiency of such datasets are calculated by SVM and MLP algorithms. The special feature of the project is that we use Spyder tool in Python. Finally, the results are compared among various machine learning algorithms.**

***Keywords*: Accuracy Generation, Classifier Algorithm, Feature Selection, Pre-processing**

1. Introduction

The Wide use of computers in all the fields results in accumulation of huge amount of data. Such kinds of data are of large size and it is tedious process to acquire the best analytic performance so that finding the pattern from the data for prediction is not accurate. Big Data mining is a widely developing field. Hence it is important to reduce all its loop holes. Dimensionality is the challenging one in the field of data mining. There are many redundant data and unwanted data will be there in the dataset. They may reduce the efficiency of process. For example, a dataset with thousands of features may not be much effective and it increase the computation time of the system. The only effective way to reduce and dimension and increase in efficiency is to apply feature selection strategy. Some of the benefits of features selection such as Reducing the number of features will also reduce the computation expense, minimises the noise to improve classification accuracy, and the generation ability of the dataset will increase.The selection strategy is done by two algorithms. Recursive Feature Elimination which reduces the weakest features recursively. Best Feature algorithm (Extra tree classifier) reduces the features by giving feature score for every individual attributes. The selected features form a new model dataset. The further computation of accuracy is calculated with the help of Support Vector Machine (SVM) and MultiLayer Perceptron (MLP) algorithm. Comparison among the feature selected dataset and non-selected dataset is done. Spyder is used here as a computation tool as the WEKA is far less flexible than the python for statistical analysis and

data exploration. In practice python and their respective packages and libraries are much more flexible and practical for data science.

2. Literature review

With the help of survey made in the area of data mining and machine learning the importance of various models to derive function are observed. In literature Yang and Zhang proposed a hybrid algorithm, called GAFF (Genetic Algorithm with embedded filter), made the feature selection process as two stages. In the first stage they used a Genetic Algorithm to select the features and after that they had implemented filter wrapper which helps to find the small feature subset makes the algorithm more accurate. The Hybrid GA/SVM approach uses fuzzy logic to reduce the dimension of the dataset by eliminating the redundant attributes. Then a decision tree rule induction technique, and fuzzy inference based on Mamdani's method. The results of feature selected dataset is compared with the machine learning algorithms like SVM, MLP. It gives accurate result with multiple machine learning algorithms. In literature Ilangovan and Vincent, the proposed method is based on Relief and Entropy GA algorithm. The filtration process uses Relief ranking for feature selection. It defines the ranking for each feature so that we can select only the features with high rank. And both of the literatures use Weka as tool for analysis.

3. Proposed work

In the process of Knowledge Discovery, data handling is considered to be an important factor for finding the data patterns for efficient analysis. The proposed feature selection is based on two algorithms such as Recursive Feature Elimination and Best Feature Algorithm (Extra Tree Classifier). The RFE eliminates the feature by recursively removing the weak features. In the Extra Tree Classifier, the attributes are ranked with the probability value. The selected features with highest probability are taken for the mining process. The efficiency of the selected dataset is calculated with the help of Support

Vector Machine (SVM) and Multilayer Perceptron (MLP) algorithms are used. The comparison of algorithms is done to verify whether the selected attributes produce an efficient result. Spyder is proposed to use here as a tool for computing the result.

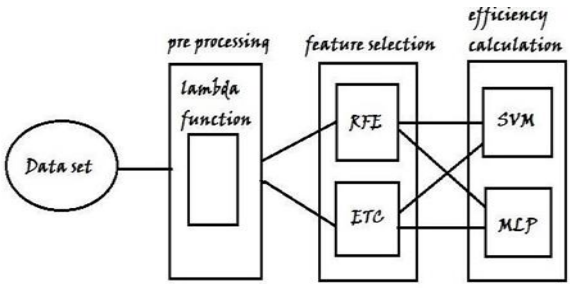

Fig. 1. Architecture diagram

A. *Lambda function*

Lambda functions creates s small, one time and anonymous function objects in python. They may have more number of arguments but at the same time they have only one expression. Such expressions are evaluated and returned. Wherever the function objects are needed, Lambda functions can be used. In our project we use lambda function for pre-processing the dataset. As its computation is based on function abstraction and application with variable binding and substitution. It is considered to be universal mode of computation. In pre-processing it eliminates all the redundant data and fill the null values for missing data. Hence the data may become errorless.

B. *Recursive feature elimination*

RFE works by recursively eliminating the attributes. It builds a model on attributes which are remaining and calculates the model accuracy. It contributes the most to predicting the target attribute. For every function, it removes the weak attributes. At the stage where I cannot eliminate any attribute it starts building a model with the escaped features.

C. *Extra tree classifier*

Extra-trees differ from classic decision trees in the way they are built. When looking for the best split to separate the samples of a node into two groups, random splits are drawn for each of the max_features randomly selected features and the best split among those is chosen. When max_features are set 1, this amounts to building a totally random decision tree.

D. *SVM*

"Support Vector Machine" (SVM) is a supervised machine learning algorithm. It is used for challenges in both regression and classification. Anyway it is widely used in classification problems. In SVM, data items are plotted as a point in n-dimensional space (where n is number of features you have) where the value of each feature is as same as the value of particular coordinate. Then the classification is done by identifying the hyper-plane which differentiates the two classes.

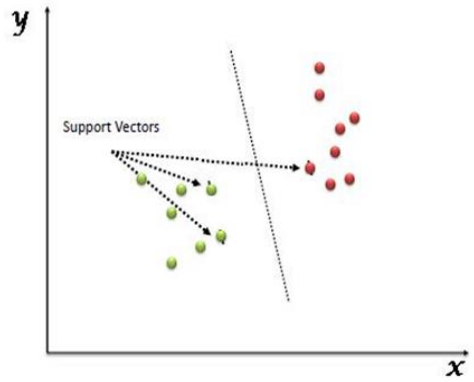

Fig. 2. General structure of SVM

Support Vectors are the co-ordinators of individual observation. SVM helps in segregating the two classes. (hyper-plane/line). For example, the inner product of the vectors [2, 3] and [5, 6] is 2*5 + 3*6 or 28. The equation for making a prediction for a new input using the dot product between the input (x) and each support vector (xi) is calculated as follows:

f(x) = B0 + sum(ai * (x,xi))

This is an equation that involves calculating the inner products of a new input vector (x) with all support vectors in training data. The coefficients B0 and ai (for each input) must be estimated from the training data by the learning algorithm.

E. *MLP*

A Multilayer Perceptron (MLP) is a class of feedforward artificial neural network. MLP contains three layers of nodes such as an input layer, a hidden layer and an output layer. In Input nodes, each node is termed as neuron which uses a function for nonlinear activation. MLP uses a Supervised learning technique called as backpropagation. Its multiple layers and such non-linear activation are distinguishing MLP from linear perceptron. It is able to distinguish data which are not linearly separable. The predictive capability of neural networks comes from the hierarchical or multi-layered structure of the networks.

MLP is sometimes referred as "vanilla" neural networks, especially when they have a single hidden layer.

1) *Activation*

The weighted inputs are summed and passed through an activation function, called a transfer function. An activation function is just a simple mapping of summed weighted input to the output of the neuron called as activation function because it governs the threshold at which the neuron is activated and strength of the output signal. In simple step activation functions were used where if the summed input was over a threshold, for example 0.5, then the neuron would output a value of 1.0, otherwise it would output a 0.0. Traditionally non-linear activation functions are used. This allows the network to combine the inputs in more complex ways and in turn provides a richer capability in the functions they can model. Non-linear functions like the logistics called the sigmoid function were

used that output a value between 0 and 1 with an s-shaped distribution, and the hyperbolic tangent function also called tanh that outputs has distribution over the range -1 to +1. More recently the rectifier activation function has shown to provide better results.

4. Application

The dependent variable has two categories hence it is treated as two classes. Every single instance is mapped as the set of positive and negative class labels. They are separated based upon the mapping instance from estimated class.

	True Class	
Estimated Class	True Positives (TP)	False Positive (FP)
	False Negatives (FN)	True Negatives (TN)

Fig. 3. Confusion matrix

As per the true class observation, True positives has the proportion of subjects called “True Positives (TP)”, which are predicted correctly as cases. The estimation class observation stated the proportion of subjects who are “False Positives (FP)”, which are predicted falsely as cases. Therefore, such precision is used for Statistical measurements. Sensitivity (True positive rate) and Specificity (True Negative rate) are also calculated.

Precision = TP / TP+FP

F-Measure= 2/((1/Precision)+(1/Recall))

Accuracy (%) = (TP+TN)/ (TP+TN+FP+FN)

Specificity (%) = FP/ (FP+TN)

Sensitivity (%) = TP/ (TP+FN)

5. Result

The comparison between various feature selection algorithms are done and their accuracy is verified with the help of SVM and MLP algorithms.

Recursive Feature Elimination algorithm selects 29 attributes from the set of 56 total attributes. The accuracy is found to be same before and after feature selection. Best Feature algorithm selects 32 features from 56 features. With SVM the Best Feature Algorithm gives accuracy of 85% before feature selection and 100% after feature selection. With MLP it gives 93% accuracy before feature selection and 100% accuracy after feature selection. The comparison tables are shown in the table.

Table 1
RFE feature selection and accuracy comparison

Recursive Feature Elimination				
Algorithm	**Actual Features**	**Accuracy**	**Selected Features**	**Accuracy**
SVM	56	85%	29	85%
MLP	56	93%	29	93%

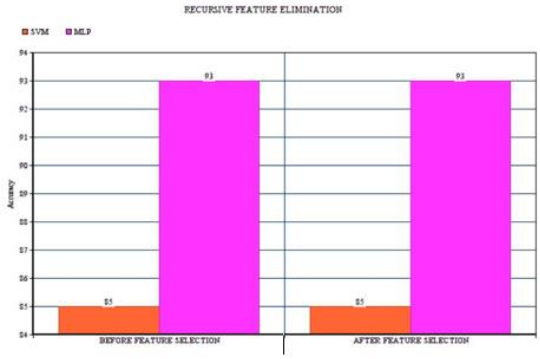

Fig. 4. comparison chart of RFE

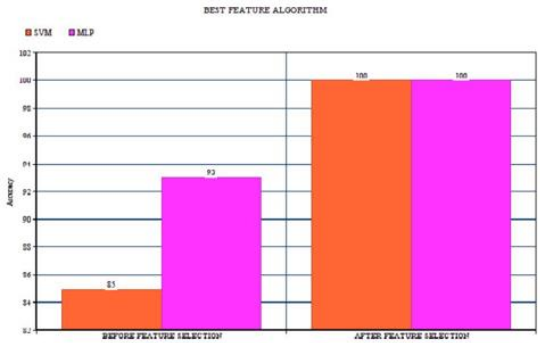

Fig. 5. comparison chart of BFA

Table 2
Best Feature Algorithm Feature Selection and accuracy comparison

Best Feature Algorithm				
Algorithm	**Actual Features**	**Accuracy**	**Selected Features**	**Accuracy**
SVM	56	85%	29	100%
MLP	56	93%	29	100%

6. Conclusion

The proposed research verifies the feature selection algorithms such as Recursive Feature Elimination and Best Feature Algorithm (Extra Tree Classifier) by using the lung cancer dataset. By using the spyder tool features selected are verified for its accuracy with both SVM and MLP. From the comparison table, it proves the analysis results gives more efficiency with such highly influential features. The results have shown that the algorithm provides high accuracy with minimum number of attributes and at the same time it reduces the dimension of the dataset which can construct the model within a seconds. The field of feature selection will have a wide scope in the near future for prediction and analysis fields of various applications.

References

[1] Ilangovan Sangaiah & Vincent Antony Kumar,“Improving medical diagnosis performance using hybrid feature selection via relief and entropy based genetic search (Rf-GEA) approach: application to breast

cancer prediction", Springer Science + Business Media, LLC, part of Springer Nature, 2018.
[2] Yang, Pengyi and Zhang, Zili 2009-12, An embedded two-layer feature selection approach for microarray data analysis, IEEE intelligent informatics bulletin, vol. 10, no. 1, pp. 24-32.
[3] M. W. Aslam, Z. Zhu, A.K. Nandi, "Feature generation using genetic programming with comparative partner selection for diabetes classification", Experts System with Applications, Vol 40, pp. 5402-5412, 2013.
[4] Han J, Kamber M, DataMining:Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, 2000.
[5] Choubey D.K, Sanchita P, "GA_MLP NN: a hybrid intelligent system for diabetes disease diagnosis", IJISA, 8(1), 49, 2016.
[6] Hualong, B, Jing, X, "Hybrid feature selection mechanism based high dimensional date sets reduction", Energy Proc., 11(1), 4973– 4978, 2011.
[7] Yilmaz N., Inan O., Uzer M.S., "A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases," J Med Syst, vol.38, no.5, 2014.
[8] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), pp. 273–297.
[9] Unler, A., Murat, A., Chinnam, R.B.: mr 2 PSO: a maximum relevance minimum redundancy approach based on swarm intelligence for support vector machine classification. Inf. Sci., 181(20), 4625– 4641, 2011.
[10] Kim, J.K., Lee, J.S., Park, D.K., Lim, Y.S., Lee, Y.H., Jung, E.Y.: Adaptive mining prediction model for content recommendation to coronary heart disease patients. Clust. Comput, 17(3), 881–891, 2014.