

# Review Paper On Various Feature Subset Selection Methods for Classification of Large Datasets

Supriya Suresh. 1<sup>a</sup>, Shwetha S M . 2<sup>b</sup>, Rajashree Bayalal . B. 3<sup>c</sup>

<sup>a</sup> Asst Prof, Department of CSE, APSCE., Bengaluru, Karnataka, India

<sup>b</sup> Asst Prof, Department of CSE, APSCE., Bengaluru, Karnataka, India

<sup>c</sup> Asst Prof, Department of CSE, APSCE., Bengaluru, Karnataka, India

## Abstract

Data mining is a technology that blends traditional data analysis with sophisticated algorithms for processing large amounts of data. Pre-processing in data mining is an important step, where feature subset selection is one of the method. In feature subset selection, we select only those subset of features, from our entire data set that optimally describes the target concept, eliminating the irrelevant features. Logistic regression tells that there can be one or more independent variables that can determine the outcome for a problem and neural network resembles the animal's central nervous system, but here information is processed at simple elements called neurons and signals are passed between these neurons over connection links. By studying various papers on feature subset selection and comparing the results from these papers with the previous feature subset selection methods, it is revealed that when logistic regression and neural network model is applied onto the feature selection methods such as forward selection and backward elimination using cross validation and percentage split as test options, neural network model with backward elimination, using percentage split gave more efficient result. To get more reduced subset of features, cuckoo search method can also be used on the selected datasets, once the optimization is done with help of cuckoo search, again forward selection and backward elimination method can be applied on the optimized set of features and again comparison can be established as whether cuckoo with forward selection or cuckoo with backward elimination, whichever out of the two yields a better result can be chosen and this has been proposed as the future work.

**Keywords:** *Cross validation, Data mining, Feature subset selection, Logistic regression, Neural network model, Percentage split.*

## 1. Introduction

Data mining is a technology where we are implementing sophisticated algorithms, extraction of data by some searching methods. For example in railway reservation, we are finding a relevant reservation from different geo-graphical areas.

Data pre-processing is a broad area and consists of a number of different strategies and techniques that are inter-related in complex ways [8].

First approach in data pre-processing is aggregation where we can combine two or more objects into one. Next is, sampling, where we are selecting subsets of data objects and they are analysed, and then they are used in preliminary analysis and final analysis. Next in, data reduction, different data sets are taken and combined into a new one thus reducing the processing time and different algorithms can be used here. Next approach is the curse of dimensionality where this tells that high dimension of data is difficult to analyse and searching of irrelevant data can be avoided. Next approach is the Feature subset selection where, new subsets are created by combining all the subsets. Hence this reduces the redundant features and irrelevant features. Next approach is feature creation where all the relevant information within all my subsets are put into a new subset. Next approach is discretization and binerisation where discretization transforms continuous attribute into a categorical one and binerization transforms discrete and continuous to binary attributes [8]. Final approach is variable transformation where the attribute values are transformed either into simple or normalized form. Here feature subset selection is mainly concentrated upon and its different approaches are seen and the best approach is selected.

Feature subset selection or variable selection is intended to select the "best" subset of predictors. Here redundant predictors should be removed. The principle of Occam's Razor states that among several plausible explanations for a phenomenon, the simplest is the best. Applied to regression analysis, this implies that the smallest model that fits the data is best. Unnecessary predictors will add noise to the estimation of other quantities that we are interested in. Feature selection is the process of identifying and removing as much of the irrelevant and redundant information as possible.

Typically feature subset selection was presented in 3 different ways, filter approach, wrapper approach and the embedded approach. In filter approach, the features are selected before the data mining algorithm is run, they are selected using some approach which is independent of the data mining task, here the less interesting features can be removed, but the drawback in this approach is, it tends to select redundant features as the relationship between the features are not established. In wrapper method, as the name suggests wraps up the data mining algorithm along with the features as a black box, to find the best subset of features, but again when the numbers of features are huge, the time for calculation of best subset of features also increases. In embedded approach, the data mining algorithm itself decides which features are best to use and which to ignore but here again the algorithm must have a pre-knowledge of what actually a good selection is.

each link is associated with a weight and each

Due to the limitations of the above principal approaches to feature subset selection and also to overcome the above limitations, from the study of various papers it is found that forward selection and backward elimination methods overcome the above limitations and forward selection tells that if entropy values are considered, for all the attributes along with a class attribute, first arrange the given attributes in ascending order according to its entropy values then classify the first attribute with its class attribute as one relation, then with this first relation, group the second attribute that had occurred in the order as the second relation and so on until all attributes are grouped.

In Backward elimination, first arrange all the attributes according to their entropy values in descending order, after arranging all the attributes in an order classify all of them with the class attribute as one set and from here eliminate from backwards that is start removing from the beginning, of all the attributes until we get one single attribute grouped with its class attribute.

From the study it is found that the analysis of the class variables is done by applying logistic regression and neural network model onto the feature selection methods such as the forward selection and backward elimination and cross validation and percentage split is used as the test option. Logistic regression tells that there can be one or more independent variables that can determine the outcome for a problem and neural network resembles the animal's central nervous system but here information is processed at simple elements called neurons and signals are passed between these neurons over connection links where

neuron applies an activation function on its net input to determine the output signal [5]. Cross validation and percentage split are standardized tools in analytics and is an important feature for helping us in developing and fine tuning data mining models.

From the study it is found that the process of data pre-processing, regression and neural network analysis, all are carried out using a data mining tool called Waikato Environment of Knowledge Analysis (Weka). The tool comprises of collection of data mining algorithms, which can be applied onto our chosen datasets.

## 2. Literature survey

An evaluation of predictive model for classification using neural networks in credit scoring is done using publically available banking data sets and classification accuracy is taken to measure the performance of neural networks model. Forward selection and backward elimination algorithm are applied on these data sets to evaluate feature selection algorithm [1].

Logistic Regression (LR) can be used in the fields like cancer diagnosis, survival prediction and many more. LR always helps to compare between a categorical outcome and a set of explanatory variables. Neural Network (NN) model can also be used in the fields like biology, business, auditing etc. Here a comparison study between LR and NN with and without hidden layers is done on publically available medical datasets and both the performances are compared, and after analysis, it is found that NN without hidden layers gives better results [2].

Logistic Regression helps us to investigate the relation between a categorical outcome and a set of explanatory variables. Artificial Neural Network are famously used as universal non-linear inference models and is a branch of Artificial Intelligence, that represents an animal nervous system. The comparison between Logistic regression and Neural network is done, on publically available medical datasets. The logistic regression and neural network with sensitivity analysis, is evaluated for the effectiveness of the classification. The classification accuracy is taken to measure the performance of both the models. But with experimental results, neural network model with sensitivity analysis model gives more efficient result [3].

Logistic Regression helps to evaluate or examine the relationship between a categorical outcome and a set of explanatory variables. The outcome of this can be either dichotomous (yes/no) or ordinal (low, medium, high). For a dichotomous outcome, the standard logistic regression is used. So here a standard logistic regression formula with feature selection using forward selection and backward elimination is applied on publically available medical datasets, the selected features from these algorithms are used to develop a predictive model for classification using logistic regression. The classification accuracy, root mean square error, mean absolute error are used to measure the performance of the predictive model [4].

The evaluation of the performance of logistic regression and neural network model with feature selection method and sensitivity analysis on publicly available medical datasets is taken. The sensitivity analysis is the one in which the output changes with changes in the input. The classification accuracy is used to measure the performance of both logistic regression and neural network models. From the experimental results it is observed that with the use of artificial neural network and backward elimination technique with sensitivity analysis yields better results [5].

Feature selection involves the selection of most useful set of features that produces compatible results as the original set of entire feature. A fast clustering based feature selection algorithm (FAST) is proposed and experimentally evaluated. FAST algorithm has 2 stages where in first stage, features are divided into clusters by using graph theoretic clustering method. In second stage, most relevant features, related to the target class are selected from each cluster to form a subset of features. To make sure there is a good efficiency of this algorithm, minimum spanning tree clustering method is used. FAST is compared with other feature selection algorithms such as FCBF, ReliefF etc, with respect to classifiers, namely the probability based Naïve-Bayes, rule based RIPPER, instance based IB1 and tree based C4.5. On 35 publically available data sets like text data, microarrays and high dimensional image, FAST relatively produces smaller subsets and also improves the performance of above four types of classifiers [6].

The granting of loans by a financial institution (bank or home loan business) is one of the important decision problems that require delicate care. This can be performed using a variety of different processing algorithms and tools. Neural networks are considered one of the most promising

approaches. Here the main goal is to find the best

tool out of the three neural network approaches like multi-layer perceptron, ensemble averaging and boosting by filtering and out of these three the multi-layer perceptron and boosting by filtering yields best results [7].

### 3. Cuckoo Search

Apart from using the feature subset selection methods like forward selection and backward elimination on the selected datasets, a new optimization algorithm called cuckoo search is also studied and by using this algorithm the reduced set of features can be obtained and this can be a further research in the field of feature subset selection. The idea of cuckoo search is as follows:

- a) Each cuckoo lays one egg at a time and dumps into a randomly chosen nest.
- b) Best eggs in best nests are carried over to next generation and eggs in worst nests are dumped for farther calculations.
- c) Number of hosts/nests is fixed and number of eggs to be laid is decided by the host bird [9] [10].

Pseudo code for the cuckoo search algorithm is as follows:

**Begin**

Objective function  $(f(x), x) = (x_1, \dots, x_d)$

Generate initial population of  $n$  host nests  $x_i (i=1, 2, \dots, n)$

**while** ( $t < \text{maxGeneration}$ ) or (stop criterion)

Get a cuckoo randomly by Levy flights and evaluate its quality/fitness  $F_i$

Choose a nest among  $n$  (say,  $j$ ) randomly

**if** ( $F_i > F_j$ ),

Replace  $j$  by a new solution;

**End**

A fraction ( $p_a$ ) of worse nests are abandoned and new ones are built;

Keep the best solutions (or nests with quality solutions);

Rank the solutions and find the current best

**End while**

Postprocess results and visualization

**End**

For the datasets taken in Table 1, forward selection and backward elimination can be applied on them by considering the idea of mean values, that is, if the attribute values in our datasets are of type real, for each and every attribute the mean values are calculated and arrange the attributes based on the mean values in ascending/descending order and then the forward selection and backward elimination method can be applied on these datasets.

As for the cuckoo search is considered this concept is considered as a new research work. When this concept is applied on the given dataset,

first it checks the values present in the dataset. If the values are of type real, then it needs to be

converted to 0's and 1's. To do this, first we need to calculate mean values of all the attributes and then fix the threshold for each attribute. The values of each attributes are compared with the threshold and changed to 0's and 1's by using the following equation:

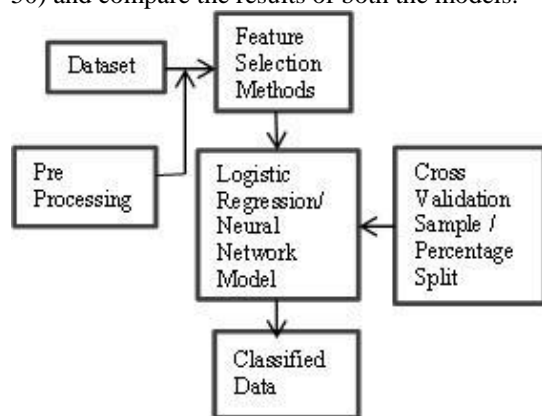
$$f(x_j) = \begin{cases} 1, & \text{if } \text{val}_i(x_j) \geq T_j \quad i < 1 < r, j < 1 < n \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where n is the number of attributes and r is the number of instances and  $x_j$  is the  $j^{\text{th}}$  attribute.

After the evaluation of the above equation (1), the value of attributes from the given dataset contains only 0's and 1's. In the next step, we need to calculate the total number of 1's in each attribute and then based on the total value arrange the attributes in ascending/descending order. The forward selection and backward elimination method is applied on these attributes to find the best set of attributes that produces the better classification accuracy.

#### 4. Design

The framework for classification using logistic regression and neural network model with feature selection methods using cross validation sample and percentage split is as shown in Figure 1. The specifications of publically available datasets are as shown in table 1. From the given dataset best subset of features are selected using forward selection and backward elimination technique. The comparative study can be made by taking the probabilities from both logistic regression and neural network model. Consider the the test option as 10-fold cross validation and percentage split (50-50) and compare the results of both the models.



**Figure 1:** Framework for classification.

**Table 1:** Specification of datasets.

Datasets	Number of Records	Number of Attributes	Class Variables
Stat log (Australian credit approval)	14,980	15	A15(0,1)
Fertility	100	10	Diagnosis Normal (N,O)
Single Proton Emission Computed Tomography (SPECT)	267	23	Overall Diagnosis of the patient (1 for normal and 0 for abnormal)
Thoracic Surgery data	470	17	Risk (T-survival, F-If died)

#### 5. Conclusion

In this review a detailed study is made and the comparison is done between feature subset selection approaches like forward selection and backward elimination. From the study, it is observed that forward selection and backward elimination methods select better subset of features from the given datasets. It is also observed that backward elimination with neural network model using percentage split gives better classification accuracy.

Apart from the above forward selection and backward elimination method, cuckoo search method can also be applied on the selected data sets, where cuckoo search is an optimization algorithm, hence when it is applied onto the attributes of the given data sets, reduced subset of features may be obtained and which in turn may increase the classification accuracy.

#### References

- [1]. Raghavendra B.K, Anil Kumar C.J and Raghavendra S, "Evaluation of Predictive Model for Classification Using Neural Networks in Credits Scoring", International Journal of Advanced Networking and Applications, December 2010 Vol 2, Issue 3, pp 714-718.
- [2]. Raghavendra B.K, S.K. Srivatsa, Raghavendra S, and Shivashankar S.K, "Comparison of Logistic Regression and Neural Network model with and without Hidden Layers", Universal Journal of Applied Computer Science and Technology, 2011 Vol 1, pp 49-53.
- [3]. Raghavendra B.K and S. K. Srivatsa

- “Evaluation of Logistic Regression and Neural Network Model with Sensitivity Analysis on Medical Datasets”, *International Journal of Computer Science and Security (IJCSS)*, 2011, Vol 5, Issue 5, pp 503-511.
- [4]. Raghavendra B.K and Jay B. Simha, “Evaluation of Logistic Regression Model with Feature Selection Methods on Medical Dataset”, *International Journal in Computational Intelligence*, December 2010, Vol 1, Issue 2, pp 35-42.
- [5]. Raghavendra B.K and Jay B. Simha, “Performance Evaluation of Logistic Regression and Neural Network Model with Feature Selection Methods and Sensitivity Analysis on Medical Data Mining”, *International Journal of Advanced Engineering Technology*, January-March 2011, Vol 2, Issue 1, pp 288-298.
- [6]. Qinbao Song, Jingjie Ni and Guangtao Wang , “A Fast Clustering Based Feature Subset Selection Algorithm for High Dimensional Data”, *IEEE Transactions on Knowledge and data engineering* 2013, Vol 25, Issue 1, pp 1-14.
- [7]. Meliha Handzic, Felix Tjandrawibawa and Julia Yeo, “How Neural Networks Can Help Loan Officers to Make Better Informed Application Decisions”, *Information Science And Information Technology Education Joint Conference*, The University of New South Wales, Sydney, Australia, June 2003.
- [8]. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, “Introduction to Data Mining”, pp 44-65, Copyright © 2006. This edition of the book is published by arrangement with Pearson Education, Inc. and Dorling Kindersley Publishing, Inc.
- [9]. Xin-She Yang, Suash Deb “Cuckoo search: Recent Advances and Applications”, *Conference on Neural Computing and applications*, 2014, Vol 24, Issue 1, pp 169-174.
- [10]. Sanket Kamat, Asha Gowda Karegowda, “ A Brief Survey on Cuckoo Search Applications”, *International Conference on Advances in Computer and Communication Engineering*, May 2014, Vol 2, Special Issue 2, pp 7-14.