[ICRTCET-2018]
ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

FEATURE SELECTION TECHNIQUES IN MACHINE LEARNING : A SURVEY

Pragya Chauhan*[1], Nirmala Sharma[2] & Harish Sharma[3]
*[1,2&3]Rajasthan Technical University, Kota, India

ABSTRACT

Feature selection is the most important and chal-lenging problem in the machine learning for selecting a subset of relevant features namely variables for the construction of models. Feature selection is used to reduce input to a feasible size for analysis. It mainly focuses on to recognize irrelevant data without influence the accuracy. It improves the accuracy of the machine learning tasks. The objective of feature selection is to improve the prediction performance, provide fast prediction and cost-effective prediction. Also, it provides a superior understanding of the underlying process to generate the data. Feature selection select relevant and redundant feature subset. This article provides a detailed survey of feature selection techniques such as Filter Method, Wrapper Method, Embedded Method and Hybrid Method and its application.

Keywords: *Feature Selection, Filter Method, Wrapper Method, Embedded Method, Machine Learning*

I. INTRODUCTION

The number of high dimensional data which is publically available on the internet has rapidly expanded in the past cou-ple of decades. In this way, machine learning techniques have difficulty in dealing with the large number of input features, which is posing an interesting challenge for researchers. In order to utilize machine learning techniques effectively, so pre-processing of the data is very essential. Feature selection technique is the most significant pre-processing step of ma-chine learning (ML) [4]. It reduces the dimension as well as eliminate inappropriate or existence of noisy, irrelevant and redundant data which becomes unfavorable so, for increasing the learning accuracy and it improves the result. Feature selection method is very important due to the same training dataset it executes better tasks with various features subset [4]. Performance of Machine learning model depends on feature selection. Major factors of feature selection are knowledge discovery, Interpretability, Insight and Curse of dimensionality.

Rest of the paper is structured as follows in section II feature selection procedure described. Feature selection general meth-ods are described in section III. Feature selection applications are described in section IV. Related survey of feature selection methods are to be done in section V. The conclusion of its work and put the idea about the future use in section VI.

II. FEATURE SELECTION PROCEDURE

The general feature selection procedure as described in four steps as discussed below fig. 1:

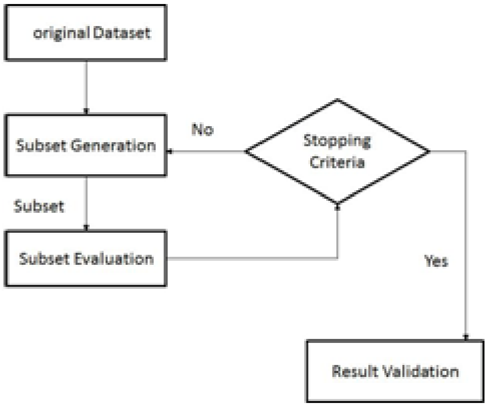

Fig. 1: Feature Selection Procedure [4]

A. Subset Generation

It is an approach of searching that generates the subsets of features by utilizing various search strategy. For searching two crucial factors are described as below:

1. *Search direction:* starting point must be selected which in turn impact the direction of search [4].

a. Forward Search:- Process of search start with an unoccu-pied set and feature is added progressively one by one.
b. Backward Search:- Process of search starts with a full set and it detaches the feature one by one.
c. Bidirectional Search:- Starts with both ends and adds and removes the features simultaneously.

2. *Search Strategy or Organization:* The search strategies are categorized into three categories sequential search, expo-nential search and random search [4].

1. Sequential Search:- It is an iterative search. It gives completeness, however not an optimal feature subset. A sequential search is easy to implement. Sequential search algorithm is such as linear forward selection, best first.
2. Exponential Search:- It is an optimal search strategy that guarantees the best result. Exponential search algorithm is such as Branch and bound, Exhaustive search.
3. Random Search:- It starts with the random selection of subsets. For next subset search, Las Vegas algorithm is used in a random search. Random search algorithm such as simulated annealing, random generation

B. Evaluation of Subset

In subset evaluation, we evaluate each newly generated subset and find the best of the feature subset based on an evaluation criterion. The correctness of search method assess-ment return by subset evaluation. It is also known as attribute selection or variable selection. The evaluation criteria are divided into three categories such as Independent, Dependent, and Hybrid criteria. Ranking of the feature is called as an Individual Evaluation [4].

C. Stopping Criteria

It is used to stop the feature selection process. Feature selection process stops only in the following condition such as when a predefined no. of the feature is selected, when it reached a predefined number of iterations, In addition, or deletion of features it fails to generate a good subset [4].

D. Result Validation

The validation process is used to measure the resultant subset using the prior knowledge about the data. The prior knowledge about the data is not available. In such case, the validation task is performed by an indirect method. For example, the classifier error rate test is used as an indirect method to validate the selected features [4].

ISSN 2348 – 8034
Impact Factor- 5.070

III. FEATURE SELECTION GENERAL METHODS

Approaches to feature selection methods are described as below:

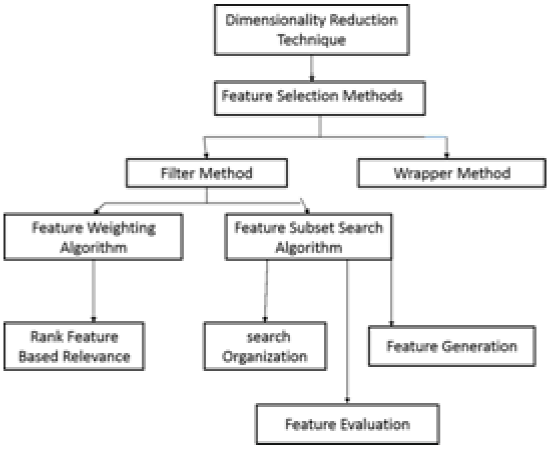
Dimensionality Reduction Technique
Feature Selection Methods
Filter Method
Wrapper Method
Feature Weighting Algorithm
Feature Subset Search Algorithm
Rank Feature Based Relevance
search Organization
Feature Generation
Feature Evaluation

Fig. 2: Feature Selection Methods

There are four approaches to feature selection are described as below:

A. Filter Method

Filter methods select feature according to a performance. It can be used only for best features. Filter method categorization based on regression, classification or clustering. Filter method includes an independent measure for evaluating a subset of features without the demand of learning method. Feature selection techniques are very efficient and fast in computation

III. Filter method is very useful when it combines with others. Filter feature selection methods are Information gain, Gain ratio, Chi-Square, Correlation feature selection (CFS), Fisher score and Inconsistency criterion [4].

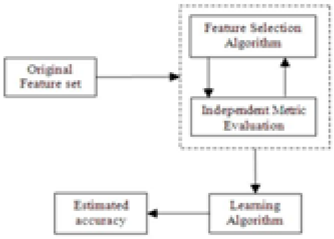

Fig. 3: Filter Method [4]

B. Wrapper Method

Wrapper methods are wrapped a classifier up in a feature selection method. Wrapper methods select subset on basis of performance measure of a classifier such as nave Bayes (NB) or support vector machine (SVM) approach. For clustering, a wrapper method select subset according to the performance of a clustering such as K-means approach. Subset generation is as same as filter method [4]. Wrapper methods are mainly slower than filter method to find better subset. Wrapper fea-ture selection methods are Sequential-forward-selection (SFS), Sequential-backward-elimination (SBE), Genetic algorithm, Estimation of distribution algorithm and Simulated annealing. The filter and wrapper method is famed by the evaluation mea-sure. Wrapper method utilize learning algorithm for evaluation of the subset. Wrapper method gives better performance [4].

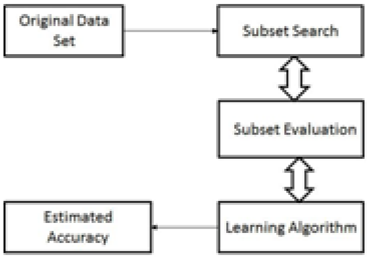

Fig. 4: Wrapper Method [4]

C. Embedded Method

Embedded method when communicates with learning method has a low cost than wrapper method. It utilizes the independent measure for deciding the optimal subset of known cardinality then, a learning algorithm is used for selection of ultimate optimal subset of different cardinality subset. Embed-ded feature selection methods are Decision trees, Weighted NB, Regularization method or Model building feature and feature selection using the weighted vector of support vector machine [4].

D. Hybrid Method

A hybrid method is developed by the combination of filter and wrapper method for handling large data set. In hybrid method, feature set is evaluated by use of independent measure and data mining technique. It selects the best subset according to diverse cardinality [12].

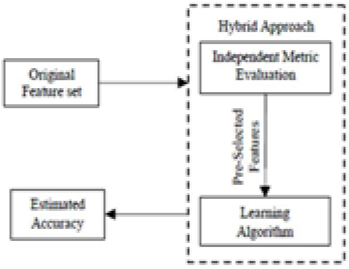

Fig. 5: Hybrid Method [12]

Algorithm 1 General Feature Selection Algorithm

```
Input:

1. Feature set of a data set having N feature SGO : Successor Generator Operator
$E_m$ : Evaluation measure (dependent or independent)
$\varphi$: Stopping Criteria Output

$X_o$: Optimal features set or weighted feature Initialization:

$X := StartP\ oint(X)$;

$X_o$:= Best of X' using $E_m$ ; Repeat:

$X := Search_S\ trategy(X , SGO(E_m), X)$;
$X_o$ := Best of X' according to $E_m$ ;
```

```
If $E_m \geq E_m X_0 or(E_m(X\ ) == E_m X_0 \&|X\ |) < |X_o|)$
Then $X_o$= X';
Until Stop criteria are not found;
```

IV. APPLICATIONS OF FEATURE SELECTION TECHNIQUES IN REAL LIFE

Application of feature selection techniques in the real world are described as below:

- Text Categorization:- Categorization of large data on internet such as email, social site, and library. Automatic text categorization and clustering are a crucial jobs.
- Remote Sensing:- In remote sensing, classification of the image by feature selection is a very important task.
- Intrusion Detection:- Information sharing, distribution, and communication task are completed by network-based computer system. So, security of the system is a very crucial problem for communication network protection from intrusion.
- Genomic Analysis:- A large number of genomic and proteomic data is build by microarray. High dimensional microarray data demands particular data analysis [4].
- Image Retrieval:- Number of images are increasing day by day. So, for accessing image, effective browsing of images, retrieval and for searching of image feature selection techniques are used [4].

V. LITERATURE REVIEW

This section reviews the various feature selection ap-proaches are discussed. This section is categorized into two different subsections. In section V-A the research that is on intermediate stage is to be discussed. In section V-B recent advancement approaches of feature selection and their applications are to be discussed.

A. Intermediate Stage

Luis Carlos Molina et al. in 2002 in this review las vegas filter (LVF) algorithm, las vegas incremental (LVI), sequential forward generation (SFG), sequential backward generation (SBG), focus algorithm, branch and bound and quick branch and bound (QBB) methods for feature selection are to be discussed. Huawen liu et al. in 2007 in this review for measuring the relevance and redundancy of features by two information criteria such as mutual information and coefficient of relevance are discussed [5].

Jasmina Novakovic et al. in 2011 in this review IB1, naive bayes, C4.5 decision tree and the RBF network machine learning approaches are used. In this paper ranking method used for feature selection with various supervised learning method [7]. Abdollah Kavousi-Fard et al. in 2012 presents, a hybrid approach which is based on teacher learning algorithm (TLA) and artificial neural network (ANN) is introduced for development of an accurate model to explore short-term load forecasting more exactly. A novel feature selection method (FSM) on basis of fuzzy cluster and fuzzy set theory are introduced. Makoto Yamada et al. in 2013 in this review, introduced a feature selection method lasso (kernelized Lasso) for capturing nonlinear input-output dependency of high di-mensional dataset.

Haytham Elghazel et al. in 2013 in this review latest ap-proach introduced named as a random cluster ensemble (RCE) for unsupervised feature selection with ensemble method. This method raise with a recursive feature elimination (RFE) method. K. Revathi et al. in 2013 in this review various FSM methods are discussed such as fast algorithm consistency the measure, agglomerative linkage algorithm, interact Algorithm, distributional clustering, relief algorithm, wrapper method and filter method. M. Akhil Jabbar et al. in 2013 in this article, proposed a ANN classification method and subset feature selection for the classification of heart disease. Cuong Nguyen et al. in 2013 in this study for diagnosing and prognosticating of breast cancer done by random forest (RF) classifier and feature selection method. To solve Diagnosis and prognosis problem via classifying Wisconsin Breast Cancer Dataset.

Xuezhi Wen et al. in 2013 in this review presents a Haar-like huge feature pool, a rapid feature selection method Ad-aBoost has been proposed and radial basis function-support vector machine (RBF-SVM) approach for vehicle detection. M.Ramya et al. in 2014 in this review text classification done by using the NB and KNN classification.

B. Recent Advancement

Yogesh Dhote et al. 2015 in this review for internet traf-fic classification feature selection techniques are described. Global Optimization Approach (GOA), Local Optimization Approach (LOA) techniques are used for best feature for traffic classification. Bryan Arguello et al. in 2015 presents a survey on feature selection technique wrapper package in Python [2]. Zena M. Hira et al. in 2015 in this review various techniques of feature selection for reducing the dimensionality of high dimensional microarray cancer data are discussed. Avinash God et al. in 2015 in this review graph-theoretic clustering technique, Density-based clustering approach, Distance-based Clustering technique, Distributed Clustering technique, model-based and grid-based clustering methods are described. Adrian Barbu et al. in 2015 in this review feature selection done by annealing method for computer vision and big data learning. Paul Bendich et al. in 2015 in this paper proposed multi-scale local shape analysis method for extraction of features from dataset. Geometric and topological features used for multi level of granularity to capture diverse types of local information from dataset. XiaoLi Zhang et al. in 2015 in this review SVM approach used intelligent fault diagnosis of rotating machinery. In this study ant colony optimization method used for synchronous feature selection and for param-eter optimization SVM approach used.

Adel Sabry Eesa et al. in 2015 in this review cuttlefish algorithm (CFA) is used as a search algorithm used for optimal subsets of feature and decision tree for judgment of the selected features that are produced by the CFA. It is used for intrusion detection system (IDS). Zhihua Cai et al. in 2015 introduced ensemble based feature selection methods. In this technique, one combine the incremental feature selection (IFS) a strategy which gives outstanding results [3]. Daniel Peralta et al. in 2015 introduced a feature selection method based on an evolutionary computation that utilizes the Map Reduce concept for subsets of features. SVM, logistic regression, and NB used as a classifier. Hadoop, spark, map-reduce techniques are introduced in this paper. Ben Hoyle et al. in 2015 in this review FSM applied to photometric redshift prediction using decision tree method with ensemble AdaBoost method. ANN used for addition features selection [14]. Pengfei Zhu et al. in 2015 in this review, introduced regularized self representation (RSR) model for unsupervised feature selection, by using RSR model in linear combination feature can be represented with its relevant features.

Muthukrishnan R et al. in 2016 introduced ridge regression and LASSO and variants of this method. This study introduces the features of the popular regression methods like OLS, Ridge, and LASSO. Vijay Bhaskar Semwal et al. in 2016 in this review feature selection based on incremental analysis of feature, classification of gait data using various machine learning approaches such as K-NN, ANN, SVM and DNN. S.Ruba Arockia Archana et al. in 2016 in this review various multi-objective genetic algorithm approach used for feature subset optimization such as a binary cuckoo search, new technique based on rough set and bat, binary ant colony algorithm, chaotic maps in binary PSO, filter based backward elimination, wrapper based PSO, GA for small and high dimensional data [1]. Wendy D.Fisher et al. 2016 in this study, introduced 2-class and 1-class SVMs and automatic feature selection. Leyi Wei et al. in 2016 in this paper present random forest predictor which is known as a MePred-RF, it is a sequence based feature selection technique [11]. Asha S Manek et al. in 2016 In this study, a Gini index based FSM with SVM classifier is introduced for sentiment analysis in large movie review dataset. K.Ramya et al. in 2017 introduced text feature selection technique, extraction of video Segmen-tation and retrieval. In this paper, a novel framework can be introduced for integrating the visual temporal information and textual distribution information. In this review solve the noisy, unavoidable video editing error and NDK detection problems

[9].Harish et al. in 2017 in this review for text feature selection five different classifiers such as naive Bayes, K-NN, Centroid-based Classifier, SVM, Symbolic Classifier used to categorize text document. Symbolic Feature Selection (SFS) a method used for text categorization discussed in this study.

Xiaokai Wei et al. in 2017 in this study introduced a new method cross-diffused matrix alignment(CDMA) feature selection, to select multi-view unsupervised feature selection by performing cross-diffused matrix alignment.

Qiang Li et al. in 2017 presents improved grey wolf optimization (IGWO) method and kernel extreme learning machine (KELM) for medical diagnosis. In this review introduced IGWO feature selection method is used for optimal feature subset for medical data. Li Ma et al. in 2017 proposed Cure smote method for the classification of imbalanced data. These methods are described in this study smote-1, safe level smote, C smote, and k means smote. Hybrid random forest (RF) method introduced for FSM. Hybrid genetic RF method, hybrid particle swarm optimization RF method, and hybrid fish swarm RF method are also described in this paper. Huseyin Polat et al. in 2017 present two FSM namely, filter and wrapper method to re-duce the dimensionality of chronic kidney disease. Abdolreza Rashno et al. in 2017 in this review, for feature selection of mars, image a latest ACO used. Mars feature vector is presented for extreme learning machine (ELM). Dijana Oreski et al. in 2017 in this paper present decision tree method for evaluation of interdependent data set characteristics and its performance [8]. Isabelle Bichindaritz et al. in 2017 in this review, present electrocardiography (ECG) signals for multilevel stress detection using feature selection method.

Bikesh Kumar Singh et al. in 2017 in this review, the latest hybrid FSM is used for subset feature selection classification of benign and malignant tumor in breast image [10]. Sepehr Abbasi Zadeh et al. in 2017 in this review, introduced scalable FS problem as a distributed diversity maximization problem by initiate a mutual information based metric distance function on the features [13]. Minnan Luo et al. in 2017 in this review, adaptive unsupervised FS used with structure regularization [6].

VI. CONCLUSION

In this study, we have presented a general overview of the feature selection techniques. We described the whole procedure of feature selection in this review. Four approaches of feature selection are described in this paper namely, fil-ter method, wrapper method, hybrid method and embedded method. Application of feature selection in the real world is also discussed in it. We have surveyed about techniques of machine learning for feature selection and their modified version. A detailed survey was done on intermediate research, and recent advancement techniques of feature selection and its application are to be discussed in this paper.

This paper helps in the understanding of feature selection and its procedure. There are several directions in which this work could be expanded. Future work in this direction is the feature selection for large dimensional data, scalability and stability issues of feature selection, to make backward elimination more efficient. This survey helps in modeling the efficient model for feature selection which gives the more accurate result.

REFERENCES

1. *S Ruba Arockia Archana and MS Thanabal. Optimiza-tion algorithms for feature selection in classification: A survey. Optimization, 4(2), 2016.*
2. *Bryan Arguello et al. A survey of feature selection methods: algorithms and software. PhD thesis, 2015.*
3. *Zhihua Cai, Dong Xu, Qing Zhang, Jiexia Zhang, Sai-Ming Ngai, and Jianlin Shao. Classification of lung can-cer using ensemble-based feature selection and machine learning methods. Molecular BioSystems, 11(3):791– 800, 2015.*
4. *Vipin Kumar and Sonajharia Minz. Multi-view ensemble learning: A supervised feature set partitioning for high dimensional data classification. In Proceedings of the Third International Symposium on Women in Computing and Informatics, pages 31–37. ACM, 2015.*
5. *Huan Liu and Hiroshi Motoda. Computational methods of feature selection. CRC Press, 2007.*
6. *Minnan Luo, Feiping Nie, Xiaojun Chang, Yi Yang, Alexander G Hauptmann, and Qinghua Zheng. Adaptive unsupervised feature selection with structure regular-ization. IEEE Transactions on Neural Networks and Learning Systems, 2017.*
7. *Jasmina Novakovic´. Toward optimal feature selection using ranking methods and classification algorithms. Yu-goslav Journal of Operations Research, 21(1), 2016.*
8. *Dijana Oreski,ˇ Stjepan Oreski,ˇ and Bozidarˇ Klicekˇ. Ef-fects of dataset characteristics on the performance of feature selection techniques. Applied soft computing, 52:109, 2017.*

9. K Ramya and M Siva Sundara Vinayagamoorthy. Text feature selection and extraction over video segmentation and retrieval. 2017.
10. Bikesh Kumar Singh, Kesari Verma, AS Thoke, and Jasjit S Suri. Risk stratification of 2d ultrasound-based breast lesions using hybrid feature selection in machine learning paradigm. Measurement, 105:146–157, 2017.
11. Leyi Wei, Pengwei Xing, Gaotao Shi, Zhi-Liang Ji, and Quan Zou. Fast prediction of protein methylation sites using a sequence-based feature selection technique.
12. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017.
13. Bing Xue, Mengjie Zhang, Will N Browne, and Xin Yao. A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation, 20(4):606–626, 2016.
14. Sepehr Abbasi Zadeh, Mehrdad Ghadiri, Vahab S Mir-rokni, and Morteza Zadimoghaddam. Scalable feature selection via distributed diversity maximization. In AAAI, pages 2876–2883, 2017.
15. Roman Zitlau, Ben Hoyle, Kerstin Paech, Jochen Weller, Markus Michael Rau, and Stella Seitz. Stacking for ma-chine learning redshifts applied to sdss galaxies. Monthly Notices of the Royal Astronomical Society, 460(3):3152– 3162, 2016.