International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Research on Feature Selection using SVM

C.Amali Pushpam, J.Gnana Jayanthi

Abstract*— A very fast and efficient classification algorithm is imperative to any application. Nowadays all kinds of applications produce a huge volume of data. Handling these 5'V characteristics data is really very crucial. While processing data, data classification simplifies the mission. Though many classification algorithms are available, they are not up to the mark to meet the fast growing challenges of current digital world. To fill this gap, feature selection is integrated with classifiers, as Feature selection has proved its impact on performance of classifiers. SVM is one of the most frequently used classifier. In this paper, different feature selection methods have been analyzed by studying 21 articles. This survey makes public that SVM based feature selection works better and widely used. Also in feature selection, filter method is widely used.

Key Words: Feature Selection, Classifier, Support Vector Machine, Ensemble

I. INTRODUCTION

Object oriented concepts brought a new dimension in research field in 1980. After its entry, researchers came to know the importance of data. Data plays a central role in solving any problem. Information or patterns mined from data are very much useful in any application like medicine, business, education, communication, agriculture...etc. After identifying the importance of data, new branches like data reduction, normalization, digitalization...etc came to exists in research field. Whenever we talk about data, its feature also should be considered. Because data and its features are inseparable. Features are characteristics or attributes of data. In pre processing technique of classification, data reduction is done. Because data is a mixture of valuable, relevant, noisy, irrelevant and redundant data. Like data reduction, feature selection also should be carried out. Because all features' contributions are not equal. Some features only are informative. Hence feature selection is done carefully to reduce research space, time and cost.

Paper Organization

The work of this paper is organized as follows: Section 2 describes feature selection and methods, in section 3 surveys on SVM based feature selection is given, section 4 research result and Section 5 conclusion and future work.

Manuscript published on November 30, 2019.
* Correspondence Author
C.Amali Pushpam*, Research Scholar, Rajah Serfoji College, (Affiliated to Bharathidasan University), Tamil Nadu, India. Email: joemarycap@gmail.com
J.Gnana Jayanthi, Assistant Professor, Dept.of Computer Science, Rajah Serfoji College, (Affiliated to Bharathidasan University), Tamil Nadu, India (Email: jgnanajayanthi@gmail.com)

II. FEATURE SELECTION

Feature selection method is based on the process of selecting subset of features from original by filtering irrelevant and redundant features. It is extracting useful and relevant features from data. As the numbers of inputs are reduced in feature selection, model will be trained faster, complexity will be reduced and accuracy will be increased. Feature selection has great impact on accuracy of classifiers [1][3][5][6][7][9]. Feature extraction is different from feature selection. It is the process of creating features for a given data or instance. In Feature extraction, input data is transformed by combining the original attributes or in another way and it can be defined as a special form of dimensionality reduction.

Various methodologies in Feature Selection
Filter, Wrapper and hybrid methods

a. Filter

It is a preprocessing step. Different types of filter methods are available. They are classifier independent. Filter methods consider individual feature and assign score based on correlation with target variables through statistical tests. Then features having high score than threshold value are selected and filtered and form feature subset.

Variance is the measurement in selecting variables in filter. Features which are having higher variance may contribute more information in prediction.

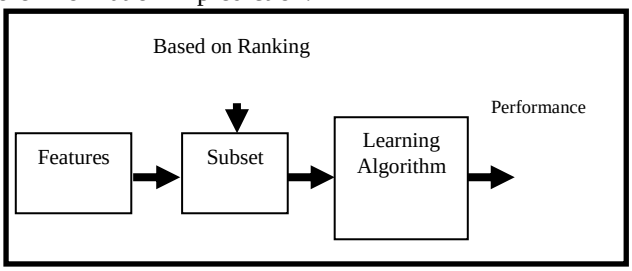

Fig.1.

Examples for Filter methods are Pearson's Correlation, LDA, ANOVA, information gain, chi-square test, fisher score, Gain Ratio, Symmetrical Uncertainty, Relief-F, One-R , Fast Correlation Based Filter (FCBF), CFS, and INTERACT, correlation coefficient, variance threshold, etc. Out of 21, 10 articles[1][2][4][6][7][8][9][14][15][17] support filter method. These articles use mutual information (information gain) metric.

b. Wrapper

In wrapper, based on performance of classifier useful features are measured. In wrapper, learning steps are repeated and best subset is identified and validated by cross validation. Hence wrapper method is more expensive than filter method

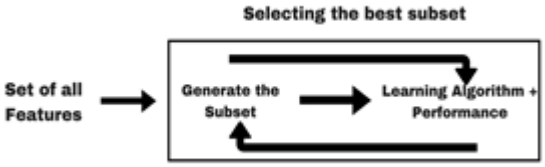

Fig.2.

Examples for wrapper methods are forward feature selection, backward feature elimination, recursive feature elimination, sequential feature selection algorithms, genetic algorithms, etc.

c. Hybrid Method

Filter methods are fast and less expensive. But wrapper methods provide best subset of features through cross validation and overcome over fitting problem. Both filter and wrapper methods have their own cons and pros. In hybrid, strengths of these two methods are merged and used to increase the performance by reducing data dimensions by filter and selecting best subset of features through wrapper methods.

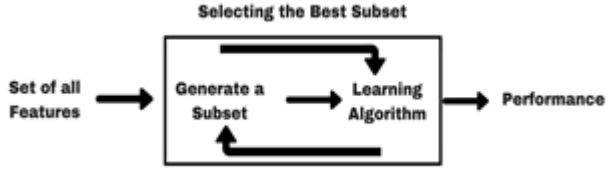

Fig.3.

Examples for hybrid methods are Regularized trees, Memetic algorithm, Random multinomial logit, LASSO and RIDGE regression, etc

III. FEATURE SELECTION USING SVM – SURVEY & RESULTS

In Ref [2], authors have used more than one SVM classifier (i.e) ensemble SVM concepts. In ensemble, to each SVM, same training datasets (D1, D2, D3…) with different feature subsets is applied. Ranking vector (R) of SVM for all features - $\{k_1, k_2, k_3, k_4, k_5\}$ is calculated. Ranking vector R is (d * j) array (d- 3 training datasets and j- 5 features, 3 * 5 = 15 elements). R_{kj}- it denotes ranking of k_{th} feature computed from the j_{th} training set. Stability criteria S_k for all features is calculated from ranking vector. Features are ranked according to stability criteria and arranged in descending order. If 'r'- no. of features is predefined, that 'r'- features will be selected and used to train SVM. Authors have proposed a new method for selecting features by estimating feature's stability. They have proved that this new method SVM-SE (Stability Evaluation) works better than SVM-RFE (Recursive Feature Elimination) method.

This method saves memory by storing aggregation sums of, not full matrix R. There is reduction in computation costs and training time and gives best results on most of the datasets.

When SVM-SE is combined with backward selection procedure to improve stability of the algorithm, Training time will be more as selection procedure is repeated with subset of features deleted at each time

Datasets: 4 datasets – Non linear toy, WDBC, usps:3 vs 5 and usps: 6 vs 8

In Ref [3], authors have proposed a new method using fisher score concept. They have applied different training data set with same feature set. Fisher score of each feature is calculated. Choose some threshold values. Features having less fisher score value than this threshold are dropped out and new feature set (fs1) is formed. Take 5 different training data sets (D1, D2 ...D5) and apply this feature subset (fs1). Calculate average validation error (VE). Repeat this with different feature subset (fs2, fs3 ...etc). Select the threshold with the lowest average validation error. Drop feature which are having less fisher score value than selected threshold value. This feature set will be considered as final feature subset. Feature score and SVM combination do feature selection effectively. This method is a useful for data reduction and feature selection and thus increases performance of classifier.

This method reduces the dimension of feature space and improves classification ability of classifier.

There is no specific method in splitting training dataset. It is done randomly. Process is repeated 5 times to calculate average validation error.

Datasets: KDD CUP '99 issued by MIT Lab

In [4] authors have discussed ensemble feature selection using filter concept. They have used the concept of same training data set with different feature set. Training data set is given to filters. Based on certain metric feature subsets are formed. These feature subsets are given as input to classifiers. Outputs are integrated. That is the final feature subset. They have applied another method. Training dataset is given to filters and various feature subsets will be output. This will be integrated and given as input to classifier. Accuracy is verified. If it is standard one, then this integrated feature subset is the final one. They have concluded that these single methods are not feasible for fast growing problems. In their proposed work, they have combined goodness of ensemble filter and hybrid wrapper methods. In ensemble filter, Fast Correlation Based Filter (FCBF) metric is applied. In wrapper, hybrids of meta-heuristic algorithms are used. Meta- heuristic algorithms worked based on inspired behaviors of birds, ants, swarms… etc. $ABACO_H$ and IBGSA are meta – heuristic algorithms in this proposed work. This proposed work gives better result in feature selection by increasing effect of meta-heuristic algorithms.

This method has higher feature reduction by increasing effect of meta-heuristic methods. It increases classification accuracy.

It does not perform well on all datasets taken for experiments. Among 5 datasets, it not works well in Leukemia dataset.

Datasets: 5 datasets- SRBCT, Leukemia, Prostate, Lung, Colon

In [5], authors proposed a new method for feature selection named HFSA (Hybrid Feature Selection Algorithm). Feature selection was done based on FMIFS where MI is evaluated to find out the dependency between features and target classes. Most relevant features are ranked high. These high scored features are selected and used to train classifier. FMIFS is outperforms than MMIFS(Modified Mutual Information Feature Selection) and MIFS(Mutual Information Feature Selection).

Its performance is higher as most informative features are scanned. But best number of features is not revealed

Dataset: KDD Cup'99

In [6], authors have presented a novel feature selection method using SVM based on RFE (Recursive Feature Elimination) and P.O (Parameter Optimization). To do P.O, GS(Grid), PSO(Particle swarm Optimization), GA(Genetic Algorithm) are used. Hence this new method contains 3 algorithms namely SVM-RFE-GS, SVM-RFE-PSO, SVM-RFE-GA. This method was compared with RFFS(Random Forest Feature Selection)-GS and mRMR. Among all these, SVM-RFE-PSO gives better result.

As this method is capable of extracting more representative and useful genes, time is saved. SVM-RFE-PSO has better prediction performance of AUC.

Datasets : DNA microarray data on GEO data set, RNA-seq data on TCGA dataset.

In [7], to increase efficiency of classifier, authors have presented a new framework to select features and reduce the dimensionality of input vector. In feature selection, out of M features, m features are selected which are having high mutual information with target classes and less mutual information with non-target classes. In order to reduce the dimensionality of input vector, DI-SVM (Diversified Input-SVM) classifier concept is applied. Here the number of SVM classifiers is equal to the number of classes in dataset. For each class, unique combination of feature set is selected. Each SVM is trained with different class of dataset and different feature set selected for that class.

In terms of accuracy, AUC & ROC (Area under the curve, Receiver operating characteristics) DI-SVM model outperforms the conventional model of SVM, NN, and K-NN.

K-fold CV (Cross Validation) technique partitions the training data into k subsets. This process is repeated. Each subset is used once for training while being used k-1 times for validating classifier. It will increase training time.

Data sets: 100 raw time signals from sensors using an Arduino Uno Micro controller board with a serial communication between the Arduino and personal computer.

In Ref [8], authors have proposed a fresh method for feature selection using 3 different metrics namely Principal components, Information Gain (IG) and chi-squared. Based on these metrics, features are ranked. Then SVM is trained with top 3 features of each metric and results are compared.

It reduces computation cost.

As this method uses a quarter of the features, it leads to 5% loss in accuracy.

Data sets : 1000 web pages.

In Ref [9], authors used Information Gain (IG) and triangle based KNN for feature selection, greedy K-means for clustering datasets and polynomial SVM for classification. Most relevant features for each attack are selected through efficient information gain and triangle based KNN. By mapping these features which are most relevant, low dimensional feature vectors are created for each data. As it is quiet easy to analyze data in clustering, datasets are reprocessed through K-means clustering.

This method provides high accuracy and high detection rate. It reduces error rate and training time.

This is not applicable to real time web analytics for intrusion Detection.

Data Sets: KDD Cup'99.

In Ref [10], authors have applied Genetic algorithm for optimizing cost and gamma parameters to enhance efficiency of SVM. Feature selection is done through Particle swarm optimization algorithm for classification model.

Feature selection by PSO is more reliable and enhances the performance of IDS.

Data sets: KDD Cup'99 issued by MIT

In Ref [11], authors have proposed an effective intrusion detection model based on SVM with feature selection and parameter optimization. SVM is classifier using Radial Basis kernel function (RBF). Particle swarm optimization optimizes parameter of RBF of SVM. In order to increase the efficiency, feature selection is done using Information Gain (IG) technique.

It gives good results in terms of accuracy, detection rate, FPR .

In order to apply IG method for feature selection, the continuous attributes of NSL –KDD data set must be first discretized using the methods like InfoGain AttributeEval from Weka. Test is not applied on different data sets in order to validate the method

Data sets: NSL – KDD data set which is enhanced version on KDD cup'99.

In Ref [12], authors have proposed a new hybrid method of SVM + GA for intrusion detection system. By using this method, 10 features are selected out of 45 features. Then these chosen 10 features are classified into 3 priorities – 4 features in first priority, next 4 features in second priority and 2 features in third priority. In these, all selected features and all selected features except third priority produce same result.

It gives true positive value of 0.973 and 0.017 as the false positive value.

In Ref [13], authors have studied various feature selection strategies and analyzed which is suitable for SVM. They tested the performance of Direct SVM, F-Score + SVM, F-Score +RF(Random Forest)+SVM, RF+RM- bound SVM on various data sets and compared the results and found that most of the feature selection strategies are classifier independent.

There is a number of feature selection strategies which are independent of classifier used. Classifier performance depends on contribution of features.

If their contribution in prediction is more, then they will be considered as important features. When all features are same type and same level, feature selection does not make any changes in classifier's performance.

In Ref [14], authors have proposed a new method for combining feature subsets which are selected by different feature selection methods using different metrics like probability distribution, entropy, correlation ..etc. They compared this with union of different feature subsets consisting redundant and irrelevant features which degrade accuracy of method. To overcome this problem, authors have applied feature-class Mutual Information for selection of relevant features and feature-feature Mutual Information for selection of non-redundant features and reduce the redundancy.

Datasets : 14 UCI and 5 gene expression and 2 network datasets.

In Ref [15], authors have proposed a new method to eliminate excess and unimportant features. Each feature is ranked through information gain technique. Keep separate subset of important features and train the model using this subset. Calculate accuracy over this subset using the samples. Train the model using different subsets of important features and samples. Based on accuracy, select the final list of features for model.

In [16], a novel classifier ensemble was proposed by using ensemble detection models. Data level and feature level detection models are generated. In detection model 1, data subsets are created from original training data and given as input to base classifiers. Results of these classifiers are combined. This is the output of ensemble detection model 1. In detection model 2 , feature subsets are created from original training data and classifiers are trained using these feature subsets. Results of classifiers are combined. The final prediction is obtained by combining output of these two detection models.

IV. SURVEY-SUMMARY

Generally this paper gives an outlook on SVM based feature selection. Introduction part briefly explains the various methods of feature selection. In this survey, 21 papers are analyzed. These papers discussed about various feature selection methods including filter, wrapper and hybrid method. Some algorithms like Genetic algorithm, heuristic algorithms like Particle swarm optimization, fireflies algorithm..etc are also used in feature selection. Filter is widely used as it is classifier independent, fast and suitable for large data set. Mainly this survey focused on SVM based feature selection. All these techniques are not providing better result to all applications. The best feature subset selected by above mentioned techniques is not good for another problem. Hence to reduce research space, cost and training time to enhance efficiency of classifier, this field invites researchers for new conception.

V. CONCLUSION

Security issue is always ever green and hot topic among researchers and all nations spent more time and finance in this area. This universe is providing so many resources to humanity. One of the important resources is data. Extracting information from these data and applying in various applications for the betterment of society is a very big and challengeable task as well as wonderful service to society also. A number of data analytic tools are available in digital market. Widely used data analytic tool is Data Mining. It provides a huge number of algorithms to analyze the data. Among them, ANN, SVM, KNN, DT and K-Means are frequently used. Researchers set their focus on SVM because of its simple and resourceful characteristics. SVM works very well in small set of data. But it suffers when data size grows larger. One of the solutions is feature selection. While analyzing data, the impact of features / variables / attributes also should be considered. Feature selection enhances efficiency and performance of classifier. This survey proves this. Still, more researches are to be carried out in this area to get better result and to meet the fast growing challenges in this field.

REFERENCES

1 Esra Mahsereci Karabuluta* , Selma Ayşe Özelb , Turgay İbrikçib,c , "A comparative study on the effect of feature selection on classification accuracy", in 2212-0173 © 2012 Published by Elsevier Ltd. doi: 10.1016/j.protcy.2012.02.068.

2 Tao Ban et.al, "Feature Subset selection by SVM Ensemble", in 978-1-5090-4240-1/16/IEEE, 2016.

3 ZHANGXue-qin, GU Chun-hua and LINJia-jun, "INTRUSION DETECTION SYSTEM BASED ON FEATURE SELECTION AND SUPPORT VECTOR MACHINE", in East China University of science and technology, 1-4244-0463-0/06/$20.00 ©2006 IEEE

4 Amirreza Rouhi ..et.al., "A hybrid feature selection approach based on ensemble method for high-dimensional data", in 2nd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC2017), 978-1-5090-4330-9/17/ IEEE, Shahid BahonarUniversity of Kerman, Iran, 2017 PP:16-20.

5 Rekha Preethi M.C1 , Mr.Chetan R2 , "Least Square Support Vector Machine based IDS, using feature selection algorithm", in International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 6, Issue 3, May- June 2017, PP: 64-68

6 Ying Zhang, et.a,., "An Efficient Feature Selection Strategy Based on Multiple Support Vector Machine Technology with Gene Expression Data", in Hindawi BioMed Research International, Volume 2018, PP:1- 11

7 Sana Ullah Jan and Insoo Koo, "A Novel Feature Selection Scheme and a Diversified-Input SVM-Based Classifier for Sensor Fault Classification" in Hindawi Journal of Sensors Volume 2018, Article ID 7467418, 21 pages

8 Nhamo Mtetwa, Mehdi Yousef et.al "Feature Selection for an SVM Based webpage classifier" in 2017 4th IEEE International Conference on Soft Computing and Machine Intelligence, 978-1-5386-1314-6/17/$31.00 ©2017 IEEE, PP:85-88

9 Venkata Suneetha Takkellapati1 , G.V.S.N.R.V Prasad2, "Network Intrusion Detection system based on Feature Selection and Triangle area Support Vector Machine" in International Journal of Engineering Trends and Technology- Volume3Issue4- 2012, PP:466-470

10 Mehdi Moukhafi, et.al "A novel hybrid GA and SVM with PSO feature selection for intrusion detection system" in International Journal of Advances in Scientific Research and Engineering (ijasre), Volume 4, Issue 5 May – 2018, PP:129-133

11 EL MOSTAPHA CHAKIR et.al, "AN EFFECTIVE INTRUSION DETECTION MODEL BASED ON SVM WITH FEATURE SELECTION AND PARAMETERS OPTIMIZATION", in Journal of Theoretical and Applied Information Technology, 30th June 2018. Vol.96. No 12, PP: 3873-3885

12 B. M. Aslahi-Shahri, et.al "A hybrid method consisting of GA and SVM for intrusion detection system", in Neural Computing & Applications- june 2015

13 Yi-Wei Chen and Chih-Jen Lin, "Combining SVMs with Various Feature Selection Strategies" - book chapter 12 , Publisher: Springer Berlin Heidelberg.

14 Nazrul Hoque1 · Mihir Singh2 · Dhruba K. Bhattacharyya2, "EFS-MI: an ensemble feature selection method for classification An ensemble feature selection method
15 Gangaprasad G.Ghungre et.al, "Intrusion Detection Using Support Vector Machine with Feature Reduction", in International Journal of Computer Engineering and Applications, volume XII, May 18, PP:1-9.
16 Uma R.Salunkhe et.al, "Security Enrichment in Intrusion Detection System using classifier Ensemble", in Journal of Electrical and Computer Engineering, Vol 2017, PP:1-6.
17 Maryam Zaffar .et.l , " A Study of Feature Selection Algorithms for Predicting Students Academic Performance" in *International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 9, No. 5, 2018, PP:541-549.*
18 Bo Chen1, Jie Yu 2*, Xiu-e Gao3*, Qing-Guo Zheng, "A human body physiological feature selection algorithm based on filtering and improved Clustering", PLOS ONE | https://doi.org/10.1371/journal.pone.0204816 October 31, 2018, PP:-15
19 Chuan Liu a , ∗, Wenyong Wang a , Qiang Zhao a , Xiaoming Shen b , Martin Konan a, " A new feature selection method based on a validity index of feature subset", in pattern recognition letters 92 by Elsevier, March,2017, PP:1-8.
20 Hui-Huang Hsu, Cheng-Wei Hsieh ⇑, Ming-Da Lu, "Hybrid feature selection by combining filters and wrappers" in Expert Systems with Applications 38 (2011) 8144–8150, 2011 Elsevier Ltd, PP:8144-8150
21 S. Sasikala a,*, S. Appavu alias Balamurugan b, S. Geetha c"Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set" in Applied Computing and Informatics (2014), 29 March 2014, PP:1-20

AUTHORS PROFILE

C.Amali Pushpam, MCA., M.Phil., is currently pursuing Ph.D. and servicing as an Assistant Professor in the Department of Information Technology, Bon Secours College for Women, Thanjavur, affiliated to Bharathidasan University, Tiruchirappalli, India since 2006. During her service, she has organized many international and nationsl conferences, seminars and symposium. Her main research work focuses on Data Mining, Big Data Analytics and Network Security.

J. Gnana Jayanthi, M.C.A., M.Phil., Ph.D., is presently servicing as an Assistant Professor in the Department of Computer Science, Rajah Serfoji Government College, Thanjavur, India. She has published more than 30 research papers in International and National conferences and Technical Journals and are cited in popular refereed publishers, IEEE, ACM and Springer. She is a life member of Computer Society of India (CSI), Member of the World Scientific and Engineering Academy and Society (WSEAS), International Association of Computer Science and Information Technology (IACSIT) and member of International Association of Engineers (IAENG). Her research interests include Distributed DBMS, Big Data Analytics and IoT.