THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

Compact Discourse on Feature Selection

Naina Handa
Research Scholar
Lovely Professional University,
Phagwara, Kapurthala (Jalandhar)
naina_pisces17@yahoo.co.in

Dr. Anil Sharma

Associate Professor

Lovely Professional University,
Phagwara,Kapurthala(jalandhar)
anil.19656@lpu.co.in

Abstract

Feature Selection is an emerging field of Machine Learning and Data Mining. Feature Selection helps to remove the inappropriate i.e. redundant and irreverent features from the dataset and develop an improve the classification accuracy. The paper has explored the different Feature Selection techniques like Filter method, Wrapper and Embedded method as basic methods.The Hybrid and Ensembled based Feature Selection have used in many papers and shown that the better Machine Learning Prediction accuracy.The different research papers have used the UCI Repository dataset for experiment purpose.Feature Selection is applicable to almost every field.

I. Introduction

Nowadays,we are at present encountering the accessibility of tremendous measure of information; trillions of gadgets produce heterogeneous information.Continuous appraisals surmise 4 − 43 exabytes of required yearly stockpiling requires in 2025 [1].There are multiple varieties and a vastquantity of data is being gathered and put away in real-world databanks at a marvelous rate. The gathered volume of warehoused data rises, the capacity to comprehend and utilize it isn't corresponding. Additionally, the clients request more sophisticated information. Therefore Feature selection isbig issue of Data Mining and Machine Learning. Feature Selection extracts the relevant highlights from the enormous information.To separate the most helpful data from datasets and to improve forecast precision, include determination is vital in Data Mining and Machine Learning as a fundamental preprocessing step.

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

There are large players of big data are bioinformatics fields YouTube, Twitter, and Social networks and the volume of information will created and should put away is considerably complex. Not exclusively is the size of information expanding, however So is the dimension of datasets; there are increasing dimensional datasets crosswise over areas.The dimensionality is rapidly creating, achieving not many or countless features and the example size is not comparable pace of addition.We experiencebig dimensional data containing only few (or considerably less) examples. These circumstancescreate it tough to make efficient models of data investigation. Taking care of high-dimensional data prompts a couple of matters that are supposed the problem of curse of dimensionality. It was first seen by Bellman in 1968[2], and the later on it was separated by Hughes [3], yet the interrelated condition is up 'til now being investigated and is a central research point in Data Mining and Machine Learning. This issue prompts overfitting, postponed the computational times and unstable Machine Learning models[4].

II. Machine Learning

Machine learning comprises a lot of methods, which enable a machine to take meaningful patterns from information straightforwardly with negligible human interaction. The quality of machine learning, to a limited extent, subject to human information. Such information can assist a machine with learning all the more efficiently through methods like Feature Selection and Learning[5]. Through this beneficial interaction, Machine Learning has been effectively applied in numerous applications and accomplishes cutting edge execution. The Machine Learningalgorithms can be categorized into Supervised,Semi-Supervised and Unsupervised[6].The Supervised technique focused on Classification and Regression of data.In this technique, the data is labeled in the training set.In Semi-Supervised the training data is not fully labeled and in unsupervised data is unlabeled and it focused on the clustering of data.The different algorithms like Random Forest, Decision tree and Support Vector Machine to name a few[7].

III. Feature Selection

Feature selection is at present utilized in various zones. Perspectives explicit to specific application space should be considered in the Feature Selection plan. Presumably, the most delegate uses of FS are bioinformatics, interactive media and social media. Feature Selection is a significant part of many ML applications managing little samples and high-dimensional information. This is a significant preprocess set before applying the Machine Learning algorithm. Picking the most significant features is a fundamental advance for Knowledge Discovery in numerous regions[8]. The accuracy of the ML model purely depends on the dataset of relevant features. The way toward expelling the redundancy and irrelevancy of features set decreases dimensionality and increases the predictive accuracy of the model.

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

There are two ways to deal with decrease the dimensionality of the dataset are Selection and Extraction of features. In Feature extraction, the new premise is picked for the information and new features are gotten from the inputs. Then again, the fundamental goal of the FS method is to lessen the effects of high dimensionality on the dataset and to get a subset of features from the whole list of the feature set that can efficiently portray the information. Decreasing the dimensionality of information helps to solve the problem of the curse of dimensionality [9] that truly corrupts the capabilities of ML algorithm to create strong models. As the diminished subset is generally significantly littler than the arrangement of the information includes, the calculation time of the ensuing investigation is incredibly decreased. Truth be told, there is no technique or methodical methodology for picking the most reasonable FS strategy for a specific issue. The expanding number of FS systems that are for sure extremely effective and advanced makes the issue of choosing the most reasonable FS significantly increasingly evident[10].

The Feature Selection can be categorized into the filter method,embedded methodand wrapper method .The primary contrast between these methods is standing out they cooperate with a model. In the Filter Feature Selection technique[11], the FS search is totally disengaged from the development of ML model. The significance score of feaures is determined to agree on some basis and the highlights that accomplish the most minimal scores are disposed of from further preparing. The benefits of filter that it is adaptable, not computationally requesting, and generally quick contrasted with different FS techniques.The filter can be additionally partitioned into univariate and multivariate methods. Though univariate strategies assess each component independently, multivariate systems additionally consider dependency between features. In spite of the fact that univariate systems are very straightforward, their presentation is aggressive with multivariate approaches and complex wrapper techniques[12].

Wrapper Feature Selection techniques use the ML model as a component of the Feature Selection procedure. Chosen includes subsets that are assessed via preparing and analysis classifiers. The most elevated assessment scoreia accomplishes by last subset. Ordinarily, the pursuit through every single imaginable mix of features isimpractical, so an alternate inquiry method must be utilized; most of the time, deterministic methodologies, for example, forward determination or backward elimination are utilized[13]. In any case, some as of late proposed wrapper FS procedures exploit evolutionarytechniques such as genetic algorithms.

The disadvantages of Wrapper technique is danger of overfitting and its great computational necessities. Also, they are classifiers explicit. A compelling technique to bringing computational prerequisites is down to consolidate a straightforward wrapper technique and Filter Feature ranking technique with two-advance FS method. In the essential advance, the Filter technique picked the most important features and generally decreasing the segment space. Now in resulting advance, the Wrapper technique checks the diminished component space for best component subset. The inserted strategy gives an elective best computational multifaceted nature over wrapper methods. Not at all like the wrapper approach, has it kept up a key decent ways from overabundance execution of the classifier and the assessment of various part subsets. The FS

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

framework is a touch of the learning estimation and utilizations its properties to assess the centrality of features.Be that as it may, correspondingly as the wrapper techniques the inserted strategies are classifier explicit. Other than these three surely understood FS draws near, another gathering of strategies has as of late risen that is based over existing FS techniques: Ensemble.Ensemble FS builds gatherings of features subsets and afterward join these subsets to create collected outcomes. The purpose of Ensemble FS is to build up a solid and stable FS execution while overseeing high dimensional dataset. The rule downside of the Ensemble technique is that it functioned as a gathering of different base models and it is complex to appreciate and comprehend.

IV. Data preprocessing

Data preprocessing is a significant and fundamental stage whose principal objective is to get the last dataset that can be viewed as right and helpful for further mining the data. To viably utilize Machine Learning and Data mining strategies, include Dimensionality Reduction and Feature Selection techniques are frequently employed as a initial step to dataprocessing.The objective of the two procedures is to compel the Dimensionality of the information to the number of features basic to depict the information.Dimensionality decline is the distinction in high-dimensional information into a basic delineation in a particular segment space of decreased dimensionality. In actuality, FS strategies lessen the first component space without change, so the first features are saved and the pertinent translation is conceivable. Different advantages of FS incorporate the expulsion of insignificant or excess featured that may deliver coincidental relationships in Machine Learning algorithms and debase the model execution; the age of models dependent on lesser features that are progressively clear, and less complex to unravel and imagine; and the necessity for less data additional room [14]. Moreover, a reduction of feature space is related to a lessening in the intriguing space that must be considered by information mining, which spares computational assets and animates information mining procedures.

V. Literature Review

A critical number of the recently anticipated FS techniques were not proposed to work with high dimensional informationand in that farthest point are not sufficient these days. It motivated to design an efficient Feature Selection procedures. This field has gotten the thought of various experts for quite a while. In writing ,the different systems have been proposed for Feature Selection. A few Researchers has influenced on filter method to join confirmation as a result of their computational cost adequacy in any case different has concentrated on the Wrapper technique and it produces progressively powerful and exact component subsets and a few have

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

concentrated on Hybrid Feature Selection technique. The table has shown the literature reviews of different papers are as follow

Reference No.	Literature Review
[15]	The researchers have implementeda novel Feature Selection method by hybridingS U ranking technique and Genetic algorithms. The Weka and MATLAB tools have used for implementing the same. The experiments have done on the UCIMachine Learning repository datasets.
[16]	This paper builds up a model named WANFIS, which is a Hybrid model. In this model, the WOA i.e. Whale Optimization Algorithm is being used for Parameter Tuning and also for the Feature Selection of ANFIS Machine Learning algorithm.Theanalyzewas done on the satellite dataset.
[17]	A wrapper made out of the Genetic Algorithm (GA), Random Forest (RF) and heuristic search tool has used in the Machine Learning model.The researchers have focused on the accuracy of Random Forests and optimality of the Genetic Algorithm. The set of Filter techniques are used to make a diminished quest space for GA-RF wrapper, which produce a decreased subset. The Breast Cancer dataset from the UCI Repository has taken for experiment. Rast Miner tool has utilized to process the outcome.
[18]	The authors propose a novel feature selection technique for phishing detection system named as, (HEFS)Hybrid Ensemble Feature Selection . The authors have explored the Ensembling technique for feature selection in this paper.The experiment datasets have taken from the (UCI) repository.
[19]	In this paper, The authors have recommended a few FS strategies centered on the Ensembling technique. The researchers have done many experiments and focused on the stability, sensitivity and classification accuracy of the model. The experiments have explored the different artificial and real-world datasets
[20]	The Researchers have proposed an improved WFS method before integration with anSVM i.e. Support Vector Machine algorithm as a full flaw finding a framework for a moving component bearing contextual analysis. The bearing vibration dataset has used for investigation.The dataset has made available by the Case Western Reserve University Bearing Data Centre.

VI. Conclusion

Feature selection is asignificant preprocessing phase in Machine learning as well as in Data Mining. There are many standard methods for feature selection like wrapper method, filter method and embedded method .but nowadays the researchers have focused on hybrid and Ensembling methods for feature selection.In the survey, it has shown that the Ensembling is

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

hybrid methods perform better. Thedifferent papers have used different implement tools like Weka,Matlab,Python and Rapid Miner to name a few. Many authors have taken the dataset from the University of California Irvine (UCI) repository for experiment purposes.Feature selection is applicable in almost every field like big data,Fault diagnosis, and Phishing detection system to name a few.

VII. References

[1]Via, M. (2017). Big Data in Genomics: Ethical Challenges and Risks. Revista de Bioética y Derecho, (41), 33-45.

[2] Fu, G. S., Levin-Schwartz, Y., Lin, Q. H., & Zhang, D. (2019). Machine Learning for Medical Imaging. Journal of healthcare engineering, 2019.

[3] Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. IEEE transactions on information theory, 14(1), 55-63.

[4] Fu, G. S., Levin-Schwartz, Y., Lin, Q. H., & Zhang, D. (2019). Machine Learning for Medical Imaging. Journal of healthcare engineering, 2019.

[5] Marinho, M., Arruda, D., Wanderley, F., &Lins, A. (2018, September). A Systematic Approach of Dataset Definition for a Supervised Machine Learning Using NFR Framework.In 2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC) (pp. 110-118).IEEE.

[6] Mafarja, M., Aljarah, I., Faris, H., Hammouri, A. I., Ala'M, A. Z., &Mirjalili, S. (2019). Binary grasshopper optimisation algorithm approaches for feature selection problems. Expert Systems with Applications, 117, 267-286.

[7] Vu, L. T., Vu, L. T., Nguyen, N. T., Do, P. T. T., & Dao, D. P. (2019). Feature selection methods and sampling techniques to financial distress prediction for Vietnamese listed companies. Investment Management & Financial Innovations, 16(1), 276.

[8] Qian, S., & Singer, Y. (2019). Fast Parallel Algorithms for Feature Selection. arXiv preprint arXiv:1903.02656.

[9] Liao, T., Wang, G., Yang, B., Lee, R., Pister, K., Levine, S., & Calandra, R. (2019). Data-efficient Learning of Morphology and Controller for a Microrobot. arXiv preprint arXiv:1905.01334.

[10] Wang, X., Guo, B., Shen, Y., Zhou, C., &Duan, X. (2019). Input Feature Selection Method Based on Feature Set Equivalence and Mutual Information Gain Maximization. IEEE Access, 7, 151525-151538.

[11] Arafat, M., Hoque, S., Xu, S., &Farid, D. M. (2019). Machine Learning for Mining Imbalanced Data. IAENG International Journal of Computer Science, 46(2).

[12] Zhu, X., Jin, X., Jia, D., Sun, N., & Wang, P. (2019). Application of Data Mining in an Intelligent Early Warning System for Rock Bursts. Processes, 7(2), 55.

[13]Alelyani, S., Tang, J., & Liu, H. (2018). Feature selection for clustering: A review. In Data Clustering (pp. 29-60).Chapman and Hall/CRC.

[14]Faris, H., Mafarja, M. M., Heidari, A. A., Aljarah, I., Ala'M, A. Z., Mirjalili, S., & Fujita, H. (2018). An efficient binary salp swarm algorithm with crossover scheme for feature selection problems. Knowledge-Based Systems, 154, 43-67.

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

[15]Sangaiya, I., & Kumar, A. V. A. (2019). A Hybrid Feature Selection Method for Effective Data Classification in Data Mining Applications. International Journal of Grid and High Performance Computing (IJGHPC), 11(1), 1-16.

[16] Bui, Q. T., Pham, M. V., Nguyen, Q. H., Nguyen, L. X., & Pham, H. M. (2019). Whale Optimization Algorithm and Adaptive Neuro-Fuzzy Inference System: a hybrid method for feature selection and land pattern classification. International Journal of Remote Sensing, 40(13), 5078-5093.

[17]Saqib, P., Qamar, U., Aslam, A., & Ahmad, A. (2019, July). Hybrid of Filters and Genetic Algorithm-Random Forests Based Wrapper Approach for Feature Selection and Prediction. In Intelligent Computing-Proceedings of the Computing Conference (pp. 190-199).Springer, Cham.

[18]Chiew, K. L., Tan, C. L., Wong, K., Yong, K. S., &Tiong, W. K. (2019). A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Information Sciences, 484, 153-166.

[19]Drotár, P., Gazda, M., &Vokorokos, L. (2019). Ensemble feature selection using election methods and ranker clustering. Information Sciences, 480, 365-380.

[20] Suresh, S., & Narayanan, A. (2019, February). Improving Classification Accuracy Using Combined Filter+ Wrapper Feature Selection Technique. In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-6).IEEE.