IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-11 E-ISSN: 2347-2693

Review On Feature Selection Techniques in Data Mining

S. Ramadass[1*], M.Gunasekaran[2]

[1*]Department of Computer Science, Government Arts College, Dharmapuri, India
[2]Department of Computer Science, Government Arts College, Dharmapuri, India

[*]*Corresponding Author: ramadass77@gmail.com*

Available online at: www.ijcseonline.org

Received: 02/Oct/2017, Revised: 16/Oct/2017, Accepted: 10/Nov/2017, Published: 30/Nov/2017

Abstract- Feature selection is a data pre-processing technique specially used for classification problems. It aims at identifying the minimal reduct with less number of features without affecting the classification accuracy of the data set. Its goal is to choose a negligible subset of features as indicated by some sensible criteria with the goal that the first undertaking can be accomplished similarly well, if worse. By picking an insignificant subset of features, unimportant and repetitive features are evacuated by the paradigm. Rough set theory is a technique that has been used for feature selection. It is utilizing to find the basic relationship from the uproarious data, which is utilizing the discritization strategy on discrete-esteemed properties and proceeds with values quality. It depends on making the equalance classes with in the given data, every one of the data tupels are making an equalance classes are indiscernalbe with the regard of the properties depicting data. Though there is many rough set based approaches like quick reduct, relative reduct entropy based reduct, these approaches are able to identify a reduct set. This paper presents a survey on various methods and techniques of feature selection and its advantages and disadvantages.

Keywords- Feature selection, PSO, ACO, GA, Data mining

I. INTRODUCTION

Data mining refers to 'mining' knowledge from large amount of data**.** The need for computational theories and techniques to assist in the extraction of useful knowledge from digital data gave rise to Knowledge Discovery in Database field (KDD). The data is modified in several ways in order to reduce its dimensionality to minimize the loss of relevant information. KDD relies heavily on known techniques from machine learning, pattern recognition and statistics to find patterns. Two different goals can be achieved by using data mining techniques-verification and discovery. The first part is from a user hypothesis needed to be checked out, while the second aims at the finding of new patterns in the data. In the biological data, discovery-oriented data mining is mostly applied and verification-oriented techniques are usually applied for validation procedures. Knowledge discovery processes explained by the following steps and figured in fig1.1

- Data integration
- Data Selection
- Data Cleaning
- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge Presentation

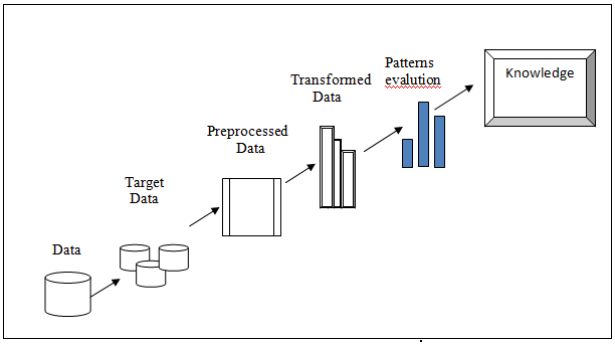

Figure 1.1: Data mining steps

Feature selection may be viewed in terms of the identification and selection of a subset of features from an original set of features forming patterns in a given dataset. The subset should be `necessary and sufficient' to describe target concepts while retaining suitably high accuracy in the representation of the original features. The selection of relevant features with the

Elimination of irrelevant features can: (1) improve classification accuracy, (2) reduce the learning period [1]. The objectives of feature selection include building easier and additional comprehensible models rising data processing performance and making ready clean comprehendible knowledge. Feature selection process explained in fig 1.2

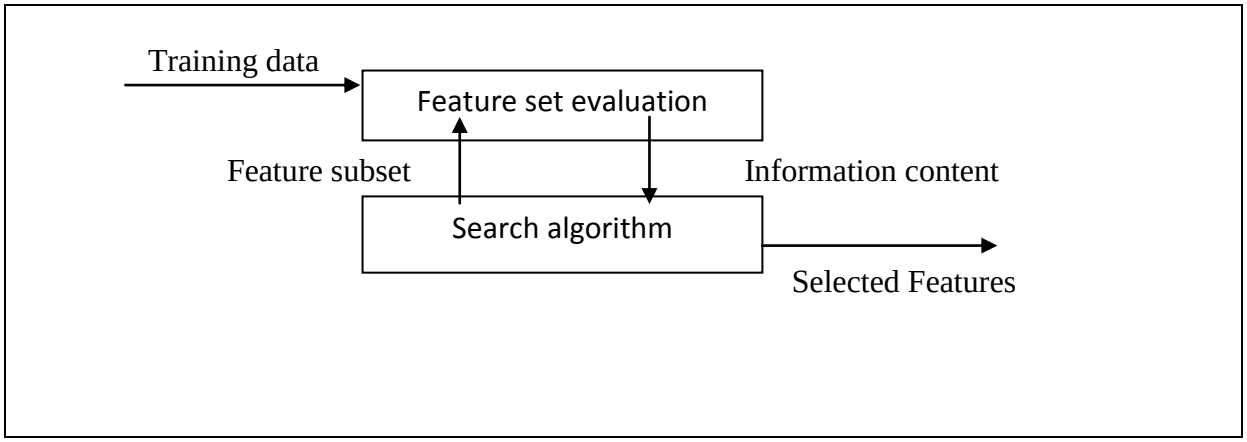

Figure 1.2: feature selection process diagram

In this paper section I contain the introduction of data mining and feature selection, Section II contain the literature survey on feature selection techniques, Section III contain the comparative study on feature selection techniques and section IV concludes the result of this paper.

II. LITERATURE SURVEY ON FEATURE SELECTION TECHNIQUES

Hannah Inbarani et.al., have proposed a new supervised feature selection using hybridization of Particle Swarm Optimization (PSO), PSO based Relative Reduct (PSO-RR) and PSO based Quick Reduct (PSO-QR) are introduced for the diagnosis of the diseases[2]. PSO-QR algorithm endeavours to compute a reduct without comprehensively creating every single conceivable subset. PSO-RR in view of a relative reliance measure was exhibited. The strategy was at first proposed to keep away from the count of discernability capacities or positive districts, which can be computationally costly without enhancements. The proposed Supervised PSO-RR algorithm ascertains a reduct set without producing all conceivable subset. It begins by choosing arbitrary esteems for every molecule and speed. A populace of particles is developed with arbitrary positions and speeds on S measurements in the issue space. The proposed techniques are thought about and demonstrated that the proposed algorithms expanded the prescient precision than existing unpleasant set based directed algorithms.

Jiye Lianget et.al., have proposed Group Incremental Approach to Feature Selection Applying Rough Set Technique (GIARC) to select useful features from a dynamically increasing data set[3]. The algorithm aims to find the new feature subset in a much shorter time when multiple objects are added to a decision table. The incremental algorithm for single object needs to be reperformed repeatedly to deal with multiple objects and it obviously gives rise to much waste of computational time. To overcome this deficiency, this section introduces a group incremental feature selection algorithm, which aims to deal with multiple objects at a time instead of repeatedly. Many real data in databases are generated as groups and Compared with existing incremental feature selection algorithms. The experiment results shown that GIARC can find a feasible feature subset of a dynamically increasing data set in a much shorter time and more efficient when multiple objects are added to a data set.

Xiaohui Lin et.al., have proposed Feature Selection Algorithm Based on Feature Overlapping and Group Overlapping(FS-FOGO) to calculate the feature importance[4]. FS-FOGO weighs features based on two aspects that overlapping degree based on the ratio of overlapping area on the effective range of each class and the overlapping degree based on the proportion of heterogeneous samples in every sample's nearest neighbours. The validation of FS-FOGO is compared with effective range based gene selection (ERGS), which calculates the feature weights based on overlapping area of the effective range, on six public biological data sets and one serum metabolomics data set about liver disease. Naive Bayes and Support Vector Machine are used as classifiers. The experiment results shown that the top ranked features by FS-FOGO are more discriminative and get higher classification accuracy rates than those by ERGS in most cases.

Guoqing Cui et.al., have proposed a new unsupervised feature selection algorithm via sparse representation (UFSSR), with respect to efficiency and effectiveness[5]. UFSSR be robust and Independent of domain knowledge based on data matrix which reconstructed via sparse representation. To reduce the reconstruction error, a new feature evaluation function is given to rank all features. When we reconstruct the entire dataset the computation cost will be increased. So, small partial data matrix reconstructed and checked is it trustworthy reconstructed part by cross validation. The experiment results shown UFFSR can obviously reduce the computation load while compared with other sparse representation based algorithms and UFSSR can select the important features easily and even get a better

performance than many classic algorithms such as max Variance, Laplacian Score and MCFS.

Hong Wang et.al., have proposed a Bacterial-Inspired Feature Selection Algorithm (BIFS) [6]. Searching process of bacteria using two main mechanisms: interactive swimming (or running) strategy used in Bacterial Colony Optimization (BCO), and random tumbling strategy embedded in Bacterial Foraging Optimization (BFO). BFO has some foraging strategies such as chemotaxis, swarming, reproduction, elimination and dispersal. BCO changes the circulation mode in BFO and employs the rule condition to control the excessive operations with the operation of reproduction, elimination and dispersal circling within the process of chemotaxis. In BCO chemotaxis has tow process that running process and tumbling process. The results shown that the proposed bacterial-inspired algorithm is capable of selecting the most sensitive sensors to detect and isolate the fault of complex structures.

Hossam et.al., have proposed a feature selection algorithm based on moth-flame optimization (MFO) [7]. MFO is implemented based on how a moth flies and maintaining a fixed angle with respect to the moon. When moths see a human-made artificial light, they try to maintain a similar angle with the light to fly in straight line. MFO is applied based on wrapper-based manner for feature selection using classification performance as fitness function. The proposed algorithm is compared to particle swarm optimization (PSO) and genetic algorithms (GA) using different evaluation criteria on 18 different data sets from UCI machine learning repository. The experiment results proved that the capability of MFO to adaptively search the feature space to find optimal feature combination and maximizing classification accuracy and the performance of MFO is significantly better than GA and PSO which are the common wrapper based feature selection.

Qian guo et.al., have proposed an Invasive Weed Optimization based Fuzzy-rough feature selection (IWO-FRFS) method for mammographic risk assessment which is used to early diagnosis of breast cancer[8]. The advantage of IWO is that the offspring individuals are randomly spread around their parents according to a Gaussian distribution during the evolution process. Such Gaussian distribution is designated with a dynamical standard deviation. The mechanism of IWO ensures a global optimal solution for the heuristic search. The performance of IWO is compared against the feature selection methods with ant colony optimization (ACO) and particle swarm optimization (PSO).

Sun jiongjiong et.al., have proposed feature selection algorithm based on Support Vector Machine (SVM) and Sequential Floating Forward Selection (SFFS) [9]. SFFS includes forward selection and backtrack steps to avoid local optimal solution and the judgement based on classification accuracy of target images. SFFS and LIBSVM classifiers were used to find an optimal feature subset. SVM classifier obtains classification accuracy quickly and accurately by training and testing target images. To increase the classification accuracy, cross- validation used during classification process.

Chunyong yin et.al., have proposed a hybrid feature selection algorithm for efficient intrusion detection [10]. The optimal feature subset chosen by combining the correlation algorithm and redundancy algorithm. Experimental results shown that the algorithm proved almost and an even better than the traditional feature selection algorithm on the different classifiers.

Kilho shin et.al., have developed two accurate and extremely fast algorithms, namely Super CWC and Super LCC [11]. Super CWC and Super LCC, which further improve the run-time performance of CWC and LCC by replacing linear search (LS) with binary search (BS). In this paper, Super CWC and Super LCC, those have excellent scalability to apply to big data analysis and exhibited excellent accuracy in the literature and do not harm the accuracy performance of the original algorithms.

III. COMPARITIVE STUDY ON FEATURE SELECTION TECHNIQUES

Table 3.1: comparative study on feature selection techniques

S.NO	AUTHORS	DATA SET	PROPOSED ALGORITHM	COMPARED ALGORITHM	RESULT
1	H. Hannah Inbarania, Ahmad Taher Azarb, G. Jothic	erythemato-squamous diseases, Breast Tissue, Prognostic, SPECTF	PSO based Relative Reduct (PSO-RR) and PSO based Quick Reduct (PSO-QR).	Supervised Quick reduct and Supervised Relative Reduct.	PSO based approach enables us to obtain the same or an even lesser number of reducts than the existing algorithm.
2	Jiye Liang, Feng Wang, Chuangyin Dang, Yuhua Qian	Breast-cancer-wisconsin(cancer), Tic-tac-toe, Kr-vs-kp, Letter-recognition(Letter), Krkopt, Shuttle, Person Activity(PA), Poker-hand	a Group Incremental Rough Feature Selection Algorithm	CAR, Classic heuristic feature selection algorithms	I. Find a feasible feature subset of a dynamically-increasing data set in a much shorter time. II. more efficient than

					existing Incremental feature selection algorithms while multiple objects are added to a data set.
3	Xiaohui Lin, Huanhuan Song, Meng Fan, Weijie Ren, Lishuang Li, Weihong Yao	Breast.2, Leukemia, Prostate, Brain_data, Leukemia2_Gems, Srbct	Feature Overlapping and Group Overlapping (FS-FOGO)	effective range based gene selection (ERGS)	FS-FOGO can measure the features more accurately and define the more informative Features.
4	Guoqing Cui, Jie Yang, Masoumeh Zareapoor, Jiechen Wang	Wine, Glass, Heart, Ionosphere, Musk	Unsupervised Feature Selection Algorithm via Sparse Representation (UFSSR)	Max Variance, Laplacian Score, MCFS	UFFSR can obviously reduce the computation load.
5	Hong Wang, Xingjian Jing, Ben Niu	9_Tumors, 11_Tumors, 14_Tumors, Brain_Tumor1, Brain_Tumor2,SRBCT	Bacterial-Inspired Feature Selection Algorithm (BIFS)	BCO, BFO, BFO-LDC, BFO-NDC, IBFS, DEFS, IG-GA, and IBPSO	Classification Accuracy of BIFS is higher than BCO and BFO.
6	Hossam M. Zawbaa, E. Emary, B. PARV , Marwa Sharawi	Breast cancer, Exactly, Exactly2, Lymphography, M-of-n, Tic-tac-toe, Vote, zoo, WineEW, SpectEW, SonarEW, PenglungEW, lonoshereEW, HeartEW, CongressEW, BreastEW, KrvskpEW, WaveformEW	Moth-Flame Optimization (MFO)	PSO, GA	proving the capability of MFO to adaptively search the feature space for optimal feature combination and its ability to avoid premature convergence than PSO and GA.
7	Qian Guo, Yanpeng Qu, Ansheng Deng, Longzhi Yang	Breast cancer dataset	Invasive Weed Optimization based Fuzzy-Rough Feature Selection (IWO-FRFS)	PSO-FRFS, ACO-FRFS	IWO-FRFS's Reduct size is less than PSO-FRFS and ACO-FRFS.
8	Sun jiongjiong,liu jun, wei xuguang	2000 sample images and 74 dimensional eigenvector	SVM based feature selection algorithm	SFFS	SVM classifier is better in efficiency and accuracy than SFFS.
9	Chunyong Yin, Luyu Ma, Lu Feng, Jin Wang, Zhichao Yin	CUP99 dataset	Hybrid Feature Selection Algorithm	Different classifiers that Naïve Bayes, Lvq1, J48, PART, Clonal	The detection precision is not lower too much, and some classifiers, the precision of the subset is superior to that of the original data set.
10	Kilho Shin, Tetsuji Kuboyama, Takako Hashimoto	ADS, ADA, ARCENE, CYLINDER, DEXTER, GINA, GISETTE, HIVA, KR-VS-KP, MADELON, MUSHROOM, NOVA, SPLICE, SYLVA	Super-CWC and Super-LCC	FRFS, CFS and FCBF, and Relief-F	Super CWC and Super LCC, that have excellent scalability to apply to big data analysis.

IV. CONCLUSION

Feature selection is an important technique for filtering necessary attributes from the data set. In this paper we have done a survey on feature selection techniques. In this survey shown that there is two techniques have a better performance. First one is SVM, It is used to avoid local optimal solution to find the better classification accuracy based on forward selection and backtrack steps. Second one is BIFS is used to select the most sensitive sensors to detect and isolate the fault of complex structures by the use of BFO and its foraging strategies such as chemotaxis, swarming, reproduction, elimination and dispersal. This Survey shown that SVM Classifier and BIFS are better and higher in efficiency and accuracy than the other algorithms.

REFRENCES

[1]. Bin Hu, Yongqiang Dai, Yun Su, Philip Moore, Xiaowei Zhang, Chengsheng Mao, Jing Chen,Lixin Xu *"Feature Selection for Optimized High-dimensional Biomedical Data using an Improved Shuffled Frog Leaping Algorithm*" IEEE 1545-5963 ©2016.

[2]. Jiye Liang, Feng Wang, Chuangyin Dang, Yuhua Qian " *A Group Incremental Approach to Feature Selection Applying Rough Set Technique*" IEEE Transactions on knowledge and data engineering, vol. 26, No. 2 , February ©2014.
[3]. Xiaohui Lin, Huanhuan Song, Meng Fan, Weijie Ren, Lishuang Li, Weihong Yao "*The Feature Selection Algorithm Based on Feature Overlapping and Group*" IEEE International Conference on Bioinformatics and Biomedicine(BIBM) ©2016.
[4]. Guoqing Cui, Jie Yang, Masoumeh Zareapoor, Jiechen Wang "*Unsupervised Feature Selection Algorithm Based on Sparse Representation*" The 2016 3rd International Conference on System and Informatics (ICSAI 2016).
[5]. Hong Wang, Xingjian Jing, Ben Niu "*Bacterial - Inspired Feature Selection Algorithm and Its Application in Fault Diagnosis of Complex Structures*" IEEE © 2016
[6]. Hossam M. Zawbaa, E. Emary, B. PARV , Marwa Sharawi "*Feature Selection Approach based on Moth-Flame Optimization Algorithm*" IEEE Congress on Evolutionary Computation (CEC) ©2016
[7]. Qian Guo, Yanpeng Qu, Ansheng Deng, Longzhi Yang "*A New Fuzzy-rough Feature Selection Algorithm for Mammographic Risk Analysis*" 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) 2016
[8]. Sun jiongjiong,liu jun, wei xuguang "*Feature Selection algorithm based on SVM*" 35th Chinese Control Conference July 27-29, 2016
[9]. Chunyong Yin, Luyu Ma, Lu Feng, Jin Wang, Zhichao Yin "*A Hybrid Feature Selection Algorithm*" 4th International Conference on Advanced Information Technology and Sensor Application 2015
[10]. Kilho Shin, Tetsuji Kuboyama, Takako Hashimoto "*Super-CWC and Super-LCC: Super Fast Feature Selection Algorithms*" IEEE International Conference on Big Data(Big Data) 2015
[11]. H. Hannah Inbarania, Ahmad Taher Azarb, G. Jothic "*Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis*" computer methods and programs in biomedicine 113(2014) 175 – 185.

Authors Profile

Mr. S. Ramadass pursed Bachelor of Computer Applications from Avs College of Arts and Science, Salem in 2013 and Master of Computer Applications from Sri Ramakrishna Engineering College, Coimbatore in year 2015. He is currently pursuing M.Phil. Computer Science in Government Arts college, Dharmapuri since 2016. His main research work focuses on Data Mining.

Mr M Gunasekaran pursed M.C.A at Bharthiar University, India in the year 2001. He received M.Phil (Computer Science) at Alagappa University, India in the year 2003. He had done his Ph.D. (Computer Science) at Anna University, Chennai, INDIA in the year 2013. He also received B.E.S (Electroics Science) at University of Madras, India in the year 1996. He has enriched himself by attending number of National and state level seminars, workshops, conferences, symposia and FDPs. He is currently working as Assistant Professor in Computer Science Department at Government Arts and Science College, Dharmapuri from June 2017. Previously, he had worked as Assistant Professor in Computer Science Department at Government Arts and Science College, Hosur from July 2015 to May 2017, he has worked as Associate Professor in the Department of Computer Applications at Dayanada Sagar Academy of Technology & Management, Bangalore-82 from August 2014 to July 2015 and he also worked at Park College of Engineering and Technology, Coimbatore, Tamilnadu, India from June 2001 to August 2014. He served as a NSS Programme Officer during the period 2002 to 2009. He has got 15 years of Teaching and 8 years of Research experience. His research interest includes applications of fuzzy logic, artificial neural network, immune system, stock market forecasting, and pattern recognition system. He has published more than 25 research papers in variuous National and International Journals and Conferences to his credit. He has been the resource person and also chaired various sessions in National and International Conferences. He is a life time member of the various professional bodies like ISTE,CSTA, UACEE, IACSIT, IAENG, etc. He has been a reviewer and Editorial board member of various International journals.

