A Review of Feature Selection Algorithms for Data Mining Techniques

K.Sutha

Research Scholar, Bharathiar University,
Coimbatore, Tamil Nadu, India
suthas@rediffmail.com

Dr.J. Jebamalar Tamilselvi

Director, Department of MCA,
Jaya Engineering College,
Chennai, Tamil Nadu, India
jjebamalar@gmail.com

Abstract

Feature selection is a pre-processing step, used to improve the mining performance by reducing data dimensionality. Even though there exists a number of feature selection algorithms, still it is an active research area in data mining, machine learning and pattern recognition communities. Many feature selection algorithms confront severe challenges in terms of effectiveness and efficiency, because of recent increase in data dimensionality (data with thousands of features or attributes or variables). This paper analyses some existing popular feature selection algorithms, addresses the strengths and challenges of those algorithms.

Keywords- feature selection; Data mining; filter; wrapper; hybrid

I. INTRODUCTION

In recent years, data collected for various research purposes are much larger. Such data set may consists of thousands of instances (records) and each of which may be represented by hundreds or thousands of features (attributes or variables) [1]. High dimensional data set is the data which contains extremely large number of features. DOROTHEA [2] is such a dataset used for drug discovery, consists of 1,950 instances and 100,000 features. Many of the features in such data set contain useful information for understanding the data, relevant to the problem, but it also contain large amount of irrelevant features, and redundant relevant features. This reduces the learning performance and computational efficiency [1]. To avoid this problem, a pre-processing step called “Feature Selection” is used to reduce the dimensionality before applying any data mining techniques such as Classification, association rules, clustering and regression.

The aim of feature selection is to determine a feature subset as small as possible. It is the essential pre-processing step prior to applying data mining tasks. It selects the subset of original features, without any loss of useful information. It removes irrelevant and redundant features for reducing data dimensionality. As a result it improves the mining accuracy, reduces the computation time and enhances result comprehensibility [3]. On applying mining tasks to the reduced feature subset produces, the same result as with original high- dimensional dataset. Feature selection offers advantages such as reducing storage requirements, avoiding over fitting, facilitating data visualization, speeding up the execution of mining algorithms and reducing training times [4].

This paper discusses about the techniques used by a collection of feature selection algorithms, compares their advantages and disadvantages, and helps to understand the existing challenges and issues in this research field.

The remainder of the paper is organized as follows, In section 2, fundamentals of feature selection is discussed. Existing feature selection algorithms are compared in section 3. Section 4 concludes our work.

II. FEATURE SELECTION PROCESS

The four key steps of a Feature selection process are feature subset generation, subset evaluation, stopping criterion and result validation. The feature subset generation is a heuristic search process which results in the selection of a candidate subset for evaluation. It uses searching strategies like complete, sequential and random search to generate subsets of features. Dunne et al. [5] stated that these searching strategies are based on step-wise addition or deletion of features.

The goodness of the generated subset is evaluated using an evaluation criterion. If the newly generated subset is better than the previous subset, it replaces the previous subset with the best subset. These two processes are repeated until the stopping criterion is reached. The final best feature subset is then validated by prior knowledge or using different tests. Fig.1 illustrates the feature selection process.

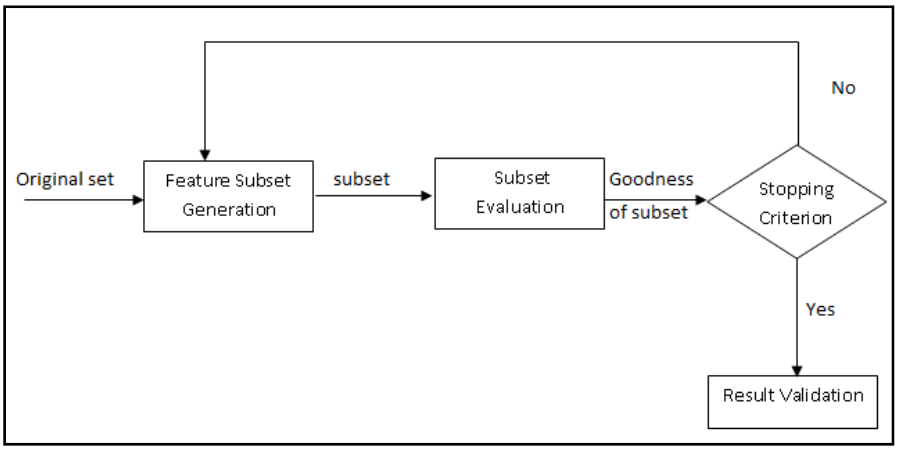

Figure 1. Feature Selection Process.

III. FEATURE SELECTION ALGORITHM

On the basis of selection strategy, feature selection algorithms are broadly classified into three categories namely Filter, Wrapper and Hybrid Method [6]. Filter Method selects the feature subset on the basis of intrinsic characteristics of the data, independent of mining algorithm. It can be applied to data with high dimensionality. The advantages of Filter method are its generality and high computation efficiency.

Wrapper Method requires a predetermined algorithm to determine the best feature subset. Predictive accuracy of the algorithm is used for evaluation. This method guarantees better results, but it is computationally expensive for large dataset. For this reason, the Wrapper method is not usually preferred [7].

Hybrid Method combines Filter and Wrapper to achieve the advantages of both the methods. It uses an independent measure and a mining algorithm to measure the goodness of newly generated subset [21]. In this approach, Filter method is first applied to reduce the search space and then a wrapper model is applied to obtain the best feature subset [8]. Fig. 2 illustrates the hybrid model.

Feature selection requires training data for learning purposes. The training data can be either labeled or unlabeled. From the perspective of utilizing label information, feature selection algorithms are classified into supervised, unsupervised and semi-supervised algorithms [9]. Supervised feature selection uses labeled data for learning purposes whereas Unsupervised feature selection uses unlabeled data.

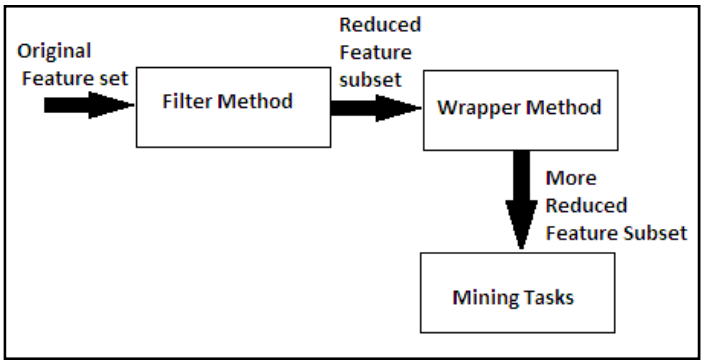

Figure 2. Hybrid Model

In supervised learning, a large amount of labeled data is required to achieve better feature selection performance [10]. If the amount of labeled data is limited, supervised learning suffers because the minimum amount of information needed to ensure that the relationships are established between the target concepts and the class labels is not available. This problem is referred as “small labeled sample problem”[22]. The Semi-supervised learning is a new concept that increases the amount of labeled data by predicting the class labels of unlabeled data which increases the learning performance

IV. COMPARISON OF FEATURE SELECTION ALGORITHMS

Feature selection is the essential preprocessing step in Data mining. Several feature selection algorithms are available. Each algorithm has its own strength and weakness. Table 1 compares some of the available algorithms.

TABLE I. COMPARISON OF SOME EXISTING FEATURE SELECTION ALGORITHMS

Algorithm	Type	Factors/ Approaches Used	Benefit	Drawback
Relief [11]	Filter	Relevance Evaluation	It is scalable to data set with increasing dimensionality.	It cannot eliminate the redundant features.
Correlation- based Feature Selection [12]	Filter	Uses Symmetric Uncertainty (for calculating Feature-Class and Feature-Feature correlation)	It handles both irrelevant and redundant features and It prevents the re-introduction of redundant features.	It works well on smaller datasets It cannot handle numeric class problems.
Fast Correlation Based Filter [6]	Filter	uses predominant correlation as a goodness measure, based on symmetric uncertainty(SU).	It hugely reduce the dimensionality	It cannot handle feature redundancy.
Interact [13]	Filter	Uses symmetric uncertainty and Backward Elimination Approach	It improves the accuracy.	Its mining performance decreases, as the dimensionality increases.

TABLE I. COMPARISON OF SOME EXISTING FEATURE SELECTION ALGORITHMS (CONTINUED)

Algorithm	Type	Factors/Approaches Used	Benefit	Drawback
Fast Clustering-Based Feature Subset Selection [8]	Filter	Graph-Theoretic Clustering Method used for clustering and a best feature is chosen from each cluster.	Dimensionality is hugely reduced	Works well only for Microarray data.
Condition Dynamic Mutual Information Feature Selection [14]	Filter	Mutual Information	Better Performance	Sensitive to noise

Affinity Propagation – Sequential Feature Selection [15]	Wrapper	Affinity Propagation clustering algorithm applied to get the clusters SFS applied for each cluster to get the best subset	Faster than Sequential Feature Selection	Accuracy is not better than SFS
Evolutionary Local Selection Algorithm [16]	Wrapper	K-Means Algorithm used for clustering	Covers a large space of possible feature combinations	As the number of features increases, the cluster quality decreases.
Wrapper Based Feature Selection using SVM [17]	Wrapper	Sequential Forward Selection for feature selection SVM for evaluation	Better Accuracy and Faster Computation	
Two-Phase Feature Selection Approach [18]	Hybrid	(Filter) Artificial Neural Network Weight Analysis used to remove irrelevant features. (Wrapper) Genetic Algorithm used to remove redundant features	Handles both irrelevant and Redundant features. Improves Accuracy	
Hybrid Feature Selection [19]	Hybrid	(Filter) Mutual Information (Wrapper)Wrapper model based feature selection algorithm which uses Shepley value	Improves Accuracy	High Computation Cost for high dimensional data set

When selecting a feature subset, Filter method make use of all the available training data, Filter methods are much faster and better than wrappers. It can be applied to large datasets having many features [12]. But Filter Method is not always enough to obtain better accuracy [18]. On the other hand, Wrapper Method also selects best feature subsets but it has proven to have high computation cost when compared to Filter for large datasets [12]. Hybrid method is less computationally intensive than wrapper methods. Relief[11] removes irrelevant data using the nearest neighbour approach, but It does not consider redundant features, whereas CFS and FCBF considers the redundant features while selecting relevant features [8]. FCBF is a fast filter method. FAST algorithm eliminates irrelevant features as well as it also handles redundant features [20]. It works well with microarray data when compared with text and image data [6]. Interact [13] and HFS[19] algorithm improves mining accuracy but it is unable to scale up with the increasing dimensionality. CDMI[14] is noise-sensitive

V. CONCLUSION

Among the existing feature selection algorithms, Some algorithms involves only in the selection of relevant features without considering redundancy. Dimensionality increases unnecessarily because of redundant features and it also affects the learning performance. And some algorithms select relevant features without considering

the presence of noisy data. Presence of noisy data leads to poor learning performance and increases the computational time. Our study concludes that there is a need for an effective unified framework for feature selection which should involve in the selection of best feature subset without any redundant and noisy data. It should be applied for all types of data and it should also able to scale up with increasing dimensionality.

REFERENCES

[1] Kashif Javed, Haroon A.Babri and Mehreen Saeed, "Feature Selection based on Class-Dependent Densities for High Dimensional Binary Data", IEEE Transactions on Knowledge and Data Engineering, Vol 24, No 3, 2012 (www.computer.org/csdl/trans/tk/2012/03/ttk2012030465-abs.html)

[2] "Feature Selection Challenge by Neural Information Processing Systems Conference (NIPS),"http://www.nipsfsc.ecs.soton.ac.uk,2003

[3] H.Liu and H.Motoda, Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, 1998.

[4] Zilin Zeng, Hongjun Zhang, Rui Zhang, Youliang Zhang, "Hybrid Feature Selection Method based on Rough Conditional MutualInformation and Naïve Bayesian Classifier", Hindawi Publishing Corporation, ISRN Applied Mathematics, Vol 2014, Articla Id 382738,11 pages .

[5] K.Dunne, Cunningham and F.Azuaje, "Solution to instability problems with sequential wrapper-based approaches to feature selection", Journal Of Machine Learning Research,2002.

[6] Lei Yu, Huan Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Department of Computer Science & Engineering, Arizone State University, Tempe, AZ 85287-5406, USA, 2003

[7] A.Blum and P.Langley, "Selection of relevant features and examples in machine learning", Artificial Intelligence, vol 97, pp 245-271, 1997

[8] 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, Vol 25, No.1, 2013

[9] Z.Zhao, H.Liu, "On Similarity Preserving Feature Selection", IEEE Transactions on Knowledge and Data Engineering, Vol 25, no 3, 2013

[10] Yongkoo Han, Kisung Park and Young-koo Lee, "Confident Wrapper-type Semi-Supervised Feature Selection Using an Ensemble Classifier", IEEE, 2011

[11] K.Kira and L.A Rendell, "The Feature Selection Problem: Traditional methods and A New Algorithm," Proc. 10th National Conference Artificial Intelligence, pp.129-134, 1992.

[12] Mark A. Hall and Lloyd A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper", Proceedings of the Twelfth International FLAIRS Conference, 1999.

[13] Zheng Zhao and Huan Liu "Searching for Interacting Features" Department of Computer Science and Engineering,Arizona State University, 2007

[14] Wang Liping, "Feature Selection Algorithm Based On Conditional Dynamic Mutual Information", International Journal O Smart Sensing and Intelligent Systems", VOL. 8, NO. 1, 2015

[15] Kexin Zhu and Jian Yang, "A Cluster-Based Sequential Feature Selection Algorihm", IEEE, 2013

[16] Y.Kim, W.Street, and F.Menczer, "Feature Selection for Unsupervised Learning Via Evolutionary Search," Proc. Sixth ACM SIGKDD International Conference, Knowledge Discovery and Data Mining, pp 365 – 369, 2000

[17] Hwang, Young-Sup, "Wrapper-based Feature Selection Using Support Vector Machine", . Department of Computer Science and Engineering, Sun Moon University, Asan, Sunmoonro 221-70, Korea, Life Science Journal 2014;11 (7)

[18] B.M Vidhyavathi, " A New Approach to Feature Selection for Data Mining", International Journal of Computational Intelligence Research, ISSN 0973-1873 Vol.7 Number 3, pp 263 – 269, 2011

[19] Jihong Liu, "A Hybrid Feature Selection Algorithm for Data sets of thousands of Variables" IEEE, 2010

[20] L.Yu and H.Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy", Machine Learning Research, Vol. 10, no. 5, pp 1205 -1224,2004

[21] Huan Liu and Lei Yu, " Towards Integrating Feature Selection Algorithms for Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering, Vol.17 No.4 2005

[22] A.Jain and D.Zongker, "Feature Selection: Evaluation, Application and small sample performance", IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(2):153-158,1997

AUTHORS PROFILE

K.Sutha is a research scholar at Bharathiar University, Coimbatore, Tamilnadu. She received the MCA degree from Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India. Her area of interests includes Data Warehousing , Data Mining , and Big Data.

Dr. J. Jebamalar Tamilselvi received her Ph.D. in 2009 from the Department of Computer Applications at Karunya University, Coimbatore, INDIA. Her area of interest includes Data cleansing approaches, Data Extraction, Data Integration, DataWarehousing and Data Mining. She is a life Member of International Association of Engineers (IAENG), International Association of Computer Science and Information Technology (IACSIT), and the Society of Digital Information and Wireless Communications. Reviewer and Member of International Journal of Engineering Science and Technology (IJEST) Member and Convergence Information Technology (JCIT).