Literature Review of Feature Selection for Mining Tasks

Muhammad Shakil Pervez
Department of Computer Science and Engineering,
United International University, Bangladesh

Dewan Md. Farid
Department of Computer Science and Engineering,
United International University, Bangladesh

ABSTRACT
During past few decades, researchers worked on data preprocessing techniques for the datasets. Data preprocessing techniques are needed, where the data are prepared for mining. The performance of data mining algorithms in most cases depends on dataset quality, since low-quality training data may lead to the construction of overfitting or fragile classifiers. Also, scientists worked on data mining areas in both algorithms section and conceptions practice section. But for better results they always used the combined or embedded or hybrid approaches. Scientists used different classifiers in different ways and also got their smoother results by arranging some modification in the algorithms. In this paper we shall describe all possible areas of attribute selection and reduction techniques. Feature selection algorithms broadly fall into three categories: filter models, wrapper models and hybrid models. Practically, scientists do the tasks in two stages for obtaining accuracy and that is, they firstly select the features and then reduce the dimensionality of feature vectors with classifiers through learning. Some promising approaches are indicated here and particular concentration is dedicated to describe different methods from raw level to experts, so that in future one can get significant instruction for further analysis.

General Terms
Embedded; hybrid; filter; wrapper; classifiers;

1. INTRODUCTION
Data mining is the process of finding hidden information and patterns from a huge database. Data mining algorithms have two major functions: classification and clustering. Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. Classification creates a function from training data. The training data consist of pairs of input objects, and desired output. The output of the function can be a continuous value, or can predict a class label of the input object. The task of the classification is to predict the value of the function for any valid input object after having seen only a small number of training examples. Clustering is similar to classification except that the groups are not predefined, but rather defined by the data alone. Clustering is alternatively referred to as unsupervised learning or segmentation. It can be thought of as patitioning or segmenting the data into groups that might or might not be disjointed. The clustering is usually accomplished by determining the similarity among the data on predefined attributes. The most similar data are grouped into clusters. Feature selection is a form of search in the training data. It selects a subset of input features d from a total of D original input features in the training data by using an optimisation of scientific theorem to improve the classification accuracy of a learning classifier [8].

The rest of the paper is organized as follows. Section 2 briefly describes about feature selection techniques used by data scientists. Section 3 presents a comprehensive literature review about different procedures of feature selection. At a glance comparison among the various techniques is depicted in section 4. Section 5 turns conclusion with a brief about this paper.

2. FEATURE SELECTION METHODS
Data mining is much matured in this modern age, as the application of this is noticeable in various regions such as robotics, machine learning and knowledge based system etc. Feature selection is a form of search in the training dataset, which involves the selection of a subset of features d from a total of D original input features from training dataset based on an optimisation principle to improve the performance of a learning classifier. It search through the subsets of features, and try to find the best one. In complex classification domains such as intrusion detection systems (IDS), feature selection is very important because irrelevant and redundant features may lead to complex classification model as well as reduce the classification accuracy. There are two main models that deal with feature selection: (a) filter methods, and (b) wrapper methods. Filter methods rely on the general characteristics of the training data to select features with independence of any learning classifier, which are usually computationally less expensive than the wrapper models, and have the ability to scale to large datasets. On the other side, wrapper methods involve optimising a learning classifier as part of the feature selection process. Wrapper models tend to give better results and the model is more precise than the filter model. However, wrapper models are very time consuming, which restricts application with some datasets[17].

The hybrid methods are based on a sequential approach where the first step is usually based on filter methods to reduce the number of features considered in the second stage. Afterwards, a wrapper method is employed to select the desired number of features using this reduced set in the second stage[2].

3. REVIEW ON FEATURE SELECTION AND REDUCTION
The voice over classification involve a large volume of data and/or a large number of features/attributes initiated around thirty years ago. Langley [18] grouped different feature selection methods into two broad groups (i.e., filter and wrapper) based on their dependence on the inductive algorithm that will finally use the selected subset. Filter methods are independent of the inductive algorithm, whereas wrapper methods use the inductive algorithm as the evaluation function.

Abdulla and Kasabov, 2003, [1] designed multi-streams paradigm where they split feature vectors in three independent continuous-density Hidden Markov Model(CHMM) frameworks. They proposed a technique that combines classifiers. Here the three HMM classifiers are applied to speech signals. HMM classifiers had done feature reduction

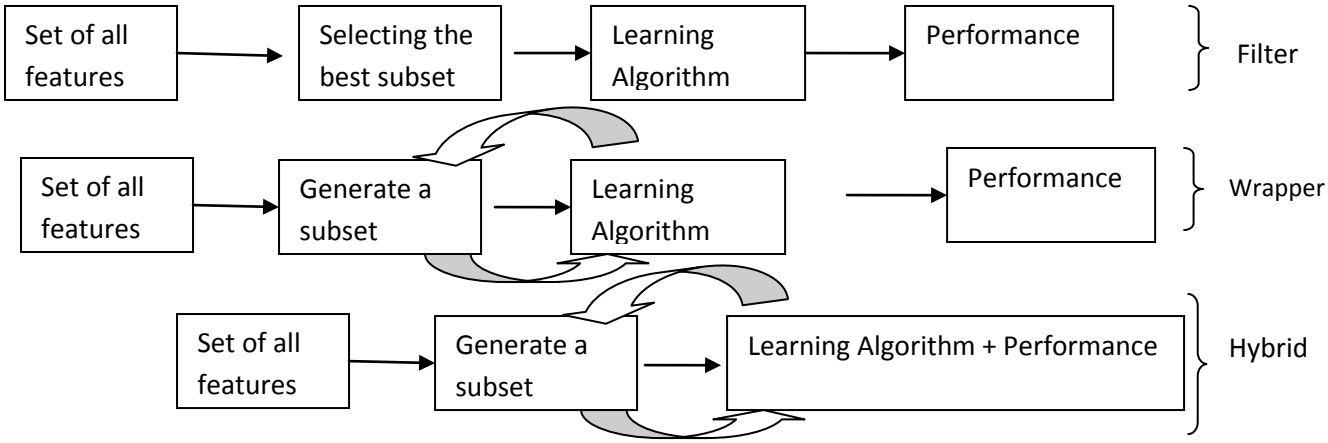

Figure 1. Different feature selection techniques

by alleviating the dominance effects of the features and in this way, they reduced the dimensionality of feature vectors.

Unler et al., 2010, [2] presented a hybrid filter–wrapper feature subset selection algorithm. The filter model is used for feature subset selection and the wrapper model was liable to use the mutual information available from the filter model. They presented a novel feature selection method, which lessen computational cost dramatically. The whole hybrid model performed feature selection and reduction.

Zhang et al., 2010, [3] redefined composite rough sets for feature selection which is a powerful mathematical tool for analyzing various types of data. Zhang et al. proposed an incremental method for dynamic data mining based on neighborhood rough sets. They accomplished a notable progress after twenty five years of the beginning of the research on feature selection using rough sets. Through rough sets they defined composite information systems that contained attributes of multiple different types, which was liable for feature selection and knowledge discovery.

Uguz, 2011, [4] worked on text categorization and here he did feature selection in two stages and all are filter methods. In the first stage, each term within the document is ranked depending on their importance for classification using the information gain (IG) method. In the next stage, genetic algorithm (GA) and principal component analysis (PCA) feature selection and feature extraction methods are applied separately to the terms which are ranked in decreasing order of importance, and a dimension reduction is carried out. Genetic algorithm is an optimization method mimicking the evolution mechanism of natural selection. GA performs a search in complex and large landscapes and provides near-optimal solutions for optimization problems.

Pacheco et al., 2013, [5] assessed the relevance of formulating the feature selection problem for classification to check and compare the efficacy with the method NSGAFS(non-dominated sorting genetic algorithm). A series of experiments was run with different databases. There they worked with financial variables and two classes: ‘‘credit-worthy’’ and ‘‘non-creditworthy’’.

Li et al., 2010, [6] worked on image annotation and for its feature selection parallel genetic algorithm was used. They investigated two methods of Genetic Algorithm feature selection.

Sun et al., 2013, [7] established a new framework of feature selection which not only selects the most relevant features and eliminates redundant features, but also tries to retain useful intrinsic feature groups. Unlike traditional frameworks of feature selection, its primary characteristic was that the features were weighted according to their interaction with the selected features. Moreover, the weight of features will be dynamically updated when each candidate feature had been selected. So, they proposed a dynamic weighting-based feature selection algorithm for ranking features based on information metric.

Carmona-Cejudo et al., 2013, [9] have focused on the comparison of several feature selection and adaptive strategies for email foldering using the Enron dataset and their proposed ABC-DynF framework for adaptation. Using naïve bayes classifier, this classification procedure was conducted. The ABC-DynF framework can work under a dynamic feature set, keeping a list of the top-k features which will be used by the learning model and allowing to add new categories using Naïve Bayes classifier.

Chen's, 2012, [10] explored an innovative method using rough sets. Their procedure utilized an integrated feature-selection approach to select 16 condition attributes of financial ratios, the CPDA to partition selected condition attributes by applying rough sets local-discretization cuts, and an rough sets LEM2 algorithm to generate decision rules set to identify important knowledge hidden in original data.

Xu et al. , 2011, [11] proposed a new dynamic attribute reduction algorithm based on a 0-1 integer programming to

deal with the dynamic data in this paper. Before employing rough sets theory, the real value data must be discrete, so K-means discrete method was employed in thier experiments.

Chen et al., 2009, [12] developed a Semantic Relationship Graph (SRG) to describe the relationship between multiple tables and guide the search within relation space. Afterwards, they optimize the Semantic Relationship Graph by avoiding undesirable joins between relations and eliminating unnecessary attributes and relations.

Mladenic and Grobelnik, 2003,[13] used classifier for each split in the text hierarchy for Feature selection on hierarchy of web documents. In the learning experiments, for each of the sub problems, naive Bayesian classifier was used on text data.

Bina et al., 2013 [14] stated that they prepared a wrapper classifier which predicted the class label. The relational Naive Bayes classifier exploits independence assumptions to achieve scalability. They introduce a weaker independence assumption to the effect that information from different data tables is independent given the class label.

Tsai and Hsiao, 2010, [15] invented a filter model to combine multiple feature selection methods to identify more representative variables for better prediction. In particular, three well-known feature selection methods, which are Principal Component Analysis (PCA), Genetic Algorithms (GA) and decision trees (CART), are used. The combination methods to filter out unrepresentative variables are based on union, intersection, and multi-intersection strategies. For the prediction model, the back-propagation neural network is developed.

Qaunz et al., 2012, [16] have explored a feature extraction perspective, starting with the popular sparse coding approach which learns a set of higher order features for the data. They presented a novel method, where they did not use any classifier. The sparse coding for knowledge transfer, they have proposed new feature generation algorithms to address those limitations and enable knowledge transfer, and verified the effectiveness of the approach on real and synthetic data.

4. COMPARISON AMONG THE TECHNIQUES

There are some attempts of using the classifiers for classification purposes but the computational complexity of the algorithms and the obtained results lead us to think that there is still much research to do in this field.

TABLE 1. Comparison among feature selection techniques

Researcher (s), Year, Reference	Description of feature selection techniques		
	Basis of procedure	***Associated Classifiers***	***Type***
Abdulla and Kasabov, 2003, [1]	continuous-density Hidden Markov Model(CHMM) frameworks	HMM	wrapper
Unler et al., 2010, [2]	hybrid filter–wrapper feature subset selection algorithm	SVM	Hybrid
Zhang et al., 2010, [3]	Through rough sets they defined composite information systems that contained attributes of multiple different types	rough sets	Wrapper
Uguz, 2011, [4]	feature selection in two stages and all are filter methods	No classifier	Filter
Pacheco et al., 2013, [5]	with the method NSGAFS((non-dominated sorting genetic algorithm)	No classifier	Filter
Li et al., 2010, [6]	for its feature selection parallel genetic algorithm was used	No classifier	Filter
Sun et al., 2013, [7]	a dynamic weighting-based feature selection algorithm	No classifier	Filter
Carmona-Cejudo et al., 2013, [9]	The ABC-DynF framework can work under a dynamic feature set	Naïve Bayes classifier	Wrapper
Chen, 2012, [10]	an integrated feature-selection approach	rough sets	Hybrid
Xu et al. , 2011, [11]	a new dynamic attribute reduction algorithm	k-means	Hybrid
Chen et al., 2009, [12]	developed a Semantic Relationship Graph (SRG)	No classifier	Filter
Mladenic and Grobelnik, 2003,[13]	naive Bayesian classifier was used on text data.	Naïve Bayes classifier	Wrapper
Bina et al., 2013 [14]	The relational Naive Bayes classifier exploits independence assumptions to achieve scalability	Naïve Bayes classifier	Wrapper
Tsai and Hsiao, 2010, [15]	Combination of Principal Component Analysis (PCA), Genetic Algorithms (GA) and decision trees (CART)	No classifier	Filter
Qaunz et al., 2012, [16]	the popular sparse coding approach	No classifier	Filter

5. CONCLUSION

Feature selection is a process that selects a subset from the original feature set according to some criteria of feature importance. In this paper, concepts of feature selection are reviewed that categorize different approaches in this ground. This literature review explore the recent trend in feature selection that comes from novice procedure to this time of computer, where data mining is used to classify. Almost all

the techniques found for feature selection have discussed here and there is hardly any research found yet to categorize with similarity measurement. We have faith that the study on feature selection is a productive region for further research. Around 15 papers have been discussed here and various key topics from other historical publication relevant with text summarization have been analyzed here from 1988 to 2014. There exist some other techniques similar with those described in this paper, the discussion of which has not been included here as it will be a large corpus. But it is expected that any researchers can get help from this literature review for better understanding of different types of procedure on feature selection. Anyone can also get direction for better perception of the diversified sorts of abstraction, which will help to construct new procedure for next generation.

6. ACKNOWLEDGMENTS

The authors wish to thank to the United International University faculty members for their valuable comments, constructive suggestion who have contributed towards development of the paper?

7. REFERENCES

[1] Waleed H. Abdulla, Nikola Kasabov, "Reduced feature-set based parallel CHMM speech recognition systems", Information Sciences, Vol. 156, 2003,pp. 21–38.

[2] Alper Unler, ,Alper Murat, Ratna Babu Chinnamb, "mr2PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification", Information Sciences, Vol. 181, 2011, pp. 4625–4641.

[3] Junbo Zhang, Tianrui Lia, Hongmei Chen, "Composite rough sets for dynamic data mining," Information Sciences", Vol. 4, 2013, pp. 129-135. http://dx.doi.org/10.1016/j.ins.2013.08.016

[4] Harun Uguz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm, " Knowledge-Based Systems, Vol. 24, 2011, pp. 1024–1032

[5] Joaquin Pacheco, Silvia Casado, Francisco Angel-Bello,Ada Álvarez, "Bi-objective feature selection for discriminant analysis in two-class classification", Knowledge-Based Systems, Vol. 44 ,2013,pp. 57–64

[6] Ran Li, Jianjiang Lu, Yafei Zhang, Tianzhong Zhao, "Dynamic Adaboost learning with feature selection based on parallel genetic algorithm for image annotation", Knowledge-Based Systems, Vol. 23 , 2010, pp. 195–201.

[7] Zhiming Zhang, "On interval type-2 rough fuzzy sets", Knowledge-Based Systems, Vol. 35, 2012, pp. 1–13

[8] Dewan Md. Farid, and Chowdhury Mofizur Rahman, "Mining Complex Data Streams: Discretization, Attribute Selection and classification," Journal of Advances in Information Technology, Vol. 4, No. 3, August 2013, pp. 129-135.

[9] José M. Carmona-Cejudo, Gladys Castillo , Manuel Baena-García , Rafael Morales-Bueno, "A comparative study on feature selection and adaptive strategies for email foldering using the ABC-DynF framework", Knowledge-Based Systems, Vol. 46, 2013, pp. 81–94

[10] You-Shyang Chen, "Classifying credit ratings for Asian banks using integrating feature selection and the CPDA-based rough sets approach," Knowledge-Based Systems, Vol. 26, 2012, pp. 259–270

[11] Yitian Xu, Laisheng Wang, Ruiyan Zhang, "A dynamic attribute reduction algorithm based on 0-1 integer programming", Knowledge-Based Systems, Vol. 24, 2011, pp. 1341–1347

[12] Hailiang Chen, Hongyan Liu, Jiawei Han, Xiaoxin Yin, Jun He, "Exploring ptimization of semantic relationship graph for multi-relational Bayesian classification", Decision Support Systems, Vol. 48, 2009, pp. 112–121

[13] Dunja Mladenic, Marko Grobelnik, "Feature selection on hierarchy of web documents", Decision Support Systems, Vol. 35, 2003, pp. 45– 87

[14] Bahareh Bina, Oliver Schulte, Branden Crawford, Zhensong Qian, Yi Xiong, "Simple decision forests for multi-relational classification", Decision Support Systems, Vol. 54, 2013), pp.1269–1279

[15] Chih-Fong Tsai, Yu-Chieh Hsiao, "Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches", Decision Support Systems,Vol. 50, 2010, pp. 258–269

[16] Brian Quanz, Meenakshi Mishra, "Knowledge Transfer with Low-Quality Data: A Feature Extraction Issue", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 24, NO. 10, OCTOBER 2012 , pp. 1789-1802

[17] Muhammad Shakil Pervez, and Dewan Md. Farid, "Feature Selection and Intrusion classification in NSL-KDD Cup 99 Dataset Employing SVMs," 8th Software, Knowledge, Information Management and Applications (SKIMA 2014), 18-20 Dec, 2014, Dhaka, Bangladesh, http://dx.doi.org/10.1109/SKIMA.2014.7083539.

[18] Langley, P., Selection of relevant features in machine learning. In: Proceedings of the AAAZ Fall Symposium on Relevance, l-5, 1994.