A Review Paper on Feature Selection Methodologies and Their Applications

Shweta Srivastava1, Nikita Joshi2, Madhvi Gaur3
Assistant Professor (CSE Department) ABES Engineering College, Ghaziabad, India.

Abstract
Feature selection is the process of eliminating features from the data set that are irrelevant with respect to the task to be performed. Feature selection is important for many reasons such as simplification, performance, computational efficiency and feature interpretability. It can be applied to both supervised and unsupervised learning methodologies. Such techniques are able to improving the efficiency of various machine learning algorithms and that of training as well. Feature selection speed up the run time of learning, improves data quality and data understanding.

Keywords
Feature Selection, Supervised, Search strategies, Unsupervised.

1. INTRODUCTION
Feature selection is used to satisfy the common goal of maximizing the accuracy of the classifier, minimizing the related measurement costs; improve accuracy by reducing irrelevant and possibly redundant features; reduce the complexity and the associated computational cost; and improve the probability that a solution will be comprehensible and realistic [13]. Feature selection is one of the stage for preprocessing the data to reduce the dimensionality [10]. It selects a subset of the existing features without any transformation. It can be said as a special case of feature extraction.

2. FEATURE SELECTION METHODOLOGIES
When we use feature selection, smaller number of features are extracted which means fewer model parameters. It improves the generalization capabilities and reduces complexity and execution time. Feature Selection methodologies can be categorized as: supervised and unsupervised feature selection methods.

Supervised Feature Selection Method: A feature selection method requires a search strategy to select the subset of attributes and an objective function to evaluate the selected subset of features.

Objective function is divided in three categories:
1. Filters: It consists of algorithms that are built in the adaptive systems for data analysis (predictors) [9]. They use an evaluation function that relies on properties of the data. Distance based and margin based criterion can be used for filters.

2. Wrapper: The algorithms of this method are wrapped around the adaptive systems providing them subsets of features and receiving their feedback (usually accuracy). These wrapper approaches are aimed at improving results of the specific predictors they work with [9]. It utilizes the classifier as black box to find the subset of features based on their predictive power [14].

3. Embedded: They are based on performance evaluation metric calculated directly from the data, without direct reference to the results of any data analysis systems.
3. **Embedded**: They are based on performance evaluation metric calculated directly from the data, without direct reference to the results of any data analysis systems.

**Heuristic search**: The number of subsets evaluated by BFF is less (even much less) than that needed by the branch and bound algorithm [18]. They require large computational time. Heuristic search is problematic when the data has highly correlated features. Heuristics help to reduce the number of alternatives from an exponential number to a polynomial number [19]. For example: Sequential forward selection, sequential backward elimination and bidirectional search.

**Forward Selection**: Forward selection considers the subset to be empty initially and keeps on adding one feature at a time until the best feature subset is obtained.

**Backward Elimination**: Backward selection takes complete set of features as input and keeps on removing one attribute at a time until the most appropriate subset of features is obtained.

**Randomized search strategy**: It performs randomized exploration of the search space where next direction is a sample from a given probability [20]. For example: genetic algorithm.

There are various challenges related to automated text classification such as:
1. An appropriate data structure is to be selected to represent the documents.
2. An appropriate objective function is to be chosen to optimize to avoid overfitting and obtain good generalization. Along with it algorithmic issues arising as a result of the high formal dimensionality of the data are to be dealt with [1].

**Genre classification**: Metadata such as filename, author, size, date, track length and genres are the common features used to classify and retrieve genre documents. On the basis of these data, the classification is infeasible, so the feature selection step is required. In case of genre classification feature selection is a process where a segment of an audio is characterized into a compact numerical representation [2]. Feature selection is done to reduce the dimensionality of the data as a preprocessing step prior to classification due to high dimensionality of the feature sets.

### Table 1: Comparison of filter, wrapper and embedded methods

<table>
<thead>
<tr>
<th>Search Strategy</th>
<th>Filter methods</th>
<th>Wrapper methods</th>
<th>Embedded methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exhaustive search</strong></td>
<td>all features are evaluated directly</td>
<td>search is restricted to small subset of features</td>
<td>search is restricted to small subset of features</td>
</tr>
<tr>
<td><strong>Heuristic search</strong></td>
<td>search is restricted to small subset of features</td>
<td>search is restricted to small subset of features</td>
<td>search is restricted to small subset of features</td>
</tr>
<tr>
<td><strong>Randomized search</strong></td>
<td>search is restricted to small subset of features</td>
<td>search is restricted to small subset of features</td>
<td>search is restricted to small subset of features</td>
</tr>
</tbody>
</table>

### Table 2: Comparison of Exhaustive, Heuristic and Randomized Search

<table>
<thead>
<tr>
<th>Search Strategy</th>
<th>Accuracy</th>
<th>Complexity</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exhaustive Search</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td>Heuristic Search</td>
<td>moderate</td>
<td>moderate</td>
<td>moderate</td>
<td>moderate</td>
</tr>
<tr>
<td>Randomized Search</td>
<td>low</td>
<td>low</td>
<td>low</td>
<td>low</td>
</tr>
</tbody>
</table>
Microarray data analysis:
1. Almost all bioinformatics problems have the number of features significantly larger than the number of samples (high feature to sample ratio datasets) [3]. For example: Breast cancer classification on the basis of microarray data. Though the information about all the genes is not required in case of such classification.
2. Content analysis and signal analysis in genomics also require feature selection.

Software defect prediction:
There are various software quality assurance attributes such as reliability, functionality, fault proneness, reusability, comprehensibility etc [6]. It is a critical issue to select most appropriate software metrics that likely to indicate fault proneness.

Sentiment analysis:
Sentiment analysis is capturing favorability using natural language processing. It is not just a topic based categorization. It deals with the computational treatment of opinion, sentiment, and subjectivity in text. It is useful in recommendation systems and question answering [8]. To decide about the positivity or negativity of the opinion on the basis of various features such as term presence, feature frequency, feature presence, term position, POS tags, syntax, topic and negation etc. All of the features are not required in each and every case. So feature selection need to be performed.

Stock market analysis:
There are hundreds of stock index futures. Along with it financial data including the stock market data is too complex to be searched easily [11]. In particular, the existence of large amount of continuous data may cause a challenging task to explicit concepts extraction from the raw data due to the huge amount of data space determined by continuous features [12]. So it is necessary to reduce the dimensionality of data and irrelevant factors before searching.

Image Retrieval
Feature selection is applied to content based image retrieval to allow efficient browsing, searching and retrieving [22]. Content based image retrieval is to index the images on the basis of their own visual contents (i.e. color, shape, texture etc.) instead of text based keyword indexing. The biggest problem for content based image retrieval is large amount of images in database [21].

IV. CONCLUSION
This paper provides a comprehensive overview of various characteristic of feature selection. Feature selection as a preprocessing step in very large databases collected from various applications. More work is required to overcome limitations imposed as it is costly to visit high dimensionality data multiple times or accessing instances at random times. In unsupervised feature selection, several clusters may exist in different subspaces of small-small dimensionality, with their sets of dimensions overlapped or non-overlapped. As most existing feature selection algorithms have quadratic or higher time complexity about N, so it is quite difficult to handle high dimensionality. Therefore, more efficient search strategies and evaluation criteria are needed for feature selection with large dimensionality.

References:


[19] Manoranjan Dash, Huan Liu; “Consistency based feature selection”; Published by Elsevier


Shweta Srivastava, birth place is Faizabad and date of birth is 2-July-1986. She has done schooling from Canossa Convent Girls Inter College, Faizabad, Uttar Pradesh, India, B. Tech. (IT) from JSS Academy of Technical Education, NOIDA, Uttar Pradesh Technical University, Uttar Pradesh, India (2008) and M. Tech. (CSE) from Jaypee Institute of Information Technology, NOIDA, JIITU, Uttar Pradesh, India (2011). She has three years of teaching experience. Currently she is working as an ASSISTANT PROFESSOR in CSE department of ABES Engineering College, Ghaziabad, Uttar Pradesh, India. She has worked as a LECTURER in JSSATE, NOIDA, Uttar Pradesh, India for one year. She has done two years TEACHING ASSISTANTSHIP in Jaypee Institute of Information Technology, JIITU, NOIDA, Uttar Pradesh, India during M. Tech.