Literature Review on Feature Selection Methods for High-Dimensional Data

AUTHOR: D. ASIR ANTONY GNANA SINGH, S. APPAVU ALIAS BALAMURUGAN, E. JEBAMALAR LEAVLINE

INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS (0975 – 8887) VOLUME 136 – NO.1, FEBRUARY 2016

Summary By: William Steimel

Motivation

- Research was written in 2016 which is fairly recent
- Cited 280 + times
- Relevant and related to my previous presentation on Machine Learning Basics, End to End Machine Learning, and Linear Regression
- Feature Selection is a topic I am trying to understand to improve my machine learning projects.

Abstract

- Feature Selection is very important for improving the performance of machine learning algorithms which is why researchers focus on it.
  - Reduces the time required to build the learning model
  - Increases accuracy in the learning process
- Therefore, Identifying the best feature selection method is very important when dealing with high-dimensional data
- This paper performs a complete literature review on various methods for feature selection

General Terms/Keywords Discussed in Research

General Terms

Literature review on feature selection methods, study on feature selection, wrapper-based feature selection, embedded-based feature selection, hybrid feature selection, filter-based feature selection, feature subset-based feature selection, feature ranking-based feature selection, attribute selection, dimensionality reduction, variable selection, survey on feature selection, feature selection for high-dimensional data, introduction to variable and feature selection, feature selection for classification.

Keywords

Introduction to variable and feature selection, information gain-based feature selection, gain ratio-based feature selection, symmetric uncertainty-based feature selection, subset-based feature selection, ranking-based feature selection, wrapper-based feature selection, embedded-based feature selection, filter-based feature selection, hybrid feature selection, selecting feature from high-dimensional data.

Introduction

- The Digital era poses some challenges to researchers as there are many data acquisition techniques, methods, and devices
- This leads to raw datasets that are massive and noisy at times which can degrade the performance of Machine Learning Algorithms
  - Leads to Overfitting of the data (A Major Machine Learning Challenge)
  - This is also known as High-Dimensionality - The High-Dimensional data often contains Irrelevant Features/Redundant Features
- These issues can be handled with feature selection
  - Feature Selection - The process of removing Irrelevant/Redundant features from a dataset to improve performance
  - Feature Selection is also known as variable or attribute selection
- Mostly feature selection is applied on supervised learning algorithms as they are more susceptible to high-dimensional data

Feature Selection

- The next section gives an overview on the Feature Selection Process which is part of the Data Pre-processing phase
- As quoted from the research paper and mentioned previously: "Feature selection is a process of removing the irrelevant and redundant features from a dataset in order to improve the performance of the machine learning algorithms in terms of accuracy and time to build the model. "
- There are two main methods for Feature Selection: (How the features are combined for evaluation in the feature selection)
  - Subset Selection
  - Feature Ranking Methods
- In addition, Feature Selection can be classified into 4 categories based on how the machine learning algorithm is utilized in the feature selection process
  - Wrapper
  - Embedded
  - Filter
  - Hybrid

Feature Selection Based on Combining the Features for Evaluation

Subset Selection and Feature Ranking

II. Feature Selection (Subset Selection)

- Subset Selection generates the possible number of combinations of feature subsets using any of the search strategies/approaches:
  - Greedy forward selection
  - Greedy backward elimination
  - Etc.
- Evaluates the individual feature subsets with a feature selection metric:
  - Correlation
  - Consistency
  - Etc.
- Requires more computational complexity due to the methods used for subset generation /evaluation

II. Feature Selection (Feature Ranking)

- Each feature is ranked by a selection metric and top ranked features are selected based on a pre-defined threshold criteria.
  - Information gain
  - Symmetric uncertainty
  - Gain ratio
  - Chi squared
  - Etc.
- Does not require as much computation complexity as subset selection but also does not deal with redundant features

Feature Selection Based on the Supervised Learning Algorithm Used

Wrapper, Embedded, Filter, and Hybrid

II. Feature Selection (Based on Supervised Learning Algorithm Application)

- Wrapper Approach:
  - Incorporates supervised learning algorithm for validating generalized feature subsets
  - High Classification accuracy only for the algorithm selected
  - Does not possess high generality and computational complexity is higher than embedded and filter methods
- Embedded Approach:
  - Uses part of a supervised learning algorithm in feature selection process
  - High classification accuracy only for the algorithm selected
  - Does not have High generality and more computationally expensive than filter less than wrapper.
- Filter Approach:
  - Selects features without influence of supervised learning algorithm and works with any classification algorithm
  - Has high generality and less computational complexity than wrapper and embedded methods.
- Hybrid Method:
  - Combines the Wrapper and Filter Approaches

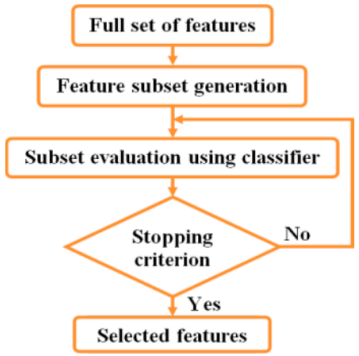

Figure 1 Feature selections with wrapper approach

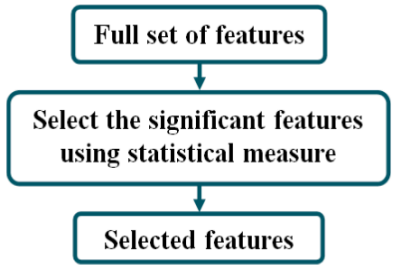

Figure 2 Feature selection with filter approach

Summary on Findings I

Subset Methods vs Feature Ranking

Feature Selection Based on Combining the Features for Evaluation

- Feature Subset Based Methods – (Searching Strategies)
  - Exhaustive/Complete Search
    - Exhaustive or Complete Search lead to the highest computational complexity
    - Exhaustive or Complete Search is considered a "Brute Force" method not suitable for high dimensional space
  - Heuristic Search [SA (simulated annealing),TS (tabu searching),ACO (ant colony optimization),GA (genetic algorithm),PSO (particle swarm optimization)]
    - Required more computational complexity because prior knowledge is required and each subset needs to develop a classification model to evaluate them
    - Heuristic Search follow a wrapper based approach which means these methods are computationally expensive and can only provide high classification to the specific algorithm

Feature Selection Based on Combining the Features for Evaluation

- Ranking based methods
  - Take less computation time and achieve higher generality since they do not use a supervised learning algorithm
  - They cannot remove redundant features as they only calculate correlations or similarities between features and the target class
  - A Redundancy Analysis Mechanism is required when using these methods

Feature Selection Based on Combining the Features for Evaluation

- (Feature Subset vs Ranking based methods)
  - **Conclusion- Ranking based methods are best for selecting features in high dimensional space as subset methods require more space and computational complexity**

Summary on Findings II

Wrapper vs Embedded vs Filter vs Hybrid

Feature Selection Based on Supervised Learning Algorithm Used

- 4 Classifications – Wrapper, Embedded, Filter, Hybrid
  - The filter approach was the most computational efficient in comparison to the wrapper, embedded, and hybrid methods
    - The Wrapper, embedded, and hybrid approaches do not have high generality as they use a supervised learning algorithm in the feature selection process
- **In summary- the Filter approach provides better generality and requires less computational complexity**

Feature Selection Literature

- State-of-the-art Feature selection methods listed in literature included use of the rule-based metric and nearest neighbors principles.
  - Both methods remove irrelevant features but fail to handle redundant features.
- Some methods used the theoretic-based metric to calculate similarity between feature and target-class as well as independency among features for redundancy analysis.
  - These methods however resulted in increased time complexity and do not have a mechanism for treating redundant features
- Hierarchical Clustering
  - Expensive and less effective in high-dimensional space due to curse of dimensionality
- K-means Clustering Algorithm
  - Simple, Scalable, and Faster for relevancy analysis in feature selection

Conclusion

- This paper analyzed many feature selection methods proposed by various researchers
- Research revealed that feature ranking-based methods are better than subset-based methods in terms of memory space and computational complexity,
  - Ranking methods however showed no decrease in redundancy (redundant features)
- Wrapper, embedded, and hybrid methods are computationally inefficient compared to the filter method and have poor generality
- **Feature Selection for high-dimensional data can be best optimized using the Filter Approach with ranking method for selecting significant features for machine learning models with a clustering approach for redundancy analysis**

Final Thoughts

- This research is incredibly useful for anyone studying supervised learning as it reviews many researcher's methods for feature selection and compares their positives and negatives in the same paper.
- This paper gave me useful insight into optimal feature selection which I will utilize when I work with more high-dimensional data in the future.
- This paper is easy to read/understand and accessible for beginner or advanced machine learning researchers.