Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 6, Issue, 06, pp.7234-7237, June, 2014

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

HYBRID APPROACH FOR FEATURE SELECTION

***Ms. Aparna Choudhary**

Galgotias University, Greater Noida, India

ARTICLE INFO

Article History:
Received 23rd March, 2014
Received in revised form
14th April, 2014
Accepted 19th May, 2014
Published online 25th June, 2014

Key words:
Feature selection, Gene selection, Term selection, Dimension Reduction, Genetic algorithm, Text categorization, Text classification.

ABSTRACT

Feature Selection (FS) is a strategy that aims at making text document classifiers more efficient and accurate. However, when dealing with a new task, it is still difficult to quickly select a suitable one from various FS methods provided by many previous studies. Feature selection, as a preprocessing step to machine learning, has been very effective in reducing dimensionality, removing irrelevant data, and noise from data to improving result comprehensibility. Researchers have introduced many feature selection algorithms with different selection criteria. However, it has been discovered that no single criterion is best for all applications. We proposed a hybrid approach for feature selection called based on genetic algorithms (GAs) that employs a target learning algorithm to evaluate features, a wrapper method. The advantages of this approach include the ability to accommodate multiple feature selection criteria and find small subsets of features that perform well for the targ*et al*gorithm. In this way, heterogeneous documents are summarized and presented in a uniform manner.

INTRODUCTION

Feature selection (known as subset selection) is a process commonly used in machine learning, wherein subsets of the features available from the data are selected for application of a learning algorithm. The best subset contains the least number of dimensions that most contribute to accuracy; one discards the remaining, unimportant dimensions. This is an important stage of preprocessing and is one of two ways of avoiding the curse of dimensionality (the other is feature extraction) (Sewell 2007).

There are two approaches

Forward selection: Start with no variables and add them one by one, at each step adding the one that decreases the error the most, until any further addition does not significantly decrease the error.

Backward selection: Start with all the variables and remove them one by one, at each step removing the one that decreases the error the most (or increases it only slightly); until any further removal increases the error significantly.

Two main models for feature selection are filtering and wrapper model (Shoushan Lui *et al.,* 2009). The filtering approach receives a set of features, and filters it independently from the induction algorithm. The wrapper model searches for good feature subsets, and evaluates them using n-fold cross-validation on the training data. This scheme may be used in conjunction with any induction algorithm, which is used for evaluating feature subsets on the validation set. The search for feature subsets can be performed using simple greedy algorithms such as backward elimination or forward selection, or more complex ones that can both add and delete features at each step. Since the wrapper model requires much more computation, filtering is the more common type of feature selection. This is especially true in the domain of textual information retrieval, where using the bag-of-words model results in a huge number of features. It was found that document frequency (DF), information gain (IG) and CHI are the most effective (reducing the feature set by 90-98% with no performance penalty, or even a small performance increase due to removal of noise). Contrary to a popular belief in information retrieval that common terms are less informative, document frequency, which prefers frequent terms (except for stop words), was found to be quite effective for text categorization.

Advantages of feature selection

It reduces the dimensionality of the feature space, to limit storage requirements and increase algorithm speed;

- It removes the redundant, irrelevant or noisy data.
- The immediate effects for data analysis tasks are speeding up the running time of the learning algorithms.
- Improving the data quality.
- Increasing the accuracy of the resulting model.

Corresponding author: Ms. Aparna Choudhary,
Galgotias University, Greater Noida, India.

- Feature set reduction, to save resources in the next round of data collection or during utilization;
- Performance improvement, to gain in predictive accuracy;
- Data understanding, to gain knowledge about the process that generated the data or simply visualize the data.

Feature Selection System Architecture

A simple three step feature selection approach is explained below. The goal of this architecture is to reduce a large set of features (on the order of thousands) to a small subset of features (on the order of tens), without significantly reducing the system's ability. The basic three steps of this system are:

- In first step the irrelevant features are removed.
- After that the redundant features are removed.
- And finally a feature selection algorithm is applied to the remaining features.

In this approach each step is working as a filter that reduces the number of candidate features, until finally only a small subset remains.

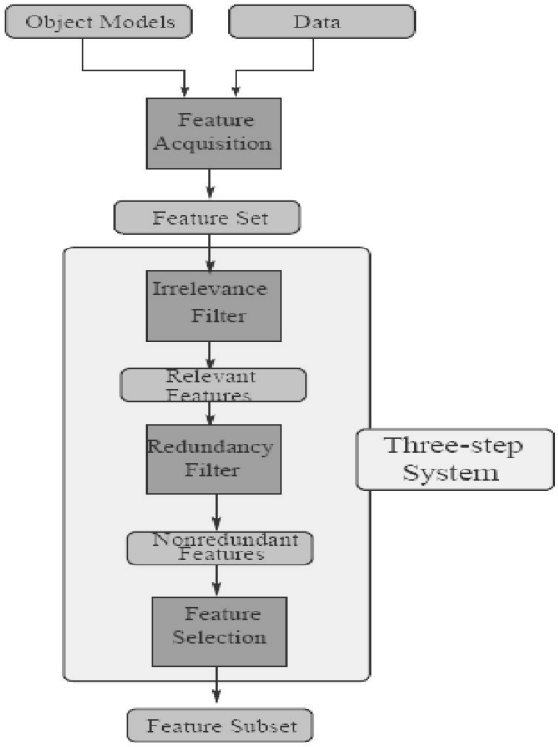


Figure 1.1. A feature selection system architecture

The first filter removes irrelevant features using a modified form of the Relief algorithm, which assigns relevance values to features by treating training samples as points in feature space. For each sample, it finds the nearest "hit" (another sample of the same class) and "miss" (a sample of a different class), and adjusts the relevance value of each feature according to the square of the feature difference between the sample and the hit and miss. There are several modifications to Relief to generalize it for continuous features and to make it more robust in the presence of noise. This system adopts Kononenko's modifications, and modifies Relief again to remove a bias against non-monotonic features, as described in (Bins 2000). Within this feature selection system, Relief is used as a relevance filter. Therefore it threshold the relevance values, to divide the feature set into relevant and irrelevant features. This can be done either by thresholding the relevance value directly, or by selecting the highest n values and discarding the remaining features. In either case, relief does not detect redundancy, so the remaining feature set still contains redundant features. The second step is a redundancy filter that uses the K means algorithm (MacQueen 1967) to cluster features according to how well they correlate to each other. When feature clusters are discovered, only the feature with the highest Relief score is kept; the other features in the cluster are removed from the feature set. This is an unusual application of K-means clustering, in that features are clustered (instead of samples), and correlation is used as the distance measure. A correlation threshold of 0.97 is used to detect when the features in a cluster are not sufficiently similar, in which case the cluster is split to make sure that potentially useful features are not removed from the feature set. The third and final filter is a combinatorial feature selection algorithm.

Problem definition and requirement analysis

Problem Definition

It has been widely observed that the huge amount of data available for text categorization and clustering often give unexpected and irrelevant results due to noisy, irrelevant or misleading features found in stored information. To overcome this problem some of the important and most relevant features should be selected which serve as the basis for further analysis of the given data. Study of unimportant features was the basic foundation of our research the filtering mechanism for removing trivial. To deal with all the aforesaid issues we tried to apply various feature selection methods including chi square, information gain, etc. on data set with the help of open source data mining tools like Weka and Rapid miner. Some researchers (Dy and Brodley 2004; Chuang *et al.,* 2004) have pointed out the problem that employing diverse feature selection criteria (either using independent evaluation criteria or using induction algorithms) often produce substantially different outcomes. This is because criteria based on different theoretic arguments introduce various biases toward some aspects. For instance, in wrapper methods, using different learning algorithms to evaluate features can produce different outcomes for this reason. Consequently, the performance of the classifiers built upon these feature selection methods varies as well. The problem leads to a dilemma:

The more algorithms available, the more challenging it is to choose a suitable one for a particular application (Liu and Yu 2005). A good understanding of the application domain and the technical details of the available Algorithms are needed to make the right choice, which is impractical in most situations. For new unknown data, it will be even more difficult to choose an appropriate method. Therefore, in our work, we proposed a hybrid approach for feature selection to solve the problem. Our objective is implementation of hybrid algorithm on datasets for

identifying the crucial feature subset that is capable of generating accurate predictions.

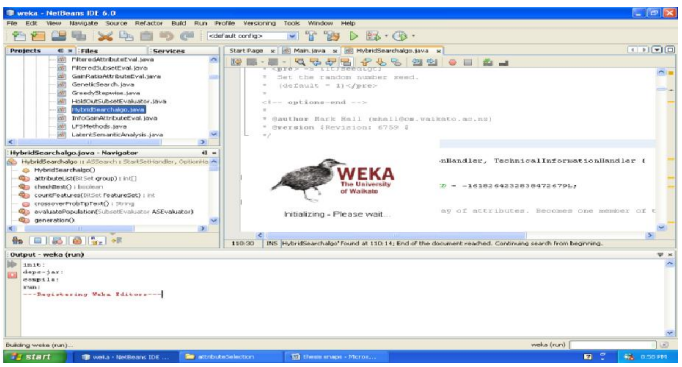
Introduction to hybrid approach for feature selection

The unique aspect of the Hybrid Algorithm is that all chromosomes and offspring's are allowed to gain some experience through a local search process before being involved in the evolutionary process (Weston *et al.,* 2000). As such, the term HA has been used to describe a GA that heavily favors local search (Liu and Setiono 1995). Similarly to GA, an initial population is randomly created by an HA. Subsequently, the local search operations move solutions towards local optima. These improvements are accumulated over all generations, resulting in a significant improvement in the overall performance (Dash and Liu 1999). Subsequently, crossover and mutation operators are applied in a fashion similar to GAs to produce offspring. These offspring are then subjected to the local search so that local optimality is always maintained.

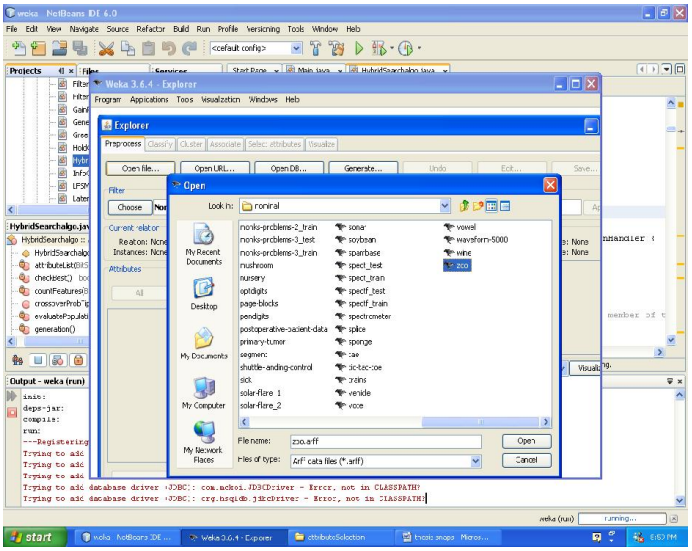
Procedure of Hybrid algorithm

1 Begin
2 Initialize: Randomly initialize population of feature subset, initialize E and others;
3 While (stop if condition is not satisfied)
4 Evaluate fitness of all feature subset encoded in the population;
5 Find E best feature subset in the population and put them into elite pop;
6 For (each subset in elite pop)
7 Perform local search and replace it with new feature subset;
8 End For
9 Evaluate fitness of new solutions which is generated by local search;
10 Select the best solution based on fitness function as global optimum;
11 Perform evolutionary operators, i.e. selection, crossover, mutation;
12 End While
13 End

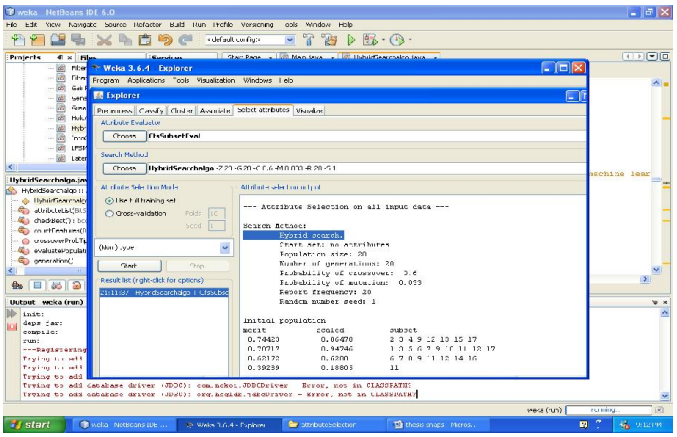
Implementation of hybrid algorithm using weka in net beans ide 6.0



HybridSearch method in weka project explorer

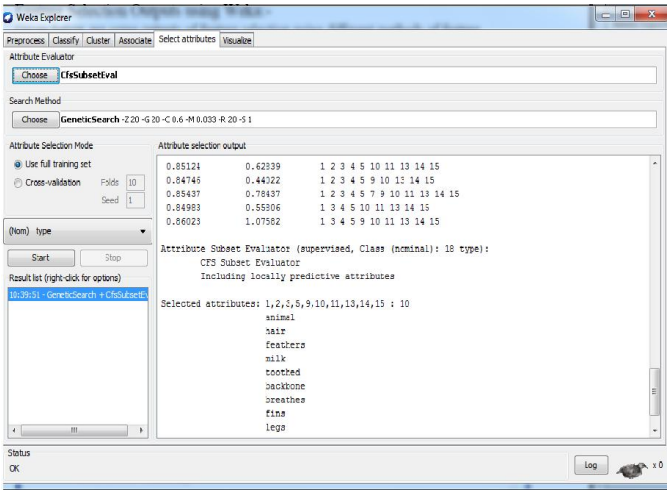


Selecting zoo.arrf file in attributeSelection Package in weka

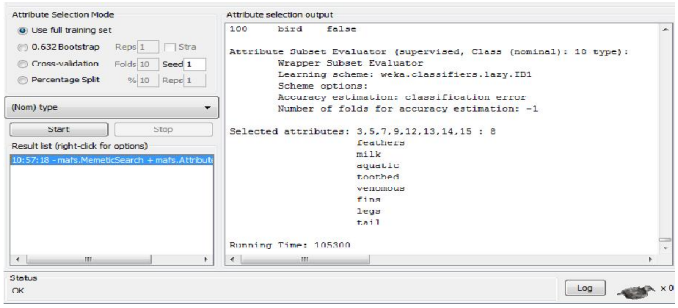


HybridSearch method in Weka Explorer

TESTING AND RESULTS



Output of Attribute Selection in Weka Using Genetic Search



Output of Attribute Selection in Weka Using Hybrid Search

Future enhancement

As a technique to reduce dimensionality of data, feature selection is fundamental to improve the efficiency and effectiveness of machine leaning algorithms. The goal of this dissertation is to improve feature selection algorithms for machine learning areas. In this dissertation, we proposed a hybrid approach for feature selection to solve the problem. We first used a simple Local Search method whose goal is to maximize the classification accuracy. Then, we designed genetic algorithm with a different goal. The genetic algorithm considers both the classification performance and the size of feature subsets. It aims to achieve a balance between the size of feature subsets and their classification performance. We tested our Hybrid algorithm over a zoo. arff file which includes 18 different attributes. Output of this process will show that which attribute among all the attributes have the best selected attribute. A limitation of our hybrid feature selection approach is that it requires much computation time because the genetic algorithms repeatedly call the induction algorithm for evaluation of feature subsets, a common drawback of wrapper method. In the future, we can speed up our method by parallelizing the genetic algorithms.

Conclusion

In this work we had projected various feature selection methods, terms, limitations, advantages and available recent innovation in feature selection. We hope, that the interested readers will have broad overview of this field and several starting point for further details. Feature selection remains and will continue to be an active field that is incessantly rejuvenating itself to answer new challenges.

REFERENCES

Bins J., "Feature Selection of Huge Feature Sets in the Context of Computer Vision", 2000.

Chuang, H.Y. *et al*. Identifying Significant Genes from Microarray Data. Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04), p. 358, 2004

Dash, M. and Liu, H. Handling large unsupervised data via dimensionality reduction. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. 1999

Dy, J. G. and Brodley, C. E. Feature Selection for Unsupervised Learning. The Journal of Machine Learning Research, vol. 5, pp.845-889, Aug. 2004

Liu, H. and Setiono, R. Chi2: Feature selection and discretization of numeric Attributes. Proc. IEEE 7th International Conference on Tools with artificial Intelligence, 338-391, 1995

Liu, H. and Yu, L. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, pp. 491-502, Apr., 2005

MacQueen J., "Some Methods for Classification and Analysis of Multivariate Observations", Fifth Berkley Symposium on Mathematics, Statistics and Probability, University of California Press, Berkeley, pp. 281-297, 1967.

Sewell M. "Feature Selection "Department of Computer Science University College Landon April 2007.

Shoushan Lui, R Xia, C Zong "A Framework of Feature Selection Methods for Text Categorization" Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 692–700, Suntec, Singapore, 2-7 August 2009.

Weston, J. *et al*. Feature selection for SVMs. Advances in Neural Information Processing Systems 13, 2000

****