

Feature Selection Methodologies In Genetic Algorithms: A Literature Review

Dr.E.CHANDRA BLESSIE

Associate Professor, Department of Computer Applications,
Nehru College of Management, Coimbatore.

Mrs.S.GNANAPRIYA

Ph.D Research Scholar, Department of Computer Applications,
Nehru College of Management, Coimbatore

Abstract

Feature selection Genetic algorithm has become an essential task in data mining for various real-world problems. This method is used to remove the redundant and irrelevant features for increasing the accuracy and prediction in classification. This selection method is also used for computing performance, efficiency and interpretability by using supervised and unsupervised learning methods. In this survey paper overview of recent advances in feature selection and classification is illustrated for various domains. In addition to feature selection filter, wrapper and embedded methods are the recent approaches for selecting subsets efficiently with hybrid method. The advantages and disadvantages are outlined to provide clear insight for computational memory and best accuracy.

Keywords:*Feature Selection, Data Mining, Classification, Genetic Algorithms.*

Introduction

Feature Selection (FS) provides effective and efficient high dimensional data for machine learning, data mining and other problems. The feature selection improves accuracy by eliminating redundant features, complexity with computation cost and measurement costs, maximizes accuracy of classifiers [1]. The features which are irrelevant provides no information

for selected feature. The feature selection technique is used in data preprocessing and machine learning process [2] for reducing dimensionality of data. The FS is also termed as attribute selection, variable selection or variable subset selection in machine learning and statistics

The feature selection has two important approaches, 1. Individual evaluation and 2. Subset evaluation. The former method provides ranking of the features and latter method constructs candidate feature subset using search strategy. The feature selection procedure has four steps, 1. Subset generation, 2. Evaluation of subset. 3. Stopping criteria and 4. Result validation.

In subset generation candidate subset is specified for each state for evaluating the search space. This generation uses heuristic search and has two issues to determine subset generation process. The first search decides the search point for influencing the search direction. The second search strategy identifies feature selection process with sequential search and exponential search [7, 8]. Each newly generated subset must be evaluated to determine the goodness of the subset for feature selection. The evaluation is categorized into 2 types namely, 1. Independent and 2. Dependent criteria [9]. The stopping criteria is used to stop the selected process in feature selection for validation procedure.

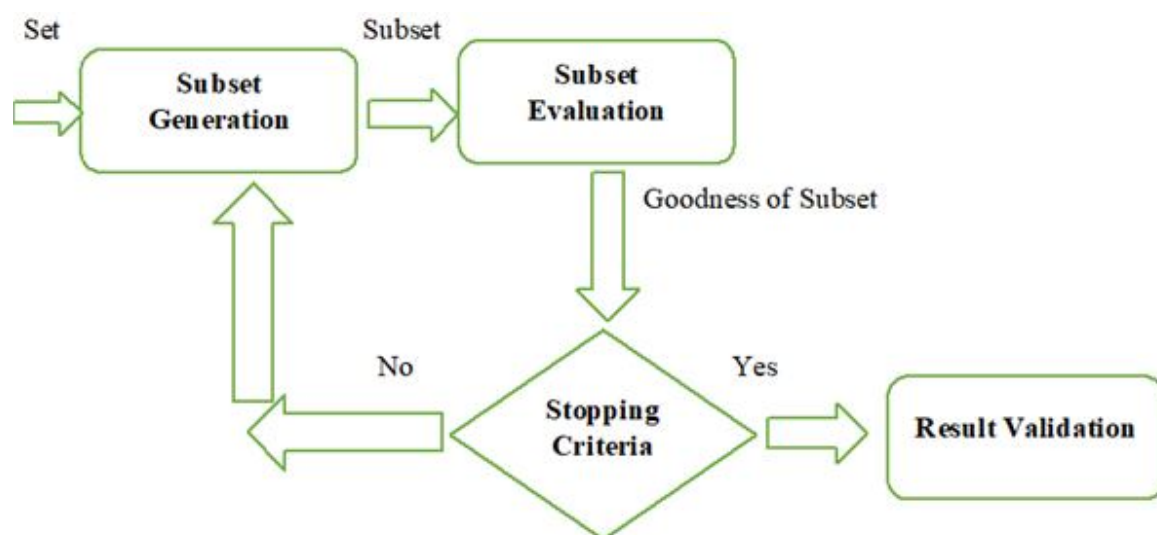


Figure 1. Feature Selection steps [3]

Feature Selection Methods for Classifications

The FS methodologies can be categorized into two types, 1. Supervised and 2. Unsupervised feature selection methods. The feature selection has three general approaches, 1. Filter approach, 2. Wrapper approach and 3. Embedded approach. The filter approach provides characteristics of training data for mining algorithms [4], wrapper approach provides relationship between optimal feature selection and relevance and embedded approach performs specific learning algorithm in the process. The FS have the advantage of improving computational efficiency, learning performance, decreases memory requirements and generates better models.

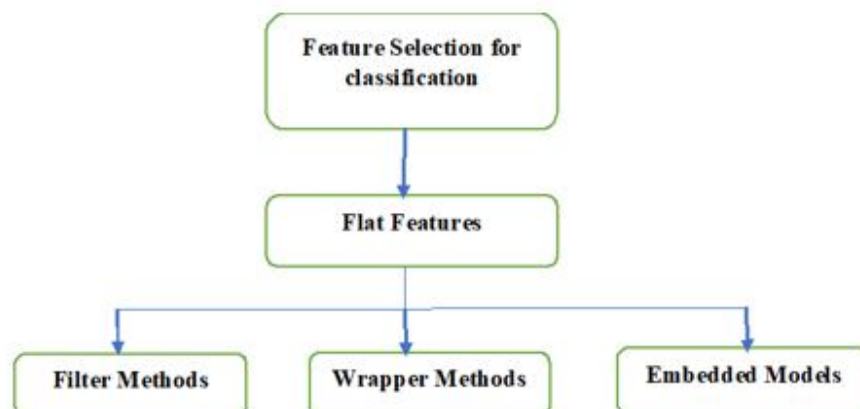


Figure 2: Feature Selection for Classification

Filter approach

This approach is used for data analysis [5] using built-in adaptive systems and extracts features from the data without involving any learning. The modeling algorithms use this approach after selecting features based on performance measure used in modeling algorithms. It uses two criteria namely, 1. Distance based and 2. Margin based for filtering. The measures for feature filtering can be classified into 1. Information, 2. Distance, 3. Consistency, 4. Similarity and 5. Statistical measures.



Figure 3. The feature filter model [14]

All filters can be used in various data mining tasks namely, 1. Classification, 2. Regression and 3. Clustering. There are many strategies for feature subset generation 1. Forward selection, 2. Back elimination, 3. Bidirectional selection and 4. Heuristic feature subset selection. The forward selection has empty feature set and includes more features to the set. Back elimination has whole feature set and eliminates one or more features from the set. Bidirectional set starts at both the ends and uses empty and whole feature set. The heuristic selection starts the subset based on genetic algorithm.

Wrapper approach

This approach is used to evaluate selected features quality using predefined learning algorithms and uses learning techniques for evaluating feature selection. It is usually wrapped around adaptive system for providing subset features and their feedback (accuracy). The wrapper approach improves results of specific predictors and uses classifiers for finding subset features using predictive power [6]. This method performs two steps: 1) subset features searching and 2) selected feature evaluation.

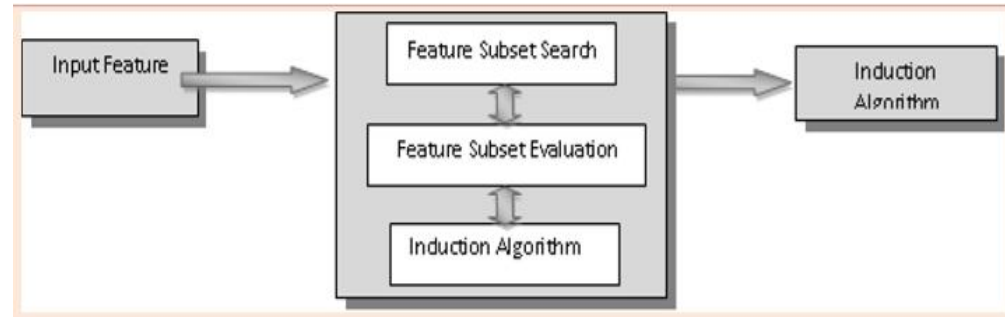


Figure 4. The wrapper model [14]

Embedded approach

This approach combines feature selection and classifier method for optimal subset of features which is built into construction and searches combined space of feature subset and hypotheses. It is a learning algorithm which has advantage of interaction with classification model.

The framework for feature selection performs the following steps:

- *Step 1: Generating a new feature;*
- *Step 2: Determining whether adding the newly generated feature to the set of currently selected features;*
- *Step 3: Determining whether removing features from the set of currently selected features;*
- *Step 4: Repeat Step 1 to Step 3.*

The feature selection uses various algorithms 1. Grafting algorithm, 2. Alpha-investing algorithm & 3. Online streaming feature selection algorithms.

Feature Selection Applications in various Domains

The feature selection methods vary among various application areas. The application domains include text classification, image processing, bioinformatics and sentiment analysis.

Text Classification: The most important problem in text classification is high dimensional documents which is increasing day by day via online text data. This dimensionality should be reduced by using efficient mining algorithms. The feature selection using Gain Ratio (GR) is used for poems [11] and lyrics[10] for text classification.

Image processing: The feature selection process in image processing is used for searching, retrieving and browsing the contents. The large amount of images in the database is the major problem for image retrieval process [12].

Bioinformatics: The feature selection in bioinformatics is used to identify the biomarker from genomics data. It analyzes output similarity with different selection methods for better predictions.

Sentiment analysis: This system analysis nature language processing for capturing the data favorability. It deals with subjectivity text, opinion and sentiment. It is used for answering and recommending the systems [13].

Approach	Advantage		Disadvantage	
	Univariate	Multivariate	Univariate	Multivariate
Filter	● Scalable ● Fast to access	● Better computation time ● dependencies	● Ignores interaction and dependencies	● Ignores interaction ● Slower variate
Wrapper	Deterministic	Randomized	Deterministic	Randomized
	● Low computation time ● Simple to use	● Dependencies ● Classifier interaction	● High risk for fitting	● Computational time
Embedded	● Low computation time ● Dependencies		● Selection dependencies	

Table 1: Advantage & disadvantages of feature selection approaches

Feature selection in genetic algorithms

Feature selection using algorithm deals with isolating the very few relevant features from the large set. This is not exactly the classical feature selection problem known in Data mining. Among the various categories of feature selection algorithms, Genetic algorithm to solve real world optimization problems. There are many applications used for genetic algorithm in data mining. Robotics is used to solve real word problems among those. Advantages of genetics algorithm:

- Perform better than common feature selection techniques.
- Genetic algorithms can manage data sets with many features and it can be easily parallelized in computer clusters.

One of the issues using Genetic algorithm for feature selection is that the optimization process can be very attractive and there is a possibility for the GA to overfill to the predictors. One of the feature selection for most algorithms is the genetic algorithm. This is a stochastic method for function optimization based on the mechanics of natural genetics and biological evolution.

In The Genetic Algorithm is an optimization method inspired by the procedures of natural evolution. In feature selection, the function to optimize is the generalization performance of a predictive model. The combination of features selection using a genetic algorithm and corresponding sentiment words can help produce accurate, meaningful, and high-quality sentiment analysis results. We present here a genetic algorithm dedicated for a particular feature selection problem encountered in genetic analysis of Cross domain opinion mining.

Conclusion

In this article the feature selection research in various domains are reviewed for reducing the high dimensional data. The feature selection includes subset generation, evaluation and stopping criteria for subsets. It also includes three general approaches for feature selection filter, wrapper and embedded methods for subset selection. Feature selection in genetic algorithm is very

important in many common applications of predictive analytics. Since it contains a large number of features, one has to select the advanced methods of genetic algorithm.

Genetic algorithm has high performance compared to the other feature selection algorithms with different classification techniques. It is the best feature selection algorithm for Cross domain opinion mining. In this paper show how genetic algorithms can be applied to optimize the performance of a cross domain by selecting the most relevant features selection. Genetic algorithms will be useful tools for data mining in many situations.

The paper shows that feature selection based on genetic algorithm along with an ensemble approach outperformed the other approaches. We conducted a comparative study experiments on multidomain dataset and Books, DVD Electronics review dataset in opinion mining.

References

1. Steppe. J., K.W. Bauer, "Feature Saliency Measures", Computers & Mathematics with Applications, Vol 33, No. 8, pp. 109-126; 1997.
2. A. Kalousis, J. Prados, M. Hilario, "Stability of Feature Selection Algorithms: a study on high dimensional spaces," Knowledge and information System, vol. 12, no. 1, pp. 95-116, 2007.
3. M. Dash, H. Liu, "Feature Selection for Classification," Intelligent Data Analysis, Elsevier, pp. 131-156, 1997.
4. Verónica Bolón-Canedo, Noelia Sánchez-Marroño, Amparo Alonso-Betanzos, "A review of feature selection methods on synthetic data," Knowledge and Information Systems, vol. 34, no. 3, pp 483-519, Mar. 2013.
5. Wlodzislaw Duch; "Filter methods", Springer- Feature Extraction Studies in Fuzziness and Soft Computing Volume 207, 2006, pp 89-117.
6. L.Ladha et al., "Feature Selection Methods and Algorithms"; International Journal on Computer Science and Engineering (IJCSSE), Vol. 3. No. 5, pp. 1787-1797; 2011.

7. P. Narendra, K. Fukunaga, "A Branch and Bound Algorithm for Feature Subset Selection," IEEE Transactions on Computer, vol. 26, no. 9, pp. 917-922, 1977. Article (CrossRef Link).
8. J. Pearl, "Heuristics," Addison-Wesley, 1983.
9. Huan Liu, Lei Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," IEEE Transaction on Knowledge and Data Engineering 2005.
10. V. Kumar, S. Minz, "Poem Classification using Machine Learning Approach," International Conference on Soft Computing for problem Solving (SocPro 2012), Dec. 2012.
11. V. Kumar, S. Minz, "Mood Classification of Lyrics using SentiWordNet", in Proc. of ICCCI-2013, Jan. 2013.
12. <http://www.di.unipi.it/~bacciu/teaching/IIA2012/lect3-exploratory-hand.pdf>.
13. Bo Pang, Lillian Lee; "Opinion Mining and Sentiment Analysis"; Foundations and Trends in Information Retrieval, Vol. 2, Nos. 1–2 pp- 1–135, 2008.
14. G. H. John, R. Kohavi, K. Pfleger, "Irrelevant feature and the subset selection problem," in Proc. of the Eleventh International Conference on Machine Learning, pp. 121-129, 1994.
15. P. Kalaivani and K. L. Shunmuganathan, "Performance evaluation of feature selection method for sentiment classification of online reviews using machine learning techniques," International Review on Computers and Software, vol. 8, no. 8, pp. 1769–1775, 2013.
16. A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: feature selection for opinion classification in Web forums," ACM Transactions on Information Systems, vol. 26, no. 3, article 12, 2008.