

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320-088X

IJCSMC, Vol. 3, Issue. 2, February 2014, pg.827 – 830

RESEARCH ARTICLE

Selection of Most Relevant Features from High Dimensional Data using IG-GA Hybrid Approach

¹Ishani Mandli, ²Prof. Mahesh Panchal

^{1,2}Computer Department & Gujarat Technological University, Gujarat, India

¹ishanimandli.90@gmail.com; ²mkhpanchal@gmail.com

Abstract— Feature selection is considered a problem of global combinatorial optimization in machine learning, which reduces the number of features, removes irrelevant, noisy and redundant data, and results in acceptable classification accuracy. In the past few decades, researchers have developed large amount of feature selection algorithms. These algorithms are designed to serve different purposes, are of different models, and all have their own advantages and disadvantages. Although there have been intensive efforts on surveying existing feature selection algorithms, to the best of our knowledge, there is still not a dedicated repository that collects the representative feature selection algorithms to facilitate their comparison and joint study. To fill this gap, in this work, an IG-GA hybrid approach with MRMR evaluation function is presented for high dimensional data set.

Keywords— Feature Selection, Classification, filter approach, wrapper approach

I. INTRODUCTION

Feature (or variable, or attribute) subset selection (FSS) is the process of identifying the input variables which are relevant to a particular learning (or data mining) problem, and is a key process in supervised classification. FSS helps to improve classification performance (accuracy, AUC, etc.) and also to obtain more interpretable classifiers or to detect outliers [1]. In the case of high-dimensional datasets, e.g., datasets with thousands of variables, FSS is even more important because otherwise the number of instances needed to obtain reliable models will be enormous (impracticable for many real applications such as microarray domains).

Classification is one of the most frequently studied problems by DM and machine learning (ML) researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes). In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how “good” the algorithm is [2].

Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information.

Most algorithms for supervised FSS can be classified as filter or wrapper approaches. In the *filter approach* an attribute (or attribute subset) is evaluated by only using intrinsic properties of the data (e.g. statistical or information-based measures). Filter techniques have the advantage of being fast and general, in the sense that the subset obtained is not biased in favour of a specific classifier. On the other hand *wrapper algorithms* are those that use a classifier (usually the one to be used later) in order to assess the quality of a given attribute subset [1]. Wrapper algorithms have the advantage of achieving greater accuracy than filters but with the disadvantage of being (far) more time-consuming and obtaining an attribute subset that is biased towards the classifier used.

II. LITERATURE REVIEW

In this section, I present the different approaches of feature subset selection like Relief, SFS, SBS, Information Gain (IG), Genetic algorithm.

1. Relief[generation=heuristic, evaluation=distance]

Relief uses a statistical method to select the relevant features. It is a feature weight-based algorithm inspired by instance-based learning algorithms ([1,8]). From the set of training instances, it first chooses a sample of instances, it first chooses a sample of instances; the user must provide the number of instances (No Sample) in this sample. Relief randomly picks this sample of instances, and for each instance in it finds Near Hit and near Miss instances based on a Euclidean distance measure. Near Hit is the instance having minimum Euclidean distance among all instances of the same class as that of the chosen instance; near Miss is the instance having minimum Euclidean distance among all instances of different class.

2. Information Gain

Information gain (IG) is a feature ranking method based on decision trees that exhibits good classification performance. The idea behind IG is to select features that reveal the most information about the classes[1]. Let S be the set of n instances and C be the set of k classes. $P(C_i, S)$ represents the fraction of the example in S that has class C_i . Then, the expected information from this class membership is given by[1]:

$$Info(S) = - \sum_{i=1}^k P(C_i, S) \times \log(P(C_i, S)) \quad (1)$$

If a particular attribute A has v distinct values, the expected information is obtained by the decision tree in which A is the root, and the weighted sum of expected information of the subsets of A is based on the distinct values. Let S_i be the set of instances and A_i the value of attribute A :

$$Info_A(S) = - \sum_{i=1}^v \frac{|S_i|}{|S|} \times Info(S_i) \quad (2)$$

3. Sequential forward selection (SFS) and sequential backward selection (SBS)

SFS starts from the empty set, and in each iteration generates new subsets by adding a feature selected by some evaluation function. SBS starts from the complete feature set, and in each iteration generates new subsets by discarding a feature selected by some evaluation function.

4. Genetic Algorithm

Genetic algorithms (GAs) are stochastic search algorithms modelled on the process of natural selection underlying biological evolution. They can be applied to many search, optimization, and machine learning problems. Genetic algorithms are good at taking larger, potentially huge, search spaces and navigating them looking for optimal combinations of things and solutions which we might not find in a life time[8].

III. MOTIVATION FROM LITERATURE

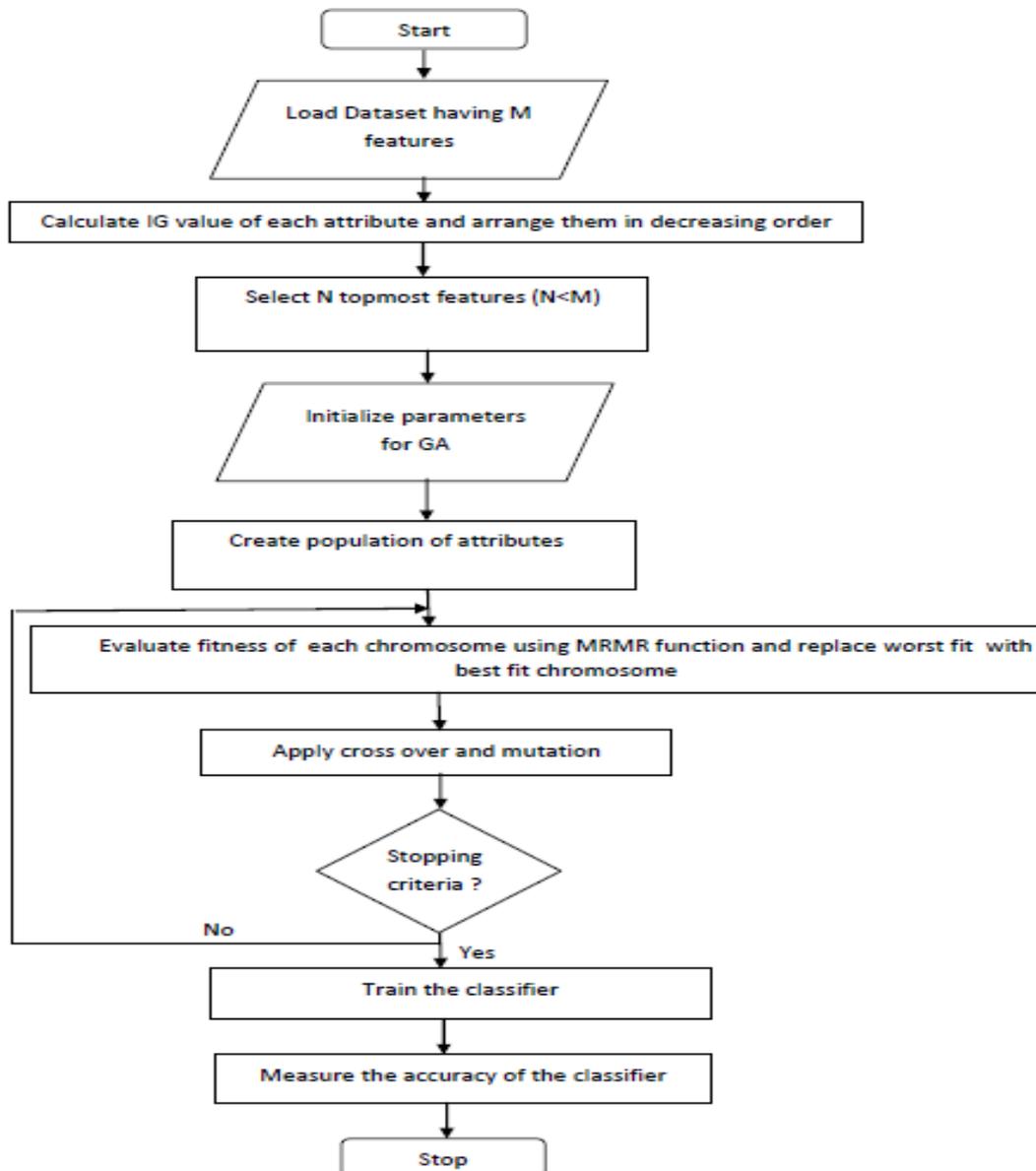
Filter methods are independent of the inductive algorithm, whereas wrapper methods use the inductive algorithm as the evaluation function.

IG is fast and also gives more accuracy. But till it generates large size of subset. So other subset reduction methods are required to perform. Wrapper approach for subset evaluation is very time consuming. So other subset measures are required to be applied.

IV. PROPOSED WORK

The proposed work uses a filter method (Information Gain, IG) and a wrapper method (Genetic Algorithm, GA) for feature selection in high dimensional data set. In the first stage, an information gain (IG) value was calculated each feature. In the second stage, all the selected features must conform to a threshold. Subsequently, feature selection was once again performed, this time capitalizing on the genetic algorithm's unique attributes to select the features. The Minimum redundancy-Maximum relevance (MRMR) served as an evaluator of the IG-GA.

1. Initially, load database having M attributes.
2. Calculate Information Gain value of each attribute using the equation given in above section.
3. Arrange attributes in decreasing order of their IG value.
4. Select N top most attributes ($N < M$) whose IG value is greater than some predefined threshold value.
5. Initialize the parameters of Genetic Algorithm like population size, crossover rate, mutation rate.
6. Create population of attribute.
7. Find the fitness value of each chromosome using MRMR fitness function.
8. Apply crossover and mutation for generation of new chromosomes.
9. Repeat step 7 and 8 until stopping criteria do not meet.
10. Train the classifier using the resulting feature subset.
11. Measure the accuracy of classifier.



V. CONCLUSIONS

Feature subset selection deals with feature subset extraction from high dimensional database containing small number of samples. This paper presents the different feature subset selection approaches including relief, SFS, SBS, Information gain, Genetic algorithm. All of these methods are implemented for high dimensional database. But in real world, many situations are occurred in which out of large number of features only few affect classification accuracy. Rest of them are not useful for classification. The proposed work is used for this type of situation. The proposed work is used to find minimal subset of features which can correctly identify the class label of a given sample.

REFERENCES

- [1] Pablo Bermejo , Luis de la Ossa, José A. Gámez, José M. Puerta , "Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking", 5 Feb, 2011
<http://www.sciencedirect.com/science/article/pii/S095070511100027X>
- [2] Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervás , "Data Mining Algorithms to Classify Students"
<http://sci2s.ugr.es/keel/pdf/specific/congreso/Data%20Mining%20Algorithms%20to%20Classify%20Students.pdf>
- [3] Jiawei Han, "Data Mining :Concept and Techniques: Chapter -6 "
http://www.mis.boun.edu.tr/gulser/index_files/DM%20Concepts%20%26%20Techniques%20_%20Han%26Kamber.pdf
- [4] Eivind Hoffmann "STANDARD STATISTICAL CLASSIFICATIONS: BASIC PRINCIPLES"
<http://unstats.un.org/unsd/class/family/bestprac.pdf>
- [5] CLASSIFICATION : BASIC ONCEPT, DECISION TREE AND MODEL EVALUATION
<http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf>
- [6] YongSeog Kim, W. Nick Street, and Filippo Menczer "Feature Selection in Data Mining "
http://www.ime.unicamp.br/~wanderson/Artigos/feature_selection_in_mining.pdf
- [7] M. Dash 1, H. Liu2, "Feature Selection for Classification "
<http://machine-learning.martinsewell.com/feature-selection/DashLiu1997.pdf>
- [8] Cheng-Huei Yang, Li-Yeh Chuang, Cheng-Hong Yang, "IG-GA: A Hybrid Filter/Wrapper Method for Feature Selection of Microarray Data", 5 Aug 2009
<http://jmbe.bme.ncku.edu.tw/index.php/bme/article/viewFile/507/734>
- [9] Chris Ding and Hanchuan Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data", 16 Jun 2004
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.6875&rep=rep1&type=pdf>
- [10] Chapter 8: Feature Selection Based on Ant Colony optimization
<http://pr.hec.gov.pk/Chapters/720S-8.pdf>
- [11] Denny Hermawanto, "Genetic Algorithm for Solving Simple Mathematical Equality Problem "
<http://arxiv.org/ftp/arxiv/papers/1308/1308.4675.pdf>