ISSN:0975-9646

Improving the life cycle of Genetic Programming using Feature Selection with Multiobjective Fitness Function

Rupali Koushal , Manoj Dhawan

Information Technology, Shri Vaishnav Institute of Technology and Science
Ujjain road, Gram Baroli, Indore, M.P.,India

Abstract— **In this paper, we proposed to select the optimum number of features from the datasets using genetic programming. For selecting the optimal number of features we run the GP life cycle for 50 percent of the generation and select the features present in the best classifiers and form their optimal set. After that we remove the other features in the classifier from the features present in the optimal set and run the process till the last generation. Then we select the features present in the best classifier after the last generation and called those features the optimal feature set.**

Keywords*— *feature selection, genetic programming, multiobjective fitness function.

I. Introduction

In feature selection [5] process subset of feature is selected from available feature. To achieve faster and more cost effective predictors and providing a better understanding of the underlying process that generated the data. In the context of classification, feature selection techniques can be organized into three categories, depending on how they combine the feature selection search with the construction of the classification model: filter methods, wrapper methods and embedded methods. Filter techniques assess the relevance of features by looking only at the intrinsic properties of the data. In most cases a feature relevance score is calculated, and low-scoring features are removed. Afterwards, this subset of features is presented as input to the classification algorithm. Advantages of filter techniques are that they easily scale to very high dimensional result, feature selection needs to be performed only once, and then different classifiers can be evaluated. A common disadvantage of filter methods is that they ignore the interaction with the classifier, and that most proposed techniques are univariate. Whereas filter techniques treat the problem of finding a good feature subset independently of the model selection step, wrapper methods embed the model hypothesis search within the feature subset search. In this setup, a search procedure in the space of possible feature subsets is defined, and various subsets of features are generated and evaluated. However, as the space of feature subsets grows exponentially with the number of features, heuristic search methods are used to guide the search an optimal subset. These search methods can be divided in two classes: deterministic and randomized search algorithms. Advantages of wrapper approaches include the interaction between feature subset search and model selection, and the ability to take into account feature dependencies. A common drawback of these techniques is that they have a higher risk of overfitting than filter techniques and are very computationally intensive, especially if building the classifier has a high computational cost. In a third class of feature selection techniques, termed embedded techniques, the search for an optimal subset of features is built into the classifier construction, and can be seen as a search in the combined space of feature subsets and hypotheses. Just like wrapper approaches, embedded approaches are thus specific to a given learning algorithm. Embedded methods have the advantage that they include the interaction with the classification model, while at the same time being far less computationally intensive than wrapper methods. Feature selection has become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. These areas include text processing of internet documents, gene expression array analysis, and combinatorial chemistry. While feature selection can be applied to both supervised and unsupervised learning, we focus here on the problem of supervised learning (classification), where the class labels are known beforehand. As many techniques were originally not designed to cope with large amounts of irrelevant features, combining them with FS techniques has become a necessity in many applications The objectives of feature selection are manifold, the most important ones being: (a) to avoid overfitting and improve model performance, i.e. prediction performance in the case of supervised classification and better cluster detection in the case of clustering, (b) to provide faster and more cost-effective models and (c) to gain a deeper insight into the underlying processes that generated the data.

GENETIC PROGRAMMING (GP) [2] introduced by Koza and his group is popular for its ability to learn relationships hidden in data and express them automatically in a mathematical manner. Genetic programming (GP) is a machine learning technique inspired by biological evolution to synthesize programs for a given computational task. GP has already spawned numerous interesting applications. GP has been used rules for two class problems. Although genetic algorithm has been used by many researchers to design classifiers for multiclass problems, only a few attempts have been made to solve the same problem using GP. Genetic programming is a method of automatically generating computer programs to perform specified tasks. It uses a genetic algorithm to search through a space of possible computer programs for one which is nearly optimal in its ability to perform a particular task. While it was not the first method of automatic programming using genetic algorithms (one earlier approach is detailed in (Cramer, 1985)), it is so far the most successful. We imposed a technique in which all the feature is not required to solve the problem. From the available set of feature, a subset of feature is finding and this feature is sufficient to solve the problem. This technique is not time consuming and less complex.

By using feature selection with multiobjective fitness function (FSMFF) [4] genetic life cycle is improved. This improvement reduced the time required to solving the problem. GP performance is improved using FSMFF. To create initial population Ramped-half-and-half method is used in this paper. After that from the available set of feature the subset of feature is selected. For the each subset of feature fitness is evaluated. Chromosomes which have high fitness are selected for next generation. To produce next generation genetic operator reproduction, crossover, mutation is applied. This process is repeated until the termination criteria are meeting or at the begging we can we can specify the number of generation we proceed.

II. Literature Review

Durga Prasad Muni et al. [8] In this paper propose a new approach for designing classifiers for a c-class ($c \geq 2$) problem using genetic programming (GP). The proposed approach takes an integrated view of all classes when the GP evolves. A multitree representation of chromosomes is used**.** Genetic algorithm comprises a set of individuals elements and a set of biologically inspired operators defining the population itself. It is an integrated evolutionary approach where classifier trees for all classes are evolved simultaneously. In computing, GA maps problems onto a set of strings each string representing a potential solution. For genetic operation, tree is selected on the basis of their unfitness. On the basis of their unfitness we proposed a modified crossover operation and a new mutation operation called non destructive directed point mutation. To optimize the classifier an OR-ing operation is introduced and a weight based scheme for conflict resolution. Then we tested our classifier with several data set and obtained satisfactory result. Bing Xue et al. [7] In the classification problem lots of feature is present in dataset but hole are not useful for classification. Unwanted and duplicate feature may minimize the performance. The objective of feature selection is to choose subset of necessary feature to achieve greater performance rather than using all features. It has two main objectives first enlarge the classification performance and second one is reduce the number of feature. However in the previous feature selection algorithm they have only task. Very first study on particle swarm optimization (PSO) for feature selection is presented by this paper. For selecting feature two PSO based multiobjective feature selection algorithm is used in this paper. The first algorithm introduces the idea of non dominated sorting into PSO to address feature selection problems. The second algorithm applies the ideas of crowding, mutation, and dominance to PSO to search for the Pareto front solutions which gives better results than the other methods mentioned previously. Peter I. Rockett et al. [9] In this paper an optimized feature extraction framework is presented. That uses multiobjective genetic programming (MOGP). Since in previous method set of inputs place into one dimensional decision space. While in MOGP method data set place into multidimensional decision space to obtain excellent classification performance. In this paper pareto dominance set is used for vector minimization By which maximal class separability is obtain. Isac Sandin et al. [10] Dealing with highly dimensional data in automatic classification is very challenging. Many feature selection technique like Information Gain and x^2 have been proposed. Whenever data is very complex this traditional technology not gives proper result. In this paper for attribute selection of misshaped data they use GP and common feature selection method. Tested his result over most complex data like p53 protein and k8 cancer rescue mutants dataset. In this paper GP is used to learn complex feature selection matrix. By which they able to understand the basic matrix easily. Since each terminal node contains a basic feature selection matrix which gives lots of information (set of feature). In this paper they combine two feature selection matrix on the basis of operations performed by their parent node (super class node). They also calculate the fitness function of each filtered input. By using combination of these technologies they improve size of data space by 98.2% (504 features to only 9). Also increases accuracy to 34% in micro F1 on p53 samples. Jianjun Ya et al. [11] Genetic programming (GP) uses evolutionary algorithm to simulate natural selection as well as population dynamics, hence leading to simple classifiers. Here we applied GP to cancer expression profiling data to select feature genes and build classifiers by mathematical integration of these genes. In this paper, we used GP to find classifiers who capable of classifying samples into different cancer types based on gene expression patterns. A classifier is discover from a training data set and then evaluated to assess its prediction capability on samples with unknown labels. A generic GP classifier based prediction is shown as: IF ’(GENE[A]/GENE[B] _ GENE[C]) > D’ THEN ’TARGET CLASS’, where the IF clause is generated by GP, ‘‘TARGET CLASS’’ is predefined in the initial configuration file, D is a constant, and GENE[A], GENE[B], and GENE[C] represent the expression levels of genes A, B, and C, respectively. A continuous prediction score is preferred for certain analysis like the ROC curve analysis. To implement that, the classifier can be converted to the form of ’GENE[A]/GENE[B] _ GENE[C]’, where the calculated mathematical expression values can be then treated as a continuous variable for ROC test.

III. Proposed Work

In our work we proposed solution for the feature selection which include following steps.

A. *Initialization*

Initialization of population is the first step of genetic programming. For this we have to construct the valid tree. For the initialization of the chromosomes we have to define the maximum tree depth. Using this parameter the size of the tree is confine. Another parameter for the initialization function required set of leaf node T also called terminal set and set of internal nodes I also called function set. Terminal set contain features, constant and function set contain arithmetic operators. In this paper ramped-half-and-half method is use for the initialization.

B. Selection of a Feature Subset

In this step we have to select those features from the all the available feature which attain proficient and superior solution of the problem. From the all the available n feature we select best subset of feature which is essential and adequate to gain the solution. There are different method for the feature selection they are filter approach and wrapper approach. In our work we used wrapper approach with some variation.

C. Fitness function

Here we present fitness function with two objective first one is select the chromosome with highest fitness value and second select the chromosome with lower number of feature. This fitness function is called feature selection with Multiobjective fitness function. From the given set of classes we will calculate the average fitness value of every subset belongs to particular class. Feature of the classifier which has highest fitness value will proceed for the next generation.

D. Genetic Operation

Here we use GP operator to evaluate next generation. Here we use reproduction, crossover and mutation. In next generation 10% population is generated by reproduction. In reproduction selected individuals copies itself into new generation. After that 80% population is generated by crossover. Crossover requires two different individuals and produce different individuals. Remaining 10% produce by mutation. The mutation operator is applied to the single individuals to make small change.

E. Termination of GP

The genetic programming life cycle is terminated when we get feature subset which can correctly classify the sample.

IV. Conclusions

Feature selection is a problem that has to be addressed in many areas. The main issues in developing feature selection techniques are selecting a small feature set in order to reduce the cost and running time of a given system, as well as achieving high fitness. This has led to the development of a variety of techniques for selecting an optimal subset of features from a larger set of possible features. The genetic programming approach exploited in this study, in particular gene expression programming, proved to be very effective for generating analytic models that can simultaneously serve two important functions: behave as classifiers in high-dimensional feature spaces, and act as general dimensionality reducers. The obtained experimental results are very promising, but preliminary. Automatically classifying data with high dimensionality is a challenge. This challenge is even more difficult when dealing with skewed datasets since methods commonly used for reducing dimensionality by feature selection are usually biased towards the larger class, when the goal is usually to maximize classification effectiveness in the smaller classes. A large and fruitful effort has been performed during the last years in the adaptation and proposal of univariate filter FS techniques. In general, we observe that many researchers in the field still think that filter FS approaches are only restricted to univariate approaches. The proposal of multivariate selection algorithms can be considered as one of the most promising future lines of work for the skewed data. The aim of the research paper here was to understand better their strengths and limitations of GA-based feature selection algorithms and to use that knowledge to develop more robust approaches. In this paper we propose a general solution for a more effective feature selection strategy, which in addition to providing a highly effective selection of important features, is also robust to skewed data.

Acknowledgment

The authors express their thanks to Assistant Prof. Manoj Dhawan, SVITS, Indore, who has helped in different ways to complete this work. Thanks are also due to the anonymous reviewers and our colleagues for their suggestions to improve the manuscript.

References

[1] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA: MIT Press, 1992.

[2]] W. Banzhaf, P. Nordin, R. E. Keller, and F. D. Francone, *Genetic Programming:An Introduction*.

[3] N. S. Chaudhari, Anuradha Purohit and Aruna Tiwari, "*A multiclass classifier using Genetic Programming*", *10th International Conference on Control, Automation, Robotics and Vision*, 17-20 Dec. 2008, pp. 1884-1887.

[4] N. S. Chaudhari, Anuradha Purohit and Aruna Tiwari, "*A multiclass classifier using Genetic Programming*", *10th International Conference on Control, Automation, Robotics and Vision*, 17-20 Dec. 2008, pp. 1884-1887.

[5] Anuradha Purohit, Narendra S. Chaudhari and Aruna Tiwari, "*Construction of Classifier with Feature Selection Based on Genetic Programming", *978-1-4244-8126-2/10/$26.00 ©2010 IEEE.

[6] F. Sebastiani and C. N. D. Ricerche, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, pp. 1–47, 2002.

[7] Bing Xue, Mengjie Zhang and Will N. Browne "*Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach*" IEEE TRANSACTIONS ON CYBERNETICS, VOL. 43, NO. 6, DECEMBER 2013.

[8] Durga Prasad Muni, Nikhil R. Pal, Senior Member, IEEE, and Jyotirmoy Das "*A Novel Approach to Design Classifiers Using Genetic Programming*" IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 8, NO. 2, APRIL 2004.

[9] Yang Zhang, Peter I. Rockett. "A generic optimising feature extraction method using multiobjective genetic programming" Applied Soft Computing 11 (2011) 1087–1097.

[10] Isac Sandin, Guilherme Andrade, Felipe Viegas, Daniel Madeira and Leonardo Rocha *Aggressive and Effective Feature Selection using Genetic Programming,* WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia.

[11] Jianjun Yu, Jindan Yu, Arpit A. Almal, Saravana M. Dhanasekaran, Debashis Ghosh, William P. Worzel and Arul M. Chinnaiyan "Feature Selection and Molecular Classification of Cancer Using Genetic Programming" Vol. 9, No. 4, April 2007, pp. 292 – 303.

[12] Yvan Saeys, Inaki Inza and Pedro Larran aga *“A review of feature selection techniques in bioinformatics”* Vol. 23 no. 19 2007.

[13] Fawaz A. Alsulaiman, Nizar Sakr, Julio J. Vald´es, Abdulmotaleb El Saddik, Nicolas D. Georganas *“Feature Selection and Classification in Genetic Programming: Application to Haptic-based Biometric data”* Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009).