Novel approach to resolve data disparity problem inGenetic Programming

Jitendra Soni

Institute of Engineering & Technology, Devi Ahilya Vishwavidyalaya, Indore, M.P., India.

ABSTRACT

The deranged dataset is one in which a class is represented with very few example, this reduces efficiency of classifier, to be a deranged dataset in any case one class should be represented with very few examples, no of solutions like data case, bagging, cost assessment of model Genetic Programming (GP) based some methods been proposed in the papers. Researchers have introduced many methods to improve the efficiency of classifier for deranged data.

In this paper, a function is planned to handle data disparity problem, by modifying learning algorithm but the original dataset remain intact. Based on Darwin's theory of natural selection, GP as a learning algorithm used to evolve classifiers by applying various GP operators. Distance Parameter is introduced to address data disparity problem. Distance parameter classification is achieved so that performance of classifier is measured in each class of datasets. Fitness value calculated for data disparity problem is taken as input along with parameter where values of parameters are predefined, to evolve in number of generations those classifier having less number of nodes with good fitness values are preferred.

This paper represents various problems with their solutions introduced by researchers to improve the performance of classifier, various advantages and disadvantage of different techniques are studied to make reasonable comparison between them.

Keywords

Dataset, Classifier, Fitness function, Data Sampling. Code Blot

1. INTRODUCTION

We Method of classification is highly used in data mining techniques, such as face detection in video images word recognition, and different biometrics and for certain medical conditions diagnosis system it is highly used [1]. Automatic classification systems are desired to classify the data in dataset of problems.

Work of Classifiers is to classify data of the datasets according to class labels. Good performance of classifier will be achieved if datasets are balanced. Dataset is deranged if it has in any case one class that do not contains sufficient no of example [3]. In deranged datasets class ratio is considerable enough that classifier became partial with few classes). Uneven distribution of class examples can leave learning algorithms with recital bias condition,[4], [5].success rate depends on important training criteria, it may depends on alternate and mainstream classes, for those problems who have minimum examples is alternate class, alternate class is equally important.

Genetic Programming (GP) [1] is an algorithm procedure stimulated from biological method to locate computer programs that perform a user-defined task. In GP, with the help of computer program a solution can be represented. GP has been used successfully in various applications such as data mining, artificial intelligence, crypt-analysis, computer automated design electronic circuit design etc. One of the popular areas where GP finds its application is in classifier design. Classifiers designed using GP classify datasets according to class labels. Good performance of classifier will be achieved if datasets are balanced. In deranged datasets class ratio is significant enough that classifier became partial with some classes. Important training criteria like total success or error rate can be influenced by the maximum number of examples of the mainstream class.

2. DATA DISPARITY

A disparity dataset[2] contains in any case one class that is represented with a handful examples and other classes contains rest of the data, these classes are called alternate and mainstream classes, this may lead to performance bias because mainstream of the data influence the decision of classifier. Performance bias means accuracy is low for alternate class and performance is high for mainstream class.

Many solutions have been provided by M.Zhang[3], M.Johnston[4], E.Smith[5] to handle data disparity problem in this paper further improvement is achieved.

3. RELATED WORK

Researchers have proposed two common approach to solve the data disparity using GP. First is to improve training criteria which is more sensitive to the data, so it can provide more appropriate result for class distribution. Overall accuracy improves and further help to improve efficiency, second approach is to assigning misclassification cost to incorrectclass prediction. Cost adjustment generally leads to creating a new fitness function that reward on its accuracy for alternate and mainstream class and penalizing those classes with poor fitness [8][9].

A way to control data disparity problem suggested by M. Zhang et al. [3] by using Area under Curve preparation criteria in GP. Area under Curve is a metric useful to assess classifier recital, generating the Area under Curve requires multiple performance points (thresholds) which are computationally costly to produce. Proposed Formula used to signify performance point is given in (1):

$$\frac{\sum_{i=0}^{N_{min}} \sum_{j=0}^{N_{maj}} I(xi,yj)}{N_{min} * N_{maj}}$$

Where,

Nmin is number of examples in alternative class;

Nmaj is number of examples in mainstream class.

Area under Curve conducts a series of pair wise comparisons on an example-by-example basis between alternative class x and mainstream class y examples collecting "rewards" (1 point) for those cases in which indicator function I(x, y) enforces constraints.

M. Zhang et al. [4] discussed a fitness function in GP as given in (2) to evaluate the performance of population of classifier based on the alternate and mainstream classes presented in dataset.

$$\frac{\sum_{i\in Min}\sum_{j\in Maj} I(P_i, P_j)}{Min * Maj}$$

Where, comparison done between the alternate and main stream classes output with the series of genetic program output.. It measures the ordering of alternate to mainstream class outputs. It calculates fitness where P i and P j represent genetic program outputs when evaluated on an example from the alternate and mainstream classes, respectively.

M. Johnston et al. [3] presented new training criteria to estimate personage class performance with complete performance. Fitness function to measure performance of classifier is given in (3):

$$\frac{hits_{min}}{N_{min}} + \frac{hits_{maj}}{N_{maj}} + \frac{hits}{N}$$

Where, hitsmin and hitsmaj represent the number of correctly classify examples in alternate and mainstream class respectively. The proposed technique is to improve the complete performance of genetic program we have to consider both the classes instead of considering single. N, Nmin, Nmaj represents the training examples in dataset, alternate class, mainstream classes respectively.

M. Zhang et al. [5] evolve varied ensembles using Genetic Programming for classification with deranged data.. If a solution gets dominated by other classes then they get affected by the result otherwise the solution result will get the importance

It is identified in the work done that multiclass data disparity problem is not addressed effectively by researchers using GP. Generally one against all approach is used in multiclass classification problems. One against all approach does not provide class separation between classifiers.

4. PROPOSED APPROCH

To improve the efficiency (performance) of classifier Genetic Problem is designed , existing method is evolved further in which, distance parameter and weightage for distance class is calculated to enhance the efficiency and performance. Cost adjustment is implemented, adopted adjustment improves performance.

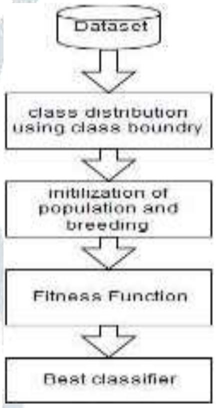

Dataset
class distribution using class boundry
initilization of population and breeding
Fitness Function
Best classifier

Fig 1 Composition architecture of planned architecture

In Fig. 1, Using dynamic boundary selection method class boundary are defined, after selection of boundary, alternate and mainstream classes are defined , that is called class distribution, using genetic programming population of classifier are evolved using reproduction, crossover and mutation then fitness function is used to assess fitness of classifier designed using genetic programming.

Class Boundary Identification: Dynamic boundary selection method also known as dynamic range selection is used to identify the boundaries of classes in datasets. Class boundaries are determined during evolution phase in Genetic Problem.

Initialization of Population: To generate the initial population of programs ramped half each approach [1] is used. In this approach half of the population is generated using grow method and half of the population is generated using full method.

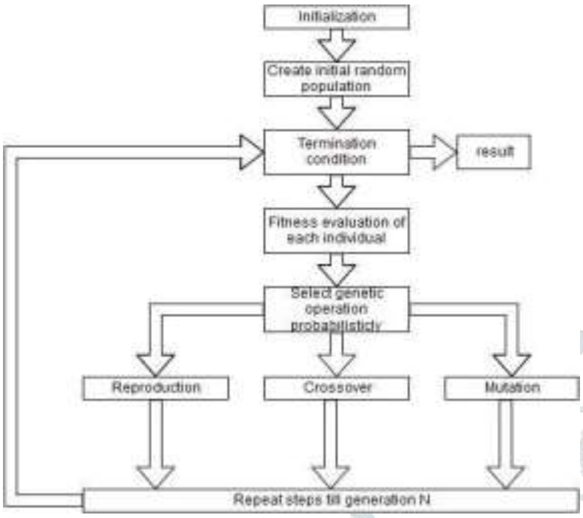
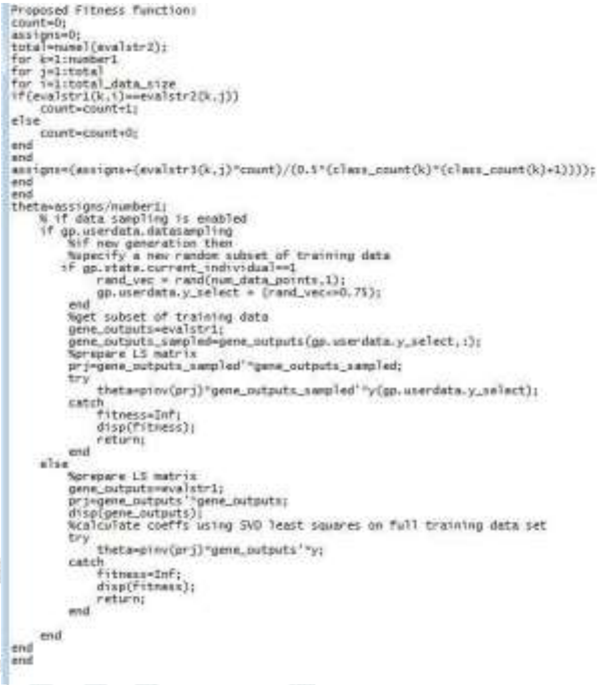

Initialization
Create initial random population
Termination condition
result
Fitness evaluation of each individual
Select genetic operation probabilisticly
Reproduction
Crossover
Mutation
Repeat steps till generation N

Fig2 Process flow

Fig 2 shows a complete process of GP of the proposed project, first initialization of population is done based on these feature random population generated using rammed half and half method. Fitness function will then calculate the fitness of classifier and for the next generation of population, reproduction, crossover mutation is done. This process is repeated till the maximum fitness is achieved.

5. PROPOSED FITNESS FUNCTION



```
Proposed Fitness function:
count=0;
assigns=0;
total=numel(evalstr2);
for k=1:number1
for j=1:total
for i=1:total_data_size
if(evalstr1(k,i)==evalstr2(k,j))
    count=count+1;
else
    count=count+0;
end
end
assigns=(assigns+(evalstr3(k,j)*count)/(0.5*(class_count(k)*(class_count(k)+1))));
end
end
theta=assigns/number1;
    % if data sampling is enabled
    if gp.userdata.datasampling
        %if new generation then
        %specify a new random subset of training data
       if gp.state.current_individual==1
            rand_vec = rand(num_data_points,1);
            gp.userdata.y_select = (rand_vec<=0.75);
        end
        %get subset of training data
        gene_outputs=evalstr1;
        gene_outputs_sampled=gene_outputs(gp.userdata.y_select,:);
        %prepare LS matrix
        prj=gene_outputs_sampled'*gene_outputs_sampled;
        try
            theta=pinv(prj)*gene_outputs_sampled'*y(gp.userdata.y_select);
        catch
            fitness=Inf;
            disp(fitness);
            return;
        end
    else
        %prepare LS matrix
        gene_outputs=evalstr1;
        prj=gene_outputs'*gene_outputs;
        disp(gene_outputs);
        %calculate coeffs using SVD least squares on full training data set
        try
            theta=pinv(prj)*gene_outputs'*y;
        catch
            fitness=Inf;
            disp(fitness);
            return;
        end

    end
end
end
```

This code is written in MATLAB, it is the proposed fitness function. It multiplies the no of individual truly identified by classifier for the alternate class to the value of the distance from the class boundary.

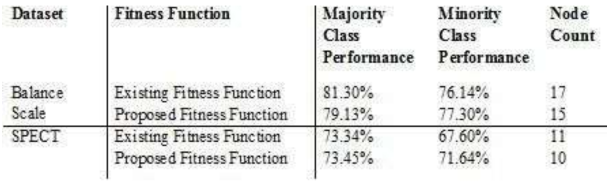
6. DATASET USED

Single-proton emission computed tomography (SPECT): It contains 267 records derived from SPECT pictures. There are unit fifty five abnormal records identified and 212 traditional records identified with ratio of 21 and 79 respectively, an disparity ratio of approximately 1:4. Each SPECT Image is processed to produce 44 recorded features. They are further processed and now contain only 22 features.

Balance scale: It contains 625 records for psychological experiments. Classified into 3 classes. The balance scale tipped to the left , right or balanced. Disparity ratio of dataset is approximately 1:12. There are 4 attributes regarding to the right and left weights and the left and right distances [12].

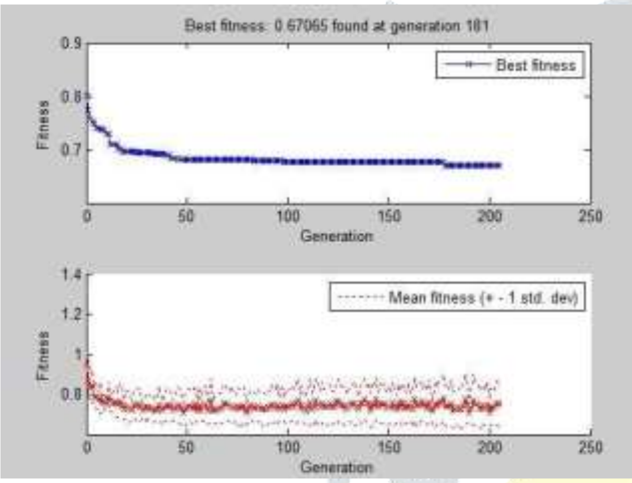
Dataset	Number of Features	Number of Classes	Number of Instances	Number of Majority Classes	Number of Minority Classes
Balance Scale	4	3	625	2	1
SPECT	22	2	267	1	1

7. RESULT



Dataset	Fitness Function	Majority Class Performance	Minority Class Performance	Node Count
Balance Scale	Existing Fitness Function	81.30%	76.14%	17
	Proposed Fitness Function	79.13%	77.30%	15
SPECT	Existing Fitness Function	73.34%	67.60%	11
	Proposed Fitness Function	73.45%	71.64%	10

The process is performed for existing fitness function and proposed fitness function on two datasets improvement is shown. Proposed fitness function perform better than the existing fitness function, this is achieved because we have separated the classes alternate and mainstream and higher weight given to those individual who are far from boundary line this is called incremental reward point. Elimination of all disadvantages of previous methods in alternate class is the cause of better performance of this proposed system.



8. CONCLUSION

This paper presents a fitness function to handle data disparity problem in classification In Genetic Programming, code blot problem in GP may exist, it is the natural phenomena of GP to increase tree size , Benchmark dataset SPECT and Balance Scale are used, to test the performance. Performance improved by 4.0% and 1.2 % respectively for SPECT and Balance Scale as compared to existing approach. In future GP operations will be analyzed to reduce the code blot and enhance the overall GP performance that will help to improve the efficiency in classification.

9. REFERENCES:

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

[2] Urvesh Bhowan, Mark Johnston and Mengije Zhang Developing New Fitness Functions in Genetic Programming for Classification With Deranged Data" IEEE Transaction on system, man and cybernetics part b, volume 42, 2012, pp 406-421.

[3] M. Zhang and W. Smart, "Multiclass object classification using genetic programming," in Proc. Applied Evolutionary Computation. vol. 3005, LNCS, 2004, pp. 369–378.

[4] U. Bhowan, M. Johnston, and M. Zhang, "A comparison of classification strategies in genetic programming with deranged data," in Proc. 23rd Australasian Joint Conference Artificial Intelligence vol. 6464, LNCS, J. Li, Ed., 2010, pp. 243–252.

[5] M. Kubat and S. Matwin, "Addressing the curse of disparityd training sets: One-sided selection," in Proc. 14th International Conference Machine Learning., 1997, pp. 179–186.

[6] A. Orriols and E. Bernado-Mansilla, "Class disparity problem in UCS classifier system Fitness adaptation," in Proc. IEEE Congress Evolutionary Computation. 2005, vol. 1, pp. 604–611.

[7] Macline, R. F., "An empirical study of boosting in nural networks," Pattern Recognition., vol. 3, 1997, pp. 1401–1405.

[8] T. Soule and R. Heckendorn, "An analysis of the causes of code growth in genetic programming," Genetic Programming Evolvable Machine., vol. 3, no. 3, 2003, pp. 283–309.

[9]M. Brameier and W. Banzhaf, "Neutral variations cause bloat in linear GP," in Proc. 6th European Conference Genetic Programming (EuroGP '03), Berlin/Heidelberg: Springer, 2003, pp. 286–296.

[10]Bhardwaj, "Controlling the problem of bloating," Machine Learning conference, vol. 4, 2011, pp. 59-68.

[11]N. Chawla, A. Kalocz, "Learning from disparity datasets," ACM, vol. 6, 2004, pp. 1–6

[12]https://archive.ics.uci.edu/ml/datasets.html