Novel ways of Improving Accuracy and Performance in Ensemble Classifiers with Multiple Unbalanced Data

Praveena Prabakaran

(Computer Science and Engineering, Anna university Regional Centre, Coimbatore/Anna University, Tamil Nadu, India)

ABSTRACT : *Imbalance classification problem is considered to be one of the emergent challenges in machine learning algorithm. This problem occurs when the number of examples that represents one of the classes of the dataset is much lower than the other classes. A multi objective genetic programming approach to evolving accurate and diverse ensembles of genetic program classifiers with good performance on both the minority and majority of classes. Six benchmark binary classification problems are taken in the existing work. The main objective of the proposed work multiclass datasets are taken to improve the accuracy of minority class and two classes can be classified and each majority and minority class has specified value. The two popular Pareto-based fitness schemes in the multi objective genetic programming algorithm, SPEA2 and NSGAII can be effective in evolving a good set of non dominated solutions in some tasks, this performance needs to be improved for difficult classification problems. The importance of developing an effective fitness evaluation strategy in the underlying MOGP algorithm to evolve good ensemble members.*

Keywords *-Classification, Class imbalance learning, Geneticprogramming, Multi-objective machine learning.*

1. Introduction

Data mining an interdisciplinary field of computer science is the process of discovering new patterns from large data sets. Data mining also used to extract knowledge from a data set in a human-understandable structure. Data mining refers to extracting or “mining” knowledge from large amounts of data. Selects data for each data set, the number of examples, number of attributes, class name of each class (minority and majority),etc.The benefit of data mining is to turn this newfound knowledge into actionable results, such as increasing a customer’s likelihood to buy, or decreasing the number of fraudulent claims.

Classification with imbalanced data-sets is considered to be a new challenge for researches. Classification with unbalance data using the accuracy of the minority and majority class as learning objectives. Classifiers have good accuracy on the majority class, but very poor accuracy on the minority class. This problem is of great interest because it turns up in many real-world classification problems, such as remote-sensing, pollution detection, risk management, fraud detection, and especially medical diagnosis.

Imbalanced data sets (IDS), also referred to as class imbalance learning, correspond to domains where there are many more instances of some classes than others. Classification on IDS always causes problems because machine learning algorithms tend to be overwhelmed by the large classes and ignore the small ones. Most classifiers operate on data drawn from the same distribution as the training data, and assume that maximizing accuracy is the principle goal. In recent years, class imbalance problem has emerged as one of the challenges in data mining community. Imbalance learning techniques shown to be less effective or even cause a negative effect in dealing with multi-class tasks. It is one of the important problem in data mining.

Genetic Programming, like many other Multilayer techniques, can evolve classifiers “biased” toward the majority class when data are unbalanced. Biased classifiers have strong classification accuracy on one class but weak accuracy on the other. This can occur because typical training criteria such as the overall accuracy or error rate can be influenced by the larger majority class. Most learning algorithms obtain a high predictive accuracy over the minority class, but predict poorly over the minority class.

Genetic programming (GP) is an evolutionary algorithm-based methodology inspired by biological evolution to find computer programs that perform a user-defined task. Genetic programming approach used to provide consistent data. This program automatically selects fitness functions using this function duplicate records are eliminated. It efficiently maximize the identification record replica while avoiding making mistakes during the process. This approach is to automatically find effective reduplication function, even when the most suitable similarity function for each record attribute is not known in advance.

In a multi-objective optimization (MOO) problem, one optimizes with respect to multiple goals or fitness functions. Multi-objective optimization (also known as multi-objective programming, vector optimization, multicriteria optimization, multiattribute optimization or Pareto optimization) is an area of multiple criteria decision making, that is concerned with mathematical optimization problems involving more than one objective function to be optimized simultaneously. Identify solution is the ultimate goal of multi-objective optimization algorithm. Due to its impossible size there is a problem in identifying the entire Pareto optimal set. In combinatorial optimization problems, proof of solution optimality is computationally infeasible. Therefore, a multi-objective optimization is to investigate a set of solutions that represent the Pareto optimal set as much as possible.

Multi-Objective Genetic Programming (MOGP) approach to evolve a Pareto front of classifiers along the optimal trade-off surface representing minority and majority class accuracy for binary class imbalance problems. MOGP framework for classification with unbalanced data using the accuracy of the minority and majority class.

Multi-objective optimization performs more than one objective function to be optimized simultaneously. In practical problems, there can be more than three objectives.

The Non-dominated Sorting Genetic Algorithm (NSGA) that solves non-convex and non-smooth single and multi-objective optimization problem. The Strength Pareto Evolutionary Algorithm (SPEA) is a relatively recent technique for finding or approximating the Pareto-optimal set for multi-objective optimization problems.

Non-dominated Sorting Genetic Algorithm-II (NSGA-II) and Strength Pareto Evolutionary Algorithm 2 (SPEA-2) are popular approaches to generating Pareto optimal solutions to a multi-objective optimization problem. NSGA-II and SPEA-2 are the evolutionary algorithms which are standard approaches; even though there are some schemes based on particle swarm optimization and simulated annealing that is significant. To solve multi-objective optimization problems, the main advantage of evolutionary algorithms is that they typically generate sets of solutions, that allows computation of an approximation of the entire Pareto front. Lower speed is the disadvantage of the evolutionary algorithms. From this it is well known that none of the generated solutions dominates the others.

2. Related Work

Our work starts with pre-processing the datasets, classification of minority and majority classes, calculate the fitness function, measure the ensemble diversity and then performance will be evaluated.

2.1 SystemArchitecture

Fig.1 shows that the Multi-objective algorithm works effectively with imbalanced domains. Dataset can be selected from the database and stored. After loading the profile data we apply preprocessing technique.

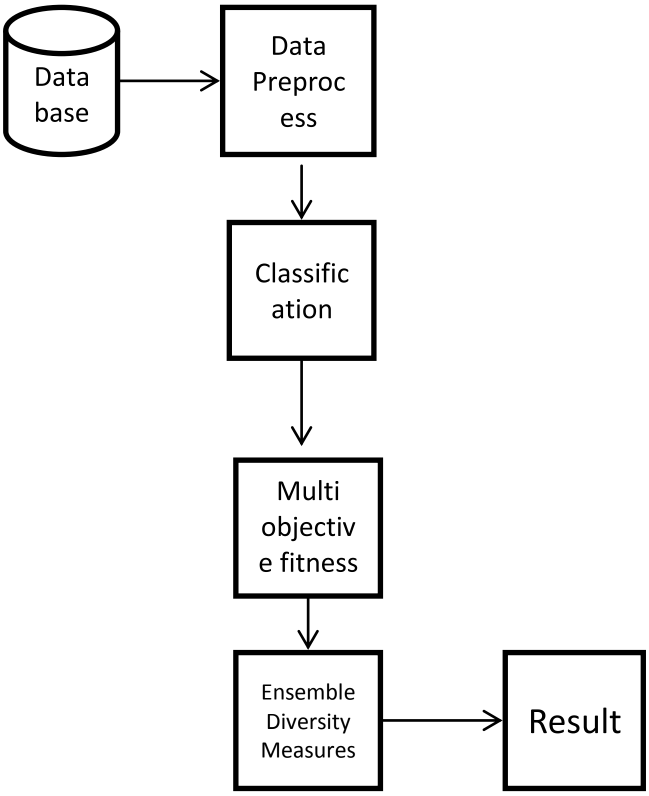


Fig 1.System Architecture

Preprocessing includes data cleaning and data reduction. Quality decisions must be based on quality data. In our project collected profile data contains errors and outliers inconsistent: containing discrepancies, null values, impossible data combinations etc. During the training phase to avoid redundant and irrelevant information it takes more processing time.So preprocess the history data and extract the actual contents.

Classification is used to classify each item in a set of data into one of predefined set of classes or groups. The attribute values of the dataset are classified based on the classes. In that each class has the majority and minority classes. Solutions to be ranked according to their performance on all the objectives with respect to all solutions in the population. This ranking is important as it affects the way selection is performed if the objectives are to be treated separately in the evolution.The strength of all its dominators is determined by the fitness value of a given solution. The final fitness value for solution is the sum of all dominance counts of other solutions in the population that are dominated.

The diversity objective in fitness to encourage the evolved solutions to make different errors on different inputs, the ensemble members are not guaranteed to be diverse with respect to their predictions. To adapt the MOGP approach to incorporate a diversity objective into the fitness function, aiming to reward solutions that have better diversity with better fitness values.

2.1.1 Data Preprocessing

The important issues related to imbalanced classification by describing the pre-processing technique. Preprocessing includes data cleaning and data reduction. Applying a pre-processing step in order to balance the class distribution is a suitable solution to the imbalanced data set problem.

Six benchmark binary classification problems are used in the experiments. For each task, half of the examples in each class were randomly chosen for the training and the test sets. Original dataset is ensured by both training and test sets that is preserved in the same class imbalance ratio. These benchmark data sets are carefully selected to encompass a varied collection of problem domains to ensure that our evaluation of the different MOGP approaches is not problem-specific. These problems have varying levels of class imbalance and complexity where some tasks are easily separable compared to others.

2.1.2 Classification Of Minority And Majority Class

To define a set of data into one of predefined set of classes or groups classification is used which it classifies the each item. Classification with unbalanced data using the accuracy of the minority and majority class as learning objectives. The attribute values of the dataset are classified based on the classes. In that each class has the majority and minority classes. Solutions to be ranked according to their performance on all the objectives with respect to all solutions in the population. This ranking is important as it affects the way selection is performed if the objectives are to be treated separately in the evolution. Evolutionary multi-objective optimization (EMO) offers a useful solution to the problem of optimizing multiple conflicting objectives. The final fitness value for solution is the sum of all dominance counts of other solutions in the population that are dominated.

2.1.3 Multi-objective Fitness Function

Evolutionary multiobjective optimization (EMO) offers a useful solution to the problem of optimizing multiple conflicting objectives.Pareto-Based Dominance Measures

Pareto-based dominance measures in fitness, i.e., dominance rank and dominance count, in different ways to evolve Pareto fronts. Two common Pareto-based dominance measures are the dominance rank and dominance count of a given solution. Dominance rank is the number of other solutions in the population that dominate a given solution (lower is better), whereas dominance count is the number of other solutions that a particular solution dominates (higher is better). Two popular EMO approaches that use these measures include SPEA2 and NSGAII; SPEA2 uses both dominance rank and dominance count, while NSGAII uses only dominance rank.

In NSGAII, the fitness value for the solution is its dominance rank, i.e., the number of other solutions in the population that dominate. A non-dominated solution will have the best fitness of 0, while high fitness values indicate poor-performing solutions, i.e., solutions dominated by many individuals.

SPEA2 uses both dominance rank and dominance count.The strength of all its dominators is determined by the fitness value of a given solution. The final fitness value for solution is the sum of all dominance counts of other solutions in the population that are dominated. Similar to NSGAII, fitness here is to be minimized where non-dominated solutions have the best fitness.

2.1.4 Evolving Ensembles Using MOGP

The diversity objective in fitness to encourage the evolved solutions to make different errors on different inputs, the ensemble members are not guaranteed to be diverse with respect to their predictions. To adapt the MOGP approach to incorporate a diversity objective into the fitness function, aiming to reward solutions that have better diversity with better fitness values. To investigate two measures to promote the evolution of diverse solutions in the population, NCL and PFC.

NCL: The first measure to encourage diversity among the individuals in the population uses NCL as a correlation penalty term in the fitness function. NCL measures the phenotypic differences between the solutions in the ensemble and the rest of the population. The NCL measure, calculates the average correlation penalty for each class, for a given solution in the population:

$$\text{NCL}=\frac{1}{2}\sum_{c=1}^{k}\left(\frac{1}{MNc}\sum_{i=1}^{Nc}(G_i^p-E_i)\left[\sum_{j=1,j\neq p}^{M}(G_i^j-E_i\right]\right)(1)$$

where

$$G_i^p=\frac{1}{1+e^{gp}}(2)$$

K is the number of classes, and Nc is the number of training examples in class c. G_i^p is the processed output and gp is the raw output of genetic program p when evaluated on the ith example in class c. Ei is the output of the ensemble on the ith example in class c, i.e., 1 or 0 to denote a minority or a majority class label, respectively. The ensemble output Ei is a majority vote of the predicted class labels of each ensemble member. The ensemble size, i.e., the number of nondominated solutions in the current generation, is given by M. The lower the NCL values, the better the diversity of the solutions.

PFC: PFC is a population level diversity measure. The second diversity measure is also used as a penalty function,but unlike the NCL, PFC is a population level diversity measure. This means that PFC measures the errors (on the training set) of each solution with respect to all other solutions in the population; whereas NCL compares the outputs of a solution to the ensemble only:

$$PFC_{c,p}=\frac{1}{T-1}\sum_{j=1,j\neq p}^{T}\frac{\sum_{i=1}^{Nc}I(gp_i^p,gp_i^j)}{Err_c^p+Err_c^j}(3)$$

where

$$\text{I}\ (gp_i^p,gp_i^j)=\begin{cases}1,\ \text{if}\ \text{pred}(gp_i^p)\neq\text{pred}(gp_i^j)\\0,\ \text{otherwise}\end{cases}(4)$$

and

$$\text{pred}\ (gp_i^p)=\begin{cases}1,\ \text{if}\ gp_i^p\geq 0\ \text{(i.e. minority class)}\\0,\ \text{otherwise}\ \text{(i.e. majority class)}\end{cases}(5)$$

Nc is the number of training examples in class c, and gp_i^p is the raw output of genetic program p when evaluated on the ith example in class c and T is population size. Indicator function I(·) returns 1 if the predicted class label between two solutions is different for a given input, or 0 otherwise; this is used to compute the Hamming distance between the predictions of two genetic programs on all inputs in class c. The errors, Err_c^p andErr_c^j, are the number of incorrect predictions in class c for two solutions p and j in the population. An incorrect prediction occurs when the predicted and actual class labels differ for a given input.

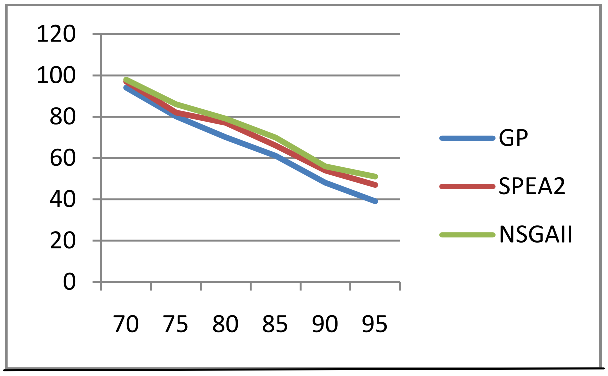
2.1.5 MOGP Evaluation Using Ensemble-Diversity Measures

The different MOGP ensemble performances, first investigate what effect the diversity objectives have on the hyper area of the evolved Pareto fronts. The average hyperarea of the Pareto-approximated fronts from the three approaches is used to statistically test the null hypothesis.

In classification the standard fitness measure is the over-all classification accuracy.The classification with unbalanced data, Acc can favor the evolution of solutions biased towards the majority class.Accuracy does not taken into account because, the smaller number of examples is present in the minority class. For example, if a classification task has a minority class represented by only 10% of available learning instances, a trivial solution can score a high fitness by assigning all the instances to the majority class.

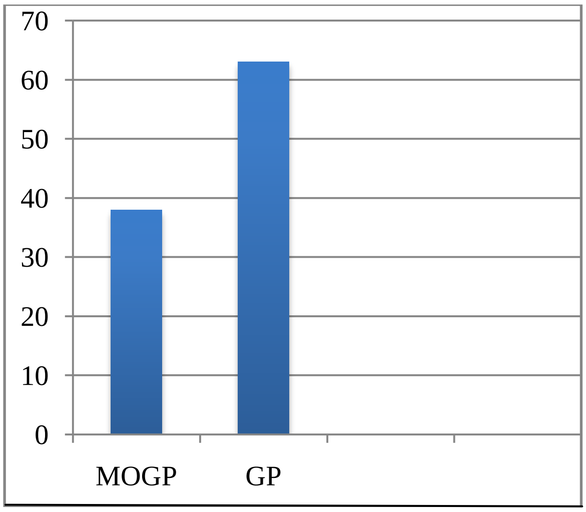
2.1.6 Performance Evaluation

The performance can be measured in terms of classification accuracy. The accuracy can be measured to calculate AUC value. Then compared to both existing and proposed work.



Graph.1 Classification performance using MOGP approaches.

Graph.1 explains that x-axis is the minority accuracy and y-axis is the majority accuracy, genetic programming has the lower accuracy and the SPEA2 and NSGAII has the higher accuracy.



Graph.2 Time comparison between MOGP and GP

Proposed work (MOGP) takes less time to execute when compared to existing work(GP).

3. Results And Discussion

To develop the Pareto-based dominance measures in fitness, i.e., dominance rank and dominance count, in different ways to evolve Pareto fronts. The strength of all its dominators is determined by the fitness value of a given solution. EMO approach is that the evolved Pareto front represents highly accurate classifiers, each with a different performance bias toward either class. To adapt the MOGP approach to incorporate a diversity objective into the fitness function, aiming to reward solutions that have better diversity with better fitness values. The hyperarea of the evolved Pareto-approximated fronts as a single figure to measure which MOGP fitness scheme is better on these tasks. Hyperarea values range between 0 and 1, where the higher the value the better the performance.

Table.1

Minority and majority values using MOGP algorithm

DATASET	MOGP	MIN VALUE	MAJ VALUE
Ionosphere	NCL PFC	85.634 84.543	86.341 92.185
Vehicle	NCL PFC	80.356 80.576	81.478 82.685
Haberman	NCL PFC	78.085 80.351	80.021 81.365
Concrete	NCL PFC	89.251 88.041	90.235 91.325
Survival	NCL PFC	83.215 87.653	85.365 89.712
Iris	NCL PFC	87.254 84.564	86.569 87.145
Magic	NCL PFC	88.542 89.356	89.475 91.658
Pima	NCL PFC	86.695 84.693	89.652 88.596
Spect heart	NCL PFC	87.653 88.845	89.354 90.352

Table.1 explains the minority and majority value of two diversity measures. Here 9 dataset are taken and calculate the values.

Table. 2

Accuracy Value for Datasets

Table.2 explains that the accuracy values of majority and minority class, minority class has the higher accuracy when compared to majority.

DATASET	ACCURACY	
	MAJORITY	MINORITY
Ionosphere	46.32	66.88
Vehicle	437.79	727.55
Haberman	218.89	258.26
Concrete	75.66	151.77
Survival	46.76	69.20
Iris	23.57	32.31
Magic	33.46	55.13
Pima	128.89	157.79
Spect heart	12.16	15.58

4. Conclusion

The Multi objective genetic programming for imbalanced datasets with the aim of tackling imbalance problems effectively and efficiently. The algorithm is carried out and their results are shown better by producing balanced datasets. The imbalanced classification using MOGP algorithm increases accuracy, performance of imbalanced classification compared than the traditional algorithm.

5. Acknowledgments

It is my proud privilege to express my profound gratitude to S.Lavanya, Assistant Professor, Department of Computer Science, Anna University Regional Centre, Coimbatore, for her invaluable guidance, constant encouragement and sincere support for this work.

REFERENCES

[1] Urvesh Bhowan, Mark Johnston, *"Evolving Diverse Ensembles Using Genetic Programming for Classification With Unbalanced Data,"IEEE transactions on evolutionary computation, vol. 17, no. 3, 2013*

[2] A.Mclntyre and M.Heywood, *"Classification as clustering: A Pareto cooperative-competitive GP approach," Evol. Comput., vol. 19, no. 1, pp. 137–166, 2011.*

[3] U.Bhowan, M.Zhang, and M.Johnston, *"Genetic programming for classification with unbalanced data," in Proc. 13th Eur. Conf. Genet.Programming, LNCS 6021. 2010.*

[4] U.Bhowan, M.Johnston, and M.Zhang, *"Multiobjective genetic programming for classification with unbalanced data," in Proc. 22nd Australasian Joint Conf. Artif. Intell., LNCS 5866. 2009, pp. 370–380.*

[5] S.Wang, K.Tang, and X.Yao, *"Diversity exploration and negative correlation learning on imbalanced data sets," in Proc. Int. Joint Conf. Neural Netw., 2009, pp. 3259–3266*

[6] N.Chawla and J.Sylvester, *"Exploiting diversity in ensembles: Improving the performance on unbalanced datasets," in Proc. 7th Int. Conf. MCS, 2007, pp. 397–406.*

[7] A.Mclntyre and M.Heywood, *"Multiobjective competitive coevolution for efficient GP classifier problem decomposition," in Proc. IEEE Int. Conf. Syst., Man, Cybern., Oct. 2007, pp. 1930–1937.*

[8] E.Alfaro-Cid, K.Sharman, and A.Esparcia-Alcazar, *"A genetic programming approach for bankruptcy prediction using a highly unbalanced database," in Applications of Evolutionary Computing (LNCS, vol. 4448), M. Giacobini, Ed. Berlin, Germany: Springer, 2007, pp. 169–178.*

[9] C.Coello Coello, G.Lamont, and D.Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic and Evolutionary Computation Series). Berlin, Germany: Springer, 2007.*

[10] A.Chandra and X.Yao, *"Ensemble learning using multiobjective evolutionary algorithms," J. Math. Modelling Algorithms, vol. 5, no. 4, pp. 417–445, 2006.*

[11] G.Batista, R.C.Prati, and M.C.Monard, *"Balancing strategies and class overlapping," in Proc. 6th Int. Adv. IDA, LNCS 3646. 2005, pp. 24–35.*

[12] N.Japcowicz and S.Stephen, *"The class imbalance problem: A systematic study," Intell. Data Anal., vol. 6, no. 5, pp. 429–450, 2002.*

[13] E.Zitzler, M.Laumanns, and L.Thiele, "*Spea2: Improving the strength Pareto evolutionary algorithm for multiobjective optimization," Dept. Electr. Eng., Swiss Federal Instit. Technol., Zurich, Switzerland, TIK Rep. 103, 2001.*

[14] M.Brameier and W.Banzhaf, *"Evolving teams of predictors with linear genetic programming," Genet. Programming Evolvable Mach., vol. 2, no. 4, pp. 381–407, 2001.*

[15] X.Yao and Y.Liu, *"Making use of population information in evolutionary artificial neural networks," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 28, no. 3, pp. 417–425, Jun. 1998.*

[16] Y.Liu and X.Yao, *"Negatively correlated neural networks can produce best ensembles," Australian J. Intell. Inform. Process. Syst., vol. 4, nos. 3–4, pp. 176–185, 1997.*

[17] D.W.Opitz and J.W.Shavlik, *"Generating accurate and diverse members of a neural-network ensemble," in Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1996, pp. 535–541.*

[18] C.Gathercole and P.Ross, *"Dynamic training subset selection for supervised learning in genetic programming," in Proc. 3rd PPSN, LNCsS 866. 1994, pp. 312–321.*