A COMPARISON BETWEEN SYNTHETIC OVER-SAMPLING EQUILIBRIUM AND OBSERVED SUBSETS OF DATA FOR EPIDEMIC VECTOR CLASSIFICATION

Authors: Terence Fusco, Yaxin Bi, Chris Nugent, Shengli Wu

Introduction

- In this work we provide results of data mining and machine learning techniques, which form the basis of our prediction model for snail density classification in relation to the Schistosomiasis epidemic disease.
- All experiments to date are cognitive components in the development of our prediction model for the epidemic disease Schistosomiasis. This disease is detrimental to the health of the communities of affected areas as well as the crop and cattle life. If detected for early warning of the disease, the local communities can be better prepared to deal with any consequences of a breakout.
- This report gives an insight into the relationship between using a snapshot sample of environment data for epidemic disease vector classification, as opposed to the construction of an increased synthetic dataset.
- The synthetic data instances used are created based on the original real-world data we have using a modified version of the Synthetic Minority Over-Sampling Technique (SMOTE). The rationale behind proposing SMOTE is based on the fact that although we potentially have access to vast sources of satellite imagery with which to perform calculations for classification and prediction, this may not be the most sufficient, to achieve the greatest performance in possible classification accuracy. We have carried out testing on each year of training and testing in which we have modified the SMOTE method to achieve an equilibrium of snail density classes in order to provide balance in the sample and eliminate the likelihood of overfitting.
- The problem we faced of partially complete data was initially addressed in the previous dragon symposium workshop looked at some initial methods of data imputation to assess the precision of the replacement values. The results of which showed that the proposed Double PreSuccession method proved to be the most accurate with replacing values.
- In this report we have also tested this method against an alternative approach which uses a regression based method for replacing missing data. Both methods were tested and compared with the standard Weka value replacement as a benchmark. The most accurate replacement method will be used as we proceed with any future missing values in a dataset.

Objectives

- To compare and improve on the replacement accuracy of existing data imputation methods.
- Discover the most appropriate Snail Density class distribution and category paradigm for classification purposes.
- Analyse and assess the efficacy of the SMOTE Equilibrium approach to address the sparse training data problem area.

Methods

- Data Imputation Methods
  - The application of the Regression CTA approach in comparison to existing methods for replacement of missing data instances from satellite imagery.
- Weighted Distribution Snail Density
  - Comparative analysis using varying ratios of snail density to discern the most appropriate balance of SD levels to pursue for future classification and prediction methods.
- SMOTE Equilibrium Application
  - The application of the SMOTE Equilibrium approach is applied to each dataset ranging from 100% to 1000% of the original instance number in order to analyse the difference made to the classification accuracy process by increasing the training potential.

Data Imputation Methods

Continuing on from our previous research into data imputation for partially complete datasets [3], we have investigated another method namely Regression CTA which uses the R^2 and Pearson's r values of the missing data then regression on the values the set in order to replace the missing value. We randomly removed data entries in each of our 5 years of data and replaced each value using WEKA missing value filter, our previous mean Double PreSuccession method and the new Regression CTA.

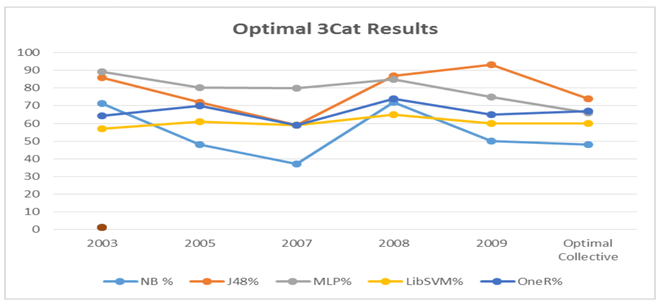
2003 Dataset			
Original Value	Weka Value	Mean Double PreSuccesion Value	Regression CTA Value
0.0348207	0.072316	**0.02367**	0.109414
0.316076	0.497184	-0.00868	**0.310440377**
-0.137255	-0.088585	0.07194	**-0.102642**
0.424546	0.497184	-0.12167	**0.4303997**
-0.0761676	-0.088585	0.11097	**-0.074591**
0.822625	0.497184	-0.14418	**0.81588**
-0.128532	-0.088585	0.09024	**-0.12002**
0.868613	0.497184	-0.16331	**0.83721**

Weighted Distribution SD

When classifying snail density levels using training data, first we must decide how to appropriately label each snail density class. The standard method which we have been using is to pre-process the data by normalising the collective SD levels and evenly distributing the classes based on either a 3 or 5 category paradigm.

The problem with this approach is that it can produce an imbalance in the spectrum of results therefore we decided to re-distribute the SD levels using a variety of weighted ratios to address this issue.

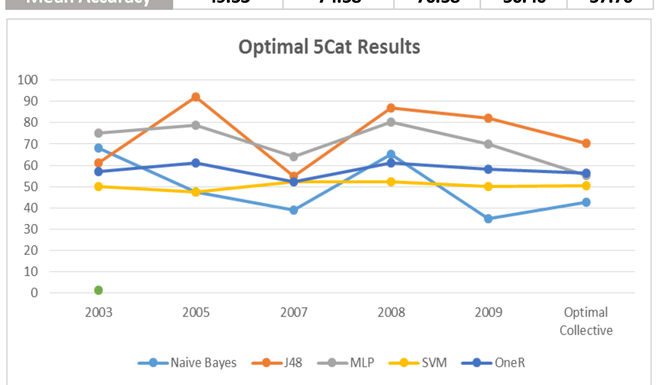
➢ 3 Category Weighted Distribution with a ratio of 20-low, 60-medium and 20-high snail density classes.



Year of Data	NB %	J48%	MLP%	LibSVM%	OneR%
2003	71.4	86	89.3	57.1	64.3
2005	48	72	80.4	61	70
2007	37	59.1	80	59.1	59.1
2008	72	87	85	65.1	74
2009	50	93.3	75	60	65
Optimal Collective	48	74	66	60	67
Mean Accuracy	**54.40**	**78.57**	**79.28**	**60.38**	**66.57**

➢ 5 Category Weighted Distribution with a ratio of 5-V.Low, 20-Low, 50-Medium, 20-High and 5-V.High snail density classes

Year Of Data	Naive Bayes	J48	MLP	SVM	OneR
2003	68	61	75	50	57.1
2005	47.4	92.1	79	47.4	61
2007	39	55	64	52.3	52.3
2008	65.2	87	80.4	52.2	61
2009	35	82	70	50	58.3
Optimal Collective	42.6	70.4	55.1	50.5	56.5
Mean Accuracy	**49.53**	**74.58**	**70.58**	**50.40**	**57.70**



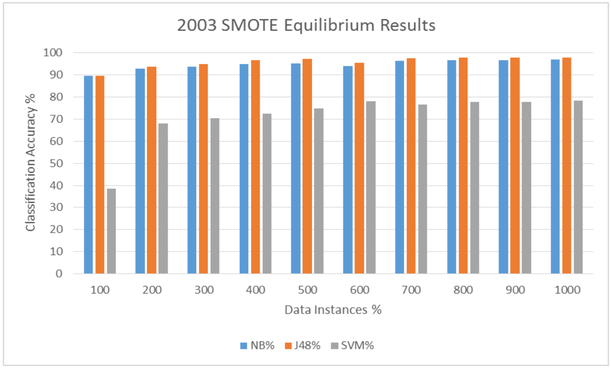
SMOTE Equilibrium

During SMOTE Equilibrium experiments conducted in this research, we aim to establish whether the classification accuracy of snail density levels can be improved with an increased dataset of environmental attributes. To convey this hypothesis we used the data provided by our partners and for each year of collected data, we applied an incremental process using the SMOTE method then equalizing the number of Snail Density classes to avoid overfitting. The process tested from 100% to 1000% of the original set size with the 3 classifiers Naïve Bayes, J48 and SVM being applied at each stage.

The main objectives of this research are to analyse and assess how well we can use synthetically created data instances for vector density prediction and disease classification. If we can prove that this SMOTE equilibrium model is a viable approach for vector density classification then the problem of data collection and sparse training data will be nullified for future research.

We can see from the SMOTE results using year 2003 data that with each multiple of synthetic data instances added, the classification accuracy increases slightly. In terms of SVM, we can see that it performs poorly in relation to the NB and J48 classifiers.

When conducting future testing using the SMOTE Equilibrium method I feel that SVM could be replaced with another classifier without having a negative impact on predictions or losing any competitive accuracy results.



Results

- Data Imputation
  - The data imputation results show that the Regression CTA method provides the most accurate data replacement when compared with the Weka model and the previously suggested Mean Double PreSuccession method. We can apply this Regression CTA method to future missing data instances when required with confidence in the accuracy level of replacement for snail density classification.
- Weighted Distribution SD
  - 3CAT results using a distribution of 20-60-20 yield the highest performing classification results with a total average of 64.15% again showing that the majority of results in the middle of the data prove to be the most fruitful for classification accuracy.
  - We can see from the optimal 5CAT results which performed best at 5-20-50-20-5 that the highest mean accuracy was present with J48 at 74.58%. This shows that we have a more balanced snail density when the higher distribution is around the middle of the dataset.
- SMOTE Equilibrium
  - The SMOTE Equilibrium results show a gradual increase in classification accuracy in correspondence to the multiplication of the synthetically added data in most cases. SVM performs poorly in relation to Naïve Bayes and J48 for classification accuracy and can be discontinued for future experiments.

Summary

- We can see that the data imputation approach using the Regression CTA has performed more favourably when compared with the alternative methods on this dataset. We now have the evidence to show that this method is viable moving forward with further research in this area.
- The weighted distribution experiments have provided us with a more balanced and appropriate ratio for snail density classification purposes when using either the 3 or 5 category combination. The most desirable results are found when using 3 categories of SD with the weighted distribution of classes being 20-60-20. This information reflects the optimum classification accuracy across the data range and can be applied to any novel environment feature dataset pertaining to Schistosomiasis vector classification.
- The SMOTE Equilibrium proposed method has yielded a slight increase with each multiple of synthetic instances that are compounded to the training dataset. The reduction of overfitting and increase of data instances has shown a gradual classification accuracy increase across the data for each year. We will now test to see what the optimum synthetic instance incremental increase is across our data and apply this to our experiments with this research.

References

[1] Chawla, N. V. et al., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1), pp.321–357
[2] Rivera, W.A. & Kincaid, J.P., 2014. OUPS : a combined approach using SMOTE and Propensity Score Matching. , 1(2), pp.4–7.
[3] Fusco, T. & Bi, Y., 2015. A Cumulative Training Approach to Schistosomiasis Vector Density Prediction.