CLASSIFICATION OF AMBROSIA POLEN USING AN ARTIFICIAL NEURAL NETWORK ON AN IMBALANCED DATA SET

Andrijana Pešić[1]

Abstract: Working with a multiclass data set often results in a class imbalance problem. This research is related to implementing a neural network for ambrosia pollen classification, and comparing in conjuction with resampling techniques, i.e. removing existing classes of the dominant class and applying the SMOTE algorithm. Experimental results show that the neural network is able to produce better results without using resampling techniques by analyzing following metrics: weighted accuracy, weighted response, weighted F1 measure, Kappa statistic and RMSE.

Key word: classification, imbalanced dataset, neural network, undersampling, SMOTE alogrithm, metrics.

1. INTRODUCTION

Unbalanced data is a problem that is ubiquitous in real life situations. The classification problems of unbalanced classes are challenging for researchers, because most algorithms are intended for classes that have an equal distribution. The solutions are focused on sampling techniques and new algorithms for solving a given problem. The usual metrics used to evaluate the model cannot be used for unbalanced classes, so other metrics are used such as: accuracy, response, F1 measure, Kappa statistics, and mean square error.

The problem of unbalanced classes is one of the significant problems in the field of data mining, [1]. Research in the field of neural networks shows that unbalanced classes are a critical factor in the performance of classifiers when working with data that contain multiple classes, ie. when the number of samples of one class is much smaller in relation to other classes, [2]. Most research is focused on the problem of imbalance of the two classes, while this problem is particularly interesting in the context of several classes, [3]. One of the reasons why the minority class is misclassified is the application of global performance measures, such as accuracy, which can give preference to the dominant class, [4]. There are many papers that deal with resampling techniques using classification. In [1] the author usees classification algorithms with sampling techniques SMOTE, SpreadSubsample, Reasample to predict adverse outcomes of albendazole, where the percentage of the majority class is 55, and of the minority is 0.4%. The application of these algorithms has improved the performance of the learning algorithm where the following performance measures were used: precision, response, F-measure and root mean square error. In [2] a multilayer perceptron on five unbalanced data sets is used as a classifier. The results show that global metrics give good results, and at the same time individual class metrics result in a poor minority class classifier, Other authors [5] state that Kappa statistics are more appropriate than accuracy on unbalanced data sets, In [6] uses the SMOTE technique is used to balance minority class instances, by applying machine learning algorithms to various data sets from the UCI repository, which has proven to be the best in classification, Other authors propose a hybrid sampling method combining the SMOTE technique of removing existing instances of the dominant class by using clusters on unbalanced classes, [7].

After a brief methodology, results and discussion give the neural network model and model evaluation on data with actual distribution, then using resampling technique and finally using SMOTE technique, comparing results by metrics.

[1] PhD student, University of Kragujevac, Faculty of Technical Sciences Čačak, e-mail:andrijana90pesic@gmail.com

2. METHODOLOGY

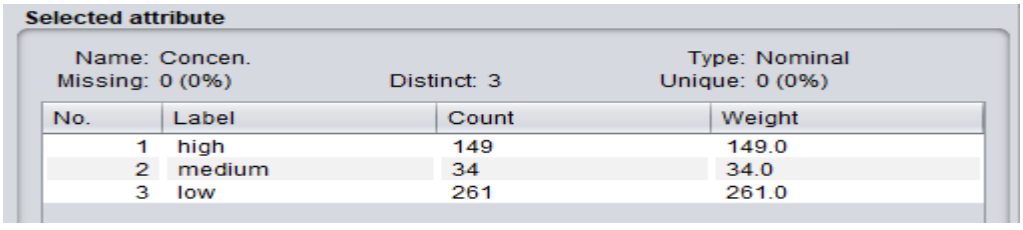
For the purpose of creating a neural network, two sets of data were collected. The first data set refers to the concentration of ambrosia pollen in the territory of Vršac for the period 2011-2016 which were downloaded from [8]. Values for the pollen concentration were already defined as high, medium and low. The second set of data is meteorological. Due to the geographical position of the Pannonian Plain and the unavailability of data for the mentioned city, meteorological data for Novi Sad were downloaded from [9]. The data contain minimum temperature, maximum temperature, wind speed and precipitation. Based on the minimum and maximum temperature, the mean temperature was also calculated. Meteorological data were matched for available days when the pollen concentration was measured creating a final set of data to be analyzed. The neural network model with defined input parameters, number of hidden layers, output parameters was tested first on actual data, then using the undersampling technique and at the end the SMOTE technique was used. 70% of the data were selected for neural network training, while the remaining data were used for testing.

3. RESULTS AND DISCUSSION

3.1. Neural network model

The data were processed and the neural network created using the WEKA program WEKA (Waikato Environment for Knowledge Analysis) is a machine learning software written in Java, developed by the University of Waikato. This is free software and is licensed from GNU General Public License, [10]. After import, the data were normalized in the range [0,1]

The multilayer perceptron was chosen for the classifier, because the used data were already classified. The initial class distribution is shown in Figure 1. The nodes in this network are sigmoid, and the neural network is created with the following parameters: hidden layers: 1, learning rate: 0.1, momentum: 0,2 and validation treshold: 20. The input data are: month, day, year, precipitation, maximum temperature, minimum temperature, mean temperature, wind speed. Based on the input data, the neural network should classify pollen concentrations according to the following classes: high, medium and low, Figure 2. All attributes are significant with respect to the concentration class.


Selected attribute

Name: Concen. Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	high	149	149.0
2	medium	34	34.0
3	low	261	261.0

Figure 1 – Initial class distribution

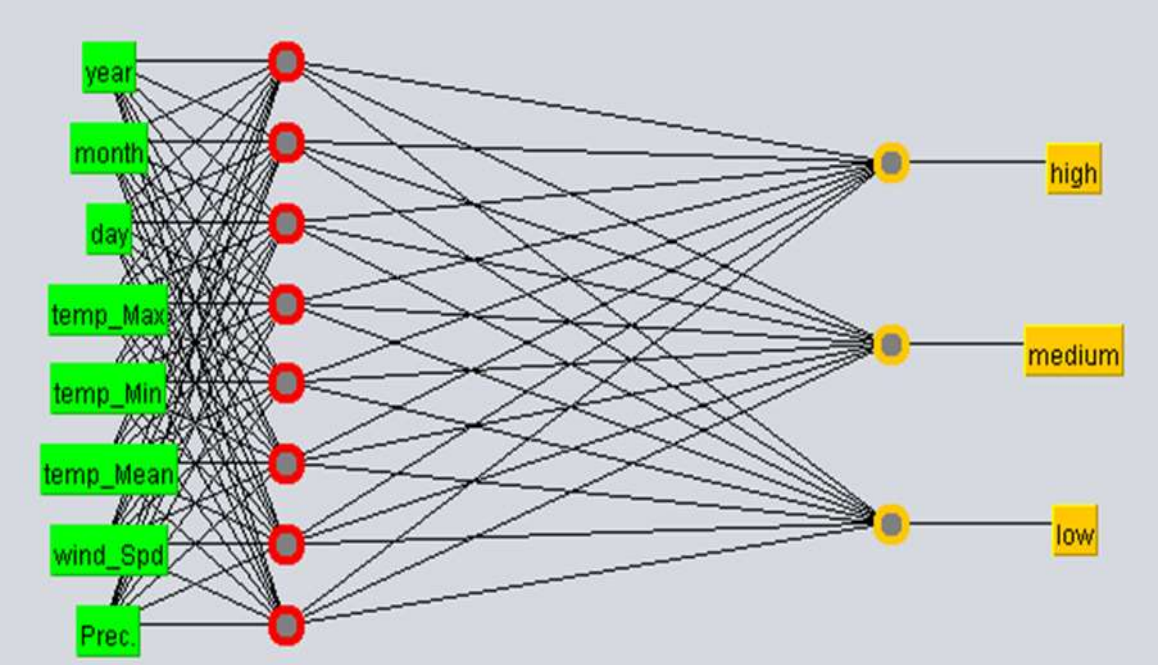


Figure 2 – Neural network model using dependent variables high, medium and low

3.2. Neural network evaluation

Undersampling is the removal of existing instances of the dominant class, while SMOTE (Synthetic Minority Over-sampling Technique) is a technique that creates synthetic data of the minority class, i.e. adds synthetic instances to the minority class. Beside evaluation on actual data (distribution of classes H:112, M:28, L:171), undersampling technique (distribution of classes H:28, M:28, L:28) and SMOTE technique (distribution of classes H:112, M:112, L:171) are used on train data, and then the neural network model is tested on unknown data.

By comparing results of all models, it is seen that weighted precision (0.875), weighted recall (0.895) and weighted F-Measure (0,883) are the greatest in the model with actual distribution. The value of Kappa statistic (0.7735) shows excellent agreement between the individual parameters examined and it has analytical significance. The root mean squared error (0.2119) is the smallest in comparison to the neural network uisng undersampling and the SMOTE technique. Also, the number of correctly classified instances is the greatest (119) of all models, Figures 3-5.

```
Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances         119               89.4737 %
Incorrectly Classified Instances        14               10.5263 %
Kappa statistic                          0.7735
Mean absolute error                      0.1175
Root mean squared error                  0.2119
Relative absolute error                 33.5708 %
Root relative squared error             52.5404 %
Total Number of Instances              133

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.946    0.094    0.795      0.946    0.864      0.812    0.989     0.976
                 0.000    0.016    0.000      0.000    0.000      -0.027   0.878     0.198
                 0.933    0.070    0.966      0.933    0.949      0.849    0.989     0.995
Weighted Avg.    0.895    0.074    0.875      0.895    0.883      0.799    0.984     0.954

=== Confusion Matrix ===

  a  b  c   <-- classified as
 35  0  2 |  a = high
  5  0  1 |  b = medium
  4  2 84 |  c = low
```

Figure 3 – Results of neural network model on actual distribution, without resampling

```
Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances          98               73.6842 %
Incorrectly Classified Instances        35               26.3158 %
Kappa statistic                          0.528
Mean absolute error                      0.2855
Root mean squared error                  0.3847
Relative absolute error                 64.2471 %
Root relative squared error             81.6153 %
Total Number of Instances              133

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.919    0.229    0.607      0.919    0.731      0.626    0.945     0.866
                 0.167    0.102    0.071      0.167    0.100      0.043    0.706     0.129
                 0.700    0.000    1.000      0.700    0.824      0.656    0.990     0.995
Weighted Avg.    0.737    0.068    0.849      0.737    0.765      0.620    0.964     0.920

=== Confusion Matrix ===

  a  b  c   <-- classified as
 34  3  0 |  a = high
  5  1  0 |  b = medium
 17 10 63 |  c = low
```

Figure 4 – Results of neural network model using undersampling

```
Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances         114               85.7143 %
Incorrectly Classified Instances        19               14.2857 %
Kappa statistic                          0.7135
Mean absolute error                      0.1312
Root mean squared error                  0.2664
Relative absolute error                 31.9644 %
Root relative squared error             60.6187 %
Total Number of Instances              133

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.973    0.135    0.735      0.973    0.837      0.778    0.947     0.818
                 0.000    0.047    0.000      0.000    0.000      -0.047   0.560     0.056
                 0.867    0.000    1.000      0.867    0.929      0.823    0.986     0.994
Weighted Avg.    0.857    0.040    0.881      0.857    0.861      0.771    0.956     0.903

=== Confusion Matrix ===

  a  b  c   <-- classified as
 36  1  0 |  a = high
  6  0  0 |  b = medium
  7  5 78 |  c = low
```

Figure 5 – Results of neural network model using SMOTE technique

4. CONCLUSION

According to the presented results the following can be concluded:

- The neural network is powerfull and can perform better results by itself, without using sampling techniques.
- The classification of ambrosia pollen can be useful for patients who have allergies.

Further research can be directed towards the application of the neural network with other classification algorithms for each season separatly.

5. REFERENCES

[1] Yıldırım, P. (2016). *Pattern classification with imbalanced and multiclass data for the prediction of albendazole adverse event outcomes*. The International Workshop on Data Mining for Decision Support (DMDMS 2016), pp. 1013-1018.

[2] R. Alejo, J. S. (2008). *An Empirical Study for the Multi-class Imbalance Problem with Neural Networks.* CIARP (pp. 479-486). Berlin Heidelberg: Springer-Verlag.

[3] Rendón, E., Alejo, R., Castorena, C., Isidro-Ortega, F., & Granda-Gutiérrez, E. (2020). *Data Sampling Methods to Deal With the Big Data Multi-Class Imbalance Problem*. Appl. Sci., pp. 10,1276.

[4] Victoria López, A. F. (2013). *An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics*. Information Sciences. 250, pp. 131-141. Elsevier Inc.

[5] Mehrdad Fatourechi, R. K. (2008). *Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets*. 2008 Seventh International Conference on Machine Learning and Applications (pp. 777-782). San Diego, CA: IEEE.

[6] Mohd F., A. J. (2019). *Improving Accuracy of Imbalanced Clinical Data Classification Using Synthetic Minority Over-Sampling.* Communications in Computer and Information Science. 1097. Springer, Cham: In: Alfaries A., Mengash H., Yasar A., Shakshuki E. (eds) Advances in Data Science, Cyber Security and IT Applications.

[7] A. Agrawal, H. L. (2015). *SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling*. 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), (pp. 226-234). Lisbon. Preuzeto sa http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7526924&isnumber=7526872

[8] "SEPA," [Online]. Available: http://data.sepa.gov.rs/dataset/koncentracija-polena-od-2011-do-2016-godine. [Accessed 5 May 2020].

[9] "Agri4Cast Resources Portal,"[Online]. Available: https://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx. . [Accessed 5 May 2020].

[10] Remco R. Bouckaert, E. F. (2017). *WEKA Manual for Version 3-8-2. Hamilto*n, New Zealand: University of Waikato.