Ensemble and Fuzzy techniques applied to Imbalanced Traffic Congestion Datasets: a Comparative Study

Pedro Lopez-Garcia[1,2], Antonio D. Masegosa[1,2,3], Enrique Onieva[1,2], Eneko Osaba[1]

[1] DeustoTech-Fundacion Deusto, Deusto Foundation, 48007, Bilbao, Spain
[2] Faculty of Engineering, University of Deusto, 48007, Bilbao, Spain
[3] IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain.
{p.lopez, ad.masegosa, e.onieva, e.osaba}@deusto.es

Abstract. Class imbalance is among the most persistent complications which may confront the traditional supervised learning task in real-world applications. Among the different kind of classification problems that have been studied in the literature, the imbalanced ones, particularly those that represents real-world problems, have attracted the interest of many researchers in recent years. In order to face this problems, different approaches have been used or proposed in the literature, between then, soft computing and ensemble techniques. In this work, ensembles and fuzzy techniques have been applied to real-world traffic datasets in order to study their performance in imbalanced real-world scenarios. KEEL platform is used to carried out this study. The results show that different ensemble techniques obtain the best results in the proposed datasets.

Keywords: Intelligent Transportation Systems, Imbalanced Data, Ensemble techniques, Fuzzy techniques, Soft Computing techniques, Classification

1 Introduction

Class imbalance is among the most persistent complications which may confront the traditional supervised learning task in real-world applications [26]. The problem appears when the number of instances in one of the classes significantly outnumbers the number of instances in the other ones. This situation is a handicap when trying to identify the minority class, as the learning algorithms are not usually adapted to such characteristics. Without the loss of generality, it can be assumed that the class of interest is the minority class, while the other ones are the majority ones. Various applications demonstrate this characteristic of high class imbalance, such as bioinformatics, e-business, information security, and national security.

Among the different kind of classification problems that have been studied in the literature, the imbalanced ones, particularly those that represents real-world problems, have attracted the interest of many researchers in recent years

[34, 35]. In particular, in traffic environments, the apparition of a particularly complicated state of the road (i.e. traffic congestion) will represent a minority class for prediction algorithms, while its proper detection in advance is a topic of interest for administrations and users.

One of the most problematic issues in the development of actual cities is road traffic. This problem is actually one of the most important study focuses of the Intelligent Transportation Systems (ITS) field. In the last decades, intelligent techniques such as those mentioned before have been applied to solve this problem. In particular fuzzy systems are used in [25] to infer the future state of the road by combining several systems in a hierarchical way. In addition different metaheuristics have been used in order to optimize systems, such as Support Vector Machines [16] (SVM); Genetic Algorithms (GA) are used in [6], while Particle Swarm Optimization (PSO)is implemented in [39], among others.

Recently, ensemble learning is a popular and significant research in data mining and machine learning area. Ensemble classifiers have received considerable attention in applied statistics and machine learning for over a decade [4]. Several studies demonstrate that the practice of combining several models into a aggregated one leads to significant gains in performance over its constituent members [10].

In this work, ensembles and fuzzy techniques have been applied to real-world traffic datasets. As principal aim, the objective is to study their performance in imbalanced real-world scenarios, comparing different and recent approaches. As other objectives, we can highlight the introduction of real-traffic datasets and its use for research purposes. Data used in this work come from two sources. The first one comes from cameras in the city of Helmond (The Netherlands) collected by TASS International company [1] and took part of the developing of different models for traffic systems in Horizon 2020 TIMON project [2] (Enhanced real time services for optimized multimodal mobility relying on cooperative networks and open data). Another data source used for the development of this work is the data obtained in Lisbon (Portugal) A5 highway, and used in the European Project ICSI (Intelligent Cooperative Sensing for Improved Traffic Efficiency).

The rest of the paper is structured as follow. Section 2 contains the state of the art of the two kind of techniques applied in this work: ensembles and metaheuristics. Section 3 is dedicated to the descriptions of the different methods used for this comparative study. In Section 4 information about the datasets used and its comparative is shown. Finally, in Section 5 the conclusions obtained for this study are collected.

2 Background

In this section, a brief study of the state of the art is presented in order to show the contributions of the community to the imbalance data problem using

[1] https://www.tassinternational.com/

[2] https://www.timon-project.eu/

ensembles (Section 2.1), in specially boosting and bagging algorithms, and metaheuristics (Section 2.2).

2.1 Ensembles

Ensemble learning is defined as the use of multiple learning algorithms to obtain better predictive performance that could be obtained from any of these algorithms alone [33]. Over the last decade, this kind of approach has been used in different themes such as optimization [28], medicine [41], or ITS [30]. Focusing in imbalance classification problems, these algorithms can be found in many articles. For example, in [24], Lim et al. propose a evolutionary cluster-based oversampling ensemble framework. This method is based on contemporary ideas of identifying oversampling regions using clusters. The evolutionary part of the ensemble is used to optimize the parameters of the data generation method and to reduce the overall computational cost. The proposal is applied to a set of 40 imbalance datasets.

Among the different ensemble techniques, two of them can be frequently found in the literature applied to several themes: bagging and boosting techniques [22]. While in bagging several models are created using different subsets of the training set [5], in boosting, a set of weak learning algorithms create a single strong learner and produce only one model [13]. Both kind of methods have been used in imbalance classification.

Authors in [10] analyze different corrective and total corrective boosting algorithms in order to present its own boosting algorithm adding a strong classifier to the linear constraints of LPBoost. Besides, in [8], an Adaboost algorithm to learn fuzzy-rule-based classifiers is proposed. Adaboost approach is applied to approximate and descriptive fuzzy-rule bases, and the performance of the proposed method is compared with other classification schemes applied on a set of benchmark classification tasks.

Other example can be found in[23]. This article presents a research about the Roughly Balanced Bagging and its basic properties that can influence its classification performance. Variables such as the number of component classifiers, their diversity, and ability to deal with difficult types of the minority examples are studied. The experiments are carried out using synthetic and real life data.

The number of articles related with this theme is wide extended in the literature, which means that it is an active issue. In this section, some interesting examples have been exposed, but, in order to give more information and related articles about the problem we are dealing with, interested reader are referred to [20], [27], and [38] for different surveys about this issue.

2.2 Soft Computing techniques applied to imbalance datasets

Soft Computing techniques have been widely used since its presentation in 90's by Zadeh [40]. Machine Learning, Fuzzy Logic, and Evolutionary Computation methods are inside the vast group of Soft Computing techniques. Techniques such as GAs, SVM, Fuzzy Rule Based Systems, PSO and so on, have been developed and applied to different themes along the years, showing their good performance and the huge range of possibilities that they offer.

Regarding Fuzzy Logic techniques, fuzzy logic methods have been used in imbalance cases of study along the years. For example, in [3], a fuzzy technique is developed to predict heart diseases. The technique is divided in three phases: first, a fuzzy c-means clustering algorithm is used. Then, rules are generated from the rough set theory, and those rules are used for prediction with the fuzzy classifier.

Another case can be found in [17], where linguistic Fuzzy Rule Based Systems have been applied to imbalance datasets to deal with the overlapping problems between the concepts to be learned. This problem is more severe in imbalance datasets due to the most of the techniques try to correctly classify the majority class and, in cases of imbalance distribution of the data, it is the minority class where the most important data can be found. Datasets used are extracted from KEEL dataset repository.

Finally, authors of this study are aware of the huge amount of related papers that can be found in the literature. In this work, we have mentioned some of the most interesting research papers, in order to give an idea of the activity that is being carried out in the community. For further information, we recommend the reading of any of the review papers that can be found in the literature, such as [21], or [32]. In this work, fuzzy methods will be used to study their performance in a real imbalance scenario.

3 Techniques used for the comparative study

As mentioned in previous sections, one of the aims of this work is to study the performance of ensembles and fuzzy meta-heuristic techniques when they are applied to imbalanced problems. A total of 10 techniques are chosen, divided in two principal groups: six ensemble techniques, and four fuzzy ones. Due to the limited space, only the name of the techniques as well as a brief description of them are listed below:

- Ensemble techniques
  1. AdaBoost (I) [12] is an adaptation of general Adaboost for imbalance datasets.
  2. MSSMOTE Bagging [14] oversamples minority class instances using MSMOTE preprocessing algorithm. In this method both classes contribute to each bag with N instances.
  3. MSSMOTE Boosting [19] introduces synthetic instances in each iteration of AdaBoost technique, using the MSMOTE data preprocessing algorithm.
  4. RUSBoost [36] removes instances from the majority class by random undersampling the data-set in each iteration.
  5. SMOTE Bagging [37] oversamples minority class instances using SMOTE preprocessing algorithm.

6. SMOTE Boosting [7] introduces synthetic instances in each iteration of AdaBoost technique, using the SMOTE data preprocessing algorithm.

  All ensemble techniques used in this work have C4.5 algorithm as base classifier.

- Fuzzy Classification techniques
  1. AdaBoost (C) [9] is a boosting algorithm, which repeatedly invokes a learning algorithm to successively generate a committee of simple, low-quality classifiers.
  2. LogitBoost [29] is a backfitting algorithm, which repeatedly invokes a learning algorithm to successively generate a committee of simple, low-quality classifiers.
  3. FARCHD-C [1] mines fuzzy association rules limiting the order of the associations in order to obtain a reduced set of candidate rules with less attributes in the antecedent.
  4. C4.5 [31] is a decision tree generating algorithm that it induces classification rules in the form of decision trees from a set of given examples. C4.5 is based on ID3 algorithm.

It is important to remark that the different between $AdaBoost(C)$ and $AdaBoost(I)$ is the base classifier. While the first one counts with fuzzy classifiers, the second one uses a C4.5 algorithm as base classifier.

4 Experimentation

This section compiles the experimentation carried out in this work. Datasets used in this work as well as the information related to them are exposed in Section 4.1 while the results, and statistic methods applied are summarized in Section 4.2.

4.1 Datasets and preprocessing

Datasets used in this work contains real data from traffic cameras in the city of Helmond (The Netherlands). This data is provided by TASS international [1] and used in the Horizon 2020 project TIMON project[2] (Enhanced real time services for optimized multimodal mobility relying on cooperative networks and open data). Congestion in the road is used as class variable. In the raw data, this variable can take four different values: Normal, Increasing, Dense and Congestion. In order to simplify and make the problem equal to the techniques mentioned in the previous section, the classes have been reduced by two: Normal (majority class) and Congestion value (minority class), which includes Increasing, Dense and Congestion instances. Each dataset counts with a total of 22 variables, which includes not only information about the speed, the number of vehicles or the occupancy of the road, but the weather when data was taken. Data used in this

[1] https://www.tassinternational.com/

[2] https://www.timon-project.eu/

work are collected during two months by four cameras, and divided in four different horizons of time (15, 30, 45 and 60 minutes respectively), which makes a total of 16 datasets.

Besides, data collected from Lisbon highway A5 used in EU project ICSI [1] have been also used. This highway is a 25 km long motorway in Portugal that connects Lisbon to Cascais. Data used in this work was collected from seven sensors displayed in the road and transformed into datasets. As well as in Helmond datasets, congestion in the road is taken as class variable. In this case, this class contains a value of congestion that appear in the next hour at a certain point and can take as values LOW, if the number of vehicles are below the percentile 15; MED (Medium), if the it is between percentiles 15 and 30; and $HIGH$ otherwise. Following the same logic applied to previous datasets, LOW and MED instances have been labeled as Normal (mayority class) while $HIGH$ instances have been changed to Congestion label (minority class). Data was collected during a month. The three first weeks are used as training data while the last week of the month is used to validate the solutions. These datasets are called BRISA datasets along the rest of the work.

Information about Imbalance Ratio (IR) and number of instances in each dataset are shown in Table 1

	Name of Dataset	N. Instances	IR
TASS datasets	C1	5333	8.1
	C2	5338	8.2
	C28	5348	8.16
	C47	5449	7
Brisa datasets	CL_{600}	721	2.04
	CL_{1980}	1441	2.26
	CL_{3600}	721	5.43
	CL_{4000}	1441	2.57
	CL_{6800}	721	2.13
	CL_{8050}	1441	2.25
	CL_{9400}	721	2.53

Table 1. Information about the datasets used in this work

4.2 Results

KEEL software [2] has been used to carry out the experiments. In the case we are dealing with, the module for imbalanced techniques are used. The experimentations have been executed in a Intel Xeon E5 2.30 GHz with a RAM memory of 32 GB. Related with the configuration of the techniques used in the experimentation, the default configuration given by KEEL has been retained. The Area Under the Curve (AUC) has been used as error metric. To show TASS dataset

[1] http://www.ict-icsi.eu

results, datasets are divided by id of the camera and horizon of time. Those results are shown in Table 2. Bold values represent the two best results obtained in each dataset. .

Techniques	C1				C2				C28				C47			
	15	30	45	60	15	30	45	60	15	30	45	60	15	30	45	60
C4.5	**.968**	.958	.956	**.963**	**.970**	**.954**	**.955**	**.955**	**.963**	.949	.958	**.962**	**.958**	**.964**	.940	.953
FARCHD	.805	.727	.642	.623	.788	.729	.612	.575	.808	.723	.667	.500	.829	.732	.642	.622
AdaBoost(C)	.549	.563	.508	.500	.564	.614	.507	.501	.560	.540	.510	.500	.566	.541	.512	.501
LogitBoost	.672	.668	.585	.537	.703	.678	.594	.548	.659	.657	.567	.531	.685	.662	.604	.560
AdaBoost(I)	.939	.950	.957	.941	.951	.940	.950	.953	.952	.952	.938	.941	.950	.945	.930	.952
MSMOTEBagging	.872	.942	.943	.941	.886	.939	.932	.929	.903	.945	.947	.936	.912	.937	.926	.927
MSMOTEBoost	.916	.948	.935	.932	.936	.939	.925	.918	.939	.942	.940	.934	.935	.952	.934	.930
RUSBoost	**.976**	**.973**	**.971**	**.968**	**.973**	**.972**	**.967**	**.965**	**.972**	**.977**	**.968**	**.971**	**.971**	**.973**	**.968**	**.969**
SMOTEBagging	.916	.953	.936	.938	.923	.941	.934	.923	.937	.947	.943	.932	.893	.945	.926	.937
SMOTEBoost	.956	**.963**	**.960**	.954	.946	**.954**	.954	.946	.946	**.961**	**.959**	.955	.952	.960	**.958**	**.955**

Table 2. AUC values obtained for each technique in each dataset and horizon of time for TASS datasets

As it can be seen, three techniques stand out from the rest: RUSBoost, SMOTEBoost, and C4.5. In case of RUSBoost, it obtains one of the two best results in every dataset used, being the first one in each one of them. For SMOTEBoost, it gets one of the two best AUC values in 7 out of 16 datasets. Finally, for C4.5, it achieves a value between the best two in 10 out of 16 datasets, especially in C2 dataset. About the rest of the techniques, in general, ensemble techniques obtain better results than fuzzy ones. Focusing in the fuzzy techniques, though FARCHD and C4.5 achieves good performance in this problem without changing anything in its execution, AdaBoost(C) and LogitBoost do not obtain a considerable performance. In fact, AdaBoost(C) obtain the lowest AUC values in every dataset in comparison with the rest of techniques. If both AdaBoost techniques presented in this experimentation are compared, ensemble version of AdaBoost (AdaBoost(I)) outperforms the fuzzy one. On the other hand, taking into account ensemble techniques, RUSBoost outperforms the rest of them, followed by SMOTEBoost. However, all the techniques obtain a good performance in every dataset and horizon of time, which always achieve an AUC value higher than 0.9. About the horizon of time, the increasing of this value does not seem to affect to the performance of the techniques significantly. Only AdaBoost(C) and LogitBoost notice the change of this value. The rest of the techniques obtains almost the same performance when the horizon of time is 15 minutes than when it takes the value 60 minutes. Some of them (SMOTEBagging, C1 dataset) even improve its performance between these two horizons.

Table 2 contains the results obtained by each one of the techniques for each BRISA dataset. As in the previous results, the two best values are highlighted in bold.

The results show that MSMOTEBagging is the best technique so far in these datasets, obtaining 4 out of 7 best values, following by RUSBoost and SMOTE-

	CL_{600}	CL_{1980}	CL_{3600}	CL_{4000}	CL_{6800}	CL_{8050}	CL_{9400}
C4.5	.893	.919	.898	.945	.872	.940	.875
FARCHD	.830	.955	.906	.928	.893	.951	.954
AdaBoost (C)	.882	.938	.808	.938	.864	.948	**.979**
LogitBoost	.853	.951	.891	.945	.884	.945	**.975**
AdaBoost(I)	.886	.954	.859	.941	.852	.957	.892
MSMOTE-Bagging	**.924**	**.961**	.928	.941	**.909**	**.962**	.867
MSMOTE-Boost	.884	.957	.899	.954	.881	**.965**	.871
RUSBoost	.902	.955	**.919**	**.955**	**.909**	.951	.896
SMOTEBagging	.914	**.958**	**.935**	**.958**	.901	.954	.875
SMOTEBoost	**.928**	.934	.915	.941	.897	.940	.921

Table 3. AUC values obtained for each technique in each Lisbon dataset

Bagging, which both obtain 3 out of 7 best results. For the rest of the techniques, about fuzzy techniques used, only AdaBoost (C) and LogitBoost obtain bold values. Although their performance is not far from those obtained by the best techniques, they do not reach the high AUC value obtained by the rest of the techniques. Adaboost (C) and LogitBoost obtain one bold value, in dataset CL_{9400}, being the two best techniques in the mentioned dataset. Comparing the results obtained in the previous datasets, in this case, bagging techniques overpass boosting techniques, being RUSBoost the only one that can be compared with the results obtained by them.

In order to assess if the differences in performance among the techniques studied here are significantly different we employed non-parametric tests following the guidelines given by Garcia et al. in [15]. The procedure carried out is described next. We first apply Friedman's non-parametric test for multiple comparison at a significance level $\alpha \leq 0.05$ to assess if we can reject the null hypothesis of similar performance among all algorithms. If so, then we evaluate if the performance of the best algorithm according to Friedman's averaged ranking versus the other classifiers is significantly better. To this end, we apply Holm's [18] and Finner's [11] post-hoc tests at a significance level $\alpha \leq 0.05$ using the best method as control algorithm. Following this procedure, we analyse the performance of the algorithms globally over the two datasets.

We do the exercise of evaluating the performance of the methods over all datasets. According to Friedman's tests there exists significant differences among algorithms. The averaged ranking displayed in Table 4 confirm that RUSBoost is the most robust classifier followed by SMOTEBoost. On the contrary, the three fuzzy algorithms are clearly the ones that show a worse performance, whereas the result of the rest of algorithms is very similar. Using RUSBoost as control algorithm for the Holm's and Finner's post-hoc tests, we observe in Table 5 that, taking into account all datasets, it obtains significantly better AUC values that the other studied methods, excepting SMOTEBoost, although even in this case the significance level is quite near to the threshold, being equal to 0.07.

Algorithm	Ranking
AdaBoost (I)	4.9783
C4.5	4.2391
FARCHD	7.5652
AdaBoost (C)	9.1739
LogitBoost	8.1087
MSMOTEBagging	5.2609
MSMOTEBoost	5.3913
RUSBoost	1.8043
SMOTEBagging	5.0652
SMOTEBoost	3.413

Table 4. Average Rankings of the algorithms provided by Friedman's non-parametric test for multiple comparisons over all datasets

Algorithm	Adjusted p-value Holm	Adjusted p-value Finner
AdaBoost (C)	0	0
LogitBoost	0	0
FARCHD	0	0
MSMOTEBoost	0.000353	0.000132
MSMOTEBagging	0.000541	0.000195
SMOTEBagging	0.001039	0.00039
AdaBoost (I)	0.001134	0.000486
C4.5	0.012778	0.007185
SMOTEBoost	0.07157	0.07157

Table 5. Adjusted p-value returned by Holm's and Finner's post-hoc tests for all datasets

5 Conclusions

In this work, ensemble and fuzzy rules techniques have been applied to imbalance real traffic datasets in order to classify correctly the state of the road in a real scenario. In this case, data collected from cameras in the city of Helmond (The Netherlands), and from A5 Highway in Lisbon are used. Data from cameras was collected by TASS international and used in H2020 TIMON project. In case of A5 highway, this data was used in ICSI project. The aim of this article is to compare the performance of ensemble and fuzzy techniques in imbalance real scenarios.

As results, in Helmond datasets, ensemble techniques outperform those fuzzy techniques used in the experimentation, with two techniques between the best ones. Three techniques stand out the rest: RUSBoost, SMOTEBoost, and C4.5. Among all, RUSBoost obtained at least one of the two best values in every dataset used. For SMOTEBoost and C4.5, they obtained 7 out of 16 and 10 out of 16 best values respectively. Regarding Lisbon datasets, ensemble techniques again, specially Bagging techniques and RUSBoost, obtain better performance than fuzzy techniques. All these results are checked using different statistical tests.

As future works, other techniques for both groups can be used. Besides, the experimentation could be applied to more datasets and other horizons of time. Regarding this, one future work to take into account is to adapt ensemble techniques to work with multiclass classification. This will increase the difficulty of the problem as well as the IR of each dataset, making the data a good real benchmark to use in comparatives like the presented in this paper.

Acknowledgements

This work has been supported by TIMON project (Enhanced real time services for optimized multimodal mobility relying on cooperative networks and open data) which received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 636220. E. Osaba performed his contribution when he was working on the University of Deusto.

Bibliography

[1] Alcala-Fdez, J., Alcala, R., Herrera, F.: A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. IEEE Transactions on Fuzzy Systems **19**(5), 857–872 (2011)

[2] Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic & Soft Computing **17** (2011)

[3] Antonelli, M., Ducange, P., Marcelloni, F.: An experimental study on evolutionary fuzzy classifiers designed for managing imbalanced datasets. Neurocomputing **146**, 125–136 (2014)

[4] Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning **36**(1), 105–139 (1999)

[5] Breiman, L.: Bagging predictors. Machine learning **24**(2), 123–140 (1996)

[6] Cervantes, J., Li, X., Yu, W.: Imbalanced data classification via support vector machines and genetic algorithms. Connection Science **26**(4), 335–348 (2014)

[7] Chawla, N., Lazarevic, A., Hall, L., Bowyer, K.: Smoteboost: Improving prediction of the minority class in boosting. Knowledge Discovery in Databases: PKDD 2003 pp. 107–119 (2003)

[8] Del Jesus, M., Hoffmann, F., Navascués, L., Sánchez, L.: Induction of fuzzy-rule-based classifiers with evolutionary boosting algorithms. IEEE Transactions on Fuzzy Systems **12**(3), 296–308 (2004)

[9] Del Jesus, M.J., Hoffmann, F., Navascués, L.J., Sánchez, L.: Induction of fuzzy-rule-based classifiers with evolutionary boosting algorithms. IEEE Transactions on Fuzzy Systems **12**(3), 296–308 (2004)

[10] Fang, Y., Fu, Y., Sun, C., Zhou, J.: Improved boosting algorithm using combined weak classifiers. Journal of Computational Information Systems **7**(5), 1455–1462 (2011)

[11] Finner, H.: On a monotonicity problem in step-down multiple test procedures. Journal of the American Statistical Association **88**(423), 920–923 (1993)

[12] Freund, Y., Schapire, R.E.: A desicion-theoretic generalization of on-line learning and an application to boosting pp. 23–37 (1995)

[13] Freund, Y., Schapire, R.E., et al.: Experiments with a new boosting algorithm **96**, 148–156 (1996)

[14] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **42**(4), 463–484 (2012)

[15] García, S., Herrera, F., Shawe-taylor, J.: An extension on —statistical comparisons of classifiers over multiple data sets‖ for all pairwise comparisons. Journal of Machine Learning Research pp. 2677–2694 (2008)

[16] Guo, L., Ge, P.S., Zhang, M.H., Li, L.H., Zhao, Y.B.: Pedestrian detection for intelligent transportation systems combining adaboost algorithm and support vector machine. Expert Systems with Applications **39**(4), 4274–4286 (2012)

[17] Harandi, F., Derhami, V.: A reinforcement learning algorithm for adjusting antecedent parameters and weights of fuzzy rules in a fuzzy classifier. Journal of Intelligent and Fuzzy Systems **30**(4), 2339–2347 (2016)

[18] Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics pp. 65–70 (1979)

[19] Hu, S., Liang, Y., Ma, L., He, Y.: Msmote: improving classification performance when training data is imbalanced. In: Computer Science and Engineering, 2009. WCSE'09. Second International Workshop on, vol. 2, pp. 13–17. IEEE (2009)

[20] Jurek, A., Bi, Y., Wu, S., Nugent, C.: A survey of commonly used ensemble-based classification techniques. Knowledge Engineering Review (2013)

[21] Kotsiantis, S., Kanellopoulos, D., Pintelas, P., et al.: Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering **30**(1), 25–36 (2006)

[22] Kotsiantis, S.B.: Bagging and boosting variants for handling classifications problems: a survey. The Knowledge Engineering Review **29**(1), 78–100 (2014)

[23] Lango, M., Stefanowski, J.: Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data. Journal of Intelligent Information Systems pp. 1–31 (2017)

[24] Lim, P., Goh, C., Tan, K.: Evolutionary cluster-based synthetic oversampling ensemble (eco-ensemble) for imbalance learning. IEEE Transactions on Cybernetics **47**(9), 2850–2861 (2017)

[25] Lopez-Garcia, P., Onieva, E., Osaba, E., Masegosa, A.D., Perallos, A.: A hybrid method for short-term traffic congestion forecasting using genetic algorithms and cross entropy. IEEE Transactions on Intelligent Transportation Systems **17**(2), 557–569 (2016)

[26] López, V., Fernández, A., Moreno-Torres, J.G., Herrera, F.: Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. Expert Systems with Applications **39**(7), 6585 – 6608 (2012)

[27] Mokeddem, D., Belbachir, H.: A survey of distributed classification based ensemble data mining methods. Journal of Applied Sciences (2009)

[28] Nama, S., Saha, A.: An ensemble symbiosis organisms search algorithm and its application to real world problems. Decision Science Letters **7**(2), 103–118 (2018)

[29] Otero, J., Sánchez, L.: Induction of descriptive fuzzy classifiers with the logitboost algorithm. Soft Computing-A Fusion of Foundations, Methodologies and Applications **10**(9), 825–835 (2006)

[30] Pescaru, D., Curiac, D.I.: Ensemble based traffic light control for city zones using a reduced number of sensors. Transportation Research Part C: Emerging Technologies **46**, 261–273 (2014)

[31] Quinlan, J.R.: C4. 5: Programming for machine learning. Morgan Kauffmann **38** (1993)

[32] Ramyachitra, D., Manikandan, P.: Imbalanced dataset classification and solutions: a review. International Journal of Computing and Business Research (IJCBR) **5**(4) (2014)

[33] Rokach, L.: Ensemble-based classifiers. Artificial Intelligence Review **33**(1), 1–39 (2010)

[34] Sardari, S., Eftekhari, M.: A fuzzy decision tree approach for imbalanced data classification. pp. 292–297 (2016)

[35] Savetratanakaree, K., Sookhanaphibarn, K., Intakosum, S., Thawonmas, R.: Borderline over-sampling in feature space for learning algorithms in imbalanced data environments. IAENG International Journal of Computer Science **43**(3), 363–373 (2016)

[36] Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: Rusboost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans **40**(1), 185–197 (2010)

[37] Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. In: Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on, pp. 324–331. IEEE (2009)

[38] Wang, S., Yao, X.: Multiclass imbalance problems: Analysis and potential solutions. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics (2012)

[39] Xu, Z., Watada, J., Wu, M., Ibrahim, Z., Khalid, M.: Solving the imbalanced data classification problem with the particle swarm optimization based support vector machine. IEEJ Transactions on Electronics, Information and Systems **134**(6), 788–795 (2014)

[40] Zadeh, L.A.: Fuzzy logic, neural networks, and soft computing. Communications of the ACM **37**(3), 77–85 (1994)

[41] Zhao, Z., Liu, Y., Li, J., Wang, J., Wang, X.: A study of fuzzy clustering ensemble algorithm focusing on medical data analysis. Lecture Notes in Electrical Engineering **422**, 383–396 (2018)