# Learning class-imbalanced problems from the perspective of data intrinsic characteristics

Kong, J.

# Appendices

# APPENDIX A

# Additional Experimental Results

Table A.1: Performance results of decision tree (C5.0) on the dataset *Contraceptive*. “1 0 1 0" represents “safe(1) borderline(0) rare(1) outlier(0)", i.e. only safe and rare samples are oversampled. $R_{min/all}$ and *TS* indicate the different rules for identifying types of samples.

| Combination | MinAcc | | MAUC | |
|---|---|---|---|---|
| | $R_{min/all}$ | TS | $R_{min/all}$ | TS |
| 1 1 1 1 | 0.4154 | 0.4154 | 0.6736 | 0.6744 |
| 1 1 1 0 | 0.3925 | 0.3503 | 0.6734 | 0.6735 |
| 1 1 0 1 | 0.3800 | 0.3846 | 0.6714 | 0.6753 |
| 1 0 1 1 | 0.3978 | 0.3690 | 0.6607 | 0.6617 |
| 0 1 1 1 | 0.3530 | 0.3695 | 0.6670 | 0.6643 |
| 1 1 0 0 | 0.4296 | **0.4360** | 0.6807 | **0.6834** |
| 1 0 1 0 | 0.3773 | 0.3518 | 0.6689 | 0.6655 |
| 0 1 1 0 | 0.3865 | 0.3882 | 0.6737 | 0.6699 |
| 1 0 0 1 | 0.3814 | 0.3932 | 0.6669 | 0.6700 |
| 0 1 0 1 | 0.3988 | 0.3950 | 0.6725 | 0.6678 |
| 0 0 1 1 | 0.3963 | 0.3605 | 0.6679 | 0.6626 |
| 1 0 0 0 | **0.4457** | 0.4360 | **0.6884** | 0.6826 |
| 0 0 1 0 | 0.3666 | 0.3688 | 0.6698 | 0.6676 |
| 0 1 0 0 | 0.3899 | 0.4207 | 0.6771 | 0.6768 |
| 0 0 0 1 | 0.4343 | 0.4040 | 0.6841 | 0.6622 |

Table A.2: Performance results of decision tree (C5.0) on the dataset *Thyroid*. “1 0 1 0" represents “safe(1) borderline(0) rare(1) outlier(0)", i.e. only safe and rare samples are oversampled. $R_{min/all}$ and *TS* indicate the different rules for identifying types of samples. “–" means that there are not enough samples to execute the $k$-nearest-neighbor algorithm in the oversampling step.

| Combination | MinAcc | | MAUC | |
|---|---|---|---|---|
| | $R_{min/all}$ | TS | $R_{min/all}$ | TS |
| 1 1 1 1 | 0.8648 | 0.8648 | 0.9813 | 0.9808 |
| 1 1 1 0 | 0.7789 | 0.7708 | 0.9829 | 0.9733 |
| 1 1 0 1 | 0.7221 | 0.7486 | 0.9726 | 0.9736 |
| 1 0 1 1 | 0.7227 | 0.7440 | 0.9703 | 0.9737 |
| 0 1 1 1 | **0.9432** | **0.9350** | 0.9831 | **0.9830** |
| 1 1 0 0 | 0.7011 | – | 0.9774 | 0.9712 |
| 1 0 1 0 | – | 0.7306 | – | 0.9765 |
| 0 1 1 0 | 0.7694 | 0.7756 | **0.9838** | 0.9815 |
| 1 0 0 1 | 0.7816 | – | 0.9735 | 0.9744 |
| 0 1 0 1 | 0.8224 | – | 0.9831 | 0.9814 |
| 0 0 1 1 | – | – | – | – |
| 1 0 0 0 | – | – | – | – |
| 0 0 1 0 | – | – | – | – |
| 0 1 0 0 | – | – | – | – |
| 0 0 0 1 | – | – | – | – |

Table A.3: Performance results of decision tree (C5.0) on the dataset *Wine*. "1 0 1 0" represents "safe(1) borderline(0) rare(1) outlier(0)", i.e. only safe and rare samples are oversampled. $R_{min/all}$ and *TS* indicate the different rules for identifying types of samples. "–" means that there are not enough samples to execute the $k$-nearest-neighbor algorithm in the oversampling step.

| Combination | MinAcc | | MAUC | |
|---|---|---|---|---|
| | $R_{min/all}$ | TS | $R_{min/all}$ | TS |
| 1 1 1 1 | 0.9385 | 0.9297 | 0.9493 | 0.9495 |
| 1 1 1 0 | 0.9578 | 0.9600 | **0.9619** | **0.9606** |
| 1 1 0 1 | 0.9232 | 0.9192 | 0.9577 | 0.9560 |
| 1 0 1 1 | 0.9500 | 0.9800 | 0.9546 | 0.9553 |
| 0 1 1 1 | – | – | – | – |
| 1 1 0 0 | 0.9068 | 0.9436 | 0.9583 | 0.9556 |
| 1 0 1 0 | 0.8986 | 0.9378 | 0.9531 | 0.9547 |
| 0 1 1 0 | – | – | – | – |
| 1 0 0 1 | **0.9618** | 0.9374 | 0.9529 | 0.9492 |
| 0 1 0 1 | – | – | – | – |
| 0 0 1 1 | – | – | – | – |
| 1 0 0 0 | 0.9532 | **0.9636** | 0.9530 | 0.9475 |
| 0 0 1 0 | – | – | – | – |
| 0 1 0 0 | – | – | – | – |
| 0 0 0 1 | – | – | – | – |

Table A.4: Performance results of decision tree (C5.0) on the dataset *Glass*. "1 0 1 0" represents "safe(1) borderline(0) rare(1) outlier(0)", i.e. only safe and rare samples are oversampled. $R_{min/all}$ and *TS* indicate the different rules for identifying types of samples. "–" means that there are not enough samples to execute the $k$-nearest-neighbor algorithm in the oversampling step.

| Combination | MinAcc | | MAUC | |
|---|---|---|---|---|
| | $R_{min/all}$ | TS | $R_{min/all}$ | TS |
| 1 1 1 1 | 0.6243 | 0.6291 | 0.8603 | 0.8605 |
| 1 1 1 0 | **0.7357** | 0.6778 | 0.8903 | **0.8958** |
| 1 1 0 1 | 0.4933 | **0.7111** | **0.9010** | 0.8925 |
| 1 0 1 1 | 0.4778 | 0.6156 | 0.8798 | 0.8840 |
| 0 1 1 1 | 0.5211 | 0.6522 | 0.8954 | 0.8952 |
| 1 1 0 0 | – | – | – | – |
| 1 0 1 0 | – | – | – | – |
| 0 1 1 0 | – | – | – | – |
| 1 0 0 1 | – | – | – | – |
| 0 1 0 1 | – | – | – | – |
| 0 0 1 1 | – | – | – | – |
| 1 0 0 0 | – | – | – | – |
| 0 0 1 0 | – | – | – | – |
| 0 1 0 0 | – | – | – | – |
| 0 0 0 1 | – | – | – | – |

# Bibliography

Abdi, L. and Hashemi, S. (2015). “To combat multi-class imbalanced problems by means of over-sampling techniques”. In: *IEEE transactions on Knowledge and Data Engineering* vol. 28, no. 1, pp. 238–251.

Acharya, U. R., Chowriappa, P., Fujita, H., Bhat, S., Dua, S., Koh, J. E., Eugene, L., Kongmebhol, P., and Ng, K. (2016). “Thyroid lesion classification in 242 patient population using Gabor transform features from high resolution ultrasound images”. In: *Knowledge-Based Systems* vol. 107, pp. 235–245.

Agrawal, A. and Menzies, T. (2018). “Is" Better Data" Better Than" Better Data Miners"?” In: *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*. IEEE, pp. 1050–1061.

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

Alcalá-Fdez, J., Sánchez, L., Garcia, S., Jesus, M. J. del, Ventura, S., Garrell, J. M., Otero, J., Romero, C., Bacardit, J., Rivas, V. M., et al. (2009). “KEEL: a software tool to assess evolutionary algorithms for data mining problems”. In: *Soft Computing* vol. 13, no. 3, pp. 307–318.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*. Vol. 463. ACM press New York.

Barua, S., Islam, M. M., Yao, X., and Murase, K. (2012). “MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning”. In: *IEEE Transactions on Knowledge and Data Engineering* vol. 26, no. 2, pp. 405–425.

Batista, G. E., Prati, R. C., and Monard, M. C. (2004). “A study of the behavior of several methods for balancing machine learning training data”. In: *ACM SIGKDD explorations newsletter* vol. 6, no. 1, pp. 20–29.

Bauer, L. (2007). *Linguistics Student’s Handbook*. Edinburgh University Press.

Baxevanis, A. D., Bader, G. D., and Wishart, D. S. (2020). *Bioinformatics*. John Wiley & Sons.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). “Algorithms for hyper-parameter optimization”. In: *Advances in neural information processing systems* vol. 24.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D. (July 2015). “Hyperopt: A Python library for model selection and hyperparameter optimization”. In: *Computational Science & Discovery* vol. 8, p. 014008.

Bergstra, J., Yamins, D., and Cox, D. (2013). “Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures”. In: *International conference on machine learning*. PMLR, pp. 115–123.

Bermejo, P., Gámez, J. A., and Puerta, J. M. (2011). “Improving the performance of Naive Bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets”. In: *Expert Systems with Applications* vol. 38, no. 3, pp. 2072–2080.

Bhowan, U., Johnston, M., Zhang, M., and Yao, X. (2012). “Evolving diverse ensembles using genetic programming for classification with unbalanced data”. In: *IEEE Transactions on Evolutionary Computation* vol. 17, no. 3, pp. 368–386.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. Vol. 4. 4. Springer.

Błaszczyński, J., Deckert, M., Stefanowski, J., and Wilk, S. (2010). “Integrating selective pre-processing of imbalanced data with ivotes ensemble”. In: *International conference on rough sets and current trends in computing*. Springer, pp. 148–157.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). “LOF: identifying density-based local outliers”. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104.

Cao, P., Yang, J., Li, W., Zhao, D., and Zaiane, O. (2014). “Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD”. In: *Computerized Medical Imaging and Graphics* vol. 38, no. 3, pp. 137–150.

Carranza-García, M., Lara-Benítez, P., García-Gutiérrez, J., and Riquelme, J. C. (2021). "Enhancing object detection for autonomous driving by optimizing anchor generation and addressing class imbalance". In: *Neurocomputing* vol. 449, pp. 229–244.

Chandola, V., Banerjee, A., and Kumar, V. (2009). "Anomaly detection: A survey". In: *ACM computing surveys (CSUR)* vol. 41, no. 3, pp. 1–58.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* vol. 16, pp. 321–357.

Chawla, N. V., Lazarevic, A., Hall, L. O., and Bowyer, K. W. (2003). "SMOTEBoost: Improving prediction of the minority class in boosting". In: *European conference on principles of data mining and knowledge discovery*. Springer, pp. 107–119.

Chen, L., Fang, B., Shang, Z., and Tang, Y. (2018). "Tackling class overlap and imbalance problems in software defect prediction". In: *Software Quality Journal* vol. 26, no. 1, pp. 97–125.

Chen, Z., Yan, Q., Han, H., Wang, S., Peng, L., Wang, L., and Yang, B. (2018). "Machine learning based mobile malware detection using highly imbalanced network traffic". In: *Information Sciences* vol. 433, pp. 346–364.

Cieslak, D. A., Hoens, T. R., Chawla, N. V., and Kegelmeyer, W. P. (2012). "Hellinger distance decision trees are robust and skew-insensitive". In: *Data Mining and Knowledge Discovery* vol. 24, no. 1, pp. 136–158.

Claesen, M. and De Moor, B. (2015). "Hyperparameter search in machine learning". In: *arXiv preprint arXiv:1502.02127*.

Cordón, I., García, S., Fernández, A., and Herrera, F. (2018). "Imbalance: oversampling algorithms for imbalanced classification in R". In: *Knowledge-Based Systems* vol. 161, pp. 329–341.

Das, B., Krishnan, N. C., and Cook, D. J. (2014). "RACOG and wRACOG: Two probabilistic oversampling techniques". In: *IEEE transactions on knowledge and data engineering* vol. 27, no. 1, pp. 222–234.

Douzas, G., Bacao, F., and Last, F. (2018). "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE". In: *Information Sciences* vol. 465, pp. 1–20.

Dua, D. and Graff, C. (2017). *UCI Machine Learning Repository*.

Elkan, C. (2001). “The foundations of cost-sensitive learning”. In: *International joint conference on artificial intelligence*. Vol. 17. 1. Lawrence Erlbaum Associates Ltd, pp. 973–978.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). “Dermatologist-level classification of skin cancer with deep neural networks”. In: *nature* vol. 542, no. 7639, pp. 115–118.

Fawcett, T. (2004). “ROC graphs: Notes and practical considerations for researchers”. In: *Machine learning* vol. 31, no. 1, pp. 1–38.

— (2006). “An introduction to ROC analysis”. In: *Pattern recognition letters* vol. 27, no. 8, pp. 861–874.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from imbalanced data sets*. Vol. 10. Springer.

Fernández, A., García, S., Herrera, F., and Chawla, N. V. (Jan. 2018). “SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary”. In: *J. Artif. Int. Res.* vol. 61, no. 1, pp. 863–905.

Fernández, A., López, V., Galar, M., Del Jesus, M. J., and Herrera, F. (2013). “Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches”. In: *Knowledge-based systems* vol. 42, pp. 97–110.

Fernández-Navarro, F., Hervás-Martínez, C., and Gutiérrez, P. A. (2011). “A dynamic over-sampling procedure based on sensitivity for multi-class problems”. In: *Pattern Recognition* vol. 44, no. 8, pp. 1821–1833.

Ferri, C., Hernández-Orallo, J., and Modroiu, R. (2009). “An experimental comparison of performance measures for classification”. In: *Pattern recognition letters* vol. 30, no. 1, pp. 27–38.

Feurer, M. and Hutter, F. (2019). “Hyperparameter optimization”. In: *Automated machine learning*. Springer, Cham, pp. 3–33.

Fürnkranz, J. (2002). “Round robin classification”. In: *The Journal of Machine Learning Research* vol. 2, pp. 721–747.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., and Herrera, F. (2011). “An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes”. In: *Pattern Recognition* vol. 44, no. 8, pp. 1761–1776.

Ganganwar, V. (2012). “An overview of classification algorithms for imbalanced datasets”. In: *International Journal of Emerging Technology and Advanced Engineering* vol. 2, no. 4, pp. 42–47.

García, V., Marqués, A. I., and Sánchez, J. S. (2019). “Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction”. In: *Information Fusion* vol. 47, pp. 88–101.

Goldstein, M. and Dengel, A. (2012). “Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm”. In: *KI-2012: Poster and Demo Track*, pp. 59–63.

Haddad, B. M., Yang, S., Karam, L. J., Ye, J., Patel, N. S., and Braun, M. W. (2018). “Multifeature, Sparse-Based Approach for Defects Detection and Classification in Semiconductor Units”. In: *IEEE Transactions on Automation Science and Engineering* vol. 15, no. 1, pp. 145–159.

Hand, D. J. and Till, R. J. (2001). “A simple generalisation of the area under the ROC curve for multiple class classification problems”. In: *Machine learning* vol. 45, no. 2, pp. 171–186.

Hart, P. (1968). “The condensed nearest neighbor rule (corresp.)” In: *IEEE transactions on information theory* vol. 14, no. 3, pp. 515–516.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, pp. 1322–1328.

He, H. and Garcia, E. A. (2009). “Learning from imbalanced data”. In: *IEEE Transactions on knowledge and data engineering* vol. 21, no. 9, pp. 1263–1284.

He, Z., Xu, X., and Deng, S. (2003). “Discovering cluster-based local outliers”. In: *Pattern Recognition Letters* vol. 24, no. 9-10, pp. 1641–1650.

Heft, A. I., Indinger, T., and Adams, N. A. (2012). “Experimental and numerical investigation of the DrivAer model”. In: *ASME 2012 Fluids Engineering Division Summer Meeting*. American Society of Mechanical Engineers Digital Collection, pp. 41–51.

Hinton, G. E. and Roweis, S. (2002). “Stochastic neighbor embedding”. In: *Advances in neural information processing systems* vol. 15.

Ho, T. K. and Basu, M. (2002). “Complexity measures of supervised classification problems”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 3, pp. 289–300.

Imam, T., Ting, K. M., and Kamruzzaman, J. (2006). "z-SVM: An SVM for improved classification of imbalanced data". In: *Australasian joint conference on artificial intelligence*. Springer, pp. 264–273.

Jo, T. and Japkowicz, N. (June 2004). "Class Imbalances versus Small Disjuncts". In: *SIGKDD Explor. Newsl.* vol. 6, no. 1, pp. 40–49.

Knupp, P. (2008). "Measurement and Impact of Mesh Quality". In: *46th AIAA Aerospace Sciences Meeting and Exhibit*, p. 933.

Kong, J., Kowalczyk, W., Menzel, S., and Bäck, T. (2020). "Improving Imbalanced Classification by Anomaly Detection". In: *International Conference on Parallel Problem Solving from Nature*. Springer, pp. 512–523.

Kong, J., Kowalczyk, W., Nguyen, D. A., Bäck, T., and Menzel, S. (2019). "Hyperparameter Optimisation for Improving Classification under Class Imbalance". In: *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, pp. 3072–3078.

Kong, J., Kowalczyk, W., Nguyen, D. A., Menzel, S., and Bäck, T. (2019). "Hyperparameter Optimisation for Improving Classification under Class Imbalance". In: *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE.

Kong, J., Rios, T., Kowalczyk, W., Menzel, S., and Bäck, T. (2020a). "On the Performance of Oversampling Techniques for Class Imbalance Problems". In: *24th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) [Accepted]*. Springer.

— (2020b). "On the performance of oversampling techniques for class imbalance problems". In: *Advances in Knowledge Discovery and Data Mining* vol. 12085, p. 84.

Krawczyk, B. (2016). "Cost-sensitive one-vs-one ensemble for multi-class imbalanced data". In: *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 2447–2452.

Krawczyk, B., Galar, M., Jeleń, Ł., and Herrera, F. (2016). "Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy". In: *Applied Soft Computing* vol. 38, pp. 714–726.

Kubat, M., Holte, R., and Matwin, S. (1997). "Learning when negative examples abound". In: *European conference on machine learning*. Springer, pp. 146–153.

Kubat, M., Holte, R. C., and Matwin, S. (1998). “Machine learning for the detection of oil spills in satellite radar images”. In: *Machine learning* vol. 30, no. 2, pp. 195–215.

Kubat, M., Matwin, S., et al. (1997). “Addressing the curse of imbalanced training sets: one-sided selection”. In: *Icml*. Vol. 97. 1. Citeseer, p. 179.

Lango, M. and Stefanowski, J. (2018). “Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data”. In: *Journal of Intelligent Information Systems* vol. 50, no. 1, pp. 97–127.

Last, M. (2002). “Online classification of nonstationary data streams”. In: *Intelligent data analysis* vol. 6, no. 2, pp. 129–147.

Laurikkala, J. (2001). “Improving identification of difficult small classes by balancing class distribution”. In: *Conference on artificial intelligence in medicine in Europe*. Springer, pp. 63–66.

Lee, T., Lee, K. B., and Kim, C. O. (2016). “Performance of machine learning algorithms for class-imbalanced process fault detection problems”. In: *IEEE Transactions on Semiconductor Manufacturing* vol. 29, no. 4, pp. 436–445.

Lertampaiporn, S., Thammarongtham, C., Nukoolkit, C., Kaewkamnerdpong, B., and Ruengjitchatchawalya, M. (2013). “Heterogeneous ensemble approach with discriminative features and modified-SMOTEbagging for pre-miRNA classification”. In: *Nucleic acids research* vol. 41, no. 1, e21–e21.

Li, J., Liu, L.-s., Fong, S., Wong, R. K., Mohammed, S., Fiaidhi, J., Sung, Y., and Wong, K. K. (2017). “Adaptive Swarm Balancing Algorithms for rare-event prediction in imbalanced healthcare data”. In: *PloS one* vol. 12, no. 7, e0180830.

Liao, T. W. (2008). “Classification of weld flaws with imbalanced class data”. In: *Expert Systems with Applications* vol. 35, no. 3, pp. 1041–1052.

Liu, B. and Tsoumakas, G. (2019). “Synthetic oversampling of multi-label data based on local label distribution”. In: *arXiv preprint arXiv:1905.00609*.

Livesu, M., Vining, N., Sheffer, A., Gregson, J., and Scateni, R. (2013). “PolyCut: Monotone Graph-Cuts for PolyCube Base-Complex Construction”. In: *Transactions on Graphics (Proc. SIGGRAPH ASIA 2013)* vol. 32, no. 6.

López, V., Fernández, A., García, S., Palade, V., and Herrera, F. (2013). “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics”. In: *Information sciences* vol. 250, pp. 113–141.

López, V., Fernández, A., Moreno-Torres, J. G., and Herrera, F. (2012). “Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics”. In: *Expert Systems with Applications* vol. 39, no. 7, pp. 6585–6608.

Lorena, A. C., Garcia, L. P., Lehmann, J., Souto, M. C., and Ho, T. K. (2018). “How Complex is your classification problem? A survey on measuring classification complexity”. In: *arXiv preprint arXiv:1808.03591*.

Lorena, A. C., Garcia, L. P., Lehmann, J., Souto, M. C., and Ho, T. K. (2019). “How Complex Is Your Classification Problem?: A Survey on Measuring Classification Complexity”. In: *ACM Computing Surveys (CSUR)* vol. 52, no. 5, p. 107.

Luengo, J., Fernández, A., García, S., and Herrera, F. (2011). “Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling”. In: *Soft Computing* vol. 15, no. 10, pp. 1909–1936.

Lusa, L. et al. (2015). “Joint use of over-and under-sampling techniques and cross-validation for the development and assessment of prediction models”. In: *BMC bioinformatics* vol. 16, no. 1, p. 363.

Mahalanobis, P. C. (1936). “On the generalized distance in statistics”. In: National Institute of Science of India.

Malina, W. (2001). “Two-parameter Fisher criterion”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* vol. 31, no. 4, pp. 629–636.

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., and Tourassi, G. D. (2008). “Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance”. In: *Neural networks* vol. 21, no. 2-3, pp. 427–436.

Menzel, S., Olhofer, M., and Sendhoff, B. (2005). “Application of Free Form Deformation Techniques in Evolutionary Design Optimisation”. In: *6th World Congress on Structural and Multidisciplinary Optimization (WCSMO6)*. Ed. by Herskovits, J., Mazorche, S., and Canelas, A. Rio de Janeiro: COPPE Publication.

Menzel, S. and Sendhoff, B. (2008). “Representing the Change - Free Form Deformation for Evolutionary Design Optimization”. In: *Evolutionary Computation in Practice*. Springer Berlin Heidelberg, pp. 63–86.

Misra, R., Wan, M., and McAuley, J. (2018). “Decomposing Fit Semantics for Product Size Recommendation in Metric Spaces”. In: *Proceedings of the 12th*

*ACM Conference on Recommender Systems*. RecSys ’18. Vancouver, British Columbia, Canada: Association for Computing Machinery, pp. 422–426.

Napierala, K. and Stefanowski, J. (2016). “Types of minority class examples and their influence on learning classifiers from imbalanced data”. In: *Journal of Intelligent Information Systems* vol. 46, no. 3, pp. 563–597.

Napierała, K., Stefanowski, J., and Wilk, S. (2010). “Learning from imbalanced data in presence of noisy and borderline examples”. In: *International conference on rough sets and current trends in computing*. Springer, pp. 158–167.

Neogi, N., Mohanta, D. K., and Dutta, P. K. (2014). “Review of vision-based steel surface inspection systems”. In: *EURASIP Journal on Image and Video Processing* vol. 2014, no. 1, pp. 1–19.

Nguyen, D. A., Kong, J., Wang, H., Menzel, S., Sendhoff, B., Kononova, A. V., and Bäck, T. (2021). “Improved automated cash optimization with tree parzen estimators for class imbalance problems”. In: *2021 IEEE 8th international conference on data science and advanced analytics (DSAA)*. IEEE, pp. 1–9.

Nguyen, H. M., Cooper, E. W., and Kamei, K. (2011). “Online learning from imbalanced data streams”. In: *2011 International Conference of Soft Computing and Pattern Recognition (SoCPaR)*. IEEE, pp. 347–352.

Olhofer, M., Bihrer, T., Menzel, S., Fischer, M., and Sendhoff, B. (2009). “Evolutionary Optimisation of an Exhaust Flow Element with Free Form Deformation”. In: *4th European Automotive Simulation Conference, Munich*.

Orriols-Puig, A. and Bernadó-Mansilla, E. (2009). “Evolutionary rule-based systems for imbalanced data sets”. In: *Soft Computing* vol. 13, no. 3, pp. 213–225.

Orriols-Puig, A., Macia, N., and Ho, T. K. (2010). “Documentation for the data complexity library in C++”. In: *Universitat Ramon Llull, La Salle* vol. 196, pp. 1–40.

Prati, R. C., Batista, G. E., and Monard, M. C. (2004). “Class imbalances versus class overlapping: an analysis of a learning system behavior”. In: *Mexican international conference on artificial intelligence*. Springer, pp. 312–321.

Radtke, P. V., Granger, E., Sabourin, R., and Gorodnichy, D. O. (2014). “Skew-sensitive boolean combination for adaptive ensembles – An application to face recognition in video surveillance”. In: *Information Fusion* vol. 20, pp. 31–48.

Ren, F., Cao, P., Li, W., Zhao, D., and Zaiane, O. (2017). “Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm”. In: *Computerized Medical Imaging and Graphics* vol. 55, pp. 54–67.

Rifkin, R. and Klautau, A. (2004). “In defense of one-vs-all classification”. In: *The Journal of Machine Learning Research* vol. 5, pp. 101–141.

Rodriguez, D., Herraiz, I., Harrison, R., Dolado, J., and Riquelme, J. C. (2014). “Preliminary comparison of techniques for dealing with imbalance in software defect prediction”. In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, pp. 1–10.

Rokach, L. (2010). “Ensemble-based classifiers”. In: *Artificial intelligence review* vol. 33, no. 1, pp. 1–39.

Sáez, J. A., Krawczyk, B., and Woźniak, M. (2016). “Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets”. In: *Pattern Recognition* vol. 57, pp. 164–178.

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., and Santos, J. (2018). “Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]”. In: *ieee ComputatioNal iNtelligeNCe magaziNe* vol. 13, no. 4, pp. 59–76.

Schaffer, J. (2015). “What not to multiply without necessity”. In: *Australasian Journal of Philosophy* vol. 93, no. 4, pp. 644–664.

Sederberg, T. W. and Parry, S. R. (1986). “Free-form deformation of solid geometric models”. In: *ACM SIGGRAPH computer graphics* vol. 20, no. 4, pp. 151–160.

Sen, A., Islam, M. M., Murase, K., and Yao, X. (2015). “Binarization with boosting and oversampling for multiclass classification”. In: *IEEE transactions on cybernetics* vol. 46, no. 5, pp. 1078–1091.

Shekar, B. and Dagnew, G. (2019). “Grid search-based hyperparameter tuning and classification of microarray cancer data”. In: *2019 second international conference on advanced computational and communication paradigms (ICACCP)*. IEEE, pp. 1–8.

Sieger, D., Menzel, S., and Botsch, M. (2015). “On shape deformation techniques for simulation-based design optimization”. In: *New Challenges in Grid Generation and Adaptivity for Scientific Computing*. Springer, pp. 281–303.

Sinclair, D. (2016). “S-hull: a fast radial sweep-hull routine for Delaunay triangulation”. In: *arXiv preprint arXiv:1604.01428v1 [cs.CG]*.

Skryjomski, P. and Krawczyk, B. (Sept. 2017). “Influence of minority class instance types on SMOTE imbalanced data oversampling”. In: *Proceedings of the First*

*International Workshop on Learning with Imbalanced Domains: Theory and Applications*. Ed. by Luís Torgo, P. B. and Moniz, N. Vol. 74. Proceedings of Machine Learning Research. PMLR, pp. 7–21.

Sleeman IV, W. C. and Krawczyk, B. (2021). “Multi-class imbalanced big data classification on Spark”. In: *Knowledge-Based Systems* vol. 212, p. 106598.

Soofi, A. A. and Awan, A. (2017). “Classification techniques in machine learning: applications and issues”. In: *Journal of Basic & Applied Sciences* vol. 13, pp. 459–465.

Sun, Y., Kamel, M. S., and Wang, Y. (2006). “Boosting for learning multiple classes with imbalanced class distribution”. In: *Sixth international conference on data mining (ICDM’06)*. IEEE, pp. 592–602.

Sun, Y., Kamel, M. S., Wong, A. K., and Wang, Y. (2007). “Cost-sensitive boosting for classification of imbalanced data”. In: *Pattern recognition* vol. 40, no. 12, pp. 3358–3378.

Tan, A. C., Gilbert, D., and Deville, Y. (2003). “Multi-class protein fold classification using a new ensemble machine learning approach”. In: *Genome Informatics* vol. 14, pp. 206–217.

Thai-Nghe, N., Busche, A., and Schmidt-Thieme, L. (2009). “Improving academic performance prediction by dealing with class imbalance”. In: *2009 Ninth International Conference on Intelligent Systems Design and Applications*. IEEE, pp. 878–883.

Thai-Nghe, N., Gantner, Z., and Schmidt-Thieme, L. (2010). “Cost-sensitive learning methods for imbalanced data”. In: *The 2010 International joint conference on neural networks (IJCNN)*. IEEE, pp. 1–8.

Tomek, I. (1976). “Two modifications of CNN”. In: *IEEE Trans. Systems, Man and Cybernetics* vol. 6, pp. 769–772.

Van den Oord, A., Dieleman, S., and Schrauwen, B. (2013). “Deep content-based music recommendation”. In: *Advances in neural information processing systems* vol. 26.

Van der Maaten, L. and Hinton, G. (2008). “Visualizing data using t-SNE.” In: *Journal of machine learning research* vol. 9, no. 11.

Wang, B. X. and Japkowicz, N. (2010). “Boosting support vector machines for imbalanced data sets”. In: *Knowledge and information systems* vol. 25, no. 1, pp. 1–20.

Wang, S. (2011a). "Ensemble diversity for class imbalance learning". PhD thesis. University of Birmingham.

— (2011b). "Ensemble diversity for class imbalance learning".

Wang, S., Chen, H., and Yao, X. (2010). "Negative correlation learning for classification ensembles". In: *The 2010 international joint conference on neural networks (IJCNN)*. IEEE, pp. 1–8.

Wang, S., Minku, L. L., and Yao, X. (2014). "Resampling-based ensemble methods for online class imbalance learning". In: *IEEE Transactions on Knowledge and Data Engineering* vol. 27, no. 5, pp. 1356–1368.

— (2016). "Dealing with Multiple Classes in Online Class Imbalance Learning." In: *IJCAI*, pp. 2118–2124.

— (2018). "A systematic study of online class imbalance learning with concept drift". In: *IEEE transactions on neural networks and learning systems* vol. 29, no. 10, pp. 4802–4821.

Wang, S. and Yao, X. (2009). "Diversity analysis on imbalanced data sets by using ensemble models". In: *2009 IEEE symposium on computational intelligence and data mining*. IEEE, pp. 324–331.

— (2012). "Multiclass imbalance problems: Analysis and potential solutions". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* vol. 42, no. 4, pp. 1119–1130.

Weng, C. G. and Poon, J. (2006). "A data complexity analysis on imbalanced datasets and an alternative imbalance recovering strategy". In: *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*. IEEE, pp. 270–276.

Wilson, D. R. and Martinez, T. R. (1997). "Improved heterogeneous distance functions". In: *Journal of artificial intelligence research* vol. 6, pp. 1–34.

Wilson, D. L. (1972). "Asymptotic properties of nearest neighbor rules using edited data". In: *IEEE Transactions on Systems, Man, and Cybernetics*, no. 3, pp. 408–421.

Wu, Z., Lin, W., and Ji, Y. (2018). "An integrated ensemble learning model for imbalanced fault diagnostics and prognostics". In: *IEEE Access* vol. 6, pp. 8394–8402.

Zhang, H. and Li, M. (2014). "RWO-Sampling: A random walk over-sampling approach to imbalanced data classification". In: *Information Fusion* vol. 20, pp. 99–116.

Zhang, X., Zhuang, Y., Wang, W., and Pedrycz, W. (2016). “Transfer boosting with synthetic instances for class imbalanced object recognition”. In: *IEEE transactions on cybernetics* vol. 48, no. 1, pp. 357–370.

— (2018). “Transfer Boosting With Synthetic Instances for Class Imbalanced Object Recognition”. In: *IEEE Transactions on Cybernetics* vol. 48, no. 1, pp. 357–370.

Zhao, Y., Nasrullah, Z., and Li, Z. (2019). “Pyod: A python toolbox for scalable outlier detection”. In: *arXiv preprint arXiv:1901.01588*.

Zhu, L., Lu, C., Dong, Z. Y., and Hong, C. (2017). “Imbalance learning machine-based power system short-term voltage stability assessment”. In: *IEEE Transactions on Industrial Informatics* vol. 13, no. 5, pp. 2533–2543.

Zięba, M., Tomczak, J. M., Lubicz, M., and Świątek, J. (2014). “Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients”. In: *Applied soft computing* vol. 14, pp. 99–108.