EPiC Series in Computing

Volume 69, 2020, Pages 141–150

Proceedings of 35th International Conference on Computers and Their Applications

On relationships between imbalance and overlapping of datasets

Waleed A. Almutairi and Ryszard Janicki

Mcmaster University, Hamilton, Ontario, Canada
almutaiw@mcmaster.ca, janicki@cas.mcmaster.ca

Abstract

The paper deals with problems that imbalanced and overlapping datasets often encounter. Performance indicators as accuracy, precision and recall of imbalanced data sets, both with and without overlapping, are discussed and compared with the same performance indicators of balanced datasets with overlapping. Three popular classification algorithms, namely, Decision Tree, KNN (k-Nearest Neighbors) and SVM (Support Vector Machines) classifiers are analyzed and compared.

1 Introduction

For many supervised learning algorithms, there is a significant difference between the prior probabilities of different classes; for example, between the probabilities of belonging to different classes of a given classification problem. This situation is known as the *class imbalance problem* [1, 2, 3]; and it occurs often in many real problems from telecommunications, web, finance, ecology, biology, medicine, oil mining, It appears that the imbalance class problem is considered one of the top current problems in data mining [4]. It is also worth to point out that the minority class is usually the one that has the highest interest from a learning point of view, so it may cost a lot if not well classified [5].

Learning and analyzing the data in order to predict class labels has been widely studied in machine learning and in artificial intelligence domains. Traditional classification algorithms assume that the data are balanced classes in the space of distributions. However, in many applications, the number of instances is some classes is substantially smaller than in other classes. For example, for credit card fraud detection, direct marketing, detecting oil spills from satellite images and network intrusion detection, the targeted class has much fewer representative data compared to other classes. Due to the increase of these kind of applications in recent years, learning in the presence of imbalanced data has become an important research topic.

It has been shown that when classes are well separated, regardless of the imbalanced ratio, instances can be correctly classified using standard learning algorithms [3]. However, having a class imbalance in complex datasets, usually results in misclassification of the data, particularly the minority classes. Other reasons for misclassification involve also overlapping classes within class imbalance, out-liars, and for example noises.

Within a given class, the imbalance occurs when a class is scattered into smaller subparts representing separate sub-concepts, especially sub-concepts with limited representatives, i.e. so called small disjuncts [2]. Classification algorithms are often not able to learn from small disjuncts examples. This problem is more noticeable in the case of under-sampling techniques. This is due to the fact that the probability of randomly choosing an instance from small disjuncts within the majority class is very low. These regions may, therefore, remain untrained [6].

In this paper, we will focus on dealing with overlapping in an imbalanced data, mainly within and between a class that is imbalanced.

It has recently been observed that the relatively huge amount of data we may have a gray area, i.e. for data points, we may have a hard time deciding to which data class they belong to. Moreover most of the algorithms that classify data in the overlapping space often provide misleading results of limited value. The problem is more severe when we have an imbalanced dataset with both majority classes and minority classes and one type of classes outnumbers the other. The minority class usually represents the most important concept to be learned from and often the data acquisition for minority class is more expensive. Often the imbalanced class problem is associated with the binary classification, but it might happen in multiclass problems. For the latter we often have some minority classes that are very difficult to classify [7, 8].

The paper is organized as follows. The related works will briefly be presented in the next section. The third section deals with case classification and data generation, while Section 4 provides the results of empirical analysis of balanced and unbalanced data sets with overlappings. The last section contains brief conclusions and description of future work.

This paper could be classified as empirical survey paper that support an urgent need for further fundamental study of the nature of overlapping.

2 Related Works

There are many algorithms for improvement the accuracy of classifying unbalanced and overlapped data. In [9] Janicki and Soudkhah have introduced a novel concept of feature domain overlappings. It can measure the feature discrimination power. The model of [9] is based on the assumption that less overlapping means more discriminatory ability, and this can be used to calculate weights characterizing the importance of particular features.

Hakime Koc proposes in [8] a new methodology of learning from examples. He modifies and extends an exemplar-based generalization technique. His technique is based on the representation of overlapping feature intervals and is called as Classification with Overlapping Feature Intervals or COFI. In this approach, learning is from projections of intervals for each dimension for each feature and these intervals correspond to the learned concepts.

Overlapping classes and ambiguous data have been studied for a long time, particularly in the area of character recognition and document analysis [10]. Tang et al. [11] proposed a 'k-nearest neighbors' (KNN) based approach to extract the ambiguous region in the data. Visa et al. [12] performed a fuzzy set representation of the concept and incorporated overlap information in the fuzzy classifiers. In Xiong et al. [13], the one class classification Support Vector Data Description algorithm (SVDD) is used to capture the overlapping regions in real time data-sets.

Handling and dealing with overlapping regions is as important as identifying such regions. Xiong et al. [13] proposed that the overlapping regions could be handled with three different schemes: discarding, merging and separating.

The scheme 'discarding' ignores the overlapping region and learn from what is left of the data that are in the non-overlapping region (cf. SMOTE and Tomek links [14, 15]).

The ‘merging’ scheme considers the overlapping area as a new class and use a 2-tier classification model. The upper tier classification is focusing on the entire data set with an additional class representing the overlapping region. If test sample is classified as belonging to the overlapping region, the lower tier classifier, which works only on the overlapping region, is used. Trappenberg et al. [16] proposed a scheme that refers to the overlapping region class as IDK (‘I Dont Know’) and predicts the class label of test data only when it is first classified as IDK. The authors claim that by losing predication accuracy on IDK, a drastic increase in confidence can be gained for the classification of the remaining data.

In ‘separating’ scheme, the data from overlapping and non-overlapping regions are treated separately to build the learning models. Tang et al. [11] proposed a multi-model classifier named Dual Rough Support Vector Machine (Dr-SVM) which combine SVM and KNN; and also uses Rough Sets paradigm [17]. First KNN is used to extract the overlapping regions and next two SVMs are then trained for the overlapping and non-overlapping regions.

Prati and Batista et al. [18, 19] analyze balancing strategies and class overlapping. They have shown that overlapping aggravates the problem of imbalance and often degrades the performance of the classifier on its own. Garcia et al. [20] analyzed the effect of the combined problems for instance-based classification scenario.

There is also a collection of data cleansing methods that tackle the problem just by cleaning up unwanted overlapping between classes. This is usually done by removing pairs of minimally distanced nearest neighbors of opposite classes, popularly known as Tomek links [14]. The algorithms SMOTE+ENN and SMOTE+Tomek [21] utilize the capability of Tomek links to clean the data. Unfortunately the cleansing techniques are not desirable for data sets that have inherent class overlap or absolutely rare minority class samples that can cause loss of highly informative data.

3 Case Classification and Generation of Synthetic Random Data Sets

In real life, the data almost always come imbalanced and skewed or overlapped. It appears that the right way to deal with this problem is to consider imbalance and overlapping together, as a single issue instead of treating them separately. In [22, 23, 18, 20], the problems of overlapping and imbalance are analyzed by using synthetic data sets. However, the problem with those datasets they used had generated the synthetic data with only one feature.

In our approach we will try to create more realistic data sets. We will try to consider all possible case, as illustrated in Figure 1 and analyzed below.

- `A`: The classes are imbalanced with no overlapping.
- `B`: The classes are imbalanced with overlapping.
- `C`: C is obtained by balancing type A dataset and it may overlap.
- `D`: D is obtained by balancing type A dataset but no overlapping.
- `E`: The classes are balanced and no overlapping (very rare in real life).
- `F`: The classes are balanced and have some overlapping.

We used the imbalanced-learn package [24] to generate synthetic random data sets, that have two classes or labels. The first type of data sets are balanced data sets that have a different level of overlapping or separation in the features, as Figure 2 indicates.

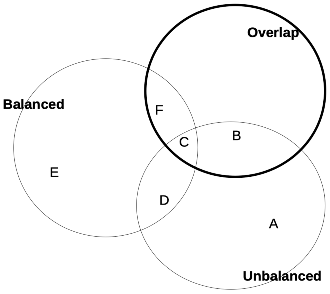
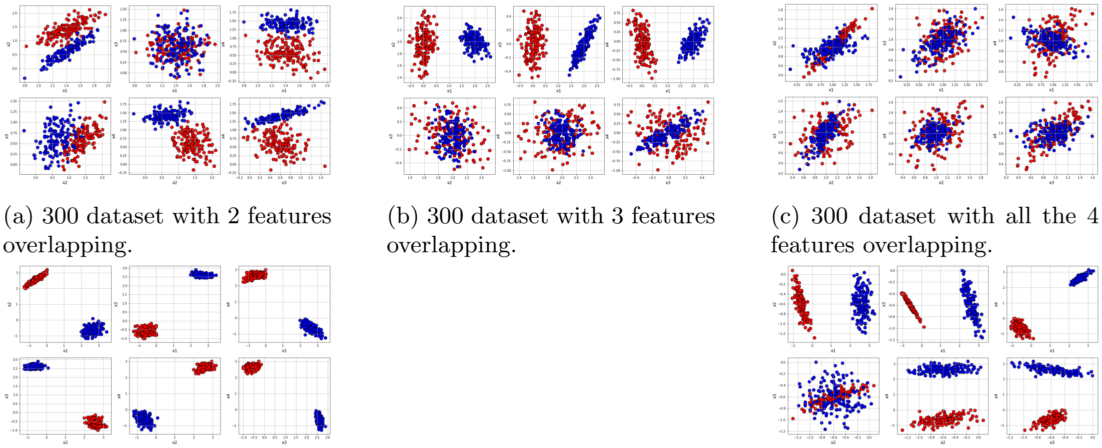


Figure 1: Case classification: balanced and imbalanced vs. overlapping



(a) 300 dataset with 2 features overlapping.

(b) 300 dataset with 3 features overlapping.

(c) 300 dataset with all the 4 features overlapping.

(d) 300 dataset with no features overlap.

(e) 300 dataset with 1 features overlapping.

Figure 2: Balanced data set total of 300, 150 for each class.

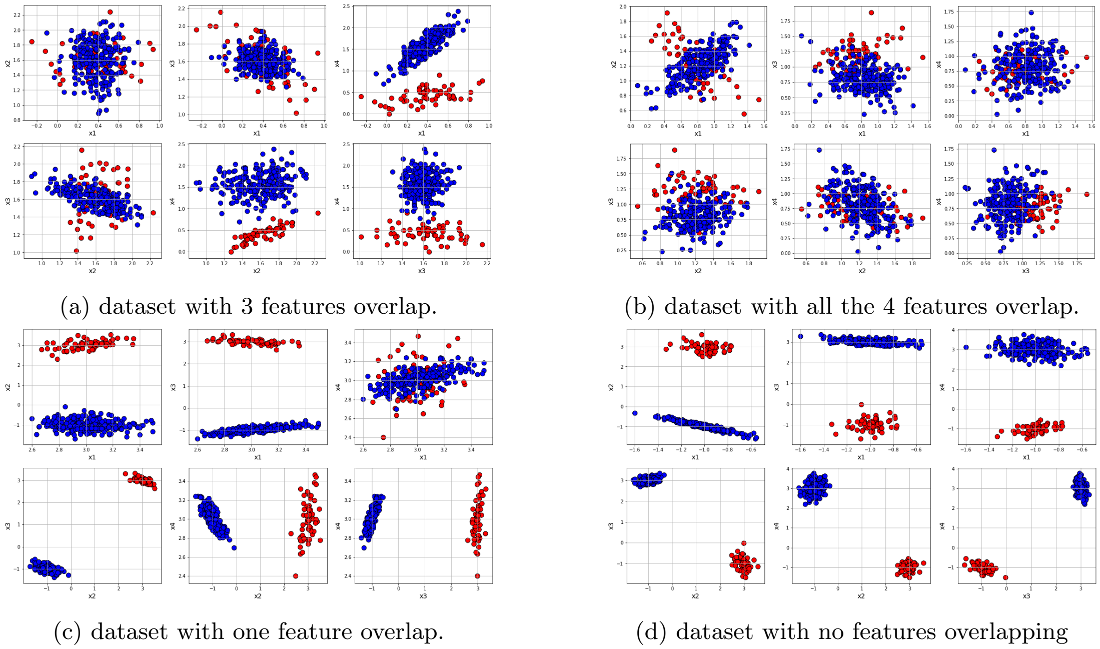
The second type are imbalanced data sets that have ratio 1:4 for class 0 versus class 1, respectively. They are presented in Figure 3 and each data set has different overlapping over four features. All data used in this paper can be found in detailed form in [25].

The first data sets consist of data that generated the features randomly by controlling the separation level. There is a dataset that has all the four features separated. Moreover, others have some variations in the overlapping. Besides, one dataset has all the features overlap.

The second type of data we have generated is for the unbalanced dataset.there are some data with a variety of overlapping or separation. Moreover, we have data that does not overlap and one that overlaps in all features.

4 Analysis of Balanced and Unbalanced Data with Overlappings

We will use three parameters for quantitative analysis of data sets: *Accuracy*, *Precision* and *Recall*. Let TN denote the number of *true negative cases*, FP denote the number of *false positive cases*, FN denote the number of *false negative cases* and TP denote the number of *true*



(a) dataset with 3 features overlap.

(b) dataset with all the 4 features overlap.

(c) dataset with one feature overlap.

(d) dataset with no features overlapping

Figure 3: Unbalanced dataset total of 300: 240 for class 1 and 60 for class 0.

positive cases.

When we have a two classes problem, the quadruple TN, FP, FN and TP form *confusion matrix* shown below:

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FP	TP

A confusion matrix shows the complete results of correctly classified and incorrectly classified examples of each class.

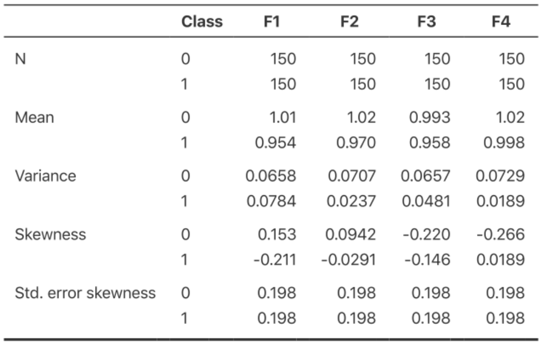
When we have the values of TN, FP, FN and TP, we can define *performance indicators* as *Accuracy*, *Precision* and *Recall* by using the following natural and intuitive formulas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

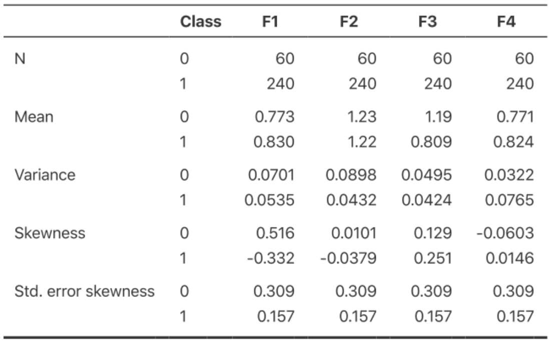
$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

We trained and evaluated the data using 10-folds and we will present and analyze the results of applying three popular machine learning algorithms: Decision Tree (DT) [26], k-Nearest Neighbors (KNN) [27] and Support Vector Machines (SVM) [28]. Additionally we will



	Class	F1	F2	F3	F4
N	0	150	150	150	150
	1	150	150	150	150
Mean	0	1.01	1.02	0.993	1.02
	1	0.954	0.970	0.958	0.998
Variance	0	0.0658	0.0707	0.0657	0.0729
	1	0.0784	0.0237	0.0481	0.0189
Skewness	0	0.153	0.0942	-0.220	-0.266
	1	-0.211	-0.0291	-0.146	0.0189
Std. error skewness	0	0.198	0.198	0.198	0.198
	1	0.198	0.198	0.198	0.198

Table 1: Consistency analysis of balanced data for all four overlapping features: F1, F2, F3 and F4.



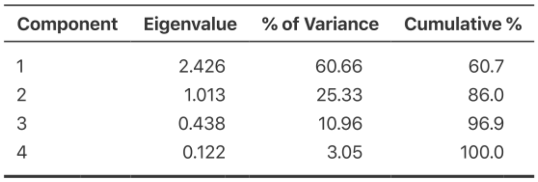
	Class	F1	F2	F3	F4
N	0	60	60	60	60
	1	240	240	240	240
Mean	0	0.773	1.23	1.19	0.771
	1	0.830	1.22	0.809	0.824
Variance	0	0.0701	0.0898	0.0495	0.0322
	1	0.0535	0.0432	0.0424	0.0765
Skewness	0	0.516	0.0101	0.129	-0.0603
	1	-0.332	-0.0379	0.251	0.0146
Std. error skewness	0	0.309	0.309	0.309	0.309
	1	0.157	0.157	0.157	0.157

Table 2: Consistency analysis of unbalanced data for all four overlapping features: F1, F2, F3 and F4.

also provide (limited) Principal Components Analysis (PCA, cf. [29]) for both balanced and unbalanced data sets.

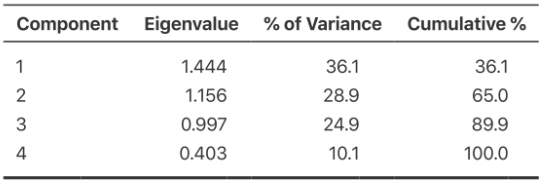
Two kind of data sets have been analyzed, balanced data sets and unbalanced data sets, both with different overlapping and separation levels. Some consistency (or inconsistency) analysis of data sets with standard tools as *mean*, *variance*, skewness and *skewness standard error* are presented in Tables 1 and 2 for all four features. In both cases the significant skewness of the overlapping features makes it harder to train the model because of the mixed boundary between the classes. Table 3 and Table 4 provide the results of Principal Components Analysis (PCA, cf. [29]) for balanced and unbalanced data sets. There is a significant difference in the results of this analysis. For balanced data, i.e. Table 3, the first component explains 60.66% of the total variance, the second component explains 25.33%, and the last one represents 3.0.5 of the variance. On the other hand, for unbalanced data sets, i.e. Table 4, the first PCA component explains only 36.1% of the variance, the second one 28.9%, and the last one explains 10.1% of the variance.

Tables 5, 6 , and 7 contain the main quantitative results of this paper. In all three tables the letter ‘F’ means ‘feature’. The upper part of all these three tables deals with balanced data sets. It turns out that for balanced data sets all three classifiers have no problem with correct classification of data that have small overlapping with some separation. Nevertheless



Component	Eigenvalue	% of Variance	Cumulative %
1	2.426	60.66	60.7
2	1.013	25.33	86.0
3	0.438	10.96	96.9
4	0.122	3.05	100.0

Table 3: PCA components for balanced data with all features overlapping.



Component	Eigenvalue	% of Variance	Cumulative %
1	1.444	36.1	36.1
2	1.156	28.9	65.0
3	0.997	24.9	89.9
4	0.403	10.1	100.0

Table 4: PCA components for unbalanced data with all features overlapping.

the results for three feature overlapping are better than for two feature overlapping, which is not expected. The reason could be explained by looking into Figure 2(a) and Figure 2(b). In Figure 2(b) there is a significant separation between the features that do not overlap (top part of (b)), while in Figure 2(a), there are rather small distances between not overlapping features.

The lower part of Tables 5, 6 , and 7 deals with unbalanced data sets. In this case, when overlapping is small with significant separation, all three classifiers have achieved high accuracy. As expected, when all features overlap we got much worse accuracy, precision and recall, again very similar for all three classifiers. However, for the case of all features overlapping, unbalanced data sets have better performance indicators than balanced data sets, again for all three classifiers. For example for Decision Tree classifier (Table 5), accuracy for balance sets with all feature overlapping is 70.67% but 87.22% for unbalance data sets., and similarly for the remaining tables and performance indicators. This could be interpreted as a fact that the unbalanced data sets have usually less overlapping than the balanced ones. Another reason could be not enough data in the area of overlapping.

Decision Tree Classifier	Accuracy%	Precision%	Recall%
Balanced-2-F-overlap	98	98.124	97.881
Balanced-3-F-overlap	100	100	100
Balanced-no-F-overlap	100	100	100
Balanced-one-F-overlap	100	100	100
Balanced-all-F-overlap	70.667	70.150	70.382
Unbalanced-3-F-overlap	98	97.213	97.016
Unbalanced-no-F-overlap	100	100	100
Unbalanced-one-F-overlap	100	100	100
Unbalanced-all-F-overlap	87.333	79.273	79.742

Table 5: Decision Tree using 10-folds

For unbalanced data sets it is always worth to the oversampling and undersampling with respect to performance indicators. The results are presented in Table 8 and Table 9, respectively.

KNN(3) Classifier	Accuracy%	Precision%	Recall%
Balanced-2-F-overlap	99	99.062	98.941
Balanced-3-F-overlap	100	100	100
Balanced-no-F-overlap	100	100	100
Balanced-one-F-overlap	100	100	100
Balanced-all-F-overlap	76.167	76.358	75.815
Unbalanced-3-F-overlap	98.833	98.514	97.883
Unbalanced-no-F-overlap	100	100	100
Unbalanced-one-F-overlap	100	100	100
Unbalanced-all-F-overlap	90.667	86.495	83.785

Table 6: K-Nearest Neighbor(k=3) using 10-folds

SVM Classifier	Accuracy%	Precision%	Recall%
Balanced-2-F-overlap	99.333	99.375	99.294
Balanced-3-F-overlap	100	100	100
Balanced-no-F-overlap	100	100	100
Balanced-one-F-overlap	100	100	100
Balanced-all-F-overlap	68.333	69.925	68.511
Unbalanced-3-F-overlap	99.222	99.009	98.589
Unbalanced-no-F-overlap	100	100	100
Unbalanced-one-F-overlap	100	100	100
Unbalanced-all-F-overlap	89	85.323	78.344

Table 7: SVM Classifier using 10-folds

Resampling: over-sampling	Accuracy%	Precision%	Recall %
Decision Tree Classifier	91.892	92.003	91.902
KNN k=3 Classifier	93.659	93.805	93.721
SVM Classifier	88.581	88.748	88.715

Table 8: Oversampling an unbalanced dataset with all features overlapping using ADASYN algorithm, then using this classifier to train the model.

Resampling: under-sampling	Accuracy%	Precision%	Recall%
Decision Tree Classifier	77.500	77.571	77.865
KNN k=3 Classifier	85.417	85.637	85.489
SVM Classifier	83.611	83.233	83.847

Table 9: Undersampling majority class in an unbalanced dataset in all features overlapping using Random Under Sampler algorithm, then using this classifier to train the model.

5 Conclusions and Future Work

It appears that for our random data all three classifiers, Decision Tree, KNN (k-Nearest Neighbors) and SVM (Support Vector Machines), have very similar performance indicators. There is no evidence that one of them was better or worse that the others.

We have also noticed that the classical PCA analysis appears to be a good measure for the degree of balancing.

When all four features overlap, the balanced data sets have worse accuracy and precision than the unbalanced data sets, which indicates that the unbalanced data set have less overlapping than the balanced data. It also indicates that for unbalanced date, using re-sampling may improve balance but does not guarantee better accuracy or precision. The reason is that by removing instances near the boundaries we might lose some critical information, information that might be helpful for the classification. The problem lies in the length of overlap among different classes of the data. For unbalanced case we need more data to distinguish the boundaries and to separate different classes. Hence re-sampling might decrease performance indicators as we might add more instances to the minority class but far away from boundaries.

To solve this problem we need to pay more attention to and analyze in detail the overlapping problem first. Then, dependently on the degree of overlapping, we may add more appropriate data to the minority class or remove some appropriate instances from the majority class.

We believe the problem of overlapping is important and still underdeveloped. For example the current over-sampling and under-sampling algorithms treat the entice class equally, while we might need to add or remove instances in the overlapping area or near that area. Moreover, we may try to restrict and make it more precise and adequate over-sampling and under-sampling by taking into account features that have smaller area of overlapping.

Acknowledgments

The authors gratefully acknowledge three anonymous referees whose comments significantly contributed to the final version of this paper. The first author would also like to thank King Abdulaziz City for Science and Technology (KACST) for supporting this work. The research was partially supported by Discovery NSERC of Canada grant.

References

[1] N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6(1):1–6, June 2004.

[2] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[3] Y. Sun, A. K. C. Wong, and M. S. Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719, 2009.

[4] Q. Yang and X. Wu. Ten challenging problems in data mining research. *International Journal of Information Technology and Decision Making*, 5(04):597–604, 2006.

[5] C. Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

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

[7] M. Lin, K. Tang, and X. Yao. Dynamic sampling approach to training neural networks for multiclass imbalance classification. *IEEE Transactions on Neural Networks and Learning Systems*, 24(4):647–660, 2013.

[8] H. Ü. Koç. *Classification with overlapping feature intervals*. PhD thesis, Bilkent University, 1995.

[9] R. Janicki and M. H. Soudkhah. On classification with pairwise comparisons, support vector machines and feature domain overlapping. *The Computer Journal*, 58(3):416–431, 2015.

[10] D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[11] T. Yaohua and G. Jinghuai. Improved classification for problem involving overlapping patterns. *IEICE Transactions on Information and Systems*, 90(11):1787–1795, 2007.

[12] S. Visa and A. Ralescu. Learning imbalanced and overlapping classes using fuzzy sets. In *Proceedings of the ICML*, volume 3, 2003.

[13] H. Xiong, J. Wu, and L. Liu. Classification with class overlapping: a systematic study. In *The 2010 International Conference on E-Business Intelligence*, pages 491–497, 2010.

[14] I. Tomek. Two modifications of CNN. *IEEE Trans. Sys., Man and Cybernetics*, 6:769–772, 1976.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[16] T. P. Trappenberg and A. D. Back. A classification scheme for applications with ambiguous data. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, volume 6, pages 296–301. IEEE, 2000.

[17] Z. Pawlak. *Rough Sets*. Kluwer, Dordrecht, 1991.

[18] M. C. Prati, G. Batista, and M. C. Monard. Class imbalances versus class overlapping: an analysis of a learning system behavior. In *Mexican International Conference on Artificial Intelligence*, pages 312–321. Springer, 2004.

[19] G. Batista, R. C. Prati, and M. C. Monard. Balancing strategies and class overlapping. In *International Symposium on Intelligent Data Analysis*, pages 24–35. Springer, 2005.

[20] V. García, R. Alejo, J Sánchez, J. Sotoca, and R. Mollineda. Combined effects of class imbalance and class overlap on instance-based classification. *Intelligent Data Engineering and Automated Learning–IDEAL 2006*, pages 371–378, 2006.

[21] G. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1):20–29, 2004.

[22] V. López, A. Fernández, S. García, V. Palade, and F. Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 2013.

[23] M. Denil and T. Trappenberg. Overlap versus imbalance. In *Canadian Conference on Artificial Intelligence*, pages 220–231. Springer, 2010.

[24] G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine lea rning. *The Journal of Machine Learning Research*, 18(1):559–563, 2017.

[25] W. A. Almutairi. *http://www.cas.mcmaster.ca/~cs3sd3/waleed-data/Data*. 2019.

[26] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[27] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[28] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proc. of* 5^{th} *Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, 1992.

[29] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.