Classification with Imbalanced Data :The Use of Binarization and Feature Selection

Tahani S. Al-Moshadak¹, Saleh M. Al-Shomrani² and Aiiad A. Albeshri¹

¹Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, and ²Department of Information Systems, Faculty of Computing and Information Technology, University of Jeddah, Jeddah, Saudi Arabia
tsalmoshadak@kau.edu.sa

Abstract. The imbalanced class problem is related to the real-world application of classification. It occurs when there is a high difference between the prior probabilities of classes in the learning phase. It is considered as a challenge since it needs to deal with uneven distribution of examples. The presence of multiple classes implies an additional difficulty since the relations between the classes tend to be complicated. This paper provides a review on multi-class imbalance problem with a focus on feature selection and problem decomposition as a solution of this problem. Also, it presents a comparative overview of the related solutions that are proposed for classification imbalanced datasets. Moreover, it provides outlines of classification in imbalanced datasets that could help any researcher in this field.

Keywords: Binary classification, Multi-classification, Imbalanced problem, Decomposition schemes, OVO, OVA, Aggregation strategies, Feature selection.

1. Introduction

Classification belongs to Supervised Machine Learning which extracts knowledge based on previously seen examples. Classification could be categorized based on the number of classes into binary classification and multi-classification. In the binary classification, instances are classified into two classes whereas in multi-classification instances are classified into more than two classes. Multi-classification is considered a general case of the binary classification. Also, it is more difficult according to the complexity of finding the decision boundaries. Many important applications use multi-classification techniques such as the classification of fingerprints, handwritten digits, microarrays, and face recognition^[1].

Imbalanced class problem is one of the problems occurred in real-world applications of classification. It occurs when there is a high difference between the prior probabilities of classes in the learning phase. The minority class which usually represents the concept of interest, is not classified well. The presence of multiple classes implies an additional difficulty since the relations between the classes tend to be complicated. One class can be a minority class for some, while a majority for others. Recently, the imbalanced problem has received an attention because it is a recurring problem in many real-world domains including high-resolution airborne imagery, forecasting of ozone levels, face recognition, and medical diagnosis^{[2][3]}.

The problem of multi-classification could be facilitated by decomposition via

binarization techniques such as One-vs-One (OVO) ^{[4][5]} and One-vs-All (OVA) ^{[6][7]}. The multi-class problem is divided into a number of binary sub-problems. Each sub-problem is solved by a classifier called base classifier and the outputs from these classifiers are aggregated to make the final decision of the class ^[8]. OVO is the most extendable scheme, established by default in several widely used software tools such as KEEL (Knowledge Extraction based on Evolutionary Learning) ^[9] and WEKA (Waikato Environment for Knowledge Analysis) ^[10]. Moreover, OVO could be better than OVA in most cases of classification multiclass datasets as discussed in ^[11] and proved in the experimental studies ^{[12][8]}. So, in this paper we focused on the use of OVO.

A large number of approaches have been proposed to deal imbalance problem including data level solutions, algorithmic level solutions and cost-sensitive solutions. In addition, the task of classification in the imbalanced dataset could be improved using a well-known data preprocessing technique called feature selection. Feature selection reduces the number of features, removes irrelevant, redundant and noisy data. As a result of that, it improves the performance of the classifiers and speeding them up, because a large number of features may degrade the discovery of the borderline areas of the problem. ^[13]. Many works proposed the use of feature selection in the classification of multiclass problems ^{[14][15][16]} due to its usefulness in the behavior of the classifiers in the multi-

In this paper, our goal is to come up with a review on multiclass imbalance problem, and we focused on the problem decomposition by OVO and the use of feature selection. In the next section, the decomposition schemes for multiclass imbalanced problems is discussed. Then, an overview of the aggregation strategies for OVO schemes is presented in

Section 3. The problem of the imbalanced datasets is discussed in Section 0 and the proposed solutions for this problem is presented in Section 0. Next, an overview of feature selection and its beneficial for the classification of multiclass imbalanced problems is showed in Section 0. After that, a comparative overview of the related solutions that suggested in the field of classification of imbalanced datasets is provided in Section 0. Finally, the conclusion in Section 0 provides outlines of multi-classification in imbalanced datasets that could help any researcher in this field.

2. Binarization for Multi-classification

Different decomposition schemes have been used to facilitate classifying multi-class problems. The most common used schemes are the OVO ^{[4][5]} and OVA ^{[6][7]}. Both of them are special cases of a general approach that uses Error Correcting Output Code ^[17]. OVO divides the problem into sub-problems that equals to all possible combinations of the pairs of the classes. Then, a classifier is learned for each binary problem. On the other hand, OVA learned classifiers that equals the number of classes since it distinguishes each class from all other classes ^[8]. Both schemes are shown in Fig. 1.

OVO is considered as the most extended scheme, established by default in several widely used software tools such as KEEL and WEKA. Several experimental studies ^{[12][8]} compared the OVO verses OVA on multi-class problems and they show that OVO methods, in general, have better behavior than OVO.

Typically, OVO produces more accurate models than OVA because it learns models for all possible combinations between pairs of classes whereas OVA learns one model for each class. Also, OVO is computationally less complex than OVA although it needs to train a quadratic number of classifiers instead of a

linear number of classifiers as in OVA because [11].

- 1) The number of training examples in the binary problems of OVO is fewer than training examples of OVA problems because all examples that do not belong to either of the two classes are ignored.
- 2) Finding the decision boundaries in OVO is easier than OVA.

As mentioned above, OVO divides a multi-class problem (m-class) problem into $\mathbf{m(m-1)/2}$ independent binary sub-problems, each of them is learned by a single classifier. In the classification phase, the instance to be classified is presented to all classifiers. The outputs these classifiers are placed on a score-matrix (1) it is used later by an aggregation method to find the final decision [1].

$$R = \begin{pmatrix} - & r_{12} & \cdots & r_{1m} \\ r_{21} & - & \cdots & r_{2m} \\ \vdots & & & \vdots \\ r_{m1} & r_{m2} & \cdots & - \end{pmatrix} \quad (1)$$

where a confidence degree $\mathbf{r_{ij}}$ and $\mathbf{r_{ji}} \in [0,1]$ in favor of their couple of classes $\mathbf{C_i}$ and $\mathbf{C_j}$ (usually $\mathbf{r_{ji}} = 1 - \mathbf{r_{ij}}$).

3. Aggregation Strategies for OVO Scheme

A critical step in OVO decomposition scheme is the aggregation of the predictions from the ensemble of binary models into a final classification. Most of the problems of OVO scheme arise at this point because the wrong class selection can lead to wrong predictions. A large number of strategies have been developed to decide the final class ranging from classical strategies such as Pair-wise Coupling, Max-Wins rule, and weighted voting, to new approaches such as binary-tree based strategies, dynamic classification

schemes and the methods that used probability estimates or preference relations [8].

The simplest and powerful aggregation strategy is voting [18], also called binary voting or Max-Wins rule. In this approach, each classifier of the binary problem gives a vote for its predicted class. Then, all votes for each class are counted. The class with the highest number of votes is considered as a final output class of the instance. The following equation represents the voting strategy [8]:

$$class = \arg \max_{i=1, \dots, m} \sum_{1 \leq j \neq i \leq m} S_{ij} \quad (2)$$

where $\mathbf{S_{ij}}$ is 1 if $\mathbf{r_{ij} > r_{ji}}$ and 0 otherwise.

Another aggregation strategy called weighted voting (WV) developed to take the advantage of the confidence values of each binary classifier. It finds the summation of confidence values instead of counting the votes. The following equation represents the weighted voting strategy [8]:

$$class = \arg \max_{i=1, \dots, m} \sum_{1 \leq j \neq i \leq m} r_{ij} \quad (3)$$

where $\mathbf{r_{ij}}$ is the confidence degree of the classifier.

Weighted voting has been proved its robustness when it compared with another aggregation approach called adaptive voting [11]. Adaptive voting is a generalized voting strategy in which the predictions of the base classifiers are adapted according to their strength. High-performance classifiers considered to be more reliable than the lower performance classifiers. Therefore, aggregation strategy becomes less sensitive to the predictions of weak classifiers.

There are some problems in the classification using OVO which make the recognition of the classes harder. One of these problems is non-competent classifiers which

lead to misclassification. The source of this problem started in the training phase when the classifier just learned by the examples that belong to its pair of classes and the other examples of other classes are completely ignored. In classification phase, the non-competent classifier gives incorrect predictions when the real class of the instance to be classified does not belong to a class of its pair of classes. These wrong predictions will affect the final decision of the system^{[1][19]}.

In order to manage and minimize the negative effect of the non-competent classifiers, the Distance-based Relative Competence Weighting for OVO (DRCW-OVO)^[19] which is a combination method was recently developed. This method consists of carrying out a dynamic adaptation of the score-matrix, altering the confidence degrees by assigning a higher weight to those classifiers whose predicted classes are closer to the query instance, assuming they are more competent than those which are farther. This distance is computed by using the standard k-nearest neighbors approach of the corresponding class.

Another problem in the classification using OVO is the difficulty of recognition of some classes. Recognize classes task is done based on their characteristics which represent the degree of class separation. These characteristics include the number of instances, the inter-class relations and the overlapping with other classes as shown in Fig. 2. By using these characteristics, some classes are easier to be recognized more than other. Difficult classes are those classes with lower correct classification rate^[20]. In order to solve this problem, a novel similarity-based

aggregation^[20] method was developed. This method generalizes the weighted voting by developing a rule that compensates the accuracy for all classes of the problem without needing to alter the underlying base classifiers. One of the advantages of this method is that it achieves a global optimization with all the classifiers at the same time which is not achieved by standard OVO combinations. The outputs of the learned classifiers could be analyzed together and then modified in such a way that globally better solutions can be obtained.

Galar M. *et al.* provided in Ref. ^[8] an experimental study about two decomposition techniques which are OVO and OVA for the multi-class problems. This study focused on the aggregation methods of the outputs of the binary classifiers. Throughout the experiments, different types of base classifiers are used with different aggregation methods. Important results are conducted. Among them:

- 1) Decomposition techniques are suitable for multi-class problems with respect to the base classifier and the dependency of the base classifier on the combination strategy.
- 2) Considering OVO vs OVA comparison, OVO scheme shows the best ranking results in the most cases when used with different classifiers as illustrated in Fig. 3 (the lowest is the best).
- 3) The performance of the aggregation method is dependent on the base classifier when OVO decomposition is used.

Fig. 1. Multi-class problem with (1) OVO decomposition scheme, (2) OVA decomposition scheme ^[12].

Fig. 2. A three class problem. The class in the center is more difficult to be correctly classified ^[20].

Fig. 3. Rankings (in accuracy) of the OVA and OVO representatives for each base classifier ^[8].

4. Imbalanced Datasets Problem

One of the problems appeared frequently in real-world applications of classification is the imbalanced class problem. It occurs when there is a high difference between the prior probabilities of classes in the learning phase. The minority class which is the most important concept to be learned is misclassified and the classifier biased toward the majority class. That's because the minority class is not represented properly in the dataset by enough examples since it may be costly, rare or an exceptional case^[2]. The imbalanced dataset could be determined by the value of imbalance ratio (IR) which computed as follows^[21]:

$$IR = \frac{N^-}{N^+} \quad (4)$$

Where N^- is the number of examples of majority classes also called negative classes and N^+ is the number of examples of minority classes also called positive classes.

Most of the standard learning algorithms do not achieve the best performance when they used for imbalanced datasets for the following reasons^{[2][17][21]}:

- 1) The use of standard metrics such as the accuracy rate which gives an advantage to the majority class since it does not distinguish between the number of correct classification of the different classes.
- 2) The classification rules which predict the majority class are not useful for minority class since it is highly specialized and their coverage is very low, while minority class needs general rules to be predicted.
- 3) The examples of minority class appear as small clusters, also called small disjuncts, in the dataset surrounded by majority class examples as shown in Fig. 4(1). These clusters may be identified as a noise and therefore discarded by the classifier
- 4) The overlapping between classes when some examples from different classes with very similar characteristics placed in the regions around decision boundaries between classes as shown in Fig. 4 (2).
- 5) The higher or lower presence of borderline examples that located around the decision boundary between the classes. These borderline examples do not contribute to a correct class prediction. Fig. 5 shows the concept of borderline, noisy and safe examples.

Fig. 4. (1) small disjuncts problem and (2) overlapping between classes problem^[21].

Fig. 5. boarder line examples (labeled as b) locator around the decision boundary, noisy examples (labeled as n) and safe examples (labeled as s) ^[21].

5. Classification in Imbalanced Datasets

Different solutions have been proposed for classification imbalance datasets for standard learning algorithms and for ensemble techniques. All these solutions aim to improve the prediction of the minority class. They are categorized into three types which are ^{[12][21]}:

- 1) Algorithmic level solutions also called internal solutions: this category of solutions includes changing the search techniques, classification decision strategies or adjusting weights for each class.
- 2) Data level solutions also called external solutions: this category of solutions works on the input dataset level. It modifies the dataset distribution to reflect the balance between classes. It considered as more versatile than algorithmic level solutions since it is independent of the classifier. In addition to that, the computation time to prepare the data is only required once. There are three well-known rebalancing methods: (1) undersampling which

eliminates some of majority class examples; (2) oversampling which replicates some of the minority class examples or creating new ones from minority class examples; (3) Hybrid methods which use both previous rebalancing methods.

- 3) Cost-sensitive learning solutions: This category of solutions incorporates both the data and algorithmic level solutions and assumes higher misclassification costs to the samples of the minority class and seeks to decrease the high-cost errors.

In addition to that, ensemble techniques could be adapted to the classification of imbalanced datasets with the use of the previous three solutions. This is done by two different ways: (1) by modifying the ensemble learning algorithm to apply the preprocessing techniques on the data before the learning stage of each classifier; or (2) by incorporating a cost-sensitive method in the ensemble learning algorithm ^[2]. Also, ensemble techniques could be combined with OVO

decomposition for multi-class imbalanced datasets as in Ref. ^[3] where the problem is divided into a number of binary sub-problems to simplify the classification task. Then, each sub-problem is handled by an ensemble classifier.

Lopez V. *et al.* provided in Ref. ^[2] an experimental study of several methods including preprocessing methods, cost-sensitive methods and ensemble-based methods for dealing with imbalanced datasets problem. Important results are conducted. Among them: (1) the cost-sensitive methods are very accurate but the process of defining an optimal cost-matrix make them hard to be used; and (2) the preprocessing methods are robust and they obtained very good global results. So, they could be considered as a standard solution for imbalanced datasets problem; and (3) the ensemble-based methods are very accurate but they may have high learning time.

Synthetic Minority Over-sampling Technique (SMOTE) ^[22] is the most well-known and widely used oversampling technique. In this technique, the minority class is over-sampled by creating synthetic examples instead of over-sampling with replacement. The performance of SMOTE is better when it combined with other methods more than using it alone as proved in ^[22]. That's because of the process of creating new positive examples in SMOTE taking into account the closeness among positive examples and the number of examples of each class, without considering other characteristics of the dataset such as the distribution of examples from the majority classes. Therefore, many efforts try to improve the performance of SMOTE either by considering datasets characteristics when creating the minority examples or by integrating it with additional techniques such as filtering or cleaning methods ^[23].

SMOTE-IPF ^[21] is an extension of SMOTE. It deals with the problems produced by noisy and borderline examples in imbalanced datasets. It uses an iterative ensemble based noise filter technique called Iterative-Partitioning Filter (IPF). The performance of SMOTE-IPF is evaluated on two-class imbalanced datasets with different levels of noise and shapes of borderline examples and it compared against SMOTE and its most well-known generalizations including the Safe-Levels-SMOTE (SL-SMOTE) ^[23], the Borderline-SMOTE (B1-SMOTE and B2-SMOTE) ^[24], SMOTE with Edited Nearest Neighbors (SMOTE-ENN) ^[25] and SMOTE with Tomek Links (SMOTE-TL) ^[25]. The results show that SMOTE-IPF outperforms the rest of the methods.

6. Feature Selection and Classification in Imbalanced Datasets

Feature selection is one of the important data preprocessing techniques that used effectively in data mining. It reduces the effect of the curse of dimensionality and gives us a better understanding of the data by removing irrelevant, redundant and noisy data. Also, it improved the performance of data mining applications by reducing computation time, improving prediction accuracy and result comprehensibility ^[13].

The task of classification in the imbalanced dataset could be improved using feature selection because a large number of features can degrade the discovery of the borderline areas of the problem ^[13], either because some of these variables might be redundant or because they do not show a good synergy among them. For this reason, some works on the topic have proposed the use of feature selection for imbalanced datasets in order to overcome this problem ^[26], and to diminish the effect of overlapping ^[2].

Feature selection process consists of four steps which are subset generation, subset evaluation, stopping criterion and result validation. The first step includes searching for a candidate subset of features. The nature of this step determined by the search starting point and searching strategy. Searching could be forward, backward or bidirectional. In the second step, the candidate subset features are evaluated among evaluation criteria which could be independent or dependent on the mining algorithm. The first two steps are repeated until the stopping criterion is satisfied which means the feature subset is the best. Finally, the selected subset of features is validated ^[13]. The four steps of feature selection process illustrated in Fig. 6.

Feature selection could be categorized into three models: filter, wrapper and hybrid model. These models differ in the methodology of incorporating the learning algorithms with feature selection process. For the imbalanced data learning, filter and wrapper models are the most commonly used ^[27].

In filter model, the variable ranking technique is used to filter out less relevant variables. The choice of these variables is independent of the learning algorithm. The filtering method is simple, computationally light, and it avoids over-fitting. Unlike filter method which relies just on the characteristics of the dataset for features selection, wrapper method incorporates the predictor performance as the evaluation criteria for feature evaluation and selection. As a result of that, the selected features are more relevant to the learning algorithm, which lead to improving the classification performance. Wrapper algorithms are more computationally expensive than the filter algorithms ^{[13][27]}.

The last category of feature selection models is the embedded model. It takes the advantages of both previous models by

applying their different evaluation criteria through various stages. The independent measure is used to decide the best subsets and then the learning algorithm is used to decide the final best subset among the best subsets ^{[13][27]}. An example of the embedded method is the study in ^[28] for binary class problems and ^[29] for multiclass problems. Both studies designed to deal with class-imbalance and high dimensionality issues with the use of SVM.

Feature selection could be combined with the ensemble techniques or the resampling methods or for classification in imbalanced datasets. Yang P. *et al.* proposed in [30] an ensemble-based wrapper approach for feature selection from highly imbalanced multi-class datasets. Also, Cao P. *et al.* proposed in ^[31] a combination of oversampling and feature selection in a unified kernel framework for the binary-class imbalanced problems. The oversampling method is applied to the subset of features that resulted from the feature selection method as illustrated in Fig. 7.

Another example of using feature selection in the imbalanced problems is the study in Ref. ^[14]. This study is dedicated to solving the multiclass problems either by using a classifier that can deal directly with the problem or by dividing the problem into multiple binary problems using OVO or OVA. A discretization method is applied over the whole dataset for solving the problems of the imbalance values. Then, a subset of features is selected using filter methods before the classification. The experimental results show that the use of feature selection with problem decomposition leads to better performance results than not using it.

7. Comparative Overview of the Related Solutions for Classification Imbalanced Datasets

Table 1 presents a comparative overview of the related solutions that suggested in the

field of classification of imbalanced datasets and discussed in the previous sections. This comparison explores the following:

- 1) The technique that used in the proposed study.
- 2) The type of the solved problem: multiclass or binary class.
- 3) The use of OVO decomposition scheme.
- 4) The use of feature selection (FS).
- 5) The level of feature selection: global (prior problem decomposition) or local (for each binary problem).

- 6) The type of the dataset: balanced or imbalanced.

From this table, we could observe that feature selection methods could be used in the classification of imbalanced problems either binary class problems as in Ref. [28] [31], or multiclass problems as in Ref. [29] [30] [14]. In all these studies, feature selection applied globally as a preprocessing step on the whole dataset in the case of binary problems and prior the problem decomposition in the case of multiclass problems. Moreover, feature selection could be used with oversampling methods [31] or with ensemble techniques [30] as a solution for the classification in imbalanced datasets.

Fig. 6. Feature selection steps.

Fig. 7. Combining feature selection with oversampling. (minority class is labeled by +1 and majority class is labeled by -1).

Table 1. Comparative overview of the related solutions that suggested in the field of classification imbalanced datasets ^[31].

Ref.	Technique	Multi-class problem	Use of OVO	Use of FS	Level of FS	Imbalance Datasets
[22]	oversampling	X	X	X	-	✓
[21]	oversampling and filtering	X	X	X	-	✓
[3]	ensemble technique	✓	✓	X	-	✓
[28]	embedded feature selection	X	X	✓	Global	✓
[29]	embedded feature selection	✓	X	✓	Global	✓
[31]	oversampling and feature selection	X	X	✓	Global	✓
[30]	ensemble technique and wrapper feature selection	✓	X	✓	Global	✓
[14]	discretization and feature selection	✓	✓	✓	Global	✓

8. Conclusion

Classification in imbalanced datasets is one of the recurring problems in real-world applications of classification. It considered as a challenge since it needs to deal with uneven distribution of examples in the training datasets that lead to generate sub-optimal classification models. The presence of multiple classes implies an additional since the relations between the classes tend to complicated. This paper shows this problem in details and provides the proposed solutions in this field with a focus on feature selection and problem decomposition for multi-class problems. At the end, we should outline the most important remarks which help any researcher in this field based on the review in this paper:

- 1) OVO and OVA used as decomposition schemes to facilitate classifying multiclass datasets via binarization ^[8].
- 2) OVO could be better than OVA in most cases of classification multiclass datasets as discussed in ^[11] and proved in ^{[12][8]}.
- 3) Aggregation of predictions is a critical step in OVO decomposition scheme and it affects the result ^{[12][8]}.

- 4) WV is an effective aggregation scheme since it has been proved its robustness theoretically ^[11] and experimentally ^[8].
- 5) DRCW-OVO aggregation scheme could manage the non-competent classifiers ^[19].
- 6) Standard metrics such as accuracy rate is not suitable to be used in imbalanced dataset problems ^[21].
- 7) Data level solutions for imbalanced data problems is more versatile than algorithmic level solutions ^{[12][2][21]}.
- 8) Filter and wrapper models are the most common feature selection models that used in imbalanced dataset problems ^[27].
- 9) Problem decomposition using OVO could be used with feature selection as in Ref. ^[14] to solve the imbalance problem for multi-class datasets.
- 10) Feature selection could be combined with oversampling methods ^{[31] [14]} or with ensemble techniques ^[30] as a solution for the classification in imbalanced datasets.

References

- [1] **Galar, M., Fernández, A., Barrenechea, E., Bustince, H. and Herrera, F.** (2013) "Dynamic classifier selection for One-vs-One strategy: Avoiding non-competent classifiers," *Pattern Recognit.*, **46**(12): 3412–3424.
- [2] **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," *Inf. Sci. (Ny)*, **250**: 113–141.
- [3] **Krawczyk, B.** (2016) "Combining One-vs-One Decomposition and Ensemble Learning for Multi-class Imbalanced Data," in: *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015. Advances in Intelligent Systems and Computing*, **403**: 27–36.
- [4] **Knerr, S., Personnaz, L. and Dreyfus, G.** (1990) "Single-layer learning revisited: a stepwise procedure for building and training a neural network," *Neurocomputing*, pp.: 41–50.
- [5] **Hastie, T.** (1998) "Classification by Pairwise Coupling," *Ann. Stat.*, **26**(2): 451–471.
- [6] **Anand, Mehrotra, R., K., Mohan, C. K. and Ranka, S.** (1995) "Efficient classification for multiclass problems using modular neural networks," *IEEE Trans. Neural Netw.*, **6**(1): 117–124.
- [7] **Clark, P. and Boswell, R.** (1991) "Rule Induction with CN2: Some Recent Improvements," in *Fifth European Conference (EWSL)*, 1991, pp.: 151–163.
- [8] **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," *Pattern Recognit.*, **44**(8): 1761–1776.
- [9] **Alcalá-Fdez, J., Sánchez, L., García, S., Jesus, M. J., Ventura, S. J., Garrell, M., Otero, J., Romero, C. Bacardit, J., Rivas, V. M., Fernández, J. C. and Herrera, F.** (2009) "KEEL: a software tool to assess evolutionary algorithms for data mining problems," *Soft Comput.*, **13**(3): 307–318.
- [10] **Hall, M., National, H., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H.** (2009) "The WEKA Data Mining Software: An Update," *SIGKDD Explor.*, **11** (1): 10–18.
- [11] **Hüllermeier, E. and Vanderlooy, S.** (2010) "Combining predictions in pairwise classification: An optimal adaptive voting strategy and its relation to weighted voting," *Pattern Recognit.*, **43**(1): 128–142.
- [12] **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," *Knowledge-Based Syst.*, **42**: 97–110.
- [13] **Liu, H., Member, S., Yu, L. and Member, S.** (2005) "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, **17**(4): 491–502.
- [14] **Sanchez-Marono, N., Alonso-Betanzos, A., Gracia-Gonzalez, P. and Bolon-Canedo, V.** (2010) "Multiclass classifiers vs multiple binary classifiers using filters for feature selection," *Ieee*, pp.: 18–23.
- [15] **Lê Cao, K. A., Bonnet, A. and Gadat, S.** (2009) "Multiclass classification and gene selection with a stochastic algorithm," *Comput. Stat. Data Anal.*, **53**(10): 3601–3615.
- [16] **Yoon, S. and Kim, S.** (2009) "Mutual information-based SVM-RFE for diagnostic classification of digitized mammograms," *Pattern Recognit. Lett.*, **30**(16): 1489–1495.
- [17] **Allwein, E. L., Schapire, R. E. and Singer, Y.** (2001) "Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers," *J. Mach. Learn. Res.*, **1**: 113–141.
- [18] **Friedman, J. H.** (1996) "Another Approach to Polychotomous Classification".
- [19] **M. Galar, A. Fernández, E. Barrenechea, and F. Herrera,** "DRCW-OVO: Distance-based relative competence weighting combination for One-vs-One strategy in multi-class problems," *Pattern Recognit.*, vol. 48, no. 1, pp. 28–42, 2015.
- [20] **Galar, M., Fernández, A., Barrenechea, E. and Herrera, F.** (2014) "Empowering difficult classes with a similarity-based aggregation in multi-class classification problems," *Inf. Sci. (Ny)*, **264**: 135–157.
- [21] **Sáez, J. A., Luengo, J., Stefanowski, J. and F. Herrera,** (2015) "SMOTE – IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Inf. Sci. (Ny)*, **291**: 184–203.
- [22] **Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P.** (2002) "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, **16**: 321–357.
- [23] **Bunkhumpornpat, C., Sinapiromsaran, K. and Lursinsap, C.** (2009) "Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, **5476** LNAI: 475–482.
- [24] **Han, H., Wang, W. Y. and Mao, B. H.** (2005) "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," pp.: 878–887.
- [25] **Batista, G. E. A. P. A., Prati, R. C. and Monard, M. C.** (2004) "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newsl.*, **6**(1): 20–29.

- [26] **Villar, P., Fernández, A., Carrasco, R. A. and Herrera, F.** (2012) “Feature Selection and Granularity Learning in Genetic Fuzzy Rule-Based Classification Systems for Highly Imbalanced Data-Sets,” *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, **20**(3): 369–397.
- [27] **Chandrashekar, G. and Sahin, F.** (2014) “A survey on feature selection methods,” *Comput. Electr. Eng.*, **40**(1): 16–28.
- [28] **Maldonado, S., Weber, R. and Famili, F.** (2014) “Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines,” *Inf. Sci. (Ny)*, **286**: 228–246.
- [29] **Maldonado, S. and López, J.** (2018) “Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification,” *Appl. Soft Comput. J.*, **67**: 94–105.
- [30] **Yang, P., Liu, W., Zhou, B. B., Chawla, S. and Zomaya, A. Y.** (2013) “Ensemble-based wrapper methods for feature selection and class imbalance learning,” in: Pei J., Tseng V.S., Cao L., Motoda H., Xu G. (eds) *Advances in Knowledge Discovery and Data Mining. PAKDD*, **7818**: 544–555.
- [31] **Cao, P., Liu, X., Zhang, J., Zhao, D., Huang, M. and Zaiane, O.** (2017) “ $\ell_{2,1}$ norm regularized multi-kernel based joint nonlinear feature selection and over-sampling for imbalanced data classification,” *Neurocomputing*, **234**(June 2016): 38–57.

تصنيف البيانات غير المتوازنة: استخدام التصنيف الثنائي واختيار الخصائص

تهاني سعد المشدق^١، و صالح محمد الشمراني^٢، و عياد أحمد البشري^١

^١ قسم علوم الحاسبات، كلية الحاسبات وتقنية المعلومات، جامعة الملك عبدالعزيز، و ^٢ قسم نظم المعلومات، كلية

الحاسبات وتقنية المعلومات، جامعة جدة، جدة، المملكة العربية السعودية

tsalmoshadak@kau.edu.sa

المستخلص. يعد تصنيف البيانات غير المتوازنة من أكثر المشاكل تكرارًا في مجال تطبيقات التصنيف، كما يعد أحد التحديات؛ لأنه يؤدي إلى نتائج تصنيف خاطئة؛ بسبب الاختلاف الشديد في عدد العناصر التي تنتمي إلى الفئات، فهناك فئات ذات أكثرية ينتمي لها العدد الأكبر من العناصر، وفئات ذات أقلية ينتمي لها عدد قليل جدًا من العناصر، إضافة إلى ذلك يمكن أن تتزامن مشكلة عدم توازن البيانات مع مشكلة تعدد الفئات مما يؤدي إلى المزيد من التعقيد؛ لأنه يمكن أن تكون هناك فئة ذات أقلية بالنسبة لبعض الفئات، وفي نفس الوقت ذات أكثرية بالنسبة لفئات أخرى. يمكن تحسين أداء التصنيف المتعدد من خلال تقسيم المشكلة إلى عدد من المشكلات الصغيرة باستخدام إحدى طرق التقسيم التي من أشهرها OVO، حيث يتم تقسيم المشكلة إلى عدد من المشكلات الصغيرة تدعى بالمشكلات الثنائية؛ لكي يسهل التعامل معها، كل مشكلة ثنائية تختص بفئتين فقط ويعين لها مصنف خاص بها. كما يمكن تحسين تصنيف البيانات غير المتوازنة عن طريق استخدام تقنية اختيار الخصائص، التي تقوم بحذف الخصائص المكررة، وحذف التشويش، واستبعاد الخصائص التي ليس لها علاقة؛ وبالتالي تزداد سرعة عملية التصنيف، ودقة المصنف. من هذا المنطلق قدمنا هذه الورقة العلمية التي تتضمن مسح شامل لطرق تقسيم مشكلة الفئات المتعددة، وأساليب التجميع المستخدمة مع OVO، وشرح لمشكلة البيانات غير المتوازنة، وطرق التعامل مع تصنيف البيانات غير المتوازنة، كما تم شرح أساليب اختيار الخصائص مع ذكر فائدتها في مجال تصنيف البيانات غير المتوازنة. وفي النهاية تم عقد مقارنة بين الدراسات السابقة التي تدرج تحت مجال تصنيف البيانات غير المتوازنة، وتم تقديم ملخص لأهم النقاط التي استخلصناها من هذا المسح والتي تفيد أي باحث في هذا المجال.

الكلمات المفتاحية: التصنيف الثنائي، التصنيف المتعدد، مشكلة عدم التوازن، طرق التقسيم، OVA، OVO، أساليب التجميع، اختيار الخصائص.