# Combining Data Mining Techniques to Enhance Cardiac Arrhythmia Detection*

Christian Gomes[1], Alan Cardoso[1], Thiago Silveira[2],
Diego Dias[1], Elisa Tuler[1], Renato Ferreira[3], and Leonardo Rocha[1]

[1] Universidade Federal de São João del-Rei, São João del-Rei, Brazil
{christian,alanc,diegodias,etuler,lcrocha}@ufsj.edu.br
[2] Tech., Tsinghua University, Beijing, China
zhuangzq16@mails.tsinghua.edu.cn
[3] Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
renato@dcc.ufmg.br

**Abstract.** Detection of Cardiac Arrhythmia (CA) is performed using the clinical analysis of the electrocardiogram (ECG) of a patient to prevent cardiovascular diseases. Machine Learning Algorithms have been presented as promising tools in aid of CA diagnoses, with emphasis on those related to automatic classification. However, these algorithms suffer from two traditional problems related to classification: (1) excessive number of numerical attributes generated from the decomposition of an ECG; and (2) the number of patients diagnosed with CAs is much lower than those classified as "normal" leading to very unbalanced datasets. In this paper, we combine in a coordinate way several data mining techniques, such as clustering, feature selection, oversampling strategies and automatic classification algorithms to create more efficient classification models to identify the disease. In our evaluations, using a traditional dataset provided by the UCI, we were able to improve significantly the effectiveness of Random Forest classification algorithm achieving an accuracy of over 88%, a value higher than the best already reported in the literature.

**Keywords:** Cardiac Arrhythmia Detection · Automatic Classification · Machine Learning.

## 1 Introduction

Cardiovascular diseases are still one of the leading causes of death in the world. One of the major abnormalities associated with these diseases is Cardiac Arrhythmia (CA), which can be detected by the specialist through a clinical analysis of the patient's electrocardiogram (ECG). Early detection of CA can aid in treatment, significantly reducing the risk of patient's life. However, their discovery in the onset of the first clues is a difficult task since they involve evaluating the several variables present in an ECG.

* This work was partially supported by CNPq, CAPES, Fapemig, INWEB and MAsWeb.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_24

In order to assist specialists in the diagnosis of cardiovascular diseases, a recent and promising line of research has been adopted, that corresponds the use of methods based on Machine Learning in the detection of Arrhythmia [18]. From a previous set of ECG examinations duly and manually classified by medical specialists, a learning technique is applied resulting in a classification model. So, this model can be used by the physician in the evaluation/classification of new patient's ECG. However, the process of creating effective classification models is challenged by two main issues: (1) each ECG consists of a very large set of attributes; and (2) datasets related to ECG assessments are very unbalanced, since the number of patients diagnosed with CA is much smaller than those classified as "normal". While the first question is related to computational cost [24], the second one limits the learning process of the smaller classes [7], which are precisely the targets of the models in this scenario.

The questions mentioned above can be solved employing some data preprocessing strategies, on which the most common ones are Feature Selection techniques (FS) [24, 2, 17] and oversampling approaches [7, 4, 8]. FS consists of techniques that can measure the importance of each attribute in the construction of the classification model for a given base, returning those attributes more relevant, aiming to solve the first question previously presented. Oversampling consists of replicate/combine samples related to smaller classes, generating new samples to compose the dataset with a smaller unbalance, increasing the amount of information associated with the smaller classes, thus relating to the second question. Regarding the techniques of oversampling, although we find significant results in the literature related to efficacy in collections of data whose unbalance is even more pronounced, as in the CA detection scenario, the excessive generation of artificial samples can generate distortions which compromise the effectiveness of the classification model generated. From this finding, recently in [26], the authors present a technique called Classification using lOcal clustering with OverSampling (COG-OS), which consists, briefly, in the application of some clustering technique in classes splitting them into other smaller classes and then applying oversampling techniques considering the new distribution of generated classes. The authors' premise is that fewer artificial samples need to be generated, thereby reducing distortions in the generation of the classification model.

Therefore, in this paper, we proposed the combination, in a coordinated way, of several data mining and data preprocessing techniques aiming at the generation of more efficient (lower computational cost) and efficient classification models for the CA detection problem. More specifically, different classification algorithms were evaluated, combined with FS, clustering and oversampling techniques. To evaluate our proposal, we consider one of the collections of data related to the CA more referenced in the literature [16]. In our experimental analysis, we demonstrate that these strategies are complementary and, when appropriately combined, can result in a more efficient classification model. For example, while a classification model constructed based on the algorithm Random Forest using the collection of data without any preprocessing results in an

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_24

accuracy of 63%, the model generated after the application of an FS technique achieve an accuracy of 72%. Moreover, the model generated after the application of clustering and oversampling strategies results in an accuracy of approximately 82%. Finally, the model that combines all these strategies achieves an accuracy of over 88%, a value higher than the best already reported in the literature.

**Roadmap.** The remaining of paper is organized as follows. Section 2 presents some related works. The work methodology is presented in Section 3. In Section 4 the results of the experimental evaluation are discussed and the conclusions and future work are presented in Section 5.

## 2 Related Work

In recent years, several investigations related to the classification of CA have been performed, with the primary objective being the detection of arrhythmia using classification models. Felipe et al. [19] developed some CA classification models using eight different sets of variables related to the onset of CA in people. These variables were collected in real time from patients hospitalized at the Hospital Center of Porto, such as vital signs, laboratory results, among others. These are well-controlled data (not public) and related only to hospitalized patients, resulting in a relatively balanced collection, different from the collection considered in our study. Using the SVM classification algorithm, the authors achieved a 95% of accuracy.

Samad et al. [22] compared three classifiers based on their accuracy for the detection of the cardiac arrhythmia in the UCI dataset [16], the same one used in our paper. The classification algorithms k-NN, Naive Bayes and Decision Tree were used. The most relevant result was obtained by k-NN, reaching an accuracy of 66.96%. Moreover, this paper provides a detailed explanation of the conversion of an ECG into numerical values for using in machine learning tasks. Shivajirao et al. [14] created an intelligent system based on artificial neural networks to determine the classification of the presence or absence of CA, also using the UCI dataset. The authors used the Multilayer Perceptron model with the Backpropagation technique, reaching an accuracy of 86.67%, the best-reported result in the literature for this collection. As we will show in Section 4, combining Feature Selection (FS), clustering and oversampling techniques, we achieve superior results (i.e., 88.8% accuracy).

An FS technique is used to designate a score for each attribute to assess its importance in the learning task. In [28], the authors compare the performance of some metrics, such as Information Gain (IG), $\chi^2$, Odds Ratio (OR) and Correlation Coefficient (CC). In our paper, we consider *CfsSubsetEval* [12]. Concerning clustering techniques, there are several proposals in the literature [5]. These are from straightforward and usable techniques in several scenarios, such as K-Means [10], to some more elaborate and accurate for certain contexts, such as subspace clustering [1] and partitioning clustering [5]. In our paper, we consider only the K-Means, but other strategies can be evaluated in the future, as detailed in Section 5.

Finally, regarding oversampling techniques, Wu et al. [26] have developed an approach to address the problem of class unbalance that overcomes the other ones to predict rare classes. The method, titled Classification using lOcal clustering (COG), applies a clustering technique to divide the major classes into smaller subclasses. A significant improvement in the efficacy of supervised classification algorithms was observed in the results. A variation of the COG was also shown by applying the local clustering method together with an oversampling technique. This change was called Classification using lOcal clustering with OverSampling (COG-OS), being one of the techniques adopted in our approach.

## 3 Methodology

In this section, we present the methodology adopted to combine different data mining techniques, such as feature selection, oversampling and automatic supervised classification algorithms to improve the process of automatically identifying Cardiac Arrhythmia. First, we present the techniques considered by each step of our methodology. Next, we present the different strategies followed by the methodology to combine the techniques, which corresponds to evaluating the classification algorithms applying different data preprocessing approaches (i.e., feature selection, clustering and oversampling). Finally, we present the metrics adopted to evaluate each one of the combinations.

### 3.1 Data Mining Techniques

In this section, we present the algorithms considered in our paper. For all of them, we adopt the implementations provided by Weka [25], an educational software package, which has several data mining algorithms implemented, including classification, feature selection, oversampling and clustering. Next, we detail these algorithms.

**Feature Selection** For this paper we considered the *CfsSubsetEval* [12], which calculates, for each subset of attributes, the correlation of each of them with the dataset classes. In this case, it is desirable the subset that has a high correlation with a class in which each attribute of the subset has a low correlation with each other. Thus, it adds/removes attributes until it reaches a subset that has only the most relevant attributes to predict the desired class.

**Clustering** The clustering algorithm chosen was the K-Means [9], which consists of partitioning the objects into $K$ groups where each object belongs to a group. The algorithm creates $K$ centers in the object space and continues to change the location of its centers until the number of objects in each center from one iteration to another does not modify. The WCSS value is determined by Equation 1, where $S_k$ is the set of observations in the kth cluster and $\bar{x}_{kj}$ is the jth variable of the cluster center for the kth cluster.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_24

To determine the number $K$, the *Within-Cluster Sum of Squares* (WCSS) value must be analyzed, which is the sum done within each cluster between its objects and its center squared. It is necessary to observe the WCSS metric ranging the value of $K$ (i.e., from 1 to 10).

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 \tag{1}$$

**Oversampling** The oversampling algorithm chosen was SMOTE [7], which consists of creating synthetic instances of rare classes. For each class that we want to create objects to make the dataset balanced, SMOTE uses $k$ neighbor objects to combine and generate a synthetic instance that is close to those $k$ objects.

**Classification Algorithms** In our analysis, supervised classification algorithms considered state-of-the-art have been chosen, which address the problem through different approaches. They are:

- **Naive Bayes:** probabilistic algorithm that calculates the probability of a given new instance belonging to each of the available classes in a collection. It is one of the most widely used learning machine methods that combine efficiency and simplicity [24, 23].
- **Random Forest:** it is an algorithm based on the bagging approach, in which a set of $m$ decision trees are trained considering different training set samples. Then, each of these trees is considered in the algorithm final decision to classify a new instance [20, 6].

  **Support Vector Machine (SVM):** this algorithm maps the training set as points in a vector space, trying to define the boundaries of the space that separates each one of the classes. New instances are mapped into this vector space and assigned to the class according to their location. It is considered the most effective algorithm in the literature [15].
- **k-Nearest Neighbor (k-NN):** this is a lazy nonlinear classification algorithm in which the classification consists of assigning a new instance for the majority class related to $k$ closest instances in a vector space [27, 21].

For all algorithms, weka default parameters were used.

### 3.2 Combination Approaches

In Figure 1 we illustrate the combination approaches proposed to be evaluated in this work. Below we detail each of them. For all them, it is important to provide a missing values treatment in the dataset, which consists of removing/replacing all missing values from the dataset attributes.

1. **Classification Without Preprocessing:** in this step, classification using the selected classification algorithms is done without any preprocessing;

2. **FS Classification:** in this step the goal is to apply the FS technique to remove attributes that do not add value to the classification, using only a subset of relevant attributes. After that, it is generated new classification models that are evaluated the quality achieved in comparison to the results without applying any preprocessing techniques;
3. **Classification with the COG-OS Technique:** for this step of the methodology, considering all the attributes, the COG-OS method mentioned in Section 2 is applied to the dataset. This method consists in applying the clustering algorithm in the majority class ("normal" ECG), redistributing its instances into $k$ smaller classes. Then, the oversampling technique is applied in the minority classes (arrhythmia ECG) to achieve a class balance in the dataset. Finally, the classification algorithms are applied again for a new round of results evaluation;
4. **Classification with FS and COG-OS Combined:** in this step the two techniques (FS and COG-OS) are applied together. The FS technique is applied to select the most relevant subset of attributes, and the COG-OS is applied to the resulting dataset. With the fully-treated dataset, all classification algorithms are executed, and the results are compared once again.

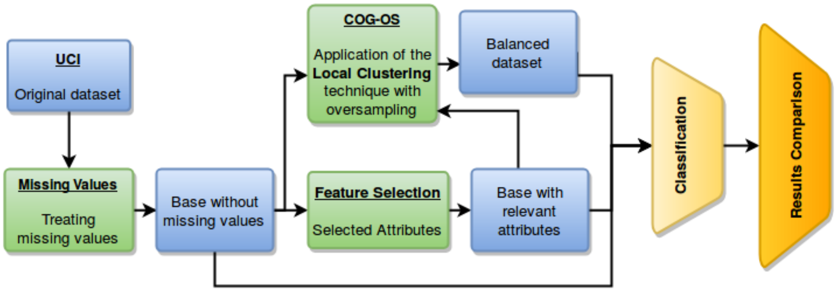

**Fig. 1.** Data Mining combination approaches for identifying Cardiac Arrhythmia.

### 3.3 Evaluation Metrics

In our evaluations, we consider two metrics: Accuracy and Macro F-Measure (Macro-F1). Accuracy measures the global effectiveness regarding all decisions made by the classifier (that is, the inverse of error rate). Macro-F1 on the other hand, measures the classification effectiveness regarding each class independently. It corresponds the mean of the F-Measure values obtained for each possible class in the dataset.

To define the F-Measure metric, we need to understand two main concepts:

- Precision: number of items classified as positive is positive;
- Recall: number of relevant items selected.

The F-Measure (**F1**) is the harmonic mean between precision and recall:

$$\text{F1} = 2 * \frac{\text{precision} * \text{recal}}{\text{precision} + \text{recall}} \tag{2}$$

We propose to use the K-fold Cross Validation Strategy [3] with K = 10, which consists of splitting the total dataset into ten mutually exclusive subsets of the same size, and from that, a subset is used for testing, and the remaining nine subsets are used for the model training. This process is repeated ten times, alternating the test subset. In the end, the reported results in the next section refer to the average of the Accuracies and Macro-F1 obtained in the ten repetitions.

## 4 Experimental Evaluation

In this section, we present the experimental results regarding each combination approach described in the previous section, considering a real dataset related to CA detection.

### 4.1 Experimental Setup

**Dataset** The dataset used was created by Guvenir et al. [11] and made available by UCI[4], being characterized by a transformation of ECGs into numerical attributes for the application of data mining algorithms. This base has missing values and ambiguous samples that need to be addressed for a more efficient use of classification algorithms. The original dataset has 280 attributes. The base has 16 classes; class 01 refers to normal ECGs; class 13 refers to ECGs that do not have a classification, and the others refer to ECGs with the presence of some arrhythmia. Three of these classes were disregarded because they did not have any associated instances. Figure 2 depicts the distribution of occurrences between classes. As we can see, this is a highly unbalanced dataset, so some classes of arrhythmia have two instances, while the normal ECG class has 245 instances.

**Treatment of Missing Values** In a previous dataset analysis, we identified that one of the attributes (V14) had 390 missing values instances, which was removed from our analyses. For the remainder of the attributes, the missing values treatment was performed using the packet *mice* provided in conjunction with the *R* language. This packet has a function for replacing incomplete values by synthetic plausible ones according to all columns, losing no data consistency. At all stages of our experimental evaluation, we used the data collection resulting from this treatment.

[4] https://archive.ics.uci.edu/ml/datasets/Arrhythmia

DOI: 10.1007/978-3-319-93701-4_24

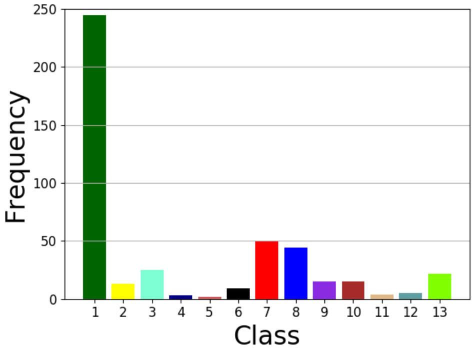

**Fig. 2.** Distributing instances between classes in the UCI dataset.

### 4.2 Analysis of Results

The first result was reached through the evaluation of classification algorithms without the use of preprocessing techniques. Table 1 presents the accuracy and Macro-F1 values achieved by each evaluated by the classification algorithm. As we can see, the algorithm Random Forest was the one that obtained the best value for Macro-F1 and accuracy in the unbalanced dataset, where Naive Bayes achieved an approximate value. The value achieved can be considerate low since in the unbalanced dataset most of the classes of arrhythmia are not classified correctly. That happens because the created models were trained on an unbalanced dataset, bias to normal class, which is the most frequent.

**Table 1.** Results achieved in the unbalanced original dataset classification.

| **Algorithm** | **Accuracy** | **Macro-F1** |
|---|---|---|
| Naive Bayes | 62.0% | 61.0% |
| Random Forest | 69.9% | 62.3% |
| k-NN | 58.1% | 45.6% |
| SVM Linear | 54.2% | 38.1% |

The second result set refers to the combination of the classification algorithms and the FS technique, and the results are presented in Table 2. The FS algorithm was able to decrease the number of attributes from the 280 ones presenting in the original dataset, selecting only the 23 most relevant attributes. We can observe that almost all classifiers, except SVM, get improvements in the classification quality considering only the 23 most relevant attributes. It is important to note that, in addition to the improving classification models, using FS techniques can also improve efficiency in the process to create classification models.

The third step consists in the use of the COG technique as a preprocessing step for classification. In the arrhythmia dataset, only the "normal" class has numerous objects, so the clustering is applied to divide it into smaller subclasses.

**Table 2.** Results achieved in classification after applying the FS technique.

| Algorithm | Accuracy | Macro-F1 |
|---|---|---|
| Naive Bayes | 68.4% | 66.2% |
| Random Forest | 75.7% | 72.7% |
| k-NN | 63.9% | 55.2% |
| SVM Linear | 54.2% | 38.1% |

The best value for WCSS was achieved for four clusters (K = 4). The last step is the application of the oversampling technique to obtain a more relevant balancing between the classes. The new class distribution achieved using this strategy is presented in Figure 3.

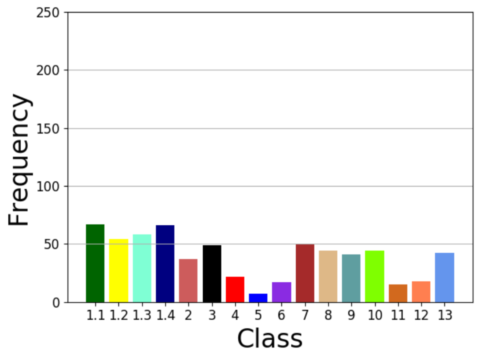

**Fig. 3.** Resulting dataset after applying the COG-OS approach.

Based on the resulting dataset from the application of the COG-OS, the classification was performed to compare with the previous results. Table 3 shows the classification results considering all the 280 original attributes of the dataset. We can observe that almost all classifiers, except SVM, get expressive improvements in the classification quality considering the COG-OS technique, demonstrating that efforts to mitigate unbalance between classes are able to improve considerably the quality of classification models for detecting Cardiac Arrhythmia.

**Table 3.** Results achieved in the classification after applying the COG-OS approach.

| Algorithm | Accuracy | Macro-F1 |
|---|---|---|
| Naive Bayes | 70.1% | 70.0% |
| Random Forest | 82.6% | 81.9% |
| k-NN | 65.6% | 62.5% |
| SVM Linear | 30.4% | 32.2% |

The fourth and final step consisted in combining the local clustering, oversampling and FS techniques as a preprocessing step, that is, COG-OS was applied to the dataset with only 23 attributes. The new class distribution achieved using this strategy is presented in Figure 4.

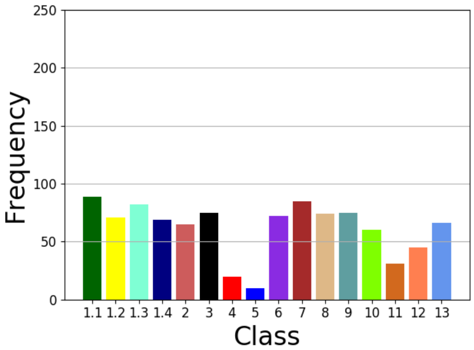

**Fig. 4.** Resulting dataset after applying the COG-OS approach considering the 23 most relevant attributes.

Table 4 presents the results achieved by the classification algorithms considering the dataset distribution shown in Figure 4. As we can see, the combination of techniques was very effective, further increasing the quality of the classifications. While the accuracy achieved by Naive Bayes in original dataset was 61%, for the preprocessed dataset the accuracy was 71.3%. A more expressive result is achieved by Random Forest algorithm, 62.3% of accuracy in original dataset and 88.8% in preprocessed dataset.

**Table 4.** Results obtained in the classification after applying the COG-OS approach considering the 23 most relevant attributes.

| **Algorithm** | **Accuracy** | **Macro-F1** |
|---|---|---|
| Naive Bayes | 71.9% | 71.3% |
| Random Forest | 88.9% | 88.8% |
| k-NN | 71.9% | 70.6% |
| SVM Linear | 29.4% | 32.2% |

### 4.3 Discussion

The FS techniques and the COG-OS method showed an excellent strategy in improving the effectiveness of the chosen classifiers, except the SVM. The algorithm that obtained the best results was the Random Forest, reaching a Macro-F1 of nearly 90%, making it the best result already reported in the literature for the

CA detection. Table 5 shows how it was possible to gradually increase the value of accuracy and, mainly, of the Macro-F1 value of the Random Forest algorithm. That is an important scientific breakthrough showing that the combination of different data mining strategies can significantly aid in the construction of classification models that assist medical specialists in the detection of CA.

**Table 5.** Random Forest result achieved at each step of our methodology.

| **Preprocessing Techniques** | **Macro-F1** |
|---|---|
| None | 62.3% |
| FS | 72.7% |
| COG-OS | 81.9% |
| **FS + COG-OS** | **88.8%** |

## 5 Conclusion and Future Works

In this paper, it has been demonstrated that the unbalance between classes of a dataset related to CA detection negatively influences the process of creating supervised classification models using traditional classifiers algorithms. The large majority of arrhythmia diagnoses are classified as normal, and cases of disease incidence are rare. In this way, several preprocessing strategies combined with automatic classification techniques were evaluated to create a more efficient classification models to assist specialists in the detection of the disease. More specifically, the results of this paper demonstrated that classification models constructed from a more relevant attributes subset, selected through an FS technique, tend to improve the quality of the models generated significantly. In an analogous and complementary way, it has been demonstrated that an oversampling strategy, combined with a clustering approach (COG-OS), also results in effective models. Besides, combining both strategies was achieved an even better classification model, surpassing the best result reported in the literature. More specifically, using the classification Random Forest algorithm, considering only the 23 most relevant attributes and applying the COG-OS oversampling strategy, a Macro-F1 of 88.8% was obtained, surpassing the 86% achieved in [14] for the same UCI dataset.

As a future work, the goal is to improve the prediction of arrhythmia further using other classification algorithms, clustering and oversampling [13] in the steps proposed in this work. Also, a detailed analysis of the 23 selected attributes can facilitate the arrhythmia detection in an ECG, finding out the relationships of these attributes in their respective ECG.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_24

## References

1. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications. In: Proc. of SIGMOD '98. pp. 94–105. ACM, New York, USA (1998)
2. Alelyani, S., Tang, J., Liu, H.: Feature selection for clustering: A review. Data Clustering: Algorithms and Applications **29**, 110–121 (2013)
3. Arlot, S., Celisse, A., et al.: A survey of cross-validation procedures for model selection. Statistics surveys **4**, 40–79 (2010)
4. Barua, S., Islam, M.M., Yao, X., Murase, K.: Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Transactions on Knowledge and Data Engineering **26**(2), 405–425 (2014)
5. Berkhin, P.: A survey of clustering data mining techniques. Grouping Multidimensional Data pp. 25–71 (2006)
6. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (Oct 2001)
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. J. Artif. Int. Res. **16**(1), 321–357 (Jun 2002), http://dl.acm.org/citation.cfm?id=1622407.1622416
8. Douzas, G., Bacao, F.: Self-organizing map oversampling (somo) for imbalanced data set learning. Expert Systems with Applications **82**, 40–52 (2017)
9. Faber, V.: Clustering and the continuous k-means algorithm. Los Alamos Science **22** (1994)
10. Farivar, R., Rebolledo, D., Chan, E., Campbell, R.H.: A parallel implementation of K-means clustering on GPUs. In: Proc. of PDPTA'08. pp. 340–345. USA (Jul 2008)
11. Guvenir, H.A., Acar, B., Demiroz, G., Cekin, A.: A supervised machine learning algorithm for arrhythmia analysis. In: Computers in Cardiology 1997. pp. 433–436. IEEE (1997)
12. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. Ph.D. thesis, University of Waikato, Hamilton, New Zealand (1998)
13. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: A new over-sampling method in imbalanced data sets learning. In: Proceedings of the 2005 International Conference on Advances in Intelligent Computing - Volume Part I. pp. 878–887. ICIC'05, Springer-Verlag, Berlin, Heidelberg (2005). https://doi.org/10.1007/11538059_91, http://dx.doi.org/10.1007/11538059_91
14. Jadhav, S.M., Nalbalwar, S., Ghatol, A.: Artificial neural network based cardiac arrhythmia classification using ecg signal data. In: Electronics and Information Engineering (ICEIE), 2010 International Conference On. vol. 1, pp. V1–228. IEEE (2010)
15. Joachims, T.: Advances in kernel methods. chap. Making Large-scale Support Vector Machine Learning Practical, pp. 169–184. MIT Press, Cambridge, MA, USA (1999), http://dl.acm.org/citation.cfm?id=299094.299104
16. Lichman, M.: UCI machine learning repository (2013), https://archive.ics.uci.edu/ml/datasets/Arrhythmia
17. Liu, H., Motoda, H.: Feature selection for knowledge discovery and data mining, vol. 454. Springer Science & Business Media (2012)
18. Özçift, A.: Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. Computers in biology and medicine **41**(5), 265–271 (2011)

19. Portela, F., Santos, M.F., Silva, Á., Rua, F., Abelha, A., Machado, J.: Preventing patient cardiac arrhythmias by using data mining techniques. In: Biomedical Engineering and Sciences (IECBES), 2014 IEEE Conference on. pp. 165–170. IEEE (2014)
20. Salles, T., Gonçalves, M., Rodrigues, V., Rocha, L.: Broof: Exploiting out-of-bag errors, boosting and random forests for effective automated classification. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 353–362. SIGIR '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2766462.2767747, http://doi.acm.org/10.1145/2766462.2767747
21. Salles, T., Rocha, L., MourÃčo, F., GonÃğalves, M., Viegas, F., Meira, W.: A two-stage machine learning approach for temporally-robust text classification. Information Systems **69**(Supplement C), 40 – 58 (2017). https://doi.org/https://doi.org/10.1016/j.is.2017.04.004, http://www.sciencedirect.com/science/article/pii/S0306437917301801
22. Samad, S., Khan, S.A., Haq, A., Riaz, A.: Classification of arrhythmia. International Journal of Electrical Energy **2**(1), 57–61 (2014)
23. Viegas, F., Gonçalves, M.A., Martins, W., Rocha, L.: Parallel lazy semi-naive bayes strategies for effective and efficient document classification. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 1071–1080. CIKM '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2806416.2806565, http://doi.acm.org/10.1145/2806416.2806565
24. Viegas, F., Rocha, L., GonÃğalves, M., MourÃčo, F., SÃą, G., Salles, T., Andrade, G., Sandin, I.: A genetic programming approach for feature selection in highly dimensional skewed data. Neurocomputing (2017). https://doi.org/https://doi.org/10.1016/j.neucom.2017.08.050, http://www.sciencedirect.com/science/article/pii/S0925231217314716
25. Weka: Weka - interface classifier. http://weka.sourceforge.net/doc.dev/weka/classifiers/Classifier.html (2016), [Online; accessed 02-December-2017]
26. Wu, J., Xiong, H., Wu, P., Chen, J.: Local decomposition for rare class analysis. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 814–823. ACM (2007)
27. Zhang, M.L., Zhou, Z.H.: A k-nearest neighbor based algorithm for multi-label classification. In: Granular Computing, 2005 IEEE International Conference on. pp. 718–721. IEEE (2005)
28. Zheng, Z., Wu, X., Srihari, R.: Feature selection for text categorization on imbalanced data. sigkddexpl **6**, 80–89 (2004)