Combinando Técnicas de Mineração de Dados para Melhorar o Processo de Detecção Automática de Arritmia Cardíaca*

Christian Gomes e Leonardo Rocha

Universidade Federal de São João del Rei, Brazil

{christian, lcrocha}@ufsj.edu.br

Abstract. *Automatic Classification Algorithms have been presented as promising tools in aid of detection of Cardiac Arrhythmia (CA). However, these algorithms suffer from two problems: (1) excessive numerical attributes generated from the decomposition of an Electrocardiogram (ECG); and (2) there are more patients diagnosed with CAs than classified as "normal"(unbalanced datasets). In this paper, we combine data mining techniques (i.e. clustering, feature selection and oversampling strategies) to create more efficient classification models. In our evaluations, using a dataset provided by the UCI, we improve significantly the effectiveness of Random Forest algorithm achieving an accuracy of over 88%, a value higher than the best already reported in the literature.*

Resumo. *Algoritmos de Classificação Automática são ferramentas promissoras no auxílio de diagnósticos de Arritmia Cardíaca (AC), entretanto sofrem com dois problemas: (1) muitos atributos numéricos gerados na decomposição de um Eletrocardiograma (ECG); e (2) o número de pacientes com ACs é muito menor do que aqueles tidos como normais (bases desbalanceadas). Nesse trabalho, combinamos técnicas de mineração de dados (i.e. clustering, feature selection e oversampling) para criar modelos de classificação mais eficazes. Em nossas avaliações, utilizando uma coleção da UCI, melhoramos significativamente a eficácia do algoritmo Random Forest, alcançando uma acurácia de 88%, valor superior ao melhor já reportado na literatura.*

1. Introdução

Doenças cardiovasculares ainda são uma das principais causas de morte no mundo. Uma das principais anormalidades associadas à essas doenças é a Arritmia Cardíaca (AC), que pode ser detectada pelo especialista por meio de uma análise clínica do Eletrocardiograma (ECG) do paciente. Uma recente e promissora linha de pesquisa que vem sendo adotada é o emprego de métodos baseados em Aprendizado de Máquina na detecção da AC [Jadhav et al. 2010]. A partir de um conjunto prévio de exames ECG's devidamente classificados por médicos especialistas, uma técnica de aprendizagem é aplicada gerando como resultado um modelo de classificação. Esse modelo então pode ser utilizado pelo médico para auxiliar na avaliação/classificação de ECG's de novos pacientes. Entretanto, a geração de modelos de classificação eficazes, nesse cenário, ainda é um desafio por duas questões principais: (1) cada ECG é composto por um conjunto muito grande de atributos; e (2) bases de dados relacionadas a avaliações de ECG são muito desbalanceadas, uma vez que o número de pacientes diagnosticados com AC é muito menor do que aqueles

*Esse trabalho foi parcialmente financiado por CNPq, CAPES, FINEP, Fapemig, e INWEB.

classificados como normais. Enquanto a primeira questão está relacionada ao custo computacional [Viegas et al. 2017], a segunda limita o processo de aprendizagem das classes menores [Chawla et al. 2002], sendo essas justamente os alvos dos modelos deste cenário.

As questões acima mencionadas podem ser resolvidas a partir da utilização de estratégias de pré-processamento de dados, sendo as mais comuns as métricas de Seleção de Atributos, ou em inglês *Feature Selection* (FS) [Viegas et al. 2017, Alelyani et al. 2013], e técnicas de Sobreamostragem, ou em inglês *Oversampling* (OVS) [Chawla et al. 2002, Barua et al. 2014, Douzas and Bacao 2017]. FS são métricas capazes de mensurar a importância de cada atributo na construção do modelo de classificação para uma determinada coleção, retornando aqueles atributos mais relevantes, visando, com isso, resolver a primeira questão previamente apresentada. OVS consiste de técnicas capazes de replicar/combinar amostras relacionadas às classes menores, gerando novas amostras para compor o conjunto de dados, reduzindo o desbalanceamento e aumentando a quantidade de informação relacionada às classes menores, estando relacionada à segunda questão. Com relação às técnicas de OVS, apesar de encontrarmos na literatura resultados importantes quanto à eficácia, em coleções cujo desbalanceamento é ainda mais acentuado, como no cenário de detecção de AC, a criação excessiva de amostras artificiais pode gerar distorções que comprometem a eficácia dos modelos de classificação.

Dessa forma, neste artigo propomos e avaliamos a combinação de técnicas de pré-processamento de dados visando a geração de modelos de classificação mais eficientes (menor custo computacional) e eficazes (melhor qualidade da classificação) para o problema de detecção da AC. Mais especificamente, foram avaliados diferentes algoritmos de classificação, combinados com técnicas de FS, técnicas de Clusterização, ou do inglês *Clustering*, e OVS. Foi utilizada uma das coleções de dados relacionadas à AC mais referenciadas na literatura, provida pela UCI [Dheeru and Karra Taniskidou 2017]. Em nossa avaliação experimental, demonstramos que se trata de estratégias complementares, que quando combinadas, resultam em modelos de classificação mais eficazes. Por exemplo, enquanto um modelo de classificação construído a partir do algoritmo *Random Forest* utilizando a coleção de dados sem nenhum pré-processamento resulta em uma acurácia de 63%, o modelo gerado após a aplicação de uma métrica de FS resulta em uma acurácia de 72%. Além disso, o modelo que combina clusterização e OVS resulta em uma acurácia de aproximadamente 82%. Por fim, o modelo que combina todas essas estratégias alcança uma acurácia de 88,8%, superior ao melhor já reportado.

A concepção da metodologia experimental, bem como todas as implementações, execuções dos experimentos e avaliação dos resultados foram realizadas pelo aluno Christian Gomes sob a orientação do professor Leonardo Rocha.

2. Trabalhos Relacionados

Nos últimos anos, várias pesquisas relacionadas à classificação de AC têm sido realizadas. Felipe et al. [Portela et al. 2014] desenvolveram modelos de classificação de AC utilizando oito conjuntos diferentes de variáveis relacionadas ao surgimento de AC. Essas variáveis foram coletadas em tempo real de pacientes internados no Centro Hospitalar do Porto, tais como sinais vitais, resultados de laboratórios, entre outros. Trata-se de dados bem controlados (não públicos) e relacionados somente a pacientes internados, resultando em uma coleção bastante balanceada, diferente da coleção considerada em

nosso trabalho. Utilizando o algoritmo SVM, os autores apresentaram modelos de classificação capazes de classificar pacientes com uma acurácia de 95%.

Samad et al. [Samad et al. 2014] compararam três classificadores com base em sua acurácia para a detecção da AC considerando a base de dados disponibilizada pela UCI [Dheeru and Karra Taniskidou 2017]. Os algoritmos de classificação k-NN, Naive Bayes e Árvore de Decisão foram utilizados. O resultado mais relevante foi obtido pelo k-NN, tendo alcançado uma acurácia de 66,96%. Esse trabalho fornece uma explicação detalhada sobre a conversão de um ECG em valores numéricos para serem usados em tarefas de aprendizagem de máquina. Jadhav et al. [Jadhav et al. 2010] criaram um sistema inteligente baseado em redes neurais artificiais para determinar a classificação da presença ou não da AC, também utilizando a base de dados da UCI. Os autores utilizaram o modelo *Multilayer Perceptron* com a técnica *Backpropagation*, alcançando uma acurácia de 86,67% o melhor resultado já reportado na literatura para essa coleção.

Métricas FS são utilizadas para designar pontuações para cada atributo, a fim de avaliar a sua importância na geração do modelo de classificação. Em [Zheng et al. 2004], os autores comparam o desempenho de diversas métricas, tais como Ganho de Informação (*Information Gain*), χ^2, *Odds Ratio* e Coeficiente de Correlação (*Correlation Coefficient*). Em nosso trabalho, consideramos *CfsSubsetEval* [Hall 1998]. No que se refere à clusterização, existem na literatura diversas propostas [Berkhin 2006] que vão desde técnicas simples e aplicáveis em diversos cenários, como o K-Means [Farivar et al. 2008], até algumas mais elaboradas e específicas para determinados contextos, tais como clusterização por subespaços [Agrawal et al. 1998] e clusterização por particionamento [Berkhin 2006]. Conforme veremos mais adiante, em nosso trabalho consideramos apenas o K-Means, entretanto outras estratégias podem ser avaliadas no futuro.

Finalmente, em relação às técnicas de OVS, Wu et al. [Wu et al. 2007] desenvolveram duas estratégias que superam as demais técnicas na tarefa de prever classes raras. Na primeira, denominada *Classification using lOcal clusterinG* (**COG**), os autores aplicam uma técnica de clusterização nas classes majoritárias, desmembrando-as em outras X classes menores, gerando, posteriormente modelos de classificação a partir da base de dados resultante. A segunda estratégia, denominada *Classification using lOcal clustering with OverSampling* (**COG-OS**), após a aplicação da clusterização das classes majoritárias, os autores propõem a aplicação de técnicas de OVS considerando a nova distribuição de classes geradas. A premissa dos autores é que menos amostras artificiais precisam ser geradas, diminuindo, assim, distorções na geração do modelo de classificação. Em ambos os casos foram observadas melhorias significativas na eficácia dos algoritmos considerados. Por apresentar resultados superiores, consideramos a estratégia **COG-OS**, bem como a combinação da mesma com métricas de FS.

3. Metodologia de Avaliação

Nessa seção apresentamos a metodologia utilizada para combinar diferentes técnicas de mineração de dados, tais como métricas de seleção de atributos (FS), estratégias de sobreamostragem (OVS) e algoritmos de classificação supervisionada automática para aprimorar o processo de identificar a presença da arritmia cardíaca em ECG's. Primeiramente, apresentamos as técnicas considerando cada etapa da metodologia. Depois, descrevemos as diferentes abordagens utilizadas para combinar as técnicas, que

corresponde à avaliação dos algoritmos de classificação aplicando diferentes abordagens de pré-processamento da base, i.e., FS, clusterização e OVS. Finalmente, apresentamos as métricas adotadas para avaliar cada uma das combinações.

3.1. Técnicas de Mineração de Dados

Nessa seção, apresentamos as métricas e técnicas utilizadas em nosso trabalho, organizadas de acordo com o objetivo de cada uma delas: Seleção de Atributos, Clusterização, Sobreamostragem e Classificação Automática.

3.1.1. Seleção de Atributos

Para esse trabalho, consideramos o algoritmo *CfsSubsetEval* [Hall 1998], que calcula, para cada subconjunto de atributos, sua correlação com as classes da coleção de dados. Neste caso, o objetivo é obter um subconjunto de atributos que possua alta correlação com uma determinada classe. Além disso, os atributos desse subconjunto devem apresentar baixa correlação entre si. Assim, o algoritmo adiciona/remove atributos até obter um subconjunto que possui apenas aqueles mais relevantes para predizer a classe desejada.

3.1.2. Clusterização

O K-Means [Faber 1994] foi o algoritmo de clusterização utilizado em nossos experimentos. Ele consiste em particionar os objetos em K grupos, onde cada objeto pertence a um grupo. O algoritmo cria K centros no espaço de objetos e muda a localização dos centros até que o número de objetos em cada centro, de uma iteração para outra, não se modifica. Para determinar o número K, a métrica *Within-Cluster Sum of Squares* (WCSS) foi analisada, que consiste na raiz quadrada da soma da distância de cada grupo entre seus objetos e seu centro. Observamos o valor de WCSS variando o K (i.e., de 1 a 10).

3.1.3. Sobreamostragem

Adotamos em nossos experimentos o algoritmo de OVS conhecido como SMOTE [Chawla et al. 2002]. Para cada classe rara que se deseja criar instâncias sintéticas para tornar o conjunto de dados balanceado, o algoritmo usa L objetos vizinhos para combinar e gerar instâncias sintéticas que devem ser próximas desses L objetos.

3.1.4. Classificação Automática

Em nossa análise, foram escolhidos algoritmos de classificação supervisionada automática considerados estado-da-arte, que tratam o problema por meio de diferentes abordagens. São eles:

- **Naive Bayes:** algoritmo probabilístico que calcula a probabilidade de uma nova instância pertencer a cada uma das classes disponíveis na coleção. É um dos métodos mais usados em aprendizado de máquina, que combina eficiência e simplicidade [Viegas et al. 2017].

- **Random Forest:** é um algoritmo baseado na abordagem *bagging*, no qual um conjunto de m árvores de decisão são treinadas considerando diferentes amostras de conjunto de treinamento. Então, cada uma dessas árvores é considerada na decisão final do algoritmo para classificar uma nova instância [Salles et al. 2015].
- **Support Vector Machine (SVM):** este algoritmo mapeia o conjunto de treinamento como pontos em um espaço vetorial, tentando definir o limite do espaço que separa cada uma das classes. Novas instâncias são mapeadas nesse espaço vetorial e atribuídas a uma classe de acordo com sua localização nesse espaço. É considerado o algoritmo mais eficaz na literatura [Joachims 1999].
- **k-Nearest Neighbor (k-NN):** este é um algoritmo de classificação não linear postergado ou, em inglês, *lazy nonlinear classification algorithm*. O processo de classificação consiste em atribuir uma nova instância para a classe majoritária relacionada com as k instâncias mais próximas no espaço vetorial.

3.2. Combinação de Técnicas

Nessa seção apresentamos as diferentes abordagens utilizadas para se combinar as métricas e algoritmos descritos nas seções anteriores com o objetivo de melhorar a efetividade dos modelos de classificação para o problema de detecção da arritmia cardíaca. Essas estratégias são detalhadas a seguir e ilustradas na Figura 1. Um passo comum a todas essas estratégias é a remoção/substituição de todos os valores faltantes.

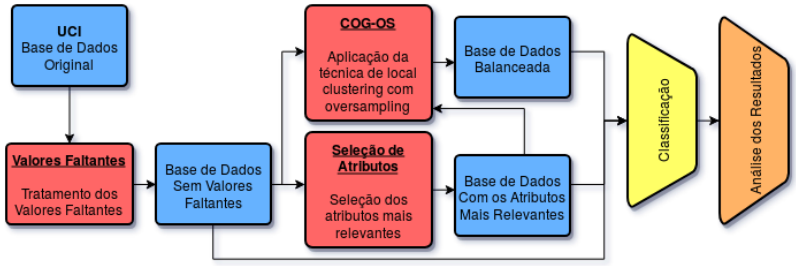

Figura 1. Abordagens de combinação de métricas e técnicas de mineração de dados na detecção automática de arritmia cardíaca.

1. **Classificação Sem Pré-Processamento:** nessa etapa, os modelos de classificação são gerados sem que nenhuma etapa de pré-processamento seja aplicada;
2. **Classificação com FS:** nessa etapa o objetivo é avaliar a métrica de FS para remover os atributos que não agregam valor no processo de geração do modelo de classificação, mantendo apenas aqueles mais relevantes. Os resultados de classificação obtidos por esses modelos são comparados aos resultados alcançados pelos modelos sem aplicação de nenhuma técnica de pré-processamento.
3. **Classificação com a Técnica COG-OS:** nessa etapa, considerando todos os atributos, utilizamos o método COG-OS mencionado na Seção 2. Esse método consiste na aplicação do algoritmo de clusterização na classe majoritária (ECG's normais), redistribuindo as instâncias em k classes menores. Então, a técnica de OVS é aplicada nas classes minoritárias (ECG's com arritmia) com objetivo de alcançar o balanceamento do conjunto de dados. Finalmente, aplicam-se novamente os algoritmos de classificação para uma nova rodada de avaliação de resultados.

4. **Classificação com a Combinação da Técnica de FS com COG-OS:** nessa etapa foram aplicadas as técnicas FS e COG-OS em conjunto. A técnica FS é aplicada para selecionar o subconjunto de atributos mais relevantes, e o COG-OS é aplicado no conjunto de dados resultante. Com a base totalmente tratada, todos os algoritmos de classificação são executados e os resultados são comparados mais uma vez.

3.3. Métricas de Avaliação

Em nossas avaliações consideramos duas métricas: acurácia e Macro F-Measure (**Macro-F1**). A acurácia mede a eficácia global em relação a todas as decisões tomadas pelo classificador (ou seja, o inverso da taxa de erro). Já a Macro-F1, mede a eficácia da classificação em relação a cada classe de forma independente. Ela corresponde à média dos valores de *F-Measure* obtido para cada classe possível no conjunto de dados. Para definir a métrica *F-Measure*, precisamos entender dois conceitos principais:

- *Precision*: quantidade itens classificados como positivos que são realmente positivos;
- *Recall*: quantidade de itens relevantes selecionados.

A *F-Measure* (**F1**) é a média harmônica entre **precision** e **recall**:

$$F1 = 2 * \frac{precision * recall}{precision + recall}. \quad (1)$$

Para avaliar e comparar a qualidade dos modelos de classificação das etapas ilustradas na Figura 1, consideramos a estratégia *10-fold Cross Validation* [Viegas et al. 2017]. Consiste em dividir o conjunto de dados total em 10 subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir disso, um subconjunto é usado para testes e os nove restantes são usados para o treinamento do modelo. Este processo é repetido 10 vezes, alternando o subconjunto de teste. No final, os resultados reportados na próxima seção referem-se à média das acurácias e Macro-F1 obtidas nas 10 repetições.

4. Avaliação Experimental

Nessa seção apresentamos os resultados referentes a cada abordagem descrita na seção anterior, considerando um conjunto de dados real relacionado à detecção de AC.

4.1. Ambiente Experimental

4.1.1. Base de Dados

A base de dados utilizada foi criada por Guvenir et al. [Guvenir et al. 1997] e disponibilizada pela UCI [1], sendo caracterizada por uma transformação de ECG's em atributos numéricos para a aplicação de técnicas de mineração de dados. Essa base possui valores faltantes e amostras ambíguas que precisam ser tratadas para uma utilização mais eficiente dos algoritmos de classificação. A base de dados original é composta de 280 atributos e 437 instâncias distribuídas desigualmente em 13 classes. A classe 01 refere-se aos ECG's normais, a classe 13 refere-se aos ECG's que não possuem classificação e as demais referem-se aos ECG's com presença de algum tipo de arritmia. A Figura 2(a) apresenta a distribuição das ocorrências entre as classes. Como podemos observar, trata-se de uma base de dados extremamente desbalanceada, de modo que algumas classes de arritmia possuem 2 instâncias, enquanto a classe de ECG's normais possui 245 instâncias.

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

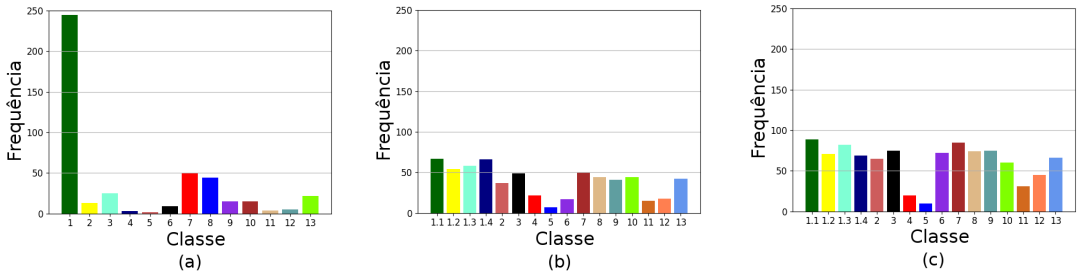

Figura 2. Distribuição das classes na base de dados: (a) Base de dados original. (b) Base de dados resultante da aplicação do COG-OS. (c) Base de dados resultante da aplicação do COG-OS considerando os 23 atributos mais relevantes.

4.1.2. Tratamento dos Valores Faltantes

Em uma análise prévia da base identificamos que um dos atributos (V14) possuía 390 instâncias com valores faltantes, o qual foi removido de nossas análises. Para o restante dos atributos, o tratamento dos valores faltantes foi efetuado utilizando o pacote *mice* disponibilizado junto com a Linguagem R. Esse pacote possui uma função para substituir valores incompletos por valores plausíveis sintéticos de acordo com todas as colunas da base, sem perder a consistência dos dados. Em todas as etapas de nossa avaliação experimental utilizamos a coleção de dados resultante desse tratamento.

4.2. Análise dos Resultados

O primeiro resultado foi obtido por meio da avaliação dos algoritmos de classificação sem o uso das técnicas de pré-processamento. Na Tabela 1 é mostrado os valores de acurácia e Macro-F1 obtidos por cada algoritmo de classificação avaliado. Como podemos observar, o algoritmo *Random Forest* foi que o obteve o melhor valor de Macro-F1 e acurácia na base desbalanceada, sendo que, o *Naive Bayes* obteve um valor aproximado. O valor obtido é considerado baixo, uma vez que na base desbalanceada a maioria das classes de arritmia não é classificada corretamente. Isso ocorre uma vez que os modelos foram treinados com uma coleção desbalanceada, enviesada pela classe normal que é mais frequente.

Algoritmo	acurácia	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%

Tabela 1. Resultados obtidos na classificação com a base desbalanceada.

O segundo conjunto de resultados diz respeito à combinação dos algoritmos de classificação com a métrica de FS, e os resultados são apresentados na Tabela 2. Por meio da métrica de FS, diminuímos significativamente o número total de atributos, considerando apenas os 23 atributos mais relevantes dos 280 atributos originais. Conforme podemos observar, para todos os algoritmos de classificação utilizados, com exceção do SVM, melhoramos a qualidade dos modelos. O destaque, novamente apresentando os melhores resultados, foi o algoritmo de Random Forest. É importante observar que, mesmo não sendo o foco desse trabalho, uma métrica FS também é capaz de contribuir para uma melhor eficiência no processo de criação dos modelos de classificação.

A terceira etapa consiste na utilização da técnica COG-OS como uma etapa de pré-processamento para a classificação. Na base de dados de arritmia, somente a classe

Algoritmo	acurácia	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%

Tabela 2. Resultados obtidos na classificação considerando apenas os 23 atributos mais relevantes definidos pela métrica de FS.

normal possui um grande número de objetos, logo, aplicamos a técnica de clusterização sob as instâncias dessa classe para que a mesma seja transformada em subclasses de tamanhos menores. O melhor valor para métrica *WCSS* foi obtido com quatro clusters ($k = 4$). A última etapa consiste na aplicação da técnica de OVS que visa criar instâncias para as classes menores e assim obter um balanceamento mais relevante entre as classes. A nova distribuição das classes obtida usando essa estratégia é mostrada na Figura 2(b). Com a base resultante da aplicação do COG-OS executamos os algoritmos de classificação selecionados para realizar uma comparação com os resultados obtidos anteriormente. A Tabela 3 apresenta o resultado, utilizando a base de dados resultante do COG-OS considerando todos os 280 atributos. Como podemos observar quase todos classificadores, exceto o SVM mais uma vez, obtiveram uma melhoria expressiva na qualidade da classificação. Esse resultado demonstra que o esforço para tratar o desbalanceamento entre as classes é capaz de melhorar consideravelmente a qualidade dos modelos de classificação na detecção da arritmia cardíaca.

Algoritmo	acurácia	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%

Tabela 3. Resultados obtidos na classificação depois da aplicação do COG-OS considerando todos os 280 atributos da coleção.

A quarta e última etapa consiste em combinar as técnicas de clusterização, OVS e FS no pré-processamento da coleção, ou seja, a técnica COG-OS é aplicado na base de dados com somente 23 atributos. Nenhum dos artigos mencionados na Seção 2 discute a combinação dessas técnicas, trata-se de uma abordagem originalmente apresentada nesse trabalho. Além disso, a avaliação dessa abordagem no cenário de detecção de arritmia cardíaca também é uma novidade. A nova distribuição das classes utilizando essa estratégia é mostrada na Figura 2(c). A Tabela 4 apresenta os resultados obtidos na aplicação dos algoritmos de classificação na base resultante dessa etapa. Como podemos observar, a combinação das técnicas foi bastante eficaz, aumentando ainda mais a qualidade dos modelos. Enquanto a acurácia obtida pelo Naive Bayes na base dados original foi 61%, com a base pré-processada a acurácia foi 71%. O melhor resultado foi obtido pelo algoritmo *Random Forest*, conseguindo obter 63% de acurácia na base de dados original e 88.8% com sua versão pré-processada.

Algoritmo	acurácia	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%

Tabela 4. Resultados obtidos na classificação depois da aplicação do COG-OS considerando os 23 atributos mais relevantes.

4.3. Discussão

A abordagem de combinar a métrica de FS e a estratégia COG-OS mostrou-se uma excelente alternativa em aprimorar a eficácia dos classificadores escolhidos, com exceção do SVM. O algoritmo que obteve os melhores resultados foi o Random Forest, tendo alcançado uma Macro-F1 de quase 90%, tornando o melhor resultado já reportado na literatura para a coleção de detecção de AC. Na Tabela 5 sumarizamos os resultados obtidos pelo Random Forest em cada uma das etapas de nossa metodologia. Conforme podemos observar, tanto a acurácia quanto a Macro-F1 foram melhorando à medida que introduzimos etapas diferentes de pré-processamento. Trata-se de um importante avanço científico que mostra que a combinação de diferentes estratégias de mineração de dados pode melhorar o processo de construção de modelos de classificação mais eficazes e que sejam capazes de auxiliar médicos especialistas na detecção de arritmia cardíaca.

Pré-Processamento	Macro-F1
Nenhum	62,3%
FS	72,7$
COG-OS	81,9%
FS + COG-OS	88,8%

Tabela 5. Resultado do Random Forest em cada etapa da nossa metodologia.

5. Conclusões e Trabalhos Futuros

Neste trabalho avaliamos diferentes técnicas de classificação supervisionada automática para o problema de Detecção Automática da Arritmia Cardíaca, uma das principais anomalias cardiovasculares causadoras de morte no mundo. A grande maioria dos diagnósticos de arritmia é classificada como normal e os casos de incidência da doença são raros, o que torna a tarefa de classificação automática ainda mais desafiadora (i.e. desbalanceamento entre classes). Dessa forma, avaliamos diferentes técnicas de pré-processamento de dados combinadas com técnicas de classificação automática no intuito de criar modelos de classificação mais eficazes para auxiliar especialistas na detecção da doença. Os resultados deste artigo demonstraram que modelos de classificação construídos a partir de um subconjunto de atributos mais relevantes, selecionado por meio de uma métrica de seleção de atributos, tendem a melhorar significativamente a qualidade dos modelos gerados. De maneira análoga e complementar, foi demonstrado que uma estratégia de sobreamostragem, combinada com uma abordagem de clusterização (COG-OS), também resulta em modelos eficazes. Além disso, a combinação de ambas as estratégias alcança um modelo de classificação ainda melhor, superando o melhor resultado reportado na literatura. Mais especificamente, utilizando o algoritmo de classificação Random Forest, considerando apenas os 23 atributos mais relevantes e aplicando a estratégia de COG-OS, foi obtida uma Macro-F1 de 88,8%, superando os 86% alcançados em [Jadhav et al. 2010] para a mesma base de dados da UCI utilizada. Como trabalho futuro, nosso o objetivo é avaliar outras métricas de seleção de atributos, bem como outros algoritmos de classificação, clusterização e sobreamostragem nas diversas etapas de nossa metodologia.

Referências

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. of SIGMOD '98*, pages 94–105, New York, USA. ACM.

Alelyani, S., Tang, J., and Liu, H. (2013). Feature selection for clustering: A review. *Data Clustering: Algorithms and Applications*, 29:110–121.

Barua, S., Islam, M. M., Yao, X., and Murase, K. (2014). Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(2):405–425.

Berkhin, P. (2006). A survey of clustering data mining techniques. *Grouping Multidimensional Data*, pages 25–71.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository.

Douzas, G. and Bacao, F. (2017). Self-organizing map oversampling (somo) for imbalanced data set learning. *Expert Systems with Applications*, 82:40–52.

Faber, V. (1994). Clustering and the continuous k-means algorithm. *Los Alamos Science*, 22.

Farivar, R., Rebolledo, D., Chan, E., and Campbell, R. H. (2008). A parallel implementation of K-means clustering on GPUs. In *Proc. of PDPTA'08*, pages 340–345, USA.

Guvenir, H. A., Acar, B., Demiroz, G., and Cekin, A. (1997). A supervised machine learning algorithm for arrhythmia analysis. In *Computers in Cardiology 1997*, pages 433–436. IEEE.

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand.

Jadhav, S. M., Nalbalwar, S., and Ghatol, A. (2010). Artificial neural network based cardiac arrhythmia classification using ecg signal data. In *Proc. of IEEE ICEIE*, volume 1, pages V1–228.

Joachims, T. (1999). Advances in kernel methods. chapter Making Large-scale Support Vector Machine Learning Practical, pages 169–184.

Portela, F., Santos, M. F., Silva, Á., Rua, F., Abelha, A., and Machado, J. (2014). Preventing patient cardiac arrhythmias by using data mining techniques. In *IEEE IECBES*, pages 165–170.

Salles, T., Gonçalves, M., Rodrigues, V., and Rocha, L. (2015). Broof: Exploiting out-of-bag errors, boosting and random forests for effective automated classification. In *Proc. of 38th ACM SIGIR*, pages 353–362.

Samad, S., Khan, S. A., Haq, A., and Riaz, A. (2014). Classification of arrhythmia. *International Journal of Electrical Energy*, 2(1):57–61.

Viegas, F., Rocha, L., Gonçalves, M., Mourão, F., Sá, G., Salles, T., Andrade, G., and Sandin, I. (2017). A genetic programming approach for feature selection in highly dimensional skewed data. *Neurocomputing*.

Wu, J., Xiong, H., Wu, P., and Chen, J. (2007). Local decomposition for rare class analysis. In *Proc. of 13th ACM SIGKDD*, pages 814–823.

Zheng, Z., Wu, X., and Srihari, R. (2004). Feature selection for text categorization on imbalanced data. *sigkddexpl*, 6:80–89.