FACULDADE DE MEDICINA DA UNIVERSIDADE DO PORTO

MESTRADO EM INFORMÁTICA MÉDICA

UNIDADE CURRICULAR DE SISTEMAS DE APOIO À DECISÃO CLÍNICA

ESTUDO COMPARATIVO

DE

TRÊS ALGORITMOS DE *MACHINE LEARNING*

NA

CLASSIFICAÇÃO DE DADOS ELECTROCARDIOGRÁFICOS

PORTO, 27 DE MARÇO DE 2009

FACULDADE DE MEDICINA DA UNIVERSIDADE DO PORTO

MESTRADO EM INFORMÁTICA MÉDICA

UNIDADE CURRICULAR DE SISTEMAS DE APOIO À DECISÃO CLÍNICA

Prof.ª Doutora Inês Dutra

ESTUDO COMPARATIVO

DE

TRÊS ALGORITMOS DE *MACHINE LEARNING*

NA

CLASSIFICAÇÃO DE DADOS ELECTROCARDIOGRÁFICOS

Autores:

António Cardoso Martins
João Miguel Marques
Paulo Dias Costa

PORTO, 27 DE MARÇO DE 2009

ÍNDICE

1. INTRODUÇÃO

Pretende-se com este trabalho conhecer ferramentas e métodos de análise e processamento, conhecer as principais técnicas de *data mining* e *machine learning*, realizar análise de dados e alterar parâmetros, utilizando a ferramenta informática WEKA[1]. Simultaneamente, pretende-se desenvolver a capacidade de pesquisa e competências na manipulação de ferramentas de *machine learning*.

Ao organizar-se como um grupo de indivíduos com *backgrounds* distintos, torna-se relevante a troca de experiências e complementaridade de ambos, por forma a potenciar um estudo transversal que vise a interacção entre os sistemas de informação e as ciências médicas.

[1] Waikato Environement for Knownledge Analysis. Disponível em www.cs.waikato.ac.nz/ml/weka.

2. CONCEITOS BÁSICOS

A necessidade de compreender os conceitos de Cardiologia subjacentes à análise do próprio *dataset*, fez com que tivessemos de pesquisar, *a priori*, suporte bibliográfico que nos permitisse compreender conceitos básicos de electrocardiografia por forma a, por um lado, identificar e analisar os conteúdos e estrutura de dados e, por outro, interpretar, compreender e extrapolar resultados.

2.1. ELECTROCARDIOGRAFIA

O Electrocardiograma (ECG) pode definir-se como a representação gráfica da actividade eléctrica do coração. A sua realização consiste na colocação de quatro eléctrodos ao nível distal dos membros superiores e inferiores e de seis eléctrodos ao nível do pré-córdio. Esta configuração permite obter 12 derivações, que traduzirão a actividade eléctrica em determinada região do coração.

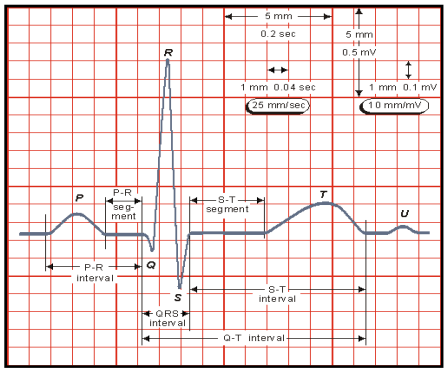

Figura 1 - Exemplificação dos diversos elementos constantes no electrocardiograma.

No ECG é possível visualizar pequenos acidentes (onda P, complexo QRS, onda T), entre os quais é possível também obter intervalos e segmentos. Por intervalo compreende-se o tempo mediado entre o início e o final de uma onda ou entre o início de uma onda e o início da onda seguinte. Já o segmento pode definir-se como o tempo compreendido entre o final de uma deflexão e o início de outra.

3. ESTUDO EXPERIMENTAL

3.1. DATASET

O *dataset* utilizado neste estudo foi obtido do repositório de dados para *machine learning* da Universidade da Califórnia[2]. O datset utilizado subdivide-se em três grupos principais, com o objectivo de facilitar o seu uso experimental e permitir a identificação da própria arritmia. Assim, a classe 1 refere-se a ECG normal, as classes 2 a 15 referem-se a ECG anormal e a classe 16 a dados não classificados. É originalmente composto por 452 instâncias, distribuídas como indicado acima, em 16 classes, e 279 atributos, sendo 206 numéricos e os restantes nominais. Existem, neste conjunto, alguns valores omissos. É fornecida também a classificação em 16 classes, realizada por um especialista da área.

Os dados foram originalmente disponibilizados num ficheiro com extensão .data, ao qual correspondia o formato *comma separated values* (.csv). Adicionalmente foi disponibilizado um ficheiro (arrhythmia.names) com a descrição dos atributos e proprietários da base de dados. Por forma a preparar os dados para utilização na ferramenta WEKA, foi necessário coligir toda a informação num só ficheiro com extensão .arff, onde etiquetamos cada atributo e o classificamos quanto ao tipo suportado pelo WEKA - numérico e nominal no nosso caso.

3.2. DESCRIÇÃO DOS ALGORITMOS UTILIZADOS

O algoritmo **OneR** cria uma regra para cada atributo dos dados de treino e selecciona a regra com menor percentagem de erro como regra única. Para criar uma regra para um atributo é necessário determinar a classe mais frequente para cada atributo. Como classe mais frequente entende-se a classe que aparece mais vezes para um dado atributo. “Uma regra” é simplesmente um conjunto de valores de atributos limitados pela sua classe maioritária. A percentagem de erro de uma regra é o número de instâncias de treino na qual a classe de um valor de atributo não é concordante com a classificação desse atributo na regra. Na eventualidade de duas ou mais regras possuirem a mesma percentagem de erro, a regra é escolhida ao acaso. Este algoritmo

é escolhido como base de comparação com outros algoritmos, devido à sua simplicidade e necessidade de apenas um atributo.

O algoritmo **J48** permite a criação de modelos de decisão em árvore. Utiliza uma tecnologia *greedy* para induzir árvores de decisão para posterior classificação. O modelo de árvore de decisão é construído pela análise dos dados de treino e o modelo utilizado para classificar dados ainda não classificados. O J48 gera árvores de decisão, em que cada nó da árvore avalia a existência ou significância de cada atributo individual. As árvores de decisão são construídas do topo para a base, através da escolha do atributo mais apropriado para cada situação. Uma vez escolhido o atributo, os dados de treino são divididos em sub-grupos, correspondendo aos diferentes valores dos atributos e o processo é repetido para cada sub-grupo até que uma grande parte dos atributos em cada sub-grupo pertençam a uma única classe. A indução por árvore de decisão é um algoritmo que habitualmente aprende um conjunto de regras com elevada acuidade. Este algoritmo é escolhido para comparar a percentagem de acerto com outros algoritmos.

O algoritmo **Naïve Bayes** é um dos mais simples classificadores probabilísticos. O modelo construído por este algoritmo é um conjunto de probabilidades. As probabilidades são estimadas pela contagem da frequência de cada valor de característica para as instâncias dos dados de treino. Dada uma nova instância, o classificador estima a probabilidade de essa instância pertencer a uma classe específica, baseada no produto das probabilidades condiconais individuas pra os valores característicos da instância. O cálculo exacto utiliza o teorema de Bayes e é por essa razão que o algoritmo é denominado um classificador de Bayes. O algoritmo é também denominado de Naïve, uma vez que todos os atributos são independentes dado o valor da variável da classe. Apesar deste pressuposto, o algoritmo apresenta um bom desempenho em muitos dos cenários de predição de classes. Estudos

[2] University of California. Machine Learning Repository. Disponível em: http://archive.ics.uci.edu/ml.

experimentais sugerem que este algoritmo tende a aprender mais rapidamente que a maioria dos algoritmos de indução e daí o seu uso na nossa análise.

3.3. METODOLOGIA

Os algoritmos utilizados para comparação foram o OneR, J48 e Naïve Bayes.

A metodologia adoptada passou, numa primeira fase, pelo substituição dos valores em falta por valores probabilísticos de acordo com a distribuição dos valores conhecidos dos atributos.

A análise dos dados foi realizada utilizando os algoritmos seleccionados em dois testes gerais: *cross-validation* (em 10 *folds*) e *percentage split* (esta sub-dividida em três configurações possíveis: *split* 50% treino / 50% teste, *split* 70% treino / 30% teste, *split* 80% treino / 20% teste).

Os dados a analisar incluiram: 1) percentagem de acerto ou número de instâncias correctamente classificadas; 2) tempo para a construção do modelo ou tempo de aprendizagem; 3) erro médio; 4) taxa de verdadeiros positivos (TVp); 5) taxa de falsos positivos (TFp); 6) sensibilidade; 7) especificidade; e 8) área de *receiver operating characteristics* (ROC).

Por forma a tentar determinar o melhor algoritmo e visto tratar-se de um teste diagnóstico, foram obtidas as áreas ROC, para cada teste. A área da curva ROC permite estabelecer uma relação entre a sensibilidade de um teste diagnóstico e a especificidade como limiar para indicação da variação positiva de um teste. É habitualmente usado para escolha de diferentes testes diagnósticos, apesar de não ter em conta a prevalência da patologia testada.

A análise estatística foi baseada nos resultados do *dataset* produzidos pelo WEKA e acima enumerados, sendo calculadas manualmente a sensibilidade e especificidade. A sensibilidade foi obtida sabendo a TVp e aplicando a fórmula abaixo, convertida posteriormente para percentagem:

$$TVp = Vp / P = Vp / (Vp+Fn) = Se$$

e em que Vp representa o número de verdadeiros positivos, Fn o número de falsos negativos e P o número total de positivos. Sabendo que a TFp é representada por Fp / (Fp+Vn) e sabendo que:

$$Sp = Vn / N = Vn / (Fp+Vn),$$

então temos que a

$$Sp = 1 - (TFp),$$

sendo esta a fórmula utilizada para o cálculo da especificidade.

4. RESULTADOS

4.1. PRINCIPAIS RESULTADOS

Os resultados obtidos, em função da percentagem de acerto, para o algoritmo J48 foram de 65,49%, 72,06% e 70,00% para uma *percentage split* de 50%/50%, 70%/30% e 80%/20% respectivamente. Já para o algoritmo OneR, utilizando a mesma característica e iguais *percentage split* os resultados foram 58,41%, 58,09% e 55,56%. Por fim, para o algoritmo Naïve Bayes, os resultados obtidos nas mesmas condições foram, respectivamente, de 64,16%, 69,85%, 74,44%.

Em relação ao tempo de aprendizagem, o algoritmo J48 apresenta valores de 1,84s, 1,50s e 1,75s para uma *percentage split* de 50%/50%, 70%/30% e 80%/20%, respectivamente. Nas mesma condições, os tempos de aprendizagem do algoritmo OneR foram, respectivamente, de 0,10s, 0,16s e 0,15s. Finalmente, e em relação ao algoritmo Naïve Bayes, os resultados em função do tempo de aprendizagem, foram de 0,15s, 0,14s e 0,14s, respectivamente para uma *percentage split* de 50%/50%, 70%/30% e 80%/20% de treino/teste. Os restantes resultados apresentam-se esquematizados na Tabela I.

Tabela I - Quadro resumo dos resultados obtidos no estudo experimental, utilizando *percentage split*.

		J48			OneR			Naïve Bayes		
	Split	*Acerto (%)*	*Tempo (s)*	*Erro (médio)*	*Acerto (%)*	*Tempo (s)*	*Erro (médio)*	*Acerto (%)*	*Tempo (s)*	*Erro (médio)*
Treino/Teste	*50%/50%*	65,49	1,84	0,0474	58,41	0,10	0,0520	64,16	0,15	0,0454
	70%/30%	72,06	1,50	0,0401	58,09	0,16	0,0524	69,85	0,14	0,0377
	80%/20%	70,00	1,75	0,0433	55,56	0,15	0,0556	74,44	0,14	0,0324

No que diz respeito ao teste de *cross-validation*, a percentagem de acerto para o J48, OneR e Naïve Bayes foram, respectivamente, de 63,27%, 57,08% e 61,50%.

Já no que diz respeito ao tempo de aprendizagem, utilizando ainda cross-validation, os resultados foram de 1,56s, 0,12s e 0,10s, respectivamente. Os restantes resultados encontram-se esquematizados na Tabela II.

Tabela II - Quadro resumo dos resultados obtidos no estudo experimental, utilizando *cross-validation*.

Cross Validation	J48			OneR			Naïve Bayes		
	Acerto (%)	*Tempo (s)*	*Erro (médio)*	*Acerto (%)*	*Tempo (s)*	*Erro (médio)*	*Acerto (%)*	*Tempo (s)*	*Erro (médio)*
Folds 10	63,27	1,56	0,0500	57,08	0,12	0,0537	61,50	0,10	0,0477

Os resultados da análise da sensibilidade e especificidade e restantes outputs do WEKA apresentam-se esquematizados, em função do *percentage split*, na Tabela III.

Tabela III - Quadro resumo dos resultados obtidos no estudo experimental, utilizando *percentage split,* em função dos *outputs* do WEKA e dados obtidos manualmente.

	Split 50%/50%		
	OneR	**J48**	**Naïve Bayes**
Vp	0,584	0,655	0,642
Fp	0,407	0,186	0,135
Se	58,40%	65,50%	64,20%
Sp	59,30%	81,40%	86,50%
Área ROC	0,588	0,728	0,811

	Split 70%/30%		
	OneR	**J48**	**Naïve Bayes**
Vp	0,581	0,721	0,699
Fp	0,467	0,187	0,117
Se	58,10%	72,10%	69,90%
Sp	53,30%	81,30%	88,30%
Área ROC	0,557	0,772	0,847

	Split 80%/20%		
	OneR	**J48**	**Naïve Bayes**
Vp	0,556	0,7	0,744
Fp	0,459	0,194	0,097
Se	55,60%	70,00%	74,40%
Sp	54,10%	80,60%	90,30%
Área ROC	0,548	0,795	0,848

Legenda: Vp-verdadeiros positivos; Fp-falsos positivos; Se-sensibilidade; Sp-especificidade; Área ROC-Área de *receiver operating characteristics; WEKA*-Waikato Environement for Knownledge Analysis .

Os resultados da análise da sensibilidade e especificidade e restantes outputs do WEKA apresentam-se esquematizados, em função do *cross-validation*, na Tabela IV.

Tabela IV - Quadro resumo dos resultados obtidos no estudo experimental, utilizando *percentage split,* em função dos *outputs* do WEKA e dados obtidos manualmente.

	Cross-validation		
	OneR	**J48**	**Naïve Bayes**
Vp	0,571	0,633	0,615
Fp	0,444	0,176	0,164
Se	57,10%	63,30%	61,50%
Sp	55,60%	82,40%	83,60%
Área ROC	0,563	0,714	0,803

Legenda: Vp-verdadeiros positivos; Fp-falsos positivos; Se-sensibilidade; Sp-especificidade; Área ROC-Área de *receiver operating characteristics; WEKA*-Waikato Environement for Knownledge Analysis .

4.2. ANÁLISE DE RESULTADOS

Quando comparados em relação à percentagem de acerto, utilizando *percentage split*, os valores dos algoritmos J48 e Naïve Bayes apresentam de uma forma geral resultados melhores que o OneR e muito semelhantes entre si. No entanto, o algoritmo Naïve Bayes apresenta uma relação directa com a percentagem de treino, o mesmo não acontecendo com o J48 e OneR que aquando da passagem de 70% para 80% de treino diminuem a sua percentagem de acerto (Figura 1).

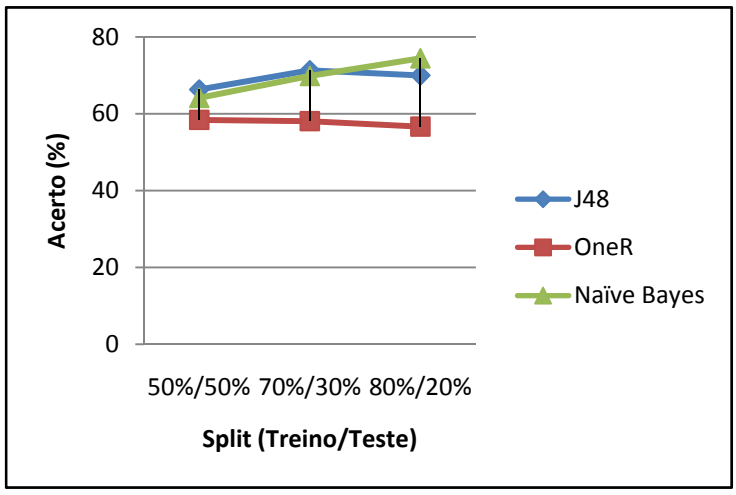

Figura 1 - Comparação da percentagem de acuidade entre os algoritmos J48, One R e Naïve Bayes em função da percentagem de treino/teste.

Em relação ao tempo de aprendizagem, igualmente com *percentage split*, as médias dos valores dos algoritmos OneR e Naïve Bayes apresentam valores inferiores ao J48 e uma vez mais com perda de relação linear entre os 70% e 80% de treino (Figura 2).

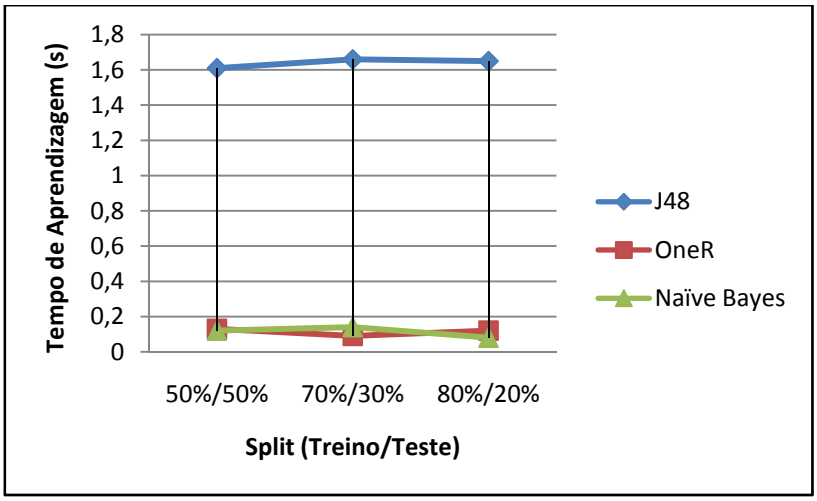

Figura 2 - Comparação do tempo de aprendizagem entre os algoritmos J48, One R e Naïve Bayes em função da percentagem de treino/teste.

Por fim, e relativamente ao erro médio utilizando *percentage split*, as médias dos valores do algoritmo OneR apresentam valores superiores aos restantes, sendo que o algoritmo Naïve Bayes, ao invés dos outros, diminui o erro médio à medida que aumenta a percentagem de treino (Figura 3).

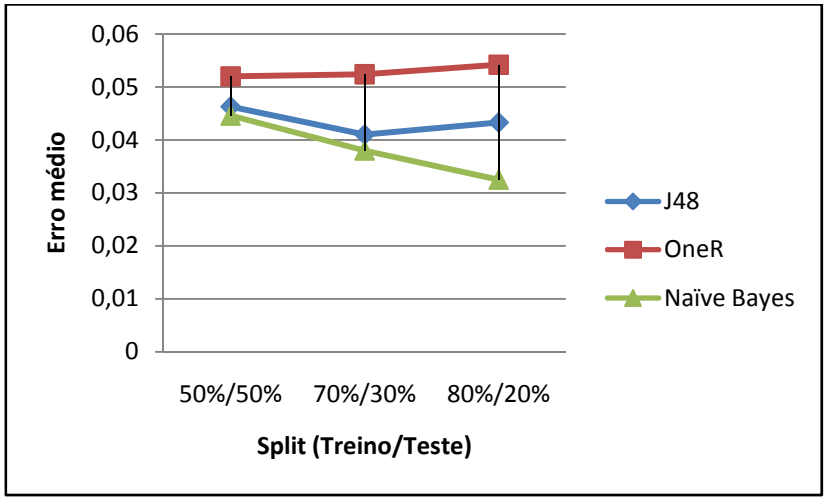

Figura 3 - Comparação do erro médio entre os algoritmos J48, One R e Naïve Bayes em função da percentagem de treino/teste.

Em relação à percentagem de acerto, utilizando *cross validation*, estas apresentam-se mais elevadas no algoritmo J48 e mais baixas no algoritmo OneR, apresentando o Naïve Bayes valores intermédios.

Já para o tempo de aprendizagem, utilizando *cross validation*, as médias dos valores apresentam-se mais baixas no algoritmo J48, sendo mais elevadas e semelhantes nos restantes dois algoritmos.

Por fim, e utilizando *cross validation* para o erro médio, as médias dos valores apresentam o seu valor mais baixo para o algoritmo Naïve Bayes, sendo os restantes mais elevados e semelhantes entre si.

A análise sumária dos dados obtidos relativos à especificidade e área ROC, são consistentemente superiores utilizando o algoritmo Naïve Bayes, utilizando *cross-validation* e *percentage split*, obtendo também o valor mais alto na sensibilidade usando *percentage split* 80%/20% (Figs. 4 e 5).

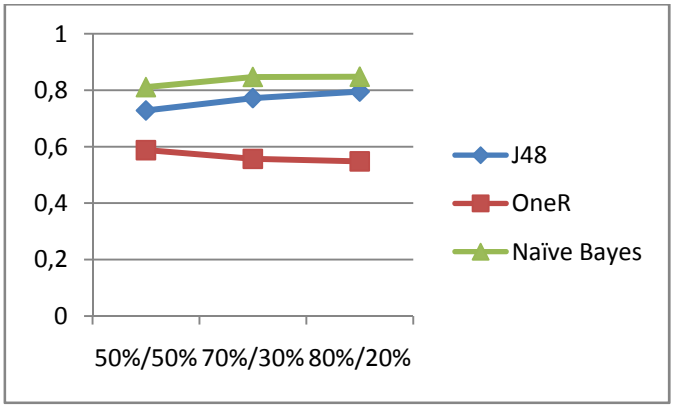

Figura 4 - Comparação da área ROC nos algoritmos J48, One R e Naïve Bayes, em função da *percentage split*.

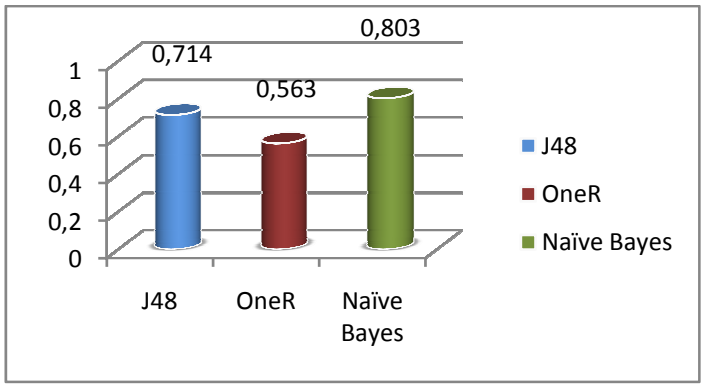

Figura 5 - Comparação da área ROC nos algoritmos J48, One R e Naïve Bayes, em função da *cross-validation*.

No que diz respeito à sensibilidade, esta é sempre superior no algoritmo J48, utlizando *cross-validation* e *percentage split*, excepto na instância referida anteriormente.

5. DISCUSSÃO E CONCLUSÃO

Na elaboração deste trabalho testamos três algoritmos de *machine learning* para classificar uma arritmia. Realizamos a análise da percentagem de acerto, tempo de aprendizagem, erro médio, sensibilidade, especificidade e área ROC.

Pela análise dos resultados acima apresentados, podemos concluir que os algoritmos OneR e Naïve Bayes apresentam características que permitem uma rápida aprendizagem, ao invés do J48. No entanto, a percentagem de acerto dos algoritmos J48 e Naïve Bayes são substancialmente superiores ao OneR.
Mais ainda, quando comparados os três algoritmos, verificamos que todas as variáveis dependem fortemente da percentagem de treino/teste, quando utilizada *percentage split*.

A interpretação dos dados sugere que o algoritmo de **Naïve Bayes**, certamente pelas suas características probabilísticas, apresenta o melhor desempenho de entre os algoritmos escolhidos, quer pelos reduzidos erro médio e tempo de aprendizagem, quer pela elevada (a mais elevada) relação diagnóstica.

Pensamos, suportados pelos resultados obtidos relativos à área ROC, que os algoritmos de *machine learning* podem auxiliar de forma importante o diagnóstico de arritmias cardíacas.

Obviamente existem limitações a este estudo, sendo mesmo provável a introdução de algum viés, na medida em que não foi possível obter os dados da área ROC em todas as classes por inexistência de dados na mesma sendo, portanto, substituídas por valores probabilísticos. Mais ainda, não foi possível determinar se as diferenças encontradas entre algoritmos são estatisticamente significativas.

Em investigações futuras pensamos que deverá ser alterada a denominação do *dataset* ou a própria classificação deste, uma vez que nem todas as classes obtidas dizem respeito, em termos formais, a arritmias propriamente ditas. Seria também interessante criar um *dataset* nacional, com denominações ajustadas, e/ou repetir o estudo com diferentes algoritmos de *machine learning*.

6. BIBLIOGRAFIA CONSULTADA

1. Bortolan G, Pedrycz W. An Interactive Framework for an Analysis of ECG Signals. Artificial Intelligence in Medicine 2002; 24:109-32.
2. Everitt B. Medical Statistics from A to Z. 2nd ed. Cambridge: Cambridge Press; 2006.
3. Haridas, M. Step by Step Tutorial for Weka. Disponível em: www.people.cis.ksu.edu/~hankley/d764/tut07/Haridas_DM.pdf
4. Lipman B, Cascio T, editors. ECG - avaliação e interpretação. Loures: Lusociência; 2001. p. 1-22.
5. Soman T, Bobbie PO. Classification of Arrhythmias Using Machine Learning Techniques. Proceedings of the 4th International Conference on System Science and Engineering 2005; Rio de Janeiro (Brasil), Abril 25-27.
6. Wagner GS, editor. Marriott's Practical Electrocardiography. 10th ed. Philadelphia: Lippincott - Raven; 2006. p. 1-69.
7. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd edition. San Francisco: Morgan Kaufmann; 2005.