Reamostragem e Preparação dos dados de Eletrocardiogramas

Gabriel Lechenco V. Pereira

[1]

1. Introdução

Neste relatório será descrito as técnicas e recursos utilizados para realizar uma normalização e breve análise de sinais de Eletrocardiograma (ECG), os quais serão utilizados, posteriormente, para treinamento e convalidação de classificadores de SVM. As duas ações se veem extremamente necessárias para que se possa realizar uma classificação coerente e que não exija um incessante processamento.

A primeira pois, apesar de todos os dados se tratarem de sinais de ECG, os mesmos se originam de banco de dados distintos, o que causa uma discrepância na frequência de amostragem e na forma como os dados foram salvos. Portanto, uma normalização dos dados se mostra indispensável.

Em seguida, para se diminuir o número de dimensões utilizadas como entrada para o classificador, uma análise e processamento dos sinais deve ser efetuada para que se possa, além de reduzir o tempo de execução do algoritmo, rearranjar os dados de uma forma mais interessante para se observar padrões e comportamentos.

2. Dados e Normalização

Os sinais utilizados tem como origem dois bancos de dados oferecidos pelo MIT, para os sinais cardíacos que apresentam características patológicas foi utilizado o *MIT-BIH Arrhythmia Database* (mitdb), que apresenta 48 sinais distintos com uma frequência de amostragem de 360 Hz; paralelamente, para representar os sinais saudáveis foram utilizados os dados do *MIT-BIH Normal Sinus Rhythm Database* (nrsdb), que apresenta 18 sinais de longa duração de pessoas sem quaisquer patologia cardíaca a uma taxa de 128 Hz.

Como cada sinal tem um arquivo separado com informações importantes para o desenvolvimento do projeto proposto, como a localização de cada batimento e se este apresenta ou não alguma anomalia, e pensando em uma manipulação mais prática destas informações o primeiro passo foi extrair estes dados e agrupá-los em um único arquivo. Permitindo que o nome do arquivo do sinal, o número de anotações, suas localidades e características pudessem ser utilizadas de forma mais efetiva.

A grande discrepância entre os sinais dos dois banco de dados descritos a cima é oferecida por suas frequências de amostragem, enquanto o mitdb apresenta uma $F_s = 360Hz$, o nrsdb possui uma frequência de apenas $128Hz$. Como os dados serão utilizados de forma conjunta, tomou-se como padrão uma $F_s = 128Hz$ e os sinais patológicos tiveram que ser re-amostrados para a frequência equivalente.

Este processo, porém, foi realizado em conjunto com o janelamento dos sinais, onde os eletrocardiogramas foram recortados em pedaços com 8 segundos de comprimento (ou 1024 amostras para $F_s = 128Hz$), outra forma de padronizar os dados que

depois serão inseridos no classificador. Para evitar sinais que saudáveis fossem classificados como patológicos neste processamento, foram filtrados por meio das anotações previamente organizadas os batimentos os quais se diferenciassem de batimentos normais ou que apresentaram algum problema durante a leitura (ex.: paciente se mexeu). Com isso, após o recorte e a re-amostragem,cada uma dessas janelas apresenta como batimento central o da patologia referente, e 4 segundos (ou 512 amostras) antes e depois deste ponto.

Como os sinais não patológicos são bem mais extensos e sem muitas anomalias, estes foram divididos levando em conta o número de recortes adquiridos previamente pelos sinais doentes, buscando um equilíbrio entre o número de exemplos de cada classe. Além disso, para se ter uma maior representatividade dos sinais saudáveis, foram escolhidos pontos aleatórios dos diferentes ECG's buscando-se obter um maior número de características.

Assim, ao todo foram processados e salvos 27.686 arquivos com 1024 amostras de um determinado ponto de seu sinal de origem, junto com o sinal também foi escrito em disco a anotação considerada ao se selecionar a janela, identificando a qual classe cada arquivo se refere.

3. Processamento e classificação

Antes de começarmos a modelar nosso classificador e ajustar seus parâmetros, temos que diminuir drasticamente o número de dimensões de nossos sinais. Apesar de termos padronizado e dividido os sinais em pedaços de 1024 amostras, treinar uma SVM com vetores com esse número de dimensões seria muito custoso, além de que muitos dos pontos seriam irrelevantes para a classificação. Para contornar este problema os sinais foram decompostos utilizando uma família de funções wavelet conhecida como 'db3', comumente utilizado em sinais de ECG. Estas funções permitem que cada um dos nossos sinais com 1024 amostras se tornem 16 índices de energia (decomposição em 4 níveis), onde cada índice representa o sinal de modos diferentes. Com isso, é possível reduzir o número de características que o nosso classificador deverá analisar para apenas 16. Todos os valores foram salvos em uma tabela com 17 colunas, sendo que as células da 17ª coluna tem o valor de ± 1, variando conforme a classe ao qual cada linha se encaixa (patológico ou saudável).

A correlação entre as duas classes para cada um dos índices de energia pode ser observado pela Tabela1, nota-se que muitos desses índices apresenta uma grande correlação o que indica que essas características podem ser relevantes para a classificação justificando novamente o uso das mesmas.

3.1. SVM

Uma SVM (*Support Vector Machine*) é um classificador que tem como objetivo encontrar, por meio de vetores de suporte, um hiperplano que divide o espaço entre as duas classes. Dessa forma, o treinamento deste modelo é definido pela equação 1.

$$f(x) =< \hat{w}, x > +w_0 \tag{1}$$

Onde $\hat{w}$ é o hiperplano que divide as duas classes e x o vetor com os dados os quais se quer classificar. A grande vantagem que a SVM encontrou nos últimos anos foi a utilização de kernels para realizar classificações não lineares.

1	-0.867312
2	0.102531
3	0.856557
4	0.681275
5	0.534810
6	0.739886
7	0.866136
8	0.837706
9	0.232222
10	0.327236
11	0.675728
12	0.401012
13	0.521165
14	0.631206
15	0.742696
16	0.716740

Table 1. Correlação de cada índice de energia wavelet entre as duas classes

	Dados testados	Class. Correta	Tx. de Acerto	Pontos entre H_- e H_+
Linear	9257	8693	93.907 %	16.615 %
Gaussiano	9257	8771	94.749 %	22.059 %
Polinomial	9257	8550	92.362 %	24.478 %

Table 2. Resultados SVM

Para o treinamento do modelo foram separados $2/3$ dos dados, se precavendo para que houvessem uma proporção equivalente entre as duas classes. O restante dos dados foram utilizados para validação do modelo e análise de sua confiabilidade.

4. Resultados Preliminares

Foram realizados alguns treinamentos utilizando de diferentes kernels para se observar a eficiência da classificação para cada um deles. Os kernels utilizados foram o linear, gaussiano e o polinomial. Após o treinamento e a predição de todos os sinais, obteve-se o número de classificações corretas, uma Taxa de acertos, além de quantos pontos pontos acabaram entre os hiperplanos de suporte H_- e H_+, que se referem à equação 1 quando ela é igualada a -1 e 1, respectivamente.

O resultado adquirido com os três kernels pode ser resumido pela tabela2, aqui percebemos a facilidade com que a SVM consegue separar as duas classes com mais de 90% de taxa de acerto, porém, ao analisar os pontos que se encontram entre os dois hiperplanos de suporte , é possível observar como os kernels mudam ligeiramente em sua confiabilidade.

Apesar disto, por terem sido utilizados kernels em seu formato padrão, espera-se que estes resultados possam ser melhorados. Mesmo que as taxas de acerto pareçam altas, por se tratar de um problema da área da saúde, características como precisão e confiabilidade são de suma importância quando se tem vidas em risco.

5. Conclusão

Com este relatório pode-se resumir as técnicas utilizadas até agora para se tentar utilizar SVM's para a classificação de sinais de eletrocardiograma que apresentam anomalias daqueles que não. O trabalho ainda está no início mas os treinamentos apontam que estamos na direção certa.

Os próximos passos para se tentar aumentar a performance do algoritmo seria tentar reduzir ainda mais o numero de dimensões ao omitir alguns dos índices de energia para se ganhar processamento. Além disso, a utilização de técnicas como *cross validation* e outras podem aumentar a taxa de precisão do algoritmo. Após a classificação entre essas duas classes, uma possível próxima abordagem seria tentar classificar entre as anomalias em si, tendo um classificador multi-classe que daria um diagnóstico mais completo ao paciente.