Análise de Dados Relacionados a Exercícios Físicos Utilizando Conceitos de Aprendizado de Máquina

Antônio G. O. Junior[1] , Victor O. Ogata[1]

[1]Instituto de Ciência e Tecnologia – Universidade Federal de São Paulo (UNIFESP) – São José dos Campos – SP – Brasil

{victoroliveiraogata,antonio.gomes.o.jr}@gmail.com

Abstract. *This paper describes the development and data analysis related to physical exercises on people who have more than 50 years with the objective of getting conclusions about the tables through algorithms seen on Machine Learning such as knn, Decision Trees and Support Vector Machine*

Resumo. *Este artigo apresenta o desenvolvimento e a análise de dados relacionados a exercícios físicos em pessoas que já passaram da meia idade com o objetivo de obter conclusões a respeito das tabelas através de algoritmos vistos em Aprendizado de Máquina , tais como: knn, Árvore de decisão e Máquina de Vetores Suporte*

1. Introdução

Aprendizado de Máquina surgiu do reconhecimento de padrões e da ideia de que máquinas podem aprender sem precisarem serem indicadas a realizar tarefas específicas. Sendo uma vertente da Inteligência Artificial , em que um sistema aprende a partir de dados e padrões a tomar decisões sem que humanos interfiram.

O objetivo do projeto é tentar prever se os exercícios físicos isométricos (intenso, de baixa frequência e baixo volume) e estímulos vibracionais, usados pelo experimento desenvolvido na monografia de referência [Oliveira, 2018], teriam um efeito positivo em uma dada paciente. Trazendo melhora no tratamento e prevenção de doenças ósseas, conhecidas por afetar pessoas a partir de uma idade avançada, como é o caso da osteoporose, que é uma doença esquelética sistêmica que tem como características a baixa densidade óssea e deterioração da microarquitetura do tecido ósseo, que leva ao aumento do risco de fraturas por fragilidade[NIH 2001].

Um dos testes mais comuns para verificar se a pessoa possui alguma doença óssea é a medida da densidade mineral óssea (DMO), sendo que resultados abaixo da média indicam problemas ósseos. Um estudo feito no Reino Unido afirma que, em mulheres, a densidade mineral óssea diminui com a idade e, durante a menopausa, tem quedas ainda maiores sendo estimado que a cada duas mulheres acima de 50 anos no país, uma irá sofrer uma fratura para o resto de sua vida [Van Staa et al. 2001].

2. Métodos utilizados de Aprendizado de Máquina

As técnicas de aprendizado de máquina são usadas para ensinar ao computador como desenvolver certa tarefa da melhor forma a partir das suas experiências passadas. No nosso projeto, tentaremos prever se uma paciente teria alguma melhora do DMO do Fêmur e da Lombar, e para isso usaremos técnicas de aprendizado preditivas, também conhecidas como supervisionadas. O algoritmo usa para treinamento o conjunto de dados sobre as características de cada pacientes e seus resultados. Esses algoritmos supervisionados usados tentam prever os resultados, classificando os pacientes se houve ou não melhora dos valores de DMO, através das características de cada paciente.

2.1. Support Vector Machines (SVM) Em português maquinas vetores de suporte, é um algoritmo de aprendizado de máquina (AM) supervisionado que pode ser utilizado tanto para classificação, quanto para regressão (Soares, 2008). A técnica constrói um hiperplano para a classificação dos dados, separando os dados de classes diferentes, com valor de separação máxima entre os exemplos. Uma SVM Linear de classificação tenta gerar uma linha em um gráfico capaz de separar os dados em dois grupos discretos, porém uma SVM linear não seria o suficiente no nosso caso para interpretar todas as características dos pacientes. Para ser feita a comparação entre os métodos foi utilizado um método de validação cruzada *leave one out*. Para poder lidar com casos mais complexos, pode-se aumentar a dimensionalidade do hiperplano gerado pelo SVM, de forma que os dados de treinamento possam ser separados pelo hiperplano. Para isso é necessário uma função kernel (Soares , 2008).

2.2. K- Nearest Neighbor (KNN) Em português k - vizinhos mais próximos, é um algoritmo de AM supervisionado baseado em distância. Ele parte da hipótese que dados similares tendem a estar concentrados em uma mesma região do espaço de entradas, rotulando os elementos de acordo com a classificação dos seus k vizinhos mais próximos. Para validação dos resultados, pode ser usado um método de validação cruzada *leave one out* em que o modelo é calibrado uma vez para cada paciente, de forma que cada vez um único paciente seja usado para o grupo de teste, e todos os outros sejam usados no grupo de treinamento.

2.3. Decision Tree Em português, árvore de decisão, é um algoritmo de AM supervisionado, que usa os dados de treinamento para gerar um fluxograma em formato de árvore onde cada nó de divisão possui dois ou mais sucessores e contém um teste condicional baseado nos valores de atributos (por exemplo, paciente sedentário Sim/Não). Os nós folhas contém um rótulo ou previsão calculado através de uma função que considera valores da variável alvo dos exemplos que chegam na folha. Uma das vantagens desse método, é que diferente de alguns algoritmos de AM, os resultados de uma árvore de decisão são simples de ser vistos e interpretados por uma pessoa leiga. Para ser possível ser feita uma comparação entre os métodos foi utilizado validação cruzada *leave one out.* Em árvores de decisão, deve ser tomado um cuidado na escolha

de quais atributos serão usados para a divisão, para esta tarefa costumam ser utilizadas duas métricas:

2.3.1 Entropia É usada como medida de impureza para medir a aleatoriedade do atributo alvo. Dado um conjunto S, com cada elemento pertencente a uma classe i, com probabilidade p_i dele pertencer aquela classe, temos:

$$\text{Entropia(S)} = \sum_{i=1}^{c} - p_i \, Log_2(p_i)$$

2.3.2 Ganho O ganho em árvores de decisão, mede a redução da entropia ao se escolher um atributo "A" para ser usado como classificação. A cada nó de decisão, o atributo com maior ganho é escolhido para divisão.

$$\text{Ganho(S, A)} = \text{Entropia (S)} - \sum_{v \varepsilon Valores(A)} \frac{|Sv|}{|S|} Entropia(Sv)$$

3. Experimento

Os dados usados por esse trabalho, foram adquiridos através de uma pesquisa feita pela Universidade Federal de São Paulo. Se trata dos dados de 85 participantes idosas pós-menopausa, que tiveram os dados coletados antes e depois de um período de intervenção de 24 semanas. Foram coletados dados da densitometria óssea, a microtomografia computadorizada de alta resolução, a mensuração da força e os testes funcionais. Segundo (Facelli at al.,2011), o desempenho dos algoritmos de aprendizado de máquina costuma ser afetado pelo qualidade dos dados disponíveis. No pré-processamento foi feito um tratamento e seleção dos dados de referência. Os pacientes com dados faltantes foram retirados, foi se selecionados quais dados seriam usados pelo algoritmo, e

3.1. Pré-Processamento

Para uma análise de dados mais concreta foi retirado da tabela os nomes das mulheres e foram feitas duas tabelas, que são usadas pelos algoritmos de aprendizado de maquina. Na primeira tabela, os atributos foram qualificados de qualitativos para quantitativos como pode ser visto na Tabela 1, e depois os dados quantitativos foram normalizados; Na segunda, o contrário foi feito, os dados quantitativos foram transformados em qualitativos como pode ser visto na Tabela 3.

Tabela 1. Exemplo tabela com dados dos pacientes normalizado, sem os resultados

T..menopausa	Altura	Peso	MOB.LP	Fumante..F	Calcio..F	Vitamina.D...F	Queda.5.anos...F	Caminhada..alongamento.ou.Danca...F
0.48717949	0.000321...	0.0000...	0.162920665	0	0	0	0	1
0.17948718	0.000385...	0.0416...	0.170612981	0	0	0	0	1
0.87179487	0.000000...	0.0679...	1.000000000	0	0	0	0	1
0.23076923	0.000771...	0.0833...	0.169897553	0	0	0	0	0
0.48717949	0.000707...	0.1184...	0.064629376	0	0	0	1	0
0.23076923	0.000707...	0.1962...	0.196099081	1	0	0	0	0

Tabela 2. Grupos de controle e resultados normalizados entre 1 e 0

Controle	Isometria	IsoVib	Vibracao	DeltaFemur	DeltaLombar
1	0	0	0	1	1
1	0	0	0	1	0
0	1	0	0	1	0
0	1	0	0	0	1

Tabela 3. Exemplo tabela com dados qualitativos, com os resultados

Grupo	Idade	T..menopausa	Altura	Peso	MOB.LP	Fumante..F	Calcio..F	Vitamina.D...F	Queda.5.anos...F	Caminhada..alongamento.ou.Danca...F	DeltaFemur
Controle	70+	9-19	1.55-1.65	60-70	3-6	0	0	0	0	0	1
Controle	50-60	0-9	1.55-1.65	70+	3-6	0	0	0	0	0	0
Isometria	70+	19+	<1.55	60-70	3-6	0	1	1	0	1	0
Isometria	60-70	19+	<1.55	<60	<3	0	0	0	0	1	1
Vibracao	60-70	0-9	<1.55	60-70	<3	0	0	1	0	0	1
Controle	70+	19+	<1.55	70+	<3	0	0	0	0	1	0
Isometria	50-60	9-19	1.55-1.65	70+	3-6	0	1	1	1	1	1
Iso+Vib	50-60	0-9	1.65+	60-70	6+	1	0	0	0	0	0

3.2. Ajuste de Parâmetros para os Algoritmos de Aprendizado de Máquina

Visando achar o melhor modelo para resolver o problema, de previsão se haverá melhora de pacientes depois dos exercícios, testamos diferentes configurações de algoritmos

SVM - Para a calibração usando o SVM, do pacote "e1071" do R, usando C =10 e gamma = 0.5, foi utilizada a mesma tabela das outras técnicas em que os dados das pacientes são normalizados em 0 e 1.

Decision Tree - Para a calibração usando as árvores de decisão, através do pacote "rpart" do R, foram-se usados uma tabela com os dados das pacientes convertidos para valores qualitativos. O método escolhido para ser usado na função foi o método "class".

Knn - Para a calibração usando o KNN com *cross validation leave one out*, do pacote "class" do R, foram-se usados os dados das pacientes normalizados entre 0 e 1. O valor de k usado pelo algoritmo foi escolhido como 1 pois obtém os melhores resultados entre os algoritmos testados

4. Análise de Resultados

4.1. Previsão de melhora do Delta Lombar e Delta Femur usando KNN

```
library(class)
train <- dataCSV[1:11]
cl <- factor(dataCSV[,12])
out <- knn.cv(dataCSV[,-12],dataCSV[,12],k = 1)
table(out)
train2 <- dataCSV[1:11]
cl2 <- factor(dataCSV[,13])
out2 <- knn.cv(dataCSV[,-12],dataCSV[,13],k = 1)
table(out2)
```

Figura 1.Código de calibração knn com cross-validation leave one out

O objetivo é prever se uma determinada paciente, teria uma melhora no Delta Lombar e Delta Fêmur depois de fazer os exercícios propostos no experimento (OLIVEIRA, 2018). Como se tem poucos dados na tabela, foi se escolhido o método *leave one out* para validação, pois se os dados dos pacientes foram separados em dois conjuntos diferentes para treinamento e teste, não haveria dados o suficiente para um treinamento eficiente.

```
>
> out <- knn.cv(dataCSV[,-13],dataCSV[,14],k = 1)
> table(out)
out
 0  1
15 22
>
> out2 <- knn.cv(dataCSV[,-13],dataCSV[,15],k = 1)
> table(out2)
out2
 0  1
13 24
```

Figura 2. Matriz de confusão do resultado da calibração usando knn

O resultado final, mostra-se longe do esperado pois não foi possível tirar conclusões relevantes. Ainda que a calibração acerte em aproximadamente 60% dos casos, ela não acerta um único caso de verdadeiro positivo.

4.2. Previsão de melhora do Delta Lombar e Fêmur usando Árvores de Decisão

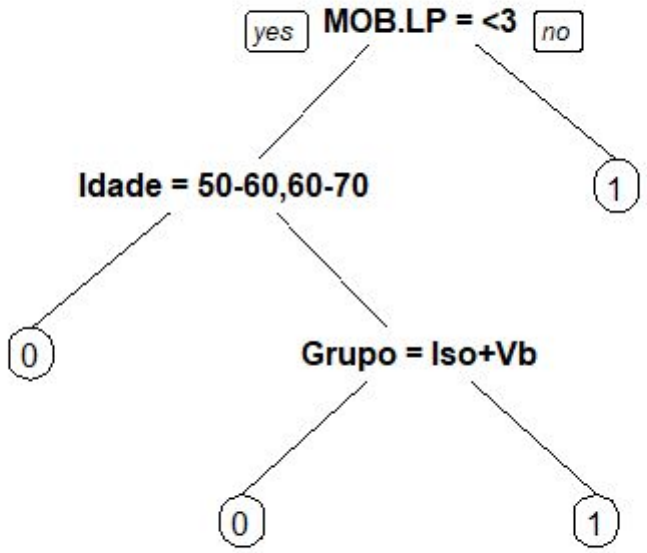

Figura 3. Árvore de Decisão Calibrada para Delta Lombar

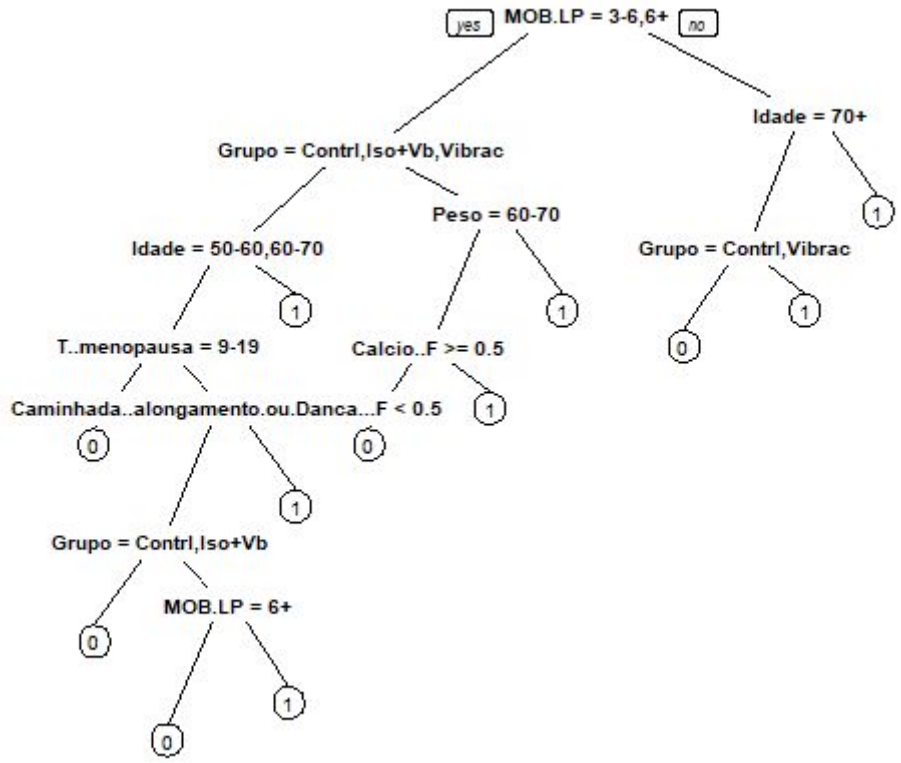

Figura 4. Árvore de Decisão Calibrada para Delta Lombar

Quando se foi feito um teste com *cross validation leave one out*, as árvores de decisão para prever o DMO Fêmur tiveram 59.459% de erro. E as árvores de decisão para prever o DMO Lombar tiveram 67.568% de erro. Mostrando que sua precisão de resultado é pior que o KNN;

4.3. Previsão de melhora do Delta Lombar e Fêmur usando Support Vector Machine

Com o teste de *cross validation leave one out, a SVM para prever o DMO do Fêmur tiveram* 23.44569 % de erro , enquanto a DMO da Lombar teve 24.81645% de erro. Com essas informações apesar de verificarmos que o erro ainda é grande , houve uma grande melhora em relação aos outros métodos utilizados.

5. Conclusão

Esse artigo foi criado como suporte para a monografia de referência, visando calibrar algoritmos de aprendizado de máquina, para tentar prever os efeitos dos exercícios

propostos na monografia. Para isso foram escolhidos a análise nas pacientes do DMO mensurados por DEXA, que é considerada padrão ouro pela academia e pela clínica médica especializada [Oliveira, 2018]. Os resultados de todos os algoritmos, depois de analisados com um validação cruzada *leave one out*, tiveram uma taxa de acerto insatisfatória para fins médicos, sendo que o SVM foi o que teve o melhor resultado com aproximadamente 75% de acerto. Entretanto existe uma dificuldade em analisar os resultados finais pois a quantia de dados de treinamento era muito pequena, e a própria conclusão da monografia de referência afirma que depois de feita uma análise probabilística ANOVA, os exercícios não tiveram efeitos significativos nos valores de DMO.

6. Referências

Facelli, K.; Lorena, A.C.; Gama, J.; Carvalho, A.C.P.L.F. Inteligência Artificial: uma abordagem de aprendizado de máquina. LTC, 2011.

Oliveira, T.S. Avaliação da indução osteogênica em mulheres pós menopausadas por um novo programa de exercícios isométricos. UNIFESP, 2018

NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy (2001). JAMA. 2001;285:785–95

Van Staa TP, Dennison EM, Leufkens HG, Cooper C. Epidemiology of fractures in England and Wales. Bone. 2001;29:517–22.

Soares, R. G. Uso de meta-aprendizado para a seleção e ordenação de algoritmos de agrupamento aplicados a dados de expressão gênica. Master's thesis, Centro de Informática- Universidade Federal de Pernambuco, Recife, 2008.