10th Brazilian Congress on Computational Intelligence (CBIC'2011), November 8 to 11, 2011, Fortaleza, Ceará Brazil

SELEÇÃO DE FATORES DE RISCO DE LESÕES EM ATLETAS, ATRAVÉS DA EXTRAÇÃO DE CONHECIMENTO DOS PESOS DE UMA REDE NEURAL

Havana Diogo Alves, Mêuser Jorge Silva Valença

Escola Politécnica - Universidade de Pernambuco

havana.alves@gmail.com, meuserv@yahoo.com.br

Abstract – The increased incidence of injuries in Brazilian athletes is of concern to then and their coaches and technical staff. The injuries are a result of various factors, and know the influence of each in favor of lesions can help prevent and treat these injuries. In health, traditional statistical methods are commonly used in the study of several cases, but it is noteworthy that such methods only detect the linear features of the data. As the real world is full of non-linear phenomena, it becomes necessary to use techniques that realize this nonlinearity. For these cases, the Artificial Neural Networks (ANN) have shown satisfactory results. This article aims to test algorithms for variable selection that extract knowledge of the weights of an ANN, to obtain the contribution of each factor in the occurrence of injuries to athletes. It is hoped that finding a new alternative needs to understand the mechanisms that favor injuries, and therefore help in prevention.

Palavras-chave – Redes Neurais,Seleção de Variáveis, Lesões, Fatores de risco, Atletas.

1 Introdução

O atletismo brasileiro vem crescendo de forma significativa nos últimos anos, conquistando, assim, posições no *ranking* mundial. O aumento na intensidade do treinamento dos atletas foi fator relevante para esse feito. Como conseqüência do aumento da demanda de exercícios cada vez maus modernos e competitivos, houve um crescimento no risco do aparecimento de lesões, o que é motivo de apreensão para atletas e treinadores, pois interrompem o processo evolutivo do treinamento. Muitos são os fatores que podem contribuir para o surgimento de tais lesões, e saber qual a relevância de cada um deles é de grande importância para os profissionais ligados ao esporte, pois sabendo os mecanismos principais dos lesionamento é possível tomar medidas preventivas.

Na área de saúde, métodos estatísticos tradicionais são comumente usados no estudo dos mais diversos casos, porém vale ressaltar que tais métodos detectam apenas as características lineares dos dados. Como o mundo real é repleto de fenômenos não-lineares, torna-se necessária a utilização de técnicas que percebam esta não-linearidade. Para estes casos, as Redes Neurais Artificiais (RNA) têm mostrado resultados bastante satisfatórios. Este artigo tem como objetivo testar algoritmos de seleção de variáveis que extraem conhecimento dos pesos de uma RNA, para obter a contribuição de cada fator no surgimento de lesões em atletas. Espera-se com isso encontrar uma nova alternativa mais precisa para se entender os mecanismos favorecedores de lesões, e conseqüentemente ajudar nas medidas preventivas.

2 Lesões no Atletismo e Fatores que contribuem para o seu surgimento

Nos últimos anos, o número de praticantes de atividades físicas cresceu vertiginosamente. Segundo Kettunen *et al.*[1], o aumento da demanda de exercícios modernos e competitivos provocou o aumento simultâneo no risco de lesões, causando grande preocupação tanto para os atletas, quantos para os profissionais responsáveis pelo condicionamento físico desses, pois interrompem o processo evolutivo de adaptações sistemáticas impostas pelo treinamento. Pastre *et al.* [2] cita que a lesões esportivas são resultantes de uma interação de fatores de risco, sendo eles intrínsecos (idade,sexo,condição física, desenvolvimento motor, alimentação e fatores psicológicos) ou extrínsecos (especificidade técnicas da modalidade, tipo de equipamento usado, organização e cargas de treino e competição, condições climáticas ou a combinação destes).

Dentre as práticas esportivas o atletismo se destaca pela diversidade de modalidades, cada qual caracterizada pela presença de condições específicas de treinamento e presença de elementos básicos, como correr, saltar, lançar ou arremessar, que são observados nos demais esportes com suas respectivas adaptações [3].

A busca em alcançar ou até ultrapassar limites que treinadores e atletas tanto necessitam, por vezes, ultrapassam a fase de adaptação, resultando em exaustão, podendo causar as tão temidas lesões[4].

No caso das lesões musculares as causas mais comuns são: excesso de treinamento; falta de controle nas tensões de exercícios e alongamentos; gestual motor (técnica) indevido nos exercícios e alongamentos; carência de exercícios de alongamento compensatórios após os exercícios físicos; excesso de força e insuficiência de flexibilidade, ou fraqueza com muita flexibilidade.

3 Redes Neurais Artificiais *MultiLayer Perceptron* com *Backpropagation*

Para resolver problemas não-lineares, as RNAs MLP com Backpropagation têm se mostrado bastante eficientes. Estas são constituídas por uma camada de entrada, pelo menos uma camada intermediária e uma camada de saída. Esta configuração pode ser vista na Figura 1. O treinamento de uma Rede MLP com Backpropagation consiste na propagação e na retro-propagação do sinal.

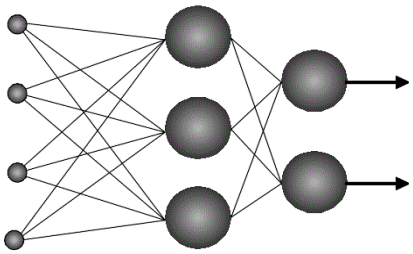


Figura 1 – **Representação de uma Rede Neural artificial MultiLayer Perceptron com uma camada de entrada, uma intermediária e uma de saída.**

Na fase de propagação há o cálculo da entrada líquida dos neurônios da camada intermediária (1) e da entrada líquida dos neurônios de saída (2 e 3). No fim da fase de propagação há o cálculo do erro (4). A entrada líquida é obtida utilizando-se os valores de entrada como parâmetros de uma função de ativação. Uma das funções de ativação mais utilizadas é a sigmoidal logística (5).

Na fase de retro-propagação os peso das conexões são ajustados. O ajuste é feito utilizando-se o valor da sensibilidade dos neurônios. Os cálculos das sensibilidades para os neurônios de saída e para os neurônios escondidos podem ser realizados utilizando-se as equações (6) e (7), respectivamente. Tendo as sensibilidades calculadas, procede-se com o ajuste dos pesos (8).

$$y^1 = f_1\left(net_1^1\right) \tag{1}$$

$$net_i^2 = \sum_{j=0}^{n} w_{ij}\, x_j \tag{2}$$

$$y^2 = f_1\left(net_1^2\right) \tag{3}$$

$$e_i = \left(d_i - y_i\right) \tag{4}$$

$$y = f(net) = \frac{1}{1 + e^{-net}} \tag{5}$$

4 Técnicas de Seleção de Variáveis

Com o propósito de medir a importância de um dado atributo em um conjunto de dados, podemos calcular a sua contribuição relativa e/ou seu perfil de contribuição. Diversos métodos já foram propostos com este fim. Para realizar a quantização da influência de cada variável no resultado do treinamento de uma RNA é necessário que a técnica de seleção tenha a capacidade de extrair conhecimento da rede. Dentre os diversos algoritmos de seleção, dois deles vêm sendo bastante utilizados: (i) o Algoritmo de Garson [5], comumente usado em artigos que tratam de dados ecológicos, como Aurelle *et al.*[5] e Gozlan *et al.*; e a (ii) Análise de Sensibilidades [6].

(i) O Algoritmo de Garson

O Algoritmo de Garson envolve essencialmente o particionamento dos pesos das conexões entre a camada escondida e a de saída de cada neurônio intermediário em componentes associados com cada neurônio de entrada. Tem como característica realizar o cálculo da importância relativa R de cada variável de entrada utilizando-se o valor absoluto dos pesos das conexões, como mostra a equação (9).

$P_{ki} = |w_{ki}|*|w_{lk}|$, onde w_{ki} são os pesos das conexões entre os neurônios de entrada e os neurônios escondidos; w_{lk} são os pesos das conexões entre os neurônios escondidos e os neurônios de saída. (9)

$$Q_{ki} = \frac{P_{ij}}{\sum_{k=1}^{\tilde{N}} \sum_{i=1}^{N} P_{ki}}$$, onde $\tilde{N}$ é a quantidade de neurônios na camada intermediária e N é o número de neurônios da camada de entrada. escondidos e os neurônios de saída. (10)

$$R_i = \frac{S_i}{\sum_{i=1}^{N} S_i} \quad (11)$$

(ii) Análise de Sensibilidades

A Análise de Sensibilidades propõe a utilização das derivadas parciais da variável de saída com relação aos pesos nas conexões, resultantes da fase de treinamento de uma RNA MLP com *Backpropagation*. O cálculo das sensibilidades de cada variável de entrada para cada exemplo de treinamento é calculado pela equação (12). Após o cálculo das sensibilidades, a contribuição de cada variável de entrada em relação a variável de saída é calculada como na equação (13).

$$Sen_{jL} = \sum_{k=1}^{N_{hid}} w_{kj} f'\left(net_k\right) \cdot \sum_{i=1}^{N_{out}} w_{ik} f'\left(net_i\right) \cdot e_i$$, onde N é a quantidade de exemplos contidos no conjunto de dados, L exemplo do conjunto de dados (L = 1 ... N), N_{inp} é a quantidade de neurônios de entrada, N_{hid} é o número de neurônios da camada intermediária e N_{out} é a quantidade de neurônios de saída. (12)

$$cont_j(\%) = \frac{\sum_{L=1}^{N} Sen_{jL}^2}{\sum_{j=1}^{N_{inp}} \sum_{L=1}^{N} Sen_{jL}} \quad (13)$$

5 Metodologia

Para realizar os experimentos foi implementada uma RNA MLP com *Backpropagation.* A quantidade de neurônios de entrada e de saída é determinada pela quantidade de variáveis do problema. O número de neurônios na camada intermediária foi escolhido por tentativa e erro e estimada inicialmente pela equação 14.. A taxa de aprendizagem foi escolhida após diversas simulações, porém o valor de 0,05, na maioria dos casos apresentou o melhor desempenho. Os números máximo e mínimmo de ciclos foram configurados para cada caso experimentado. A função de ativação escolhida foi a função sigmoidal logística, representada em (5).

$N_{hid} = (N_{out}+N_{inp})/2$ (14)

Na fase de pré-processamento, o conjunto de dados foi particionado de forma que 50% foi utilizada para a fase de treinamento, 25% para a fase de validação cruzada e 25% para os testes de verificação. Também no processamento, foi feita a normalização dos dados utilizando-se a transformação linear em (15). Em casos em que o formato da variável era nominal, procedeu-se com a binarização, que consiste em atribuir valores binários aos dados nominais.

$$y = (b-a)\frac{x_i - x_{min}}{x_{max} - x_{min}} + a \quad (15)$$

Para testar os algoritmos de seleção, a fim de encontrar o que pode melhor se aplica aos dados dos atletas, foram utilizadas algumas funções e algumas bases de dados, as quais possuem características bem conhecidas, por terem sido utilizadas em diversos estudos anteriores. As funções utilizadas foram o polinômio multivariado (16) o polinômio de Kolmogorov-Gabor (17) e a função com interação complexa (18). As bases utilizadas foram, a base das espécies-íris e a de vinhos, obtidas no site da UCI *Machine Learning* [7].

$$f(x_1, x_2, x_3, x_4, x_5) = 2 + 3x_1 x_2 + 4x_3 x_4 x_5 \tag{16}$$

$$f(x_1, x_2) = \frac{x_1^4 - 16x_1^2 + 5x_1}{2} + \frac{x_2^4 - 16_2^4 + 5x_2}{2} \tag{17}$$

$$f(x_1, x_2) = 1{,}9\left(1{,}35 + e^{x_1}\, seno\left(13\left(x_1 - 0{,}6\right)^2\right) e^{-x_2}\, seno\left(7x_4\right)\right) \tag{18}$$

O teste utilizando as funções consistiu em gerar, aleatoriamente, valores para as variáveis que faziam parte da função e também valores para variáveis que não faziam parte, e verificar se o algoritmo conseguiu identificar quais as variáveis que realmente têm importância para a função. Exemplo: como o polinômio de Kolmogorov-Gabor é uma função de duas variáveis x_1 e x_2, foram gerados valores aleatórios para estas duas variáveis e foram gerados valores para mais outras variáveis (x_1,x_2,x_3,...) que em nada influenciavam o resultado da função; o algoritmo de seleção, para estar correto, deverá atribuir maior importância às variáveis x_1 e x_2. Foram realizadas 30 simulações para cada caso e tirado o valor médio.

A base das espécies-íris é constituída por 4 atributos de entrada (*sepallenght,sepalwidth,petallenght, petalwidth*) e 150 registros. Os registros estão distribuídos entre três clases: íris-setosa, íris-virginica e íris-versicolor. A base de vinhos possui 13 atributos de entrada (*Alcohol, Malic-acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Nonflavanoid phenols, Flavanoids, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline*) , 178 registros distribuídos entre as classes 1,2 e 3.

Após os testes, o algoritmo de seleção que obtiver melhor resultado será utilizado para calcular a contribuição de cada fator de risco no surgimento de lesões em atletas.

5 Resultados dos Testes com os Algoritmos de Seleção de Variáveis

(i) Resultados obtidos nos testes com as funções

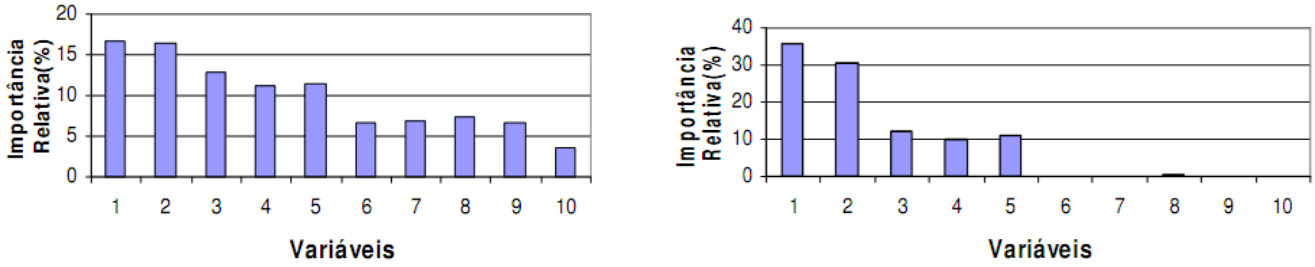
- O resultado dos testes com o polinômio multivariado pode ser visto na Tabela 1 e melhor visualizado na Figura 2. Lembrando que as variáveis não apresentavam esta ordem durante o treinamento, pois estavam em ordem aleatória, e foram ordenados na tabela, apenas para melhor visualização.

Tabela 1 – Resultado da aplicação dos algoritmos de seleção aos dados obtidos do treinamento da RNA para o polinômio multivariado

	Importância Relativa(%)	
Variável	**Algoritmo de Garson**	**Análise de Sensibilidades**
x_1	*16.71*	*35.69*
x_2	*16.51*	*30.65*
x_3	*12.83*	*12.14*
x_4	*11.21*	*9.96*
x_5	*11.48*	*10.93*
x_6	*6.68*	*0.0041*
x_7	*6.91*	*0*

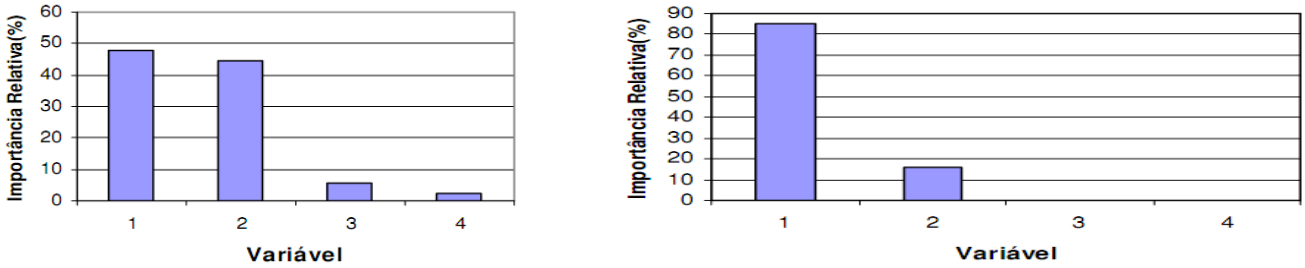
x_8	*7.38*	*0.545*
x_9	*6.64*	*0.03*
x_{10}	*3.61*	*0.04*

- Os testes com o polinômio de Kolmogorov-Gabor resultaram nos gráficos ilustrados na Figura 3. Ambos os algoritmos apresentaram valores de importância mais altos para as variáveis corretas, no caso x_1 e x_2.
- A Figura 4, ilustra os gráficos obtidos com os testes aplicados à função com interação complexa. É possível ver que o Algoritmo de Garson não apresentou um bom resultado na seleção das variáveis para essa função, dando maiores valores de importância às variáveis de x_3 a x_5 do que à variável x_1. Já a Análise de Sensibilidades não só deu maior importância às variáveis x_1 e x_2, como podou as variáveis que não influenciavam no resultado da função.



(a) **(b)**

Figura 2 – **Gráficos ilustrando os valores da importância relativa das variáveis resultantes do treinamento com o polinômio multivariado. Valores atribuídos pelo Algoritmo de Garson (a) e pela Análise de Sensibilidades (b).**



(a) **(b)**

Figura 3 – **Gráficos ilustrando os valores da importância relativa das variáveis resultantes do treinamento com a função de Kolmogorov-Gabor. Valores atribuídos pelo Algoritmo de Garson (a) e pela Análise de Sensibilidades (b).**

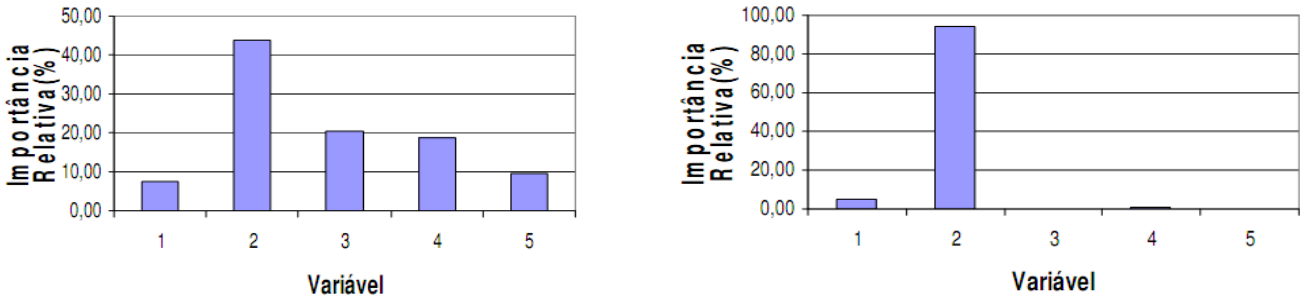
Observando os resultados dos testes com os algoritmos de seleção observamos que nos três casos apresentados, a Análise de Sensibilidades não apenas atribuiu maior percentual de influência às variáveis corretas, como também praticamente podou as variáveis que em nada influenciavam as funções.

(ii) Resultados obtidos nos testes com as bases de dados

- O treinamento da RNA MLP que gerou melhor resultado com a base das espécies íris teve uma taxa de acerto de 98,5%.

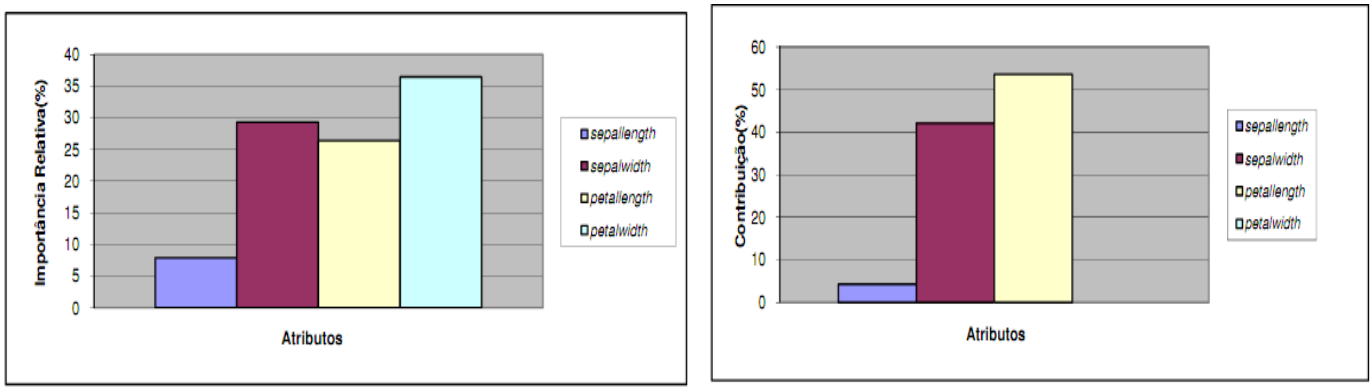
  A Figura 5 exibe o gráfico com os resultados da seleção de variáveis para a base de espécies-íris.

Percebe-se que a Análise de Sensibilidades podou o atributo *petalwidth*, assim realizou-se um novo treinamento da RNA MLP retirando este atributo. O novo treinamento resultou em uma taxa de acerto de 95%. Assim retirar o atributo *petalwidth* não gerou diferença significativa na taxa de acerto.



(a) **(b)**

Figura 4 – **Gráficos ilustrando os valores da importância relativa das variáveis resultantes do treinamento com a função com interação complexa. Valores atribuídos pelo Algoritmo de Garson (a) e pela Análise de Sensibilidades (b).**



(a) **(b)**

Figura 5 – **Gráficos ilustrando os valores da importância relativa das variáveis resultantes do treinamento com a base das espécies-íris. Valores atribuídos pelo Algoritmo de Garson (a) e pela Análise de Sensibilidades (b).**

6 Resultados da Seleção de atributos

Por ter mostrado melhores resultados nos testes, a Análise de Sensibilidades foi o algoritmo escolhido para realizar a quantização da contribuição de cada fator no favorecimento de lesões em atletas.

A base de dados utilizada para o estudo da contribuição dos fatores de risco de lesões foi resultado da pesquisa de campo realizada por Pastre. Consiste em 215 registros de atletas da elite brasileira, sendo estes distribuídos em cinco diferentes categorias: velocistas (50 registros), fundistas (38 registros), nadadores de categoria borboleta (44 registros), nadadores de categoria peito(40 registros) e nadadores de categoria costas (43 registros). Os atributos de entrada são idade, sexo, Índice de Massa Corporal (IMC),peso, estatura, tempo de prática, categoria, frequência de ocorrência de lesões e horas de treino diário. A saída de cada registro é a existência ou não de lesões (0 ou 1).

O treinamento que de melhor resultado obteve uma taxa de acerto de 61%. Esta taxa de acerto poderia ser maior no caso a base utilizada contivesse uma maior quantidade de dados.

Aplicando-se a Análise de Sensibilidades aos dados dos atletas, obtemos os valores de contribuição constantes na Tabela 2, e podem ser visualizados pelo gráfico da Figura 6.

Tabela 2 – Contribuição dos fatores para o favorecimento de lesões.

Fator	**Contribuição(%)**

Idade	*14,42*
Sexo	*13,3*
Peso	*4,86*
Estatura	*10,86*
IMC	*1,52*
Tempo de Prática	*36,34*
Categoria	*0,39*
Horas de Treino Diário	*18,33*
Frequência de Ocorrência de lesões	*0*

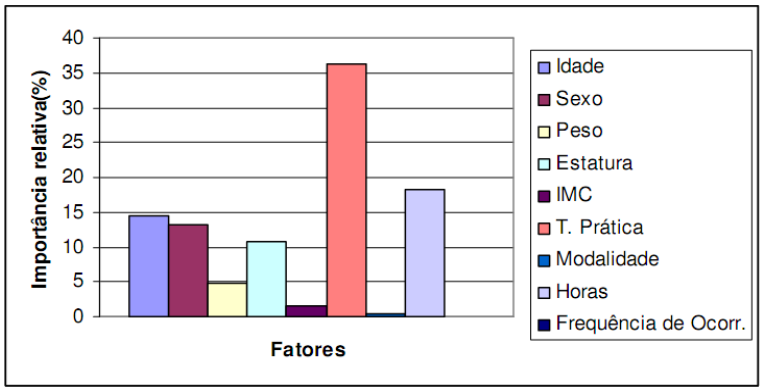


Figura 6 – **Gráfico ilustrando os valores da importância relativa das variáveis resultantes do treinamento com a base de dados dos atletas. Valores atribuídos pela Análise de Sensibilidades .**

7 Conclusões

Observando os dados resultantes da seleção de fatores de risco de lesões, podemos fazer as seguintes considerações:

- Ao fator freqüência ocorrida de lesões não foi atribuído qualquer valor de importância, e o fator categoria também foi considerado ter pouca contribuição no surgimento de lesões. O fato de a categoria ter recebido baixo valor de importância pode estar ligado ao fato de os atletas lesionados estarem distribuídos de forma aproximada em cada categoria. Como mostra a tabela 4.
- O atributo de maior importância foi o tempo de prática da modalidade, seguido do número de horas de treino diário. Tais resultados fazem sentido segundo Hootman [8] e Hino *et al.* [9], os quais afirma que à medida que se aumenta o tempo de treino, o risco de lesões tende a crescer.
- Aos fatores peso e IMC foi atribuído menos importância que aos fatores Sexo e Idade. Isto pode ser explicado pelo fato de que apenas 8% dos atletas apresentam sobrepeso.

Diante do observado, os resultados de seleção dos fatores que incidem sobre lesões em atletas pode ser considerado satisfatório. Isto mostra que a Análise de Sensibilidades é uma boa alternativa para tratar problemas deste tipo. A existência de uma base de dados com maior quantidade de registros poderia resultar em uma melhor seleção destes fatores.

Dos algoritmos de seleção utilizados podemos faze alguma considerações:

- Como explanado anteriormente, o Algoritmo de Garson por utilizar os valores absolutos dos pesos das conexões, não permite uma análise das direções dos valores na saída em relação às modificações nas entradas. Em alguns casos determinar esta direção é de grande importância.

- A Análise de Sensibilidades, por utilizar as sensibilidades resultantes do treinamento da RNA MLP com *Backpropagation*, impede que este tipo de teste seja feito em outro tipo de algoritmo de treinamento sem que posteriormente estas sensibilidades tenham que ser calculadas.

Assim, como trabalho futuro, fica o desenvolvimento de um técnica de seleção que não tenha a limitação do Algoritmo de Garson e que não dependa da técnica de treinamento, como a Análise de Sensibilidades.

8 References

[1] J.A. Kettunen, U.M. Kujala, J. Kaprio, M. Koskenvuo, S. Sarna. Lower-limb function among former elite male athletes. **Am J Sports Med** (2001);29:2-8.

[2] C.M . Pastre, G.C. Filho, H.L. Monteiro, J.N. Júnior, C.R. Padovani. Lesões desportivas na elite do atletismo brasileiro: estudo a partir de morbidade referida. Rev Bras Med Esporte. (2005);11:43-7.

[3] J. Weineck. Biologia do esporte. **3a ed. São Paulo: Manole**,(1991).

[4] B.D.O. Hernandes Jr.. Treinamento desportivo**. Rio de Janeiro: Sprint**, (2002)

[5] GARSON, G.Dx Interpreting neural-network connection weights**. Artif. Intell. Expert** 6 (1991), 47–51.

[6] M.J.S. Valença, T. B. Ludemir. Explicando a relação entre as variáveis de uma rede neural – Iluminando a "Caixa Preta". **XVII Simpósio Brasileiro de Recursos Hídricos, São Paulo**, Novembro, (2007).

[7] http://archive.ics.uci.edu/ml/datasets

[8] J.M. Hootman, C.A. Macera, B.E. Ainsworth, M. Martin, C.L. Addy, S.N. Blair. Association among physical activity level, cardiorespiratory fitness, and risk of musculoskeletal injury. **Am J Epidemiol**. 2001;154(3):251-8.

[9] A.A.F. Hino, R.S. Reis, C.R. Rodriguez-Añez,R, R.C. Fermino. Prevalência de Lesões em corredores de rua e Fatores Associados. **Rev Bras Med Esporte** – Vol. 15, No 1 – Jan/Fev, 2009.