DOI: 10.18605/2175-7275/cereus.v10n2p50-59

<<Recebido: 15/06/2018.Aceito: 03/07/2018 >>

ARTIGO ORIGINAL

REVISTA CEREUS ISSN: 2175-7275

MÉTODO DE ANÁLISE COMBINATÓRIA APLICADO À EXTRAÇÃO DE CARACTERÍSTICAS VIA REDES NEURAIS ARTIFICIAIS

COMBINATORIAL ANALYSIS METHOD APPLIED TO FEATURE EXTRACTION VIA ARTIFICIAL NEURAL NETWORKS

Guilherme Fernandes Marchezini[1], Gabriel Alves de Campos Mattar[2], Daniel de Paula Braga Lopes[2], Rogério Martins Gomes[3], Bruno André Santos[3]

[1] Graduando em Engenharia da Computação do CEFET-MG.

E-mail:

guilhermefmar@gmail.com

[2] Graduando em Engenharia da Computação do CEFET-MG

[3] Professor do Departamento de Computação do CEFET-MG.

RESUMO

As Redes Neurais Artificiais (RNA) têm sido muito utilizadas na solução de problemas de reconhecimento de padrões. Entretanto, poucos são os trabalhos que abordam o problema de extração de características. Sendo assim, este artigo propõe um método de análise combinatória na verificação da relevância das características aplicadas ao diagnóstico de diabetes utilizando uma RNA de múltiplas camadas. Os experimentos foram realizados utilizando uma rede neural com uma camada escondida e uma base de dados com oito características de uma população de mulheres com pelo menos 21 anos de idade descendentes da tribo indígena Pima, Arizona. Após o treino, uma análise combinatória com todas as características foi realizada a fim de determinar a relevância de cada uma delas no processo de classificação. Este artigo concluiu que a característica [b] (concentração plasmática de glicose em um teste oral de tolerância) apresentou a maior relevância para o processo de classificação com uma acurácia de 60,99% em relação a predição real. A característica [g] por sua vez, apresentou a menor influência na taxa de classificação, enquanto as combinações [b, c, f] se mostraram relevantes em toda a fase de teste.

Palavras-chave: Classificação, Extração de Características, Redes Neurais Artificiais

ABSTRACT

Artificial Neural Networks (ANNs) have been widely used in the solution of pattern recognition problems. However, few studies have addressed the problem of feature extraction. Thus, this article proposes a method of combinatorial analysis in the verification of the relevance of the features applied to the diagnosis of diabetes using a multilayer RNA. The experiments were performed using a neural network with a hidden layer and a database with eight features of a population of women at least 21 years of age descended from the Pima indigenous tribe of Arizona. After the training, a combinatorial analysis with all the features was carried out to determine the relevance of each of them in the classification process. This article concluded that the feature [b] (plasma glucose concentration in an oral tolerance test) presented the highest relevance for the classification process with an accuracy of 60.99% in relation to the real prediction. The feature [g], on the other hand, had the lowest influence on the classification rate, while the combinations [b, c, f] were relevant throughout the test phase.

Keywords: Classification, Feature Extraction, Artificial Neural Networks

1. INTRODUÇÃO

As técnicas de *Machine Learning* (ML) usualmente trabalham com base no raciocínio indutivo no reconhecimento de padrões. Neste raciocínio, a forma de inferência lógica que permite se chegar a determinadas conclusões é realizada a partir de um conjunto de exemplos. Estes exemplos são, na verdade, medidas ou características extraídas de um determinado objeto que sejam capazes de representá-lo. Além disso, no processo de reconhecimento de padrão, torna-se imprescindível a escolha de um método de classificação, que deve ser treinado a partir do conjunto de exemplos obtidos (DUDA et al., 2012). Entretanto, na maioria das vezes torna-se difícil ou praticamente impossível encontrar uma explicação de quais características (*features*) são relevantes para um sistema de classificação realizar a sua tarefa de forma adequada. Além das características, *per si*, é possível também que a interação entre elas possa interferir de maneira positiva ou negativa no desempenho do sistema.

Para resolver esse problema surgiram duas vertentes dentro deste contexto. A primeira é a utilização de árvores de decisão, adequadas para classificação, que facilitam a extração de regras de decisão. A segunda vertente seria o uso de alguma técnica ou algoritmo que seja capaz de extrair ou descobrir quais características são relevantes a partir dos exemplos fornecidos, como por exemplo, as redes neurais artificiais (RNA).

As árvores de decisão são representações gráficas que consistem de: nodos que representam os atributos; arcos que correspondem ao valor de um atributo; e nodos folha que designam uma classificação. Neste método, o espaço de entrada é dividido em regiões disjuntas com o objetivo de construir uma fronteira de decisão. Estas regiões são escolhidas com base em uma otimização heurística no qual, a cada passo, os algoritmos selecionam a variável que provê a melhor separação de classes de acordo com alguma função custo.

O algoritmo ID3, proposto por (QUINLAN, 1986), é um dos algoritmos usados na construção das árvores de decisão. Este algoritmo constrói uma árvore de decisão em que cada vértice (nodo) corresponde a um atributo e cada aresta da árvore a um valor possível do atributo. Uma folha da árvore, por sua vez, corresponde ao valor esperado da decisão, segundo os dados de treinamento utilizados (classe). Uma característica importante deste algoritmo é que a seleção dos nodos, a serem utilizados na árvore, é baseada na Teoria da Informação de Shannon (SHANNON, 1998), mais especificamente nos conceitos de entropia e ganho de informação. Assim, os atributos que farão parte dos nodos são escolhidos entre os atributos que possuem o maior ganho de informação. Uma extensão

deste algoritmo é o algoritmo C 4.5, também proposto por (QUINLAN, 1986), que constrói árvores de decisão com valores desconhecidos para alguns atributos. Este algoritmo, além de trabalhar com atributos que apresentam valores contínuos, é capaz de utilizar o conceito de poda (*pruning*) de árvores.

As Redes Neurais Artificiais (RNA), por sua vez, devido a sua capacidade de aprendizado, têm sido muito utilizadas no reconhecimento de padrões, além de permitir a inferência das relações existentes entre as variáveis de entrada (*features*) e saída da rede. Um exemplo de aplicação das RNA seria a utilização das características de um paciente na determinação da existência ou não de uma doença. As RNA têm como vantagens: rapidez no processo de classificação; capacidade de generalização; e tolerância a ruído, oferecendo boas respostas mesmo na falta de dados (HAYKIN, 2009).

Diversos métodos utilizando redes neurais artificiais na determinação da quantidade mínima de características necessárias à solução de problemas de classificação de padrões têm sido propostos na literatura (DA CUNHA CALVACANTI, 2000). Neste artigo, por sua vez, é proposto um método de análise combinatória na verificação da relevância de cada uma das características (*features*), bem como das suas inter-relações, a serem usadas por uma rede neural artificial de múltiplas camadas (*MultiLayer Perceptron* - MLP) no diagnóstico de diabetes. Os experimentos foram realizados com uma base de dados que possuem 8 características de uma população de mulheres com pelo menos 21 anos de idade descendentes da tribo indígena Pima, de Phoenix, Arizona. Este conjunto de dados é originalmente do Instituto Nacional de Diabetes e Doenças Digestivas e renais dos Estados Unidos da (KAGGLE, 2018).

Este artigo está organizado da seguinte forma: a Seção 2 apresenta a metodologia utilizada no processo de verificação da relevância das características de um problema de diagnóstico de doença. A Seção 3 apresenta os resultados obtidos e a Seção 4 os analisa. Finalmente, a Seção 5 conclui o artigo e apresenta perspectivas de trabalhos futuros.

2. MATERIAIS E MÉTODOS

Um sistema completo de reconhecimento de padrões é composto por diversas etapas. A primeira etapa é constituída por um módulo de aquisição de dados seguido de um extrator de características, também chamado de módulo de pré-processamento. A seguir, os dados obtidos da etapa anterior são enviados a um módulo classificador que, utilizando algum algoritmo ou técnica computacional, define a classe referente dos dados de entrada.

Neste trabalho não é abordado os módulos de aquisição de dados e pré-processamento, mas somente o módulo classificador. Os experimentos foram realizados com uma base de dados já existe na literatura (KAGGLE, 2018) que possuem oito características de uma população de mulheres com pelo menos 21 anos de idade descendentes da tribo indígena Pima, de Phoenix, Arizona. Esta base de dados tem por objetivo diagnosticar a incidência de diabetes e possui as características descritas a seguir:

1- Características extraídas de cada um dos pacientes:
    a. Número de vezes que engravidou;
    b. Concentração plasmática de glicose há 2 horas em um teste oral de tolerância à glicose;
    c. Pressão sanguínea diastólica (mm Hg);
    d. Espessura da dobra da pele do tríceps (mm);
    e. Insulina sérica de 2 horas (muU/ ml);
    f. Índice de massa corporal (peso_em_kg /(altura_em_m))^2
    g. Função de pedigree de diabetes
    h. Idade (Anos)
2- Definição da Classe: 1 ou 0, sendo 1 para pacientes que apresentam teste positivo para diabetes e 0 para teste negativo.
3- 768 Intâncias, sendo:
    a. 500 da classe 0
    b. 268 da classe 1
4- Alguns atributos da base estão incompletos.

O módulo classificador utilizado neste trabalho é responsável, não somente pelas etapas de classificação, mas também tem por objetivo determinar as características que são mais relevantes na discriminação das classes existentes na base de dados utilizada. Este classificador, mostrado na Figura 1, é constituído por uma rede neural artificial com uma camada escondida.

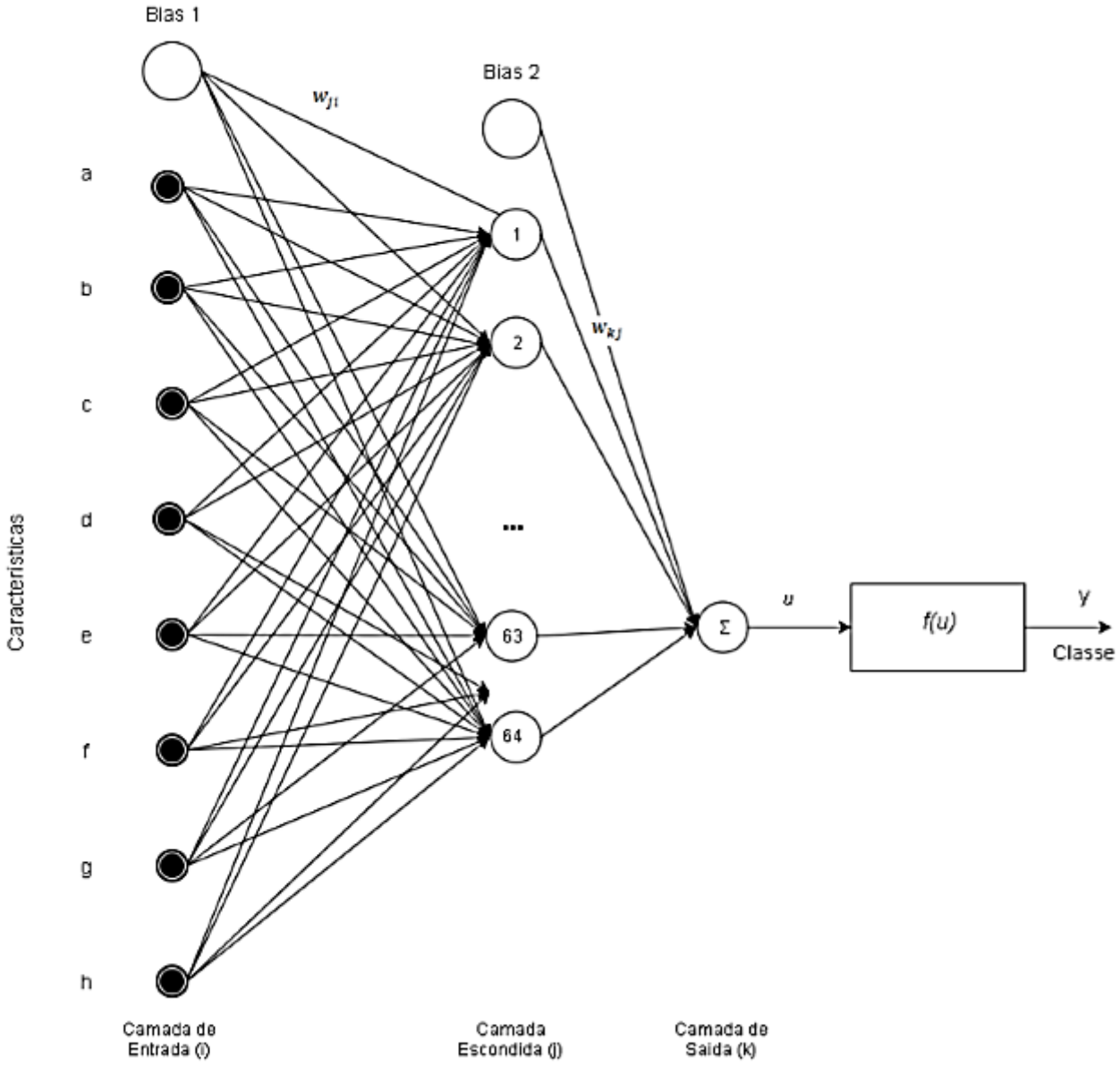

Figura 1. Classificador - RNA

Esta rede (Figura 1) possui 8 entradas (características), 64 neurônios na camada escondida, com função de ativação sigmoidal e 1 neurônio na camada de saída (classe: 0 ou 1).

O nível de ativação da camada escondida y_j é definido por:

$$y_j = f(\sum_{i=0}^{h} x_i w_{ji})$$

onde x_i representa as características de entrada da rede; $i = 0$ representa o bias 1; e y_j é a saída do neurônio j da camada escondida com função de ativação sigmoidal.
O nível de ativação u é definido por:

$$u = \sum_{j=0}^{64} y_j w_{jk}$$

onde y_j é a saída do neurônio j da camada escondida; e $j = 0$ representa o bias 2.

A função de ativação final da rede $f(u)$ é também definida por uma função sigmoidal e é expressa da seguinte forma:

$$y = f(u) = \begin{cases} 1 \; se \; u \geq 0.5 \\ 0 \; se \; u < 0.5 \end{cases}$$

O algoritmo de treinamento utilizado é o *backpropagation* e os exemplos da base de dados são divididos em dois conjuntos: treinamento (70%) e teste (30%). Primeiramente, essa divisão é realizada com o objetivo de verificar a ocorrência de *overfit* ou *underfit* durante o processo de aprendizado e predição das classes. Além disso, a base de teste, usada no processo de predição das classes, é utilizada na verificação da acurácia nos passos posteriores do algoritmo.

Inicialmente a rede neural é treinada considerando as oito características presentes na base de dados utilizada. Uma vez obtida a acurácia de classificação, este valor é tomado como referência para as futuras avaliações. Dessa forma, para analisar a importância de cada uma das características um processo iterativo com todas as possíveis combinações das características é realizado, começando com combinações 1 a 1 e terminando com combinações $N - 1$ a $N - 1$. Para cada uma das possíveis combinações o algoritmo irá zerar os pesos referentes às características que não estão sendo analisadas, sendo a nova rede utilizada na medição da taxa de acerto de classificação das amostras de teste. Dessa forma, a medição da acurácia é realizada para todas as combinações possíveis das características. Caso seja utilizado apenas combinações 1 a 1 é possível analisar as características, de forma isolada, que mais influenciaram na predição da rede. Outra análise realizada é a verificação da relevância total de uma característica na capacidade de classificação da rede. Esta tarefa é realizada somando-se a taxa de acerto de cada uma das características para todas as combinações em que ela aparece, desde 1 a 1 até $N - 1$ a $N - 1$. A característica que apresenta a maior soma é a que possui a maior relevância considerando que as características não são independentes. O pseudocódigo (Figura 2) sumariza todas as etapas referentes à realização dos testes.

Algorithm 1 Classificador

1 [Início] Separar os dados e criar uma rede neural artificial com *N* neurônios na camada escondida.
2 [Treino] Treinar a rede com a base de dados de treinamento, utilizando todas as 8 características.
3 [Predição] Executar a tarefa de predição, armazenando todas as classes das entradas da base de dados de treinamento.
4 [Arranjo] Iniciar com uma característica e a cada interação aumentar o número de características que serão combinadas até chegar a *N-1* por *N-1* :
4.1 Combinações Selecionar cada uma das possíveis combinações.
4.11 Rede Selecionar a rede treinada no passo 2.
4.12 Pesos Utilizar a rede selecionada e zerar todos os pesos das ligações da entrada com a camada escondida referente às características que não fazem parte da combinação selecionada.
4.13 Predição Executar a tarefa de predição para todas as classes utilizando a base de dados de treinamento e a rede com os novos pesos.
4.14 Aceitação Calcular a acurácia do sistema utilizando como referência as classes preditas no passo 3.
5 [Resultado] Ordenar em ordem crescente de acurácia e exibir as combinações que apresentaram os maiores valores.

Figura 2. Pseudo-Código

3. RESULTADOS

O resultado obtido para o problema é uma lista com os números das características presentes nas combinações que apresentaram maior relevância na classificação de uma rede treinada. A escolha do número de características presentes nas combinações é extremamente importante, pois a medida que este número cresce e se aproxima da quantidade total da rede, ou seja, das 8 características, a acurácia tende a aumentar, chegando a valores bem próximos de 100%. Apesar disso, é possível, dependendo da base de dados, descobrir a ausência de relevância de uma determinada característica, caso não seja observado alteração significativa da taxa de acerto do sistema na presença ou ausência da mesma. Assim, a rede neural poderia ser simplificada caso o sistema chegasse ao mesmo valor de acurácia utilizando um número menor de características.

A Tabela 1 mostra as características que apresentaram maior relevância na classificação combinadas de 1 a 1 até 7 a 7 em relação a rede neural treinada com todas as 8 características. Isto significa que o valor da acurácia da rede para as 8 características, foi tomada como referência e todas as outras medições foram realizadas em relação a este valor e não em relação ao valor real de classificação.

Tabela 1. Combinações *versus* acurácia relativa.

Combinações	(Acurácia de cada combinação)/(Acurácia para as 8 características) (%)
[b]	60.99
[b,e]	62.73
[b,c,f]	75.55
[b,c,f,g]	75.18
[a,b,c,f,g]	85.56
[a,b,c,d,f,g]	89.65
[a,b,c,d,e,f,g]	97.61

4. DISCUSSÃO

Podemos observar na Tabela 1 que, isoladamente, a característica [b] (*concentração plasmática de glicose há 2 horas em um teste oral de tolerância à glicose*) apresentou a maior relevância para a predição da rede neural artificial. É interessante observar que quando somente duas características foram observadas, a combinação que apresentou o melhor resultado foi a [b, e], enquanto que para três ou mais características a combinação [b, c, f] sempre esteve presente.

Outra observação importante foi em relação a característica [e] *(Insulina sérica de 2 horas*), que somente foi encontrada quando duas características foram combinadas, voltando a ser observada novamente somente quando sete características foram utilizadas. O interessante desta característica é que para duas combinações a melhoria da acurácia foi relativamente pequena ($\approx 2\%$) em relação a combinação anterior, enquanto para 7, a melhora na acurácia foi praticamente de 8% em relação a 6 combinações, mostrando que as características se interagem, influenciando mutualmente na classificação.

A Figura 3 mostra que, na maioria dos casos, a acurácia cresce à medida em que as características vão sendo adicionadas. A exceção ocorre na transição entre a combinação de duas características ([b, e]) para três ([b, c, f]). Da mesma forma, é possível observar que característica [h] (*Idade da paciente*) apresentou pouca relevância na classificação, já que o seu uso implicaria em uma melhoria de somente 2,39% na acurácia.

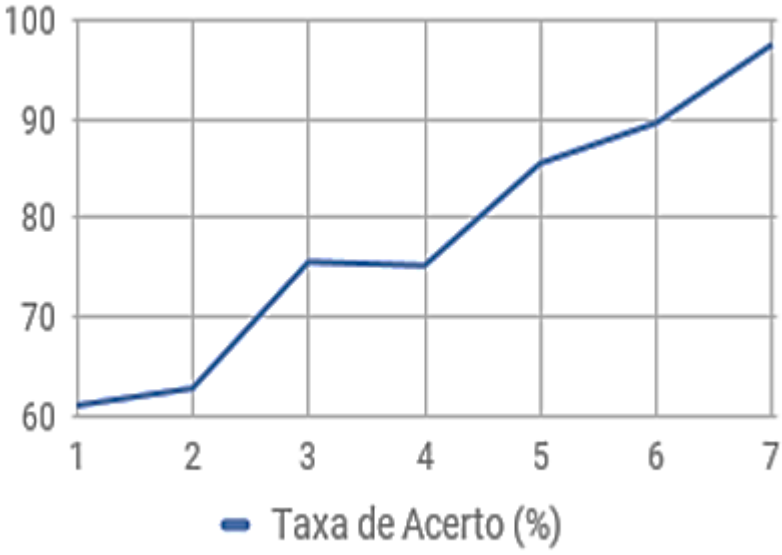

Figura 3. Taxa de acerto X Quantidade de *features* combinadas

5. CONSIDERAÇÕES FINAIS

Este artigo propôs um método de análise combinatória na verificação da relevância das características a serem usadas por uma RNA de múltiplas camadas no diagnóstico de diabetes. Como a complexidade do algoritmo utilizado neste trabalho é não polinomial por usar força bruta, uma condição para a aplicabilidade do método seria a utilização de bases de dados que possuam poucas características.

Os resultados encontrados mostraram que a característica [b] (*concentração plasmática de glicose há 2 horas em um teste oral de tolerância à glicose*) apresentou a maior relevância para o processo de classificação apresentando uma acurácia de 60,99% em relação a predição real. A característica [g] (*função de pedigree de diabetes*), por sua vez, apresentou a menor influência na taxa de classificação, enquanto as combinações [b, c, f] (*concentração plasmática de glicose há 2 horas em um teste oral de tolerância à glicose; pressão sanguínea diastólica; e índice de massa corporal*) se mostraram relevantes em toda fase de teste.

A importância deste tipo de análise está na identificação dos principais indicadores de uma determinada doença. Dessa forma, variáveis menos relevantes poderiam ser retiradas de forma a reduzir a complexidade do problema, facilitando a ação dos especialistas.

Uma proposta de trabalho futuro seria o estudo sobre o uso de heurísticas computacionais na categorização das características relevantes de uma base, aliado ao processo de otimização do classificador.

REFERÊNCIAS

DA CUNHA CALVACANTI, H. M. V. **Extração de características via redes neurais**. Dissertação (Mestrado) - FEEC/UNICAMP, 2000.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. 2. ed. USA: John Wiley & Sons, 2012.

HAYKIN, S. O. **Neural networks and learning machines**. 3. ed. Upper Saddle River, NJ, USA: Pearson, 2009.

QUINLAN, J. R. **Induction of decision trees.** Machine learning, v. 1, n. 1, 1986, p. 81-106.

QUINLAN, J. R. **C4. 5**: Programming for machine learning. USA: Morgan Kauffmann, 1993.

SHANNON, C. E.; WEAVER, W. **The mathematical theory of communication**. USA: University of Illinois press, 1998.

KAGGLE. **The Home of data Science & Machine Learning**. https://www.kaggle.com/uciml/pima-indians-diabetes-database. Acesso em: 01 de junho de 2018.