Método Computacional para o Diagnóstico Precoce da Granulomatose de Wegener

Computational Method for early Diagnosis Wegener's Granulomatosis

José do Nascimento Linhares [1 2]
Lúcio Flávio A. Campos [1 3]
Ewaldo Eder Carvalho Santana [1 4]
Jardiel Nunes Almeida [1 5]
Flávia Larisse da Silva Fernandes [1 6]

Data de submissão: 29/12/2015, Data de aceite: 25/04/2016

Resumo: Neste trabalho é apresentado um sistema de reconhecimento de padrões proteômicos com o objetivo de auxiliar o diagnóstico precoce da Granulomatose de Wegener (GW), uma vasculite idiopática rara de difícil detecção e alta taxa de mortalidade para indivíduos não tratados. O método consiste em extrair características de sinais proteômicos e classificá-las como sendo de indivíduos portadores ou não portadores de GW. Para tanto, utiliza-se Análise de Componentes Independentes para extrair características dos sinais, Algoritmo de Máxima Relevância e Mínima Redundância para reduzir o número de características e custos computacionais e Máquina de Vetores de Suporte para classificar. A qualidade do método foi avaliada utilizando uma base de dados com 335 sinais proteômicos, composta por 75 casos ativos, 101 casos negativos e 159 em remissão. O melhor resultado obtido foi para um vetor de vinte características cuja acurácia, especificidade e sensibilidade foram, respectivamente, de: $98,24\%$, $99,73\%$ e $99,50\%$.

***P*alavras-chave:** diagnóstico, granulomatose de Wegener, método computacional, padrões proteômicos

[1] Universidade Estadual do Maranhão (UEMA), Centro de Ciências Tecnológicas, Programa de Pós-Graduação em Engenharia de Computação e Sistemas - São Luís - Maranhão - Brasil
[2] {linhares.jose@yahoo.com.br}
[3] {lucioflavio@gmail.com}
[4] {ewaldoeder@gmail.com}
[5] {jardieliguaiba@gmail.com}
[6] {larisse.nandes@gmail.com}

Abstract: This paper presents a recognition system of proteomic patterns in order to assist in the early diagnosis of Wegener's Granulomatosis (WG), a rare idiopathic vasculitis difficult to detect and of high mortality rate for untreated individuals. The method consists of extracting features of proteomic signs and classifying them as being of bearers individuals or non-carriers of GW. For this purpose, we use Independent Components Analysis to extract characteristics of these signals, Algorithm of Maximum Relevance and Minimum Redundancy to reduce the number of features and computational costs and Support Vector Machine to qualify them. The performance of the method was evaluated using a database of 335 proteomic signals, comprising 75 active cases, 101 negative cases and 159 in remission. The best result was obtained for a vector with twenty characteristics whose accuracy, sensitivity and specificity were, respectively: 98.24%, 99.73% and 99.50%.

***K*eywords:** diagnosis, Wegener's granulomatosis, computational method, proteomic patterns

1 Introdução

A Granulomatose de Wegener (GW) é uma vasculite granulomatosa autoimune multissistêmica rara de difícil detecção, que atinge 3 em cada 100.000 pessoas no mundo (1, 2, 3). Esta doença afeta os vasos sanguíneos de pequeno e médio calibre e vênulas do sistema respiratório superior, pulmões e rins, causando inflamação e consequente necrose dos tecidos desses órgãos. Em alguns casos, pode atingir também o coração, o sistema nervoso, olhos, pele, trato gastrointestinal e musculoesquelético (4, 2). A GW é uma patologia que quando não diagnosticada e tratada precocemente, pode levar o paciente a óbito em apenas um ano.

Atualmente a GW é diagnosticada através de sintomas, exames clínicos, radiológicos e sorológicos que seguem critérios propostos pelo *American College of Rheumatology* (5). Se dois dos seguintes achados: inflamação oral ou nasal, nódulos ou opacidades na radiografia de tórax, hematúria microscópica, inflamação granulomatosa na biópsia da parede de vasos e a presença do anticorpo Anti Citoplasma de Neutrófilos (ANCA-c) positivo forem encontrados, tem-se até 90% de especificidade. Porém, outras doenças da classe das vasculites sistêmicas também apresentam o ANCA-c positivo (6). Vale ressaltar, que os sintomas iniciais da GW são praticamente inespecíficos, o que não permite sua diferenciação em estágios iniciais.

O tratamento é feito com uso de citotóxicos e imunossupressores para combater as reações imunológicas do organismo. O sucesso da terapia está diretamente relacionado com a detecção precoce da enfermidade, pois isto influencia na dosagem dos medicamentos. Se o tratamento for iniciado de forma tardia, doses maiores de medicamentos são aplicadas o que pode potencializar seus efeitos colaterais, trazendo complicações cardíacas, infertilidade, obesidade, osteoporose, hipertensão arterial, diabetes e infecções oportunistas (7). Verifica-se

assim, a necessidade do desenvolvimento de métodos de diagnósticos para a GW que sejam precisos e que permitam a detecção precoce da mesma.

Recentemente a comunidade científica vem aplicando técnicas de CAD (*Computer Aided Diagnosis*) em várias doenças (8, 9, 10, 11). Araújo (8), por exemplo, utiliza a Análise de Componentes Independentes (ICA) para extrair características de sinais proteômicos com o objetivo de diagnosticar o câncer de ovário . Áurea (9) propõe um método de diagnóstico precoce da Diabetes utilizando ICA e Máquina de Vetor de Suporte (SVM). Yu (10) aplica sinais proteômicos e bioinformática para detecção do câncer de colo retal. Mantini (11) usa ICA e padrões proteômicos para identificação de biomarcadores e sua possível associação com doenças.

Neste trabalho, a partir do estudo da espectrometria de massa, especificamente de sinais proteômicos, combinado com métodos computacionais, propõe-se uma metodologia de detecção precoce da GW. O método proposto consiste basicamente em extrair caraterísticas de sinais proteômicos para classificá-los como sendo de indivíduos portadores ou não portadores de GW.

2 Metodologia Proposta

O método proposto é constituído de três submétodos que consistem em: extrair características de sinais proteômicos utilizando Análise de Componentes Independentes (ICA), reduzir a quantidade de características com a técnica de Máxima Relevância e Mínima Redundância (mRMR), afim de diminuir os custos computacionais e classificar com a Máquina de Vetores de Suporte (SVM). A figura 1 mostra um diagrama do método proposto. A seguir descreveremos cada um desses métodos.

2.1 Espectrometria de Massa e Sinais Proteômicos

De acordo com Araújo (12), a ciência tem procurado e desenvolvido formas de diagnosticar doenças precocemente. Nesse sentido o estudo de sinais proteômicos, que é o conjunto de proteínas expressas a partir de um determinado genoma, tem se mostrado promissor, pois o proteoma está em constante mudança devido as respostas que podem ser obtidas aos estímulos externos e internos . Assim, a presença de uma doença pode mudar de forma significativa as características das proteínas e consequentemente do proteoma, indicando qual a patologia que acomete o paciente ou possíveis biomarcadores que possam indicar a presença da enfermidade (13).

Atualmente um dos métodos mais utilizados para obtenção de sinais proteômicos é a espectrometria de massa, que é uma técnica analítica física que permite detectar e identificar moléculas por meio de sua razão massa/carga (m/z). Para a aplicação dessa técnica,

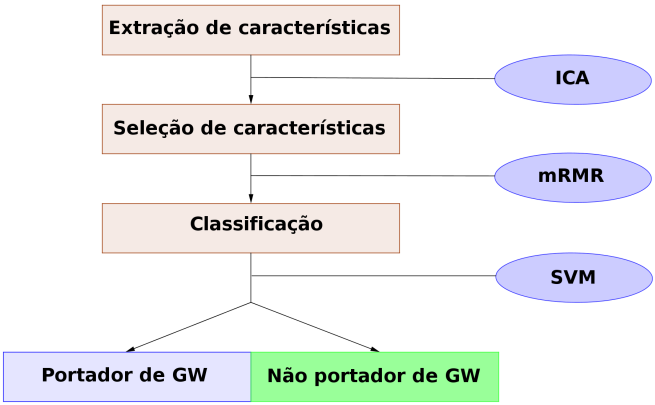

Figura 1. Diagrama da metodologia proposta.

utiliza-se um espectrômetro de massa que é composto basicamente por uma fonte de íons, um analisador de massas, um detector de íons e uma unidade de aquisição de dados.

Neste trabalho utilizamos uma base de dados com sinais proteômicos obtidos a partir de um espectrômetro de massa que utiliza a técnica de ionização *Surface-enhanced laser desorption/ionization* (SELD) e um analisador de massas do tipo *Time of Flight* (TOF) (14). Em SELD, a ionização é feita depositando-se a mistura de proteínas, que se deseja analisar, sobre uma superfície com afinidade química, em seguida, essa superfície é lavada restando apenas as moléculas que se ligaram a ela. Após a lavagem, uma matriz é posta sobre a superfície e deixada cristalizar. Logo após, o analito é excitado por laser para formar os íons em fase gasosa.

No analisador TOF, os íons são acelerados por um potencial elétrico em um tubo de vácuo e detectados de acordo com seu tempo de voo (15), que é proporcional a m/z. O resultado ao final de todo o processo é um espectro de massas. O espectro obtido é um gráfico que mostra a intensidade relativa de cada íon que aparece como picos com m/z definidos. A figura 2 mostra um espectro de massa obtido com a técnica SELD-TOF.

2.2 Extração de Características pela Análise de Componentes Independentes

A análise de componentes independentes (ICA-*Independent Component Analysis*) é um modelo estatístico usado em processamento de sinais para recuperar fontes estatisticamente independentes ou extrair características de um sinal (16). No modelo ICA linear é considerado que um dado vetor aleatório $\mathbf{X}$ de sinais observados, por exemplo, o sinal pro-

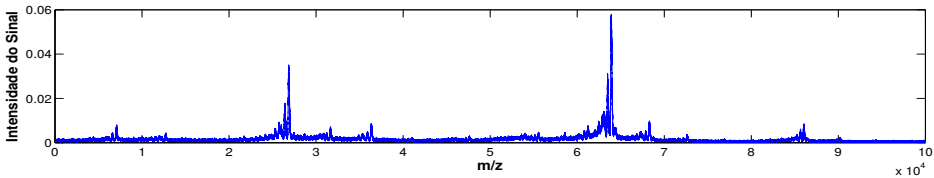

Figura 2. Espectro de massa obtido de um espectrômetro de massas.

teômico, é gerado a partir da atuação de um operador linear $\mathbf{A}$ sobre um vetor $\mathbf{S}$, cujas componentes são mútua e estatisticamente independentes e não gaussianas. Matematicamente pode-se escrever

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

Sendo: $\mathbf{X} = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n} \end{pmatrix}$, $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$ e $\mathbf{S} = \begin{pmatrix} s_{11} \\ s_{12} \\ \vdots \\ s_{1n} \end{pmatrix}$.

A matriz $\mathbf{A}$ é vista como uma matriz de mistura e a equação 1 (modelo ICA) mostra como os sinais observados $\mathbf{X}$ são gerados a partir da mistura das componentes independentes de $\mathbf{S}$.

O problema principal em ICA é encontrar $\mathbf{A}$ e $\mathbf{S}$ conhecendo apenas o vetor $\mathbf{X}$ e dependendo da aplicação que se queira fazer, a matriz de interesse poderá ser $\mathbf{A}$ ou $\mathbf{S}$. Na extração de características de sinais proteômicos, por exemplo, a matriz utilizada é $\mathbf{A}$, pois suas colunas representam as características de cada um dos sinais.

Na prática é impossível resolver com exatidão a equação 1 e obter a matriz de características $\mathbf{A}$, porém estimativas podem ser obtidas utilizando a informação mútua ou explorando a propriedade de não gaussianidade das componentes de $\mathbf{S}$. Essa segunda abordagem, tem como alicerce o teorema do limite central, que diz que a soma de variáveis aleatórias estatisticamente independentes e identicamente distribuídas tende a uma distribuição gaussiana (17). Assim, $\mathbf{X}$ tem distribuição de probabilidade mais próxima de uma distribuição gaussiana, uma vez que é gerada pela soma dos elementos de $\mathbf{S}$ ponderados pelos elementos de $\mathbf{A}$.

Para estimar as componentes independentes e a matriz de características $\mathbf{A}$ utiliza-se a equação 1. Nessa equação basta multiplicar os dois lados por $\mathbf{W} = \mathbf{A^{-1}}$ para encontrar $\mathbf{Y} = \mathbf{WX}$, sendo $\mathbf{Y}$ a estimativa de $\mathbf{S}$. Como $\mathbf{X}$ é mais gaussiano que $\mathbf{S}$, uma componente independente é estimada quando se encontra um $\mathbf{W}$ que projeta os elementos de $\mathbf{X}$ em uma

distribuição de probabilidade não gaussiana.

Dentre os algoritmos utilizados para estimar a matriz de características $\mathbf{A}$ e as componentes independentes destaca-se o algoritmo fastICA, por ter rápida convergência, e, comparado com algoritmos baseados em gradiente, é mais simples, pois não necessita de ajuste no passo de adaptação (18). O fastICA usa como medida de não gaussianidade uma versão aproximada da negentropia dada pela equação 2

$$J(y) \approx \sum_{i=1}^{N} k_i[E(G_i(y)) - E(G_i(y_{gaus}))]^2. \tag{2}$$

Sendo os k_i constantes positivas, E é o operador esperança, y_{gaus} variáveis gaussianas com variância unitária e média zero e os G_i são funções não quadráticas. Segundo (19), as funções G_1 e G_2, representadas nas equações 3 e 4, garantem boas aproximações da negentropia e melhoram a convergência do algoritmo fastICA.

$$G_1(y) = \frac{1}{\beta} log(cosh(\beta y)), \text{ com } 1 \leq \beta \leq 2 \tag{3}$$

$$G_2(y) = -exp\left(-\frac{y^2}{2}\right) \ . \tag{4}$$

Os passos de execução do fastICA são:

1. inicializa-se aleatoriamente uma estimativa $\mathbf{W}$ para $\mathbf{A}^{-1}$;
2. ajusta-se $\mathbf{W}$
$$\mathbf{W}_{n+1} \longleftarrow E\{\mathbf{X}G_1(\mathbf{WX}) - G_1'(\mathbf{WX})\}\mathbf{W};$$
G_1' é a derivada de G_1.
3. normaliza-se $\mathbf{W}$
$$\mathbf{W}_{n+1} \longleftarrow \frac{\mathbf{W}_{n+1}}{\|\mathbf{W_{n+1}}\|};$$
4. se não convergir repete-se o passo 2.

Implementações do fastICA nas linguagens R, C++, Python e MATLAB podem ser encontradas em (20).

2.3 Seleção de Características mais Discriminativas

Definir o número de características a serem utilizadas em um sistema de reconhecimento de padrões é de suma importância, pois permite melhorar a performance do classificador, diminuir os custos computacionais e reduzir o tempo na etapa de classificação.

A redução de características consiste na escolha de um subconjunto das características mais informativas produzidas a partir dos sinais originais sem que se perca sua capacidade discriminante (21), isto é, o subconjunto selecionado deve ser capaz de descrever o conjunto como um todo.

Nesse trabalho, foi utilizado o algoritmo de Máxima Relevância e Mínima Redundância (mRMR) para reduzir o conjunto de características. O mRMR seleciona do conjunto $\mathbf{A}$ as características mais relevantes e menos redundantes. Para tanto, utiliza a medida de máxima relevância, dada pela informação mútua I entre a variável de classe c e cada característica x_i, como mostra equação 5,

$$max\ D(A,c),\ D = \frac{1}{\mid A \mid} \sum_{x_i \in A} I(x_i;c), \tag{5}$$

e minimiza a medida de redundância, uma vez que é possível que entre as características selecionadas via máxima relevância tenham informações redundantes (21, 22) e estas não acrescentam nenhuma informação nova, por isso, podem ser removidas do conjunto de características sem comprometê-lo. A mínima redundância é dada em termos da informação mútua I por 6

$$min\ R(A),\ R = \frac{1}{A^2} \sum_{x_i,x_j \in A} I(x_i;x_j). \tag{6}$$

Em resumo, o mRMR combina as equações 5 e 6 para encontrar a equação 7 que fornece conjuntamente, após um processo de otimização, as características mais relevantes e menos redundantes. Essa equação foi utilizada por Ding e Peng (22) para implementar o algoritmo de máxima relevância e mínima redundância. Tal algoritmo foi testado com varias bases de dados e em todas mostrou-se ser o mais eficiente (22).

$$max\ \Phi(D,R),\ \Phi(D,R) = D - R \tag{7}$$

2.4 Classificação com a Máquina de vetores de suporte

Como etapa final, foi realizada a classificação das amostras utilizando a Máquina de Vetor de Suporte (SVM), que é uma técnica de aprendizado de máquina baseada na teoria do aprendizado estatístico, criado por Vapnick em 1965 para resolver problemas de regressão e classificação (23).

Essa técnica estabelece princípios que permitem induzir um classificador para separar duas ou mais classes de forma que a distância das margens seja máxima. Isso faz com que a SVM seja robusta diante de dados com grandes dimensões, tenha boa capacidade de generalização e suporte ruídos nos dados (24). Aplicações de SVMs podem ser encontradas em categorização de textos, análise de imagens e bioinformática (25).

Para dados linearmente separáveis, um classificador SVM toma como entrada um conjunto de dados e prediz através de uma função de decisão (hiperplano), induzida a partir de um conjunto de treinamento, a que classe cada dado pertence. Em geral o conjunto usado para o treino é um subconjunto das características escolhidas mediante algum critério de seleção como o mRMR. No treino da máquina apenas os dados localizados às margens das classes são utilizados, tais dados são denominados vetores de suporte.

Nas situações em que os elementos do conjunto de dados não sejam linearmente separáveis, a SVM faz o mapeamento desses dados para um espaço de dimensão maior. Nesse espaço, existe uma alta probabilidade que sejam classificados por um hiperplano (26). As funções que realizam a mudança do espaço de representação dos dados do conjunto a ser classificado são chamadas de kerneis.

A tabela 1 mostra as funções kerneis mais utilizadas e que apresentam bons resultados em processos de classificação. Nesse trabalho foi utilizado o kernel definido pela função de base radial (kernel gaussiano).

Tabela 1. Kernel.

Tipo de função	Forma matemática
Função de base radial	$k(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2}$
Função polinomial	$k(x_i, x_j) = (1 + x_i \cdot x_j)^n$
Função sigmoidal	$k(x_i, x_j) = tanh(ax_i \cdot x_j + b)$

2.5 Métricas de Desempenho

A avaliação da qualidade de testes diagnósticos é feita, em geral, calculando-se as medidas de acurácia, sensibilidade e especificidade. A acurácia é a taxa de acertos do teste. A sensibilidade é a capacidade que o teste diagnóstico apresenta de detectar os indivíduos verdadeiramente positivos, isto é, de diagnosticar corretamente os doentes. A especificidade informa a eficácia do método em diagnosticar corretamente os indivíduos sadios.

Essas medidas dependem da quantidade de indivíduos classificados correta e incorretamente. Os resultados da classificação podem ser divididos em: verdadeiro positivo, falso positivo, verdadeiro negativo ou falso negativo. Um resultado é definido como verdadeiro

positivo ou verdadeiro negativo se a classificação é feita de forma correta e falso positivo ou falso negativo se ela apresenta resultado incorreto.

As equações para calcular a sensibilidade, a especificidade e a acurácia são, respectivamente (27):

$$Acurácia = \frac{V_P + V_N}{V_P + V_N + F_P + F_N} \tag{8}$$

$$Sensibilidade = \frac{V_P}{V_P + F_N} \tag{9}$$

$$Especificidade = \frac{V_N}{V_N + F_P} \tag{10}$$

Sendo: V_P o número de verdadeiros positivos, V_N o número de verdadeiros negativos, F_P o número de falsos positivos e F_N o número de falsos negativos identificados pelo método.

3 Resultados e Discussão

3.1 Base de dados

Para testar a eficiência desse método, utilizou-se uma base de dados com 335 sinais proteômicos, que pode ser encontrada em (28). Esses sinais foram obtidos por meio da técnica SELDI-TOF e estão divididos em 75 casos com diagnóstico positivo (grupo ativo), 101 casos com diagnóstico negativo (grupo controle) e 159 casos com a doença em fase de remissão. Cada vetor dessa base possui dimensão de 380000. Nesse trabalho, foram utilizados o grupo ativo e o grupo controle.

A figura 3 mostra um sinal proteômico dessa base de dados. O eixo horizontal corresponde aos valores de razão *massa/carga* e o eixo vertical equivale a intensidade do sinal. Cada pico observado dá uma ideia da abundancia das moléculas que compõem esse espectro de massa.

3.2 Extração de características

Como primeira etapa, antecedendo a extração de características, foi realizado um preprocessamento sobre o conjunto de sinais da base de dados, com o objetivo de reduzir os ruídos verificados que certamente degradariam o desempenho do classificador SVM. Nesse

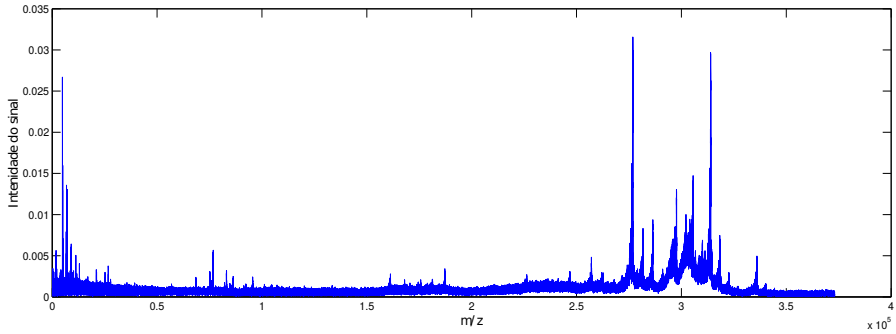

Figura 3. Espectro de massa de um sinal proteômico retirado da base de dados de forma aleatória.

primeiro processo, foi selecionado de cada amostra os pontos no intervalo $[250000; 350000]$, pois verificou-se que a maior parte da informação de todos espectros encontravam-se nesse intervalo.

A figura 4 ilustra os resultados obtidos para dois sinais proteômicos com diagnósticos positivo e negativo, respectivamente, antes e depois desse processo.

O processo de extração de características consistiu em unir os vetores de casos ativos com os vetores de casos negativos, já reduzidos, para gerar a matriz $\mathbf{X}$ de ordem 176×100001 a ser utilizada como entrada no modelo ICA. Cada linha dessa matriz corresponde a um caso e cada coluna a um nível de intensidade do sinal proteômico. Na etapa seguinte, foi utilizado o algoritmo FastICA para extrair as características dos sinais da matriz $\mathbf{X}$. Assim, obteve-se a matriz de características $\mathbf{A}$ de ordem 176×176. As linhas dessa matriz correspondem aos vetores de características dos sinais e permitem identificar cada uma das amostras entre presença ou ausência de Granulomatose de Wegener.

3.3 Redução de dimensionalidade

A redução da dimensionalidade da matriz de características $\mathbf{A}$ foi feita utilizado o algoritmo de Máxima Relevância e Mínima Redundância. Como resultado foi obtido a matriz $\mathbf{A_R}$ com as características organizadas da mais relevante para a menos redundante. Isso significa que as entradas dessa matriz possuem os dados distribuídos em ordem decrescente de representatividade, o que possibilita definir o número de características a serem utilizadas no classificador SVM para obter o seu melhor desempenho.

Para determinar quantas características permitiam um melhor desempenho do classifi-

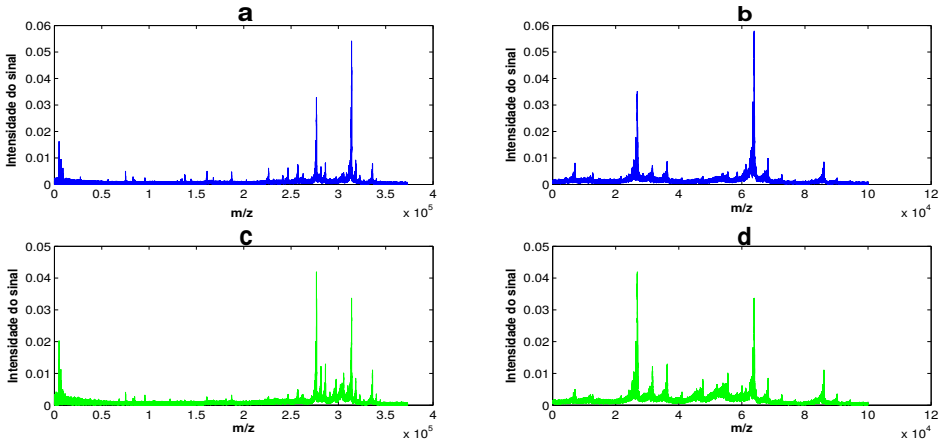

Figura 4. Espectros de massa. A figura (a) corresponde a uma amostra da base de dados com diagnóstico negativo de dimensão 380000 e a figura (b) mostra essa mesma amosta já reduzida para o intervalo $[250000; 350000]$. De forma semelhante, afigura (c) apresenta uma amostra com diagnóstico positivo e a figura (d) equivale a essa amostra com dimensão menor.

cador para cada amostra, foram realizados testes incrementando de cinco em cinco o número de características até um total 175 e cada vetor gerado foi testado com o classificador SVM.

3.4 Classificação das amostras e avaliação do método

Como etapa final, as linhas da matriz $\mathbf{A_R}$, que correspondem aos casos de pacientes portadores e não portadores da GW, foram classificadas por meio da máquina de vetores de suporte, utilizando o kernel dado pela função de base radial representada na tabela 1, com $\gamma = 0,5$.

Por último, foi avaliada a eficácia do método proposto calculando a acurácia, a sensibilidade e a especificidade do classificador com a técnica de validação cruzada *10-fold-cross validation*, que consistiu em dividir a base de dados em dez partes, usar nove para treino e uma para teste. Esse processo foi repetido permutando circularmente as divisões até que todas fossem usadas.

A tabela 2 mostra os melhores resultados obtidos no processo de classificação pela SVM. Estes foram alcançados com vetores de 5, 10, 15 e 20 características. Da observação desses dados é possível ver que o melhor desempenho do classificador e, consequentemente,

do método proposto foi obtido para um vetor com 20 características (linha 4 da tabela 2). Para esse vetor, obteve-se $98,24\%$ de acurácia, $99,73\%$ de sensibilidade e $99,50\%$ de especificidade, com desvios padrão respectivamente de $0,174$, $0,035$ e $0,073$. Isso significa que dos 176 indivíduos portadores e não portadores de GW, 173 foram diagnosticados corretamente (soma dos verdadeiros positivos V_P com os verdadeiros negativos V_N) e 3 de forma incorreta (soma dos falsos positivos F_P com os falsos negativos F_N). Apenas um indivíduo foi diagnosticado como normal (falso negativo) sendo portador de GW.

Tabela 2. Desempenho da SVM para 5, 10, 15 e 20 características. A acurácia, a sensibilidade e a especificidade são apresentadas com seus respectivos desvios padrões.

Carac	VP	FP	VN	FN	Acurácia	Especificidade	Sensibilidade
5	73	3	98	2	(97,22±1,94)%	(97,93±3,24)%	(96,33±2,43)%
10	73	2	99	2	(97,75±2,07)%	(98,28±2,66)%	(98,70±2,30)%
15	73	2	99	2	(97,75 ±1,97)%	(94,85±3,86)%	(99,10±1,62)%
20	**74**	**2**	**99**	**1**	(**98,24** $\pm 1,74$)%	(**99,73** $\pm 0,35$)%	(**99,50** $\pm 0,73$)%

Para implementação da metodologia proposta foi utilizada a linguagem de programação *MatLab*, utilizando os pacotes fastICA e mRMR, disponíveis em (20) e (29), respectivamente, e o pacote SVM, foi adquirido de (8).

4 Considerações Finais

Neste trabalho foi apresentado um método computacional que utiliza Análise de Componentes Independentes, técnica de seleção de atributos Máxima Relevância e Mínima Redundância e Máquina de Vetores de Suporte para diagnosticar precocemente a Granulomatose de Wegener, uma doença rara com complicações multissistêmica que quando não diagnosticada e tratada rapidamente pode levar o paciente a morte. Esse método foi usado para classificar 176 sinais proteômicos de pacientes e os resultados corroboram estudos anteriores quanto à eficiência da técnica ICA para extrair características de sinais proteômicos, a mRMR permite selecionar as melhores características que identificam os portadores de GW, além de reduzir custos computacionais e a SVM implementada com um kernel gaussiano tem um bom desempenho num cenário de classificação não linear.

Para um vetor com apenas vinte características o método proposto obteve $98,24\%$ de acurácia, $99,73\%$ de sensibilidade e $99,50\%$ de especificidade. Das 176 amostras apenas 3 foram classificadas incorretamente, sendo duas falso positivo e uma falso negativo.

Apesar dos bons resultados, para um aumento da confiabilidade do método apresentado novos testes devem ser realizados em diferentes bases de dados.

Diante dos resultados apresentados, espera-se que em um futuro bem próximo o método desenvolvido neste trabalho possa ajudar profissionais da saúde no diagnóstico da Granulomatose de Wegener. Isso possibilitará um aumento da sobrevida do paciente com diagnóstico positivo, uma vez que a completa remissão dessa doença está relacionada com a precocidade do tratamento.

Contribuição dos autores:

Os autores contribuíram de forma equivalente na construção do presente artigo.

Referências

[1] REZENDE, C. E. B. et al. Granulomatose de wegener: relato de caso. *Revista Brasileira de Otorrinolaringologia*, v. 69, n. 2, p. 261–265, 2003. ISSN 1809-4570. Disponível em: <http://www.scielo.br/pdf/rboto/v69n2/15634.pdf>. Acesso em: 2 mar. 2014.

[2] FIGUEIREDO, S. et al. Granulomatose de wegener: Envolvimento otológico, nasal, laringotraqueal e pulmonar. *Revista Portuguesa de Pneumologia*, v. 15, n. 5, p. 929–935, 2009. ISSN 0873-2159. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2173511509701630>. Acesso em: 27 abr. 2014.

[3] SANTOS, S. K. J. dos et al. Granulomatose de wegener: importância do diagnóstico precoce. relato de caso. *Revista Brasileira Clinica Medica*, v. 7, p. 427–433, 2009. ISSN 1679-1010. Disponível em: <http://www.sbcm.org.br/revista/completas.php>. Acesso em: 02 set. 2014.

[4] GOMIDES, A. P. M. et al. Perda auditiva neurossensorial em pacientes com granulomatose de wegener: Relato de três casos e revisão de literatura. *Revista Brasileira de Reumatologia*, v. 46, n. 3, p. 234–236, 2006. ISSN 1809-4570. Disponível em: <http://www.scielo.br/pdf/rbr/v46n3/31356.pdf>. Acesso em: 2 mar. 2014.

[5] RHEUMATOLOGY, A. C. of. *Granulomatosis with Polyangiitis (Wegener's)*. 2014. Disponível em: <http://www.rheumatology.org/I-Am-A/Patient-Caregiver/Diseases-Conditions/Granulomatosis-with-Polyangitis-Wegners>. Acesso em: 2 mar. 2014.

[6] RADU, A. S.; LEVI, M. Anticorpos contra o citoplasma de neutrófilos. *Jornal Brasileiro de Pneumologia*, v. 1, n. 31, p. 16–20, 2009. ISSN 1806-3756. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1806-37132005000700006>. Acesso em: 21 abr. 2014.

[7] STONE, J. H. et al. A serum proteomic approach to gauging the state of remission in wegeners granulomatosis. *American College of Rheumatology*, v. 52, n. 3, p. 902–910, 2005. ISSN 2175-2745. Disponível em: <http://seer.ufrgs.br/index.php/rita/article/view/rita_v14_n2_p43-67/3543>. Acesso em: 21 jun. 2014.

[8] ARAUJO, W. B. D.; CAMPOS, L. F. A.; ALINE, S. F. Método de detecção de câncer de ovário utilizando padrões proteômicos, análise de componentes independentes e máquina de vetores de suporte. In: XIV WORKSHOP DE INFORMÁTICA MÉDICA, 14. *Anais do congresso da sociedade brasileira de computação*. Brasília: CSBC, 2014. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/wim/2014/011.pdf>. Acesso em: 2 dez. 2014.

[9] RIBEIRO, A. C. et al. Diabetes classification using a redundancy reduction preprocessor. *Research on Biomedical Engineering*, v. 31, n. 2, p. 97–106, 2015. ISSN 2446-4740. Disponível em: <http://www.rebejournal.org/files/v31n2/v31n2a02.pdf>. Acesso em: 3 jul. 2015.

[10] YU, J. K.; CHEN, Y. D.; ZHENG, S. An integrated approach to the detection of colorectal cancer utilizing proteomics and bioinformatics. *World journal of gastroenterology: WJG*, Baishideng Publishing Group Inc, v. 10, n. 21, p. 3127–3131, 2004. ISSN 2219-2840.

[11] MANTINI, D. et al. Independent component analysis for the extraction of reliable protein signal profiles from maldi-tof mass spectra. *Bioinformatics*, Oxford Univ Press, v. 24, n. 1, p. 63–70, 2008.

[12] ARAUJO, W. B. D. *Método de detecção de câncer de ovário utilizando análise de componentes independentes, algoritmo de máxima relevâcia e mínima redundância e máquina de vetores de suporte*. Dissertação (Mestrado em Engenharia de Computação e Sistemas) — Universidade Estadual do Maranhão, São Luís, 2014.

[13] GALDOS-RIVEROS, A. C. et al. Proteômica: novas fronteiras na pesquisa clínica. *Enciclopédia Biosfera*, v. 6, n. 11, p. 1–24, 2010.

[14] AFONSO, C. et al. Activated surfaces for laser desorption mass spectrometry: application for peptide and protein analysis. *Current pharmaceutical design*, Bentham Science Publishers, v. 11, n. 20, p. 2559–2576, 2005.

[15] WILSON, K.; WALKER, J. *Principles and techniques of biochemistry and molecular biology*. [S.l.]: Cambridge university press, 2010.

[16] DENNER, R. R. G. *Compressão de Sinais de Eletrocardiograma Utilizando Análise de Componentes Independentes*. Dissertação (Programa de Pós-Graduação em Engenharia de Eletricidade) — Universidade Federal do Maranhão, São Luís, 2006.

[17] PAPOULIS, A. (Ed.). *Probability, Random Variables and Stochastic Processes*. New York, USA: McGraw-Hill, 1991.

[18] LEITE, V. C. M. N. *Separação Cega de Sinais: análise comparativa de algoritmos*. Dissertação (Programa de Pós-Graduação em Engenharia Elétrica) — Universidade Federal de Itajubá, Itajubá, 2004.

[19] HYVARINEN, A.; KARHUNEN, J.; OJA, E. (Ed.). *Independent component analysis*. New York: John Wiley e Sons, 2001.

[20] AAPO. *Independent Component Analysis (ICA) and Blind Source Separation (BSS)*. Disponível em: <http://research.ics.aalto.fi/ica/fastica/>. Acesso em: 2 mar. 2014.

[21] CATARINO, F. M. I. F. *Segmentação da íris em imagens com ruído*. Dissertação (Dissertação de Mestrado) — Universidade da Beira Interior, Covilhã, 2009.

[22] DING, C.; PENG, H. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, Imperial College Press, v. 3, n. 2, p. 185–205, 2005. ISSN 1757-6334. Disponível em: <http://penglab.janelia.org/papersall/docpdf/2004_JBCB_feasel-04-06-15.pdf>.

[23] GUNN, S. *Support Vector Machines for Classification and Regression*. 1998. Disponível em: <http://users.ecs.soton.ac.uk/srg/publications/pdf/SVM.pdf>. Acesso em: 2 set. 2014.

[24] RODRIGUES, T. A. O. et al. Predição de função de proteínas através da extração de características físico-químicas. *Revista de Informática Teórica e Aplicada*, v. 22, n. 1, p. 29–51, 2015. ISSN 2175-2745. Disponível em: <http://seer.ufrgs.br/index.php/rita/article/view/RITA-VOL22-NR1-29/33912>. Acesso em: 2 jul. 2015.

[25] LORENA, A. C.; CARVAHO, A. C. P. L. F. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007. ISSN 2175-2745. Disponível em: <http://seer.ufrgs.br/index.php/rita/article/view/rita_v14_n2_p43-67/3543>. Acesso em: 21 abr. 2014.

[26] HAYKIN, S. (Ed.). *Redes neurais: princípios e prática*. Porto Alegre: Bookman, 2007.

[27] NEVES, S. C. F. *Classificação de câncer de ovário através de padrão proteômico e análise de componentes independentes*. Dissertação (Programa de Pós-Graduação em Engenharia de Eletricidade) — Universidade Federal do Maranhão, São Luís, 2012.

[28] PROGRAM, C. P. *Biomarker Profiling, Discovery and Identification*. 2015. Disponível em: <http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>. Acesso em: 2 mar. 2015.

[29] MATWORKS. *minimum-redundancy maximum-relevance feature selection*. 2015. Disponível em: <http://www.mathworks.com/matlabcentral/fileexchange/14916-minimum-redundancy-maximum-relevance-feature-selection>. Acesso em: 6 mar. 2015.