UNIVERSIDADE ESTADUAL DO MARANHÃO
CENTRO DE CIÊNCIAS TECNOLÓGICAS
MESTRADO PROFISSIONAL EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS

JOSÉ DO NASCIMENTO LINHARES

MÉTODO COMPUTACIONAL PARA AUXILIAR O DIAGNÓSTICO PRECOCE DA GRANULOMATOSE DE WEGENER

SÃO LUÍS
2016

JOSÉ DO NASCIMENTO LINHARES

MÉTODO COMPUTACIONAL PARA AUXILIAR O DIAGNÓSTICO PRECOCE DA GRANULOMATOSE DE WEGENER

Dissertação apresentada ao programa de Pós-Graduação em Engenharia de Computação e Sistemas da Universidade Estadual do Maranhão como parte dos requisitos para a obtenção do título de Mestre em Engenharia da Computação.

Orientador: Prof. Dr. Lúcio Flávio A. Campos

SÃO LUÍS
2016

.

Linhares, José do Nascimento.
Método computacional para auxiliar o diagnóstico precoce da Granulomatose de Wegener/ José do Nascimento Linhares.- São luís, 2016.
85f.
Dissertação(Mestrado) - Curso de Mestrado Profissional em Engenharia de Computação e Sistemas, Universidade Estadual do Maranhão, 2016.
Orientador: Prof^0 Lúcio Flávio Albuquerque Campos.
1. Diagnóstico.2.Granulomatose de Wegener.3.Método computacional.4.Sinais proteômicos.

CDU: 004.932:616.833-002

JOSÉ DO NASCIMENTO LINHARES

MÉTODO COMPUTACIONAL PARA AUXILIAR O DIAGNÓSTICO PRECOCE DA GRANULOMATOSE DE WEGENER

Dissertação apresentada ao programa de Pós-Graduação em Engenharia de Computação e Sistemas da Universidade Estadual do Maranhão como parte dos requisitos para a obtenção do título de Mestre em Engenharia da Computação.

Orientador: Prof. Dr. Lúcio Flávio A. Campos

Aprovado em 22 de julho 2016

BANCA EXAMINADORA

Prof. Dr. Lúcio Flávio de A. Campos (Orientador)

Doutor em Engenharia Elétrica-Universidade Estadual do Maranhão (UEMA)

Prof. Dr. Fernando Jorge Cutrim Demétrio

Doutor em Engenharia de Produção-Universidade Estadual do Maranhão (UEMA)

Prof. Msc. Wesley Batista Dominices de Araujo

Mestre em Engenharia de Computação e Sistemas-Universidade Estadual do Maranhão (UEMA)

SÃO LUÍS
2016

"Dedico este trabalho a minha mãe , Antonia do Nascimento.
Ao meu pai, João Vieira Linhares e a todos meus irmãos:
Antonio, Francisco, Angélica, João e Cacilda Linhares

Agradecimentos

A Deus, pelo dom da vida.

Agradeço aos meus familiares pelo apoio.

Ao professor e orientador Lúcio F. A. Campos pela orientação e paciência.

Ao grande amigo, Jardiel, pelo companheirismo e conversas frutíferas ao longo da pesquisa.

Ao casal Elizângela e Sidiney que estiveram sempre presentes ao longo de toda a jornada do mestrado.

Agradeço imensamente a minha companheira de todas as horas, Camila Correia Soares.

A todos amigos do Mestrado.

Aos professores do Programa de Pós-Graduação em Engenharia de Computação e Sistemas: Fenando Demétrio, Ewaldo Eder Carvalho Santana, Henrique Mariano Costa do Amaral, Ivanildo Silva Abreu, Luís Carlos Costa Fonseca, Rogério Moreira Lima Silva, Carlos Henrique Rodrigues de Oliveira, Reinaldo de Jesus da Silva, Áurea Celeste da Costa Ribeiro, José Bello Salgado Neto e João Coelho Silva Filho.

Aos professores do departamento de física da UEMA: José Clet Brito, Paulo Sérgio, Passinho, Márcio Tavares e Edvam Moreira.

Uma coisa de cada vez.
(Josilberto Mesquita Lindoso)

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 proposto consiste em extrair características de sinais proteômicos e classificá-los 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 $99,50\%$. Estes resultados mostram que o sistema proposto é eficiente para diagnosticar GW e supera a metodologia utilizada atualmente, que é baseada em exames clínicos, sorológicos e radiológicos propostos pelo *American College of Rheumatology*.

Palavras-chave: Diagnóstico, Granulomatose de Wegener, Método computacional, Sinais proteômicos.

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%. These results show that the proposed system is effective to diagnose GW and overcomes the methodology currently used, which is based on clinical, serological and radiological examinations proposed by the American College of Rheumatology.

Keywords: Diagnosis, Wegener's Granulomatosis, Computational method, Proteomic patterns.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ACR	*American College of Rheumatology*
SVM	*Support Vector Machine*
mRMR	Máxima Relevância Mínima Redundância
ICA	*Independent Component Analysis*
f.d.p	Função densidade de probabilidade
SELDI	*Surface-enhanced laser desorption ionization*
TOF	*Time of flight*
PCA	*Principal Component Analysis*
GW	Granulomatose de Wegener
TAE	Teoria do Aprendizado Estatístico
R.B.F	*Radial basis function*
MATLAB	*Matrix Laboratory*

Sumário

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 (REZENDE et al., 2003; PROGRAM, 2015b). 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 (GOMIDES et al., 2006; FIGUEIREDO et al., 2009). A GW é uma patologia que quando não diagnosticada e tratada precocemente, pode levar o paciente a óbito em apenas um ano (SANTOS et al., 2009).

Atualmente a GW é diagnosticada através de sinais, sintomas, exames clínicos, radiológicos e sorológicos que seguem critérios propostos pelo *American College of Rheumatology*-ACR (RHEUMATOLOGY, 2014). Segundo o ACR, 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 (RADU; LEVI, 2009). Vale ressaltar, que os sintomas iniciais da GW são praticamente inespecíficos, o que não permite sua diferenciação em estágios iniciais.

O tratamento de pacientes diagnosticados com GW 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 e tratamento precoce da enfermidade, pois isto influencia na dosagem dos medicamentos. Se o tratamento for iniciado de forma tardia, doses maiores de fármacos são aplicadas ou seu uso é prolongado, o que pode potencializar seus efeitos colaterais, trazendo complicações cardíacas, infertilidade, obesidade, osteoporose, hipertensão arterial, diabetes e infecções oportunistas (STONE et al., 2005). Verifica-se assim, a necessidade do desenvolvimento de métodos de diagnósticos para a GW que sejam precisos e que permitam a sua detecção precoce.

Recentemente a comunidade científica vem aplicando técnicas de CAD (*Computer Aided Diagnosis*) para diagnosticar várias doenças (ARAUJO; CAMPOS; ALINE, 2014; RIBEIRO et al., 2015; YU; CHEN; ZHENG, 2004; MANTINI et al., 2008). Araújo (ARAUJO; CAMPOS; ALINE, 2014), por exemplo, utilizou a Análise de Componentes Independentes (ICA) para extrair características de sinais proteômicos com o objetivo de diagnosticar o câncer de ovário. Áurea (RIBEIRO et al., 2015) propôs um método de

diagnóstico precoce da Diabetes utilizando ICA e Máquina de Vetor de Suporte (SVM). Yu (YU; CHEN; ZHENG, 2004) aplicou sinais proteômicos e bioinformática para detecção do câncer de colo retal. Mantini (MANTINI et al., 2008) usou 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 em extrair caraterísticas de sinais proteômicos para classificá-los como sendo de indivíduos portadores ou não portadores de GW. Para tanto, utiliza-se Análise de Componentes Independentes na extração de características, Algoritmo de Máxima Relevância e Mínima Redundância para selecionar as melhores características e Máquina de Vetores de Suporte na classificação.

1.1 Estrutura da dissertação

A presente dissertação está dividida em 5 capítulos: no capítulo 1 é apresentado o problema e os objetivos do trabalho. No capítulo 2 faz-se uma breve discussão sobre a Granulomatose de Wegener, sinais proteômicos e o processo de obtenção destes sinais: espectrometria de massa com a técnica SELDI-TOF. Também são apresentados a Análise de Componentes Independentes, o algoritmo de Máxima Relevância e Mínima Redundância e a Máquina de Vetores de Suporte, destacando classes linearmente separáveis e não linearmente separáveis. No capítulo 3 são apresentados os materiais e métodos utilizados na presente pesquisa. No capítulo 4 apresentam-se os resultados obtidos. Finaliza-se com o capítulo 5, onde são apresentadas as conclusões obtidas e perspectivas para trabalhos futuros.

2 Fundamentação Teórica

Nesta seção são abordados os fundamentos teóricos necessários para o desenvolvimento da presente pesquisa.

2.1 Vasculites e a Granulomatose de Wegener

O sistema circulatório humano é o conjunto formado pelo coração e vasos sanguíneos que tem a função de fornecer, carregar e manter o fluxo de sangue aos órgãos do corpo conforme a necessidade metabólica para o bom desempenho de suas funções. O sangue transporta oxigênio, substâncias nutritivas e hormônios aos tecidos e retira também elementos nocivos a sobrevivência das células, tais como dióxido de carbono e ureia (CIÊNCIAS, 2009).

Os vasos sanguíneos são tubos por onde o sangue circula e dividem-se em artérias, veias e capilares: (WEBCIENCIA, 2015):

- **Artérias**: São tubos cilíndricos com elasticidade que levam sangue do coração ao corpo e se dividem em artérias de grande, médio e pequeno calibre com espessuras, respectivamente, de $(2,5$ a $7\,mm)$, $(0,5$ a $2,5\,mm)$ e arteríolas com espessura menor que $0,5\,mm$;

- **Veias:** São tubos de paredes finas que conduzem o sangue que passou pelas trocas com os tecidos ao coração, podem ser de grande, médio e pequeno calibre, além de vênulas;

- **Capilares**: São vasos microscópicos que fazem a comunicação entre as artérias e as veias. Os capilares são responsáveis pelas trocas entre o sangue e os tecidos. A sua distribuição é praticamente no corpo todo, sendo rara sua ausência em tecidos ou órgãos (EDUCAÇÃO, 2015).

Vasculites são doenças causadas pela inflamação das paredes dos vasos sanguíneos (veias, artérias e capilares) (DRAUZIOVARELA, 2015). Nestas doenças, o sistema imunológico é ativado e invade o sistema circulatório atacando-o como se este fosse um intruso. O processo inflamatório torna a parede do vaso mais grossa o que pode reduzir sua espessura diminuindo o fluxo de sangue (estenose) ou até mesmo bloqueá-lo (oclusão). Em outras situações, a parede pode tornar-se mais fina com dilatações localizadas (aneurismas) que podem romper causando hemorragia (LIBANES, 2014).

A gravidade da lesão causada pela doença depende do tamanho do vaso e a que órgão ele pertence, se o vaso atingido for, por exemplo, da pele observam-se manchas cutâneas e pequenas zonas de pele necrosada, por outro lado, se o vaso afetado for dos rins haverá comprometimento do correto funcionamento deste órgão, o que exige rápida intervenção terapêutica.

As vasculites são classificadas como primárias ou secundárias (DRAUZIOVARELA, 2015; LIBANES, 2014). As primárias são doenças raras e de causas pouco conhecidas, e ocorrem quando o vaso sanguíneo é o alvo principal da doença. Já as secundárias, acontecem em decorrência de alguma doença de base causada por vírus, doença autoimune[1] como lúpus, artrite reumatoide, esclerodermia, reações alérgicas a medicamentos e alguns tipos de câncer, tais como, leucemia e os linfomas. As principais vasculites são: Doença de Behcet, Doença de Buerger, Vasculite do Sistema Nervoso Central, Síndrome de Churg Strauss, Crioglobulinemia, Arterite de Células Gigantes ou Arterite Temporal, Púrpura de Henoch-Shönlein, Vasculite de Hipersensibilidade, Doença de Kawasaki, Poliangeite Microscópica, Poliarterite Nodosa, Polimialgia Reumática, Arterite de Takayasu e Vasculite Granulomatosa de Wegener. No presente trabalho dar-se-á ênfase à granulomatose de Wegener.

2.1.1 Granulomatose de Wegener

A granulomatose de Wegener (GW) é uma vasculite incomum que se apresenta em estágios iniciais como problemas nasais e sinusais que evoluem levando o indivíduo à morte. Essa doença foi mencionada pela primeira vez no ano de 1931, quando Heinz Karl Ernst Klinger descreveu um caso de morte de um paciente que apresentava sinusite, glomerulonefrite e vasculite disseminada. Em 1936, o médico patologista Friedrich Wegener também mencionou um caso de morte com características semelhantes a relatada por Klinger e cinco anos mais tarde, escreveu sobre 6 mortes provocadas pela doença que seria intitulada de granulomatose de Wegener, em sua homenagem (FIGUEIREDO et al., 2009; DOURADO, 2015).

A GW causa inflamação granulomatosa nos vasos sanguíneos dos tecidos dos órgãos afetados, um tipo de tumoração microscópica em forma de grânulos. Os granulomas destroem os vasos e impedem o transporte de sangue para os órgãos (oclusão), dessa forma, as células dos tecidos ficam sem nutrientes e não realizam as trocas necessárias a sua sobrevivencia.

Estudos mostram que a GW pode afetar o corpo como um todo, no entanto existe prevalência pelo sistema respiratório superior, pulmões e rins (SANTOS; SILVA; LOTUFO, 2008). O acometimento das vias respiratórias superior está presente em até 95% dos

[1] Qualquer doença causada pelo sistema imunológico, quando este ataca e destrói tecidos saudáveis do corpo por "engano".

casos, os sintomas comuns são sinusite, rinorréia purulenta, úlceras mucosas, crostas nasais, epistaxe e obstrução nasal. Devido o aparecimento de ulcerações, existe ainda predisposição do paciente à infecção crônica por Staphylococcus aureus e Pseudomonas aeruginosa. As manifestações pulmonares, ocorrem em cerca de 45% dos casos no início da doença e entre 66% e 85% no seu decorrer. Os sintomas genéricos são tosse e hemoptise, seguidos de dispnéia. Comumente, são encontrados infiltrados pulmonares e nódulos nas radiografias de tórax. As manifestações renais estão presentes entre 70 e 77% dos pacientes diagnosticados com GW, frequentemente o atingimento renal evolui para insuficiência renal dialítica (ANTUNES; BARBAS, 2005). A figura 1 mostra os três orgãos atingidos e os resultados da ação da GW neles.

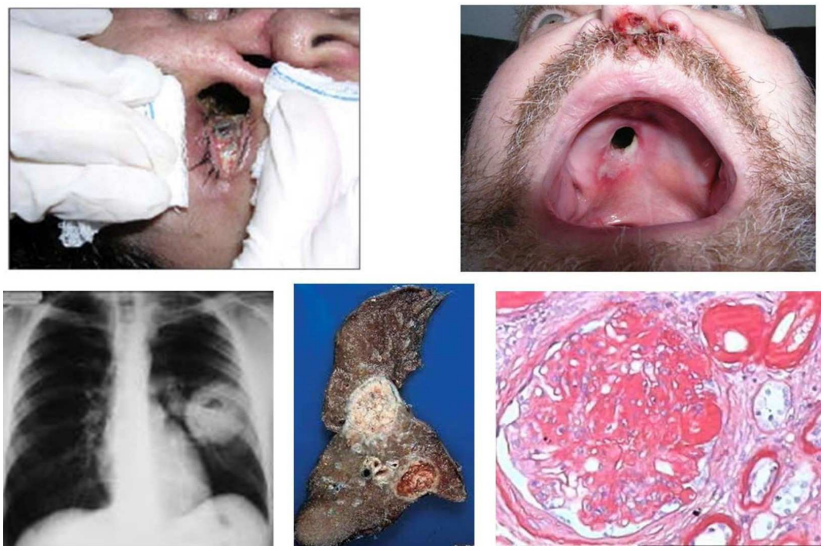

Figura 1 – Órgãos atingidos pela GW. A figura 1(a) e (b) mostram necroses da região do globo ocular, nariz e palato. A figura 1(c) mostra a radiografia do pulmão de um paciente com GW, onde se pode ver uma opacidade. Já a figura 1(d) apresenta um pulmão necrozado por GW e a figura 1(e) traz tecido atingido por GW visto de um microscópio, onde se pode ver os granulomas. Fonte: Modificado de (WEGENER, 2014; PEREIRA et al., 2007; GRANULOMATOSEDEWEGENER, 2016; SHINJO et al., 2011)

A Tabela 1 traz um resumo dos diversos órgãos que podem ser acometidos pela GW e as características de sua manifestação.

Tabela 1 – Principais manifestações clínicas de granulomatose de Wegener.

Órgão	**M**anifestações
Nasal	Rinorreia, epistaxe, obstrução nasal, perfuração do septo nasal, nariz em sela
Seios nasais	Sinusite de repetição
Ouvidos	Hipoacusia
Oral	Úlceras orais
Olhos	Pseudotumor orbitário, esclerite, episclerite, ceratite, uveíte anterior
Traqueia	Estenose subglótica
Pulmões	Nódulos, lesões cavitárias, infiltrado pulmonar não específico, hemorragia alveolar, lesões brônquicas
Coração	Lesão valvular, pericardite
Trato gastrintestinal	Vasculite mesentérica, infarto esplênico
Rins	Glomerulonefrite
Pele	Púrpuras palpáveis, nódulos subcutâneos
Articulações	Oligo ou poliartrite migratória, artralgias
Nervos periféricos	Mononeurites multiplex motora ou sensitiva

2.2 Espectrometria de Massa e Sinais Proteômicos

A ciência tem procurado e desenvolvido formas de diagnosticar doenças precocemente, uma vez que um número significativo de enfermidades podem ser curadas ou terem seus efeitos amenizados se forem descobertas e tratadas em fase inicial (ARAUJO, 2014). Nesse sentido, o estudo do proteoma, que é o conjunto de proteínas e variantes de proteínas encontradas em uma determinada célula, expressas a partir do genoma tem se mostrado promissor. O proteoma, ao contrário do genoma que é relativamente estático, está em constante mudança, estimulada por fatores externos: ambientais, nutricionais e de estresse, ou fatores internos, os quais podem estar ligados a atividades de enfermidades.

A presença de uma doença, por exemplo, pode mudar de forma significativa as características das proteínas de um determinado tecido e consequentemente do proteoma deste. Essa modificação pode ser utilizada para investigar a atividade da patologia que acomete o paciente ou possíveis biomarcadores que possam indicar a sua presença (GALDOS-RIVEROS et al., 2010). Desse modo, a proteômica é um poderoso método de diagnóstico.

Atualmente uma das técnicas mais utilizados para o estudo do proteoma 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, é utilizado um espectrômetro de massa que é composto basicamente por uma fonte de íons, onde uma amostra em estado sólido, líquido ou gasoso é introduzida para ser ionizada, um analisador de massas, um detector de íons e uma unidade de aquisição de dados. A figura

2 apresenta a estrutura de um espectrômetro de massa.

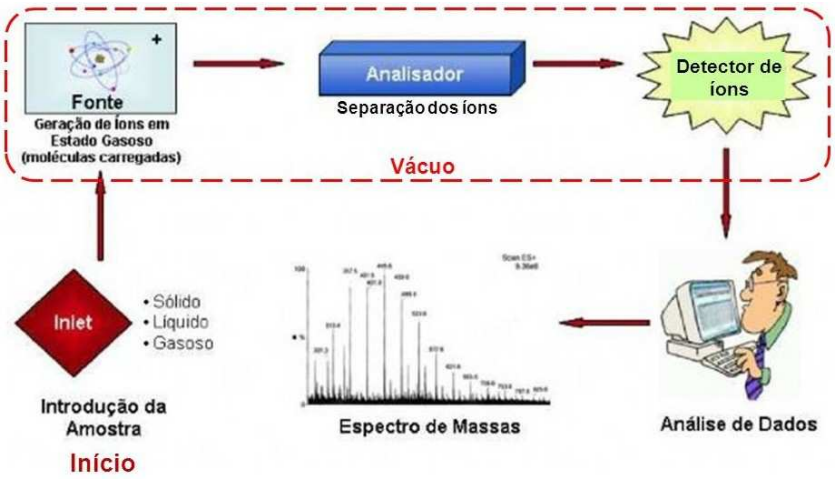

Figura 2 – Espectrômetro de massas. Fonte: (PORTELA, 2014)

Neste trabalho utilizou-se 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* (SELDI) e um analisador de massas do tipo *Time of Flight* (TOF) (AFONSO et al., 2005). Em SELDI, 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 íons em fase gasosa.

No analisador TOF, os íons são acelerados por uma diferença de potencial elétrico em um tubo de vácuo e detectados de acordo com seu tempo de voo (WILSON; WALKER, 2010), que é proporcional a m/z. O resultado ao final de todo o processo é um espectro de massas (sinal proteômico). O espectro obtido é um gráfico que mostra a intensidade relativa de cada íon que aparece como picos com m/z definidos. A figura 3 mostra um espectro de massa obtido com a técnica SELDI-TOF.

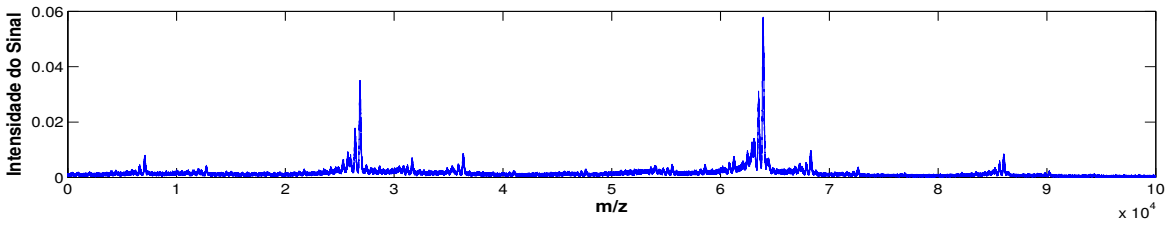

Figura 3 – Espectro de massa. Fonte: (PROGRAM, 2015a)

2.3 Análise de Componentes Independentes

A análise de componentes independentes (ICA-*Independent Component Analysis*) é um modelo estatístico e computacional usado para recuperar fontes estatisticamente independentes ou revelar componentes ocultas de um conjunto de dados, medições ou sinais aleatórios multivariados (GUILHON, 2006; LEITE, 2013). Por isso, é amplamente utilizado no reconhecimento de padrões em eletrocardiograma, análise financeira, telecomunicações, processamento de áudio, imagem e processamento de sinais biomédicos (LEITE, 2004).

No modelo de análise de componentes independentes é considerado que um dado vetor aleatório $\mathbf{X}$ de sinais observados, por exemplo um sinal proteômico, 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. Em outras palavras, um sinal observado $\mathbf{X_i}$ pode ser representado como uma combinação linear dos elementos de $\mathbf{S}$.

Essas suposições permitem escrever cada sinal $\mathbf{X_i}$ como

$$X_i = a_{i1}s_{1i} + a_{i2}s_{12} + ... + a_{ij}s_j = \sum_{i=1}^{n} a_{i1}s_1, \tag{2.1}$$

em que os a_{ij} são coeficientes reais e $i, j = 1, 2, 3, ..., n$ (HYVÄRINEN; KARHUNEN; OJA, 2004; GUILHON, 2006; LEITE, 2013; LEITE, 2004).

Numa notação compacta, a equação 2.1 do modelo ICA pode ser escrita na forma

$$\mathbf{X} = \mathbf{AS}, \tag{2.2}$$

com $\mathbf{X} = (x_{11} \;\; x_{12} \cdots x_{1n})$, $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix}$ e $\mathbf{S} = (s_{11} \;\; s_{12} \cdots s_{1n})$. Onde a matriz de coeficientes $\mathbf{A}$ é conhecida como matriz de mistura ou matriz de características. A equação 2.2 mostra como os elementos do vetor $\mathbf{S}$ são combinados para gerar os sinais $\mathbf{X}$ observados.

A figura 4 ilustra o proposto na equação 2.2 para um sinal proteômico. Nessa figura, as componentes s_1, s_2 e s_3 são ponderadas, respectivamente, por w_{11}, w_{12} e w_{13} para gerar o espectro observado.

O problema da análise de componentes independentes consiste em encontrar uma transformação inversa, que permita obter os elementos de $\mathbf{S}$ a partir dos sinais observados segundo o modelo proposto na equação 2.2. Isso é equivalente a resolver a equação 2.3, na qual $\mathbf{W}$ é uma matriz que "desmistura" $\mathbf{X}$ para obter $\mathbf{S}$. Matematicamente representa-se por (LEITE, 2013):

$$\mathbf{Y} = \mathbf{WX}. \tag{2.3}$$

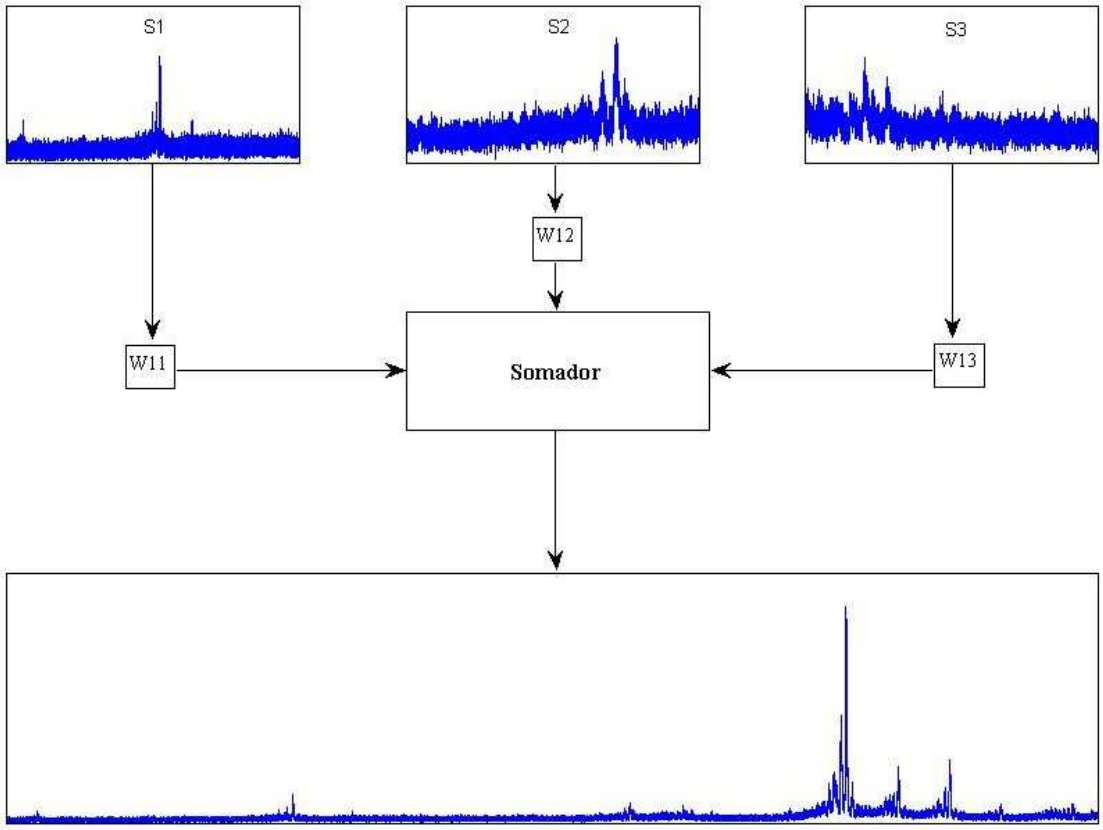

Figura 4 – Espectro de massa gerado pela mistura (combinação linear) das componentes indepentes de **S**

Analisando as equações 2.2 e 2.3 nota-se que se **W** for igual a inversa de **A**, **Y** será igual ao vetor de componentes independentes. Como na prática **A** e **S** não são conhecidos, **W** é apenas uma estimativa dessa matriz, que deve ser encontrada a partir das considerações de independência estatística ou não gaussianidade dos componentes do vetor **S**.

2.3.1 Pressuposições

Para que o modelo ICA fique bem definido é necessário supor que as componentes independentes S_i sejam estatisticamente independentes e não possuam distribuição de probabilidade gaussiana (LEITE, 2013).

2.3.1.1 Independência estatística

O conceito de independência estatística é uma ideia chave em ICA e é o que a diferencia da Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*). Enquanto a PCA considera que um sinal observado é resultado da mistura de componentes descorrelacionadas em primeira ordem, ICA busca um conjunto de componentes estatisticamente independentes para representá-lo, o que significa dados descorrelacionados em qualquer ordem, sendo assim muito mais geral.

Duas ou mais variáveis aleatórias x_1, x_2,...,x_n são estatisticamente independentes se a densidade de probabilidade conjunta $p(x_1, x_2, ..., x_n)$ puder ser escrita como um produto das distribuições marginais $p_1(x_1), p_2(x_2), ..., p_n(x_n)$ tal como representado na equação 2.4. Em outras palavras, o conhecimento do valor de uma das n variáveis não traz informação acerca das demais variáveis.

$$p(x_1, x_2) = p_1(x_1)p_2(x_2) \cdot ... \cdot p_n(x_n) = \prod_{i=1}^{n} p_i(x_i) \tag{2.4}$$

Por outro lado, diz-se que duas variáveis x_i e x_j são descorrelacionadas se a covariância entre elas é igual a zero, ou seja:

$$cov_{x_i,x_j} = \mathbf{E}[x_i x_j] - \mathbf{E}[x_i]\mathbf{E}[x_j] = 0, \tag{2.5}$$

que resulta em

$$\mathbf{E}[x_i x_j] = \mathbf{E}[x_i]\mathbf{E}[x_j], \tag{2.6}$$

Sendo $\mathbf{E}[.]$ o operador esperança ou média.

Uma observação importante deve ser feita com repeito as equações 2.6 e 2.4. Elas mostram que a descorrelação é um caso particular da independência estatística, o que comprova a generalidade da Análise de Componentes Independentes.

2.3.1.2 Variáveis não gaussianas

Outro conceito importante, sendo uma restrição fundamental da ICA, é o de não-gaussianidade das componentes independentes, pois se mais de uma componente possuir distribuição gaussiana estas não poderão ser estimadas pelo modelo ICA.

Considerando que duas variáveis s_1 e s_2 sejam componentes independentes e tenham função densidade de probabilidade (fdp) gaussiana, com média zero e descorrelacionadas, sua fdp é:

$$f(s_1, s_2) = \frac{1}{2\pi} exp\left[\frac{-(s_1^2 + s_2^2)}{2}\right] = \frac{1}{2\pi} exp\left(\frac{-|\mathbf{s}|^2}{2}\right). \tag{2.7}$$

Assumindo ainda que a matriz $\mathbf{A}$ seja ortogonal e usando a fórmula de transformação jacobiana de funções de densidade de probabilidade, tem-se que a densidade conjunta das variáveis x_1 e x_2 obtidas a partir da equação 2.2, é dada por:

$$f(x_1, x_2) = \frac{1}{2\pi} exp\left(\frac{-|\mathbf{A}^T\mathbf{x}|^2}{2}|det\mathbf{A}^T|\right). \tag{2.8}$$

Onde $\mathbf{A}^T$ é a transposta da matriz $\mathbf{A}$ e $det\mathbf{A}^T$ é o determinante de $\mathbf{A}^T$.

Como $\mathbf{A}$ é ortogonal, sua transposta $\mathbf{A}^T$ também é, consequentemente, $|\mathbf{A}^T\mathbf{x}|^2 = |\mathbf{x}|^2$ e $|det\mathbf{A}^T| = 1$. Assim,

$$f(x_1, x_2) = \frac{1}{2\pi} exp\left(\frac{-|\mathbf{x}|^2}{2}\right). \tag{2.9}$$

Essa equação mostra que uma transformação ortogonal não modifica a fdp conjunta de variáveis gaussianas, isto porque esta distribuição é simétrica. A não alteração de uma fdp gaussiana sob rotações denota a impossibilidade de se obter a matriz de mistura $\mathbf{A}$ na equação 2.2 se os elementos de $\mathbf{S}$ forem gaussianos (LEITE, 2013).

2.3.2 Ambiguidades do ICA

Em virtude da definição de ICA em (2.2) observam-se ambiguidades intrínsecas ao modelo (MORETO, 2008; HYVÄRINEN; KARHUNEN; OJA, 2004):

- Não é possível determinar a variância das componentes independentes.

  Isso decorre do fato de $\mathbf{A}$ e $\mathbf{S}$ não serem conhecidas a priore. Se alguma das componentes S_i for multiplicada por um escalar α_i, este poderá sempre ser cancelado, dividindo-se pelo mesmo α_i a correspondente coluna a_i de $\mathbf{A}$:

  $$X = \sum_{i=1}^{n} \frac{1}{\alpha_i} \mathbf{a}_i S_i \alpha_i \tag{2.10}$$

  Segundo (MORETO, 2008), esse problema é facilmente contornado considerando que os componentes independentes tenham variância unitária, isto é, $E(S_i^2) = 1$. Contudo, a ambiguidade de sinal permanece, pois, sempre se pode multiplicar um componente independente por -1 sem afetar o modelo.

- Não é possível determinar a ordem das componentes independentes.

  Aplicando uma matriz de permutação $\mathbf{P}$ e a sua inversa ao modelo, este resulta em $\mathbf{X} = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{S}$ em que os elementos de $\mathbf{PS}$ são componentes independentes em outra ordem, assim como $\mathbf{AP^{-1}}$ é uma outra matriz de mistura a ser estimada.

2.3.3 Estimação das componentes independentes e da matriz de características pela maximização da não gaussianidade

Segundo o teorema do limite central, a soma de variáveis aleatórias identicamente distribuídas e estatisticamente independentes tende a uma distribuição gaussiana (PAPOULIS, 1991). Logo, as componentes de $\mathbf{X}$ na equação 2.2 possuem distribuição de probabilidade que está mais próxima de uma gaussiana do que os elementos de $\mathbf{S}$. Tal propriedade pode ser usada para estimar as componentes de $\mathbf{S}$ e em seguida os elementos

de **A**. Na figura 5 essa importante propriedade pode ser vista. Nela tem-se uma mistura de duas fontes estatisticamente independentes: s_1 e s_2, sendo a primeira uniforme e a segunda laplaciana. O resultado da mistura x_1 e x_2 tem distribuição de probabilidade que se aproxima de uma distribuição gaussiana (SUYAMA, 2007).

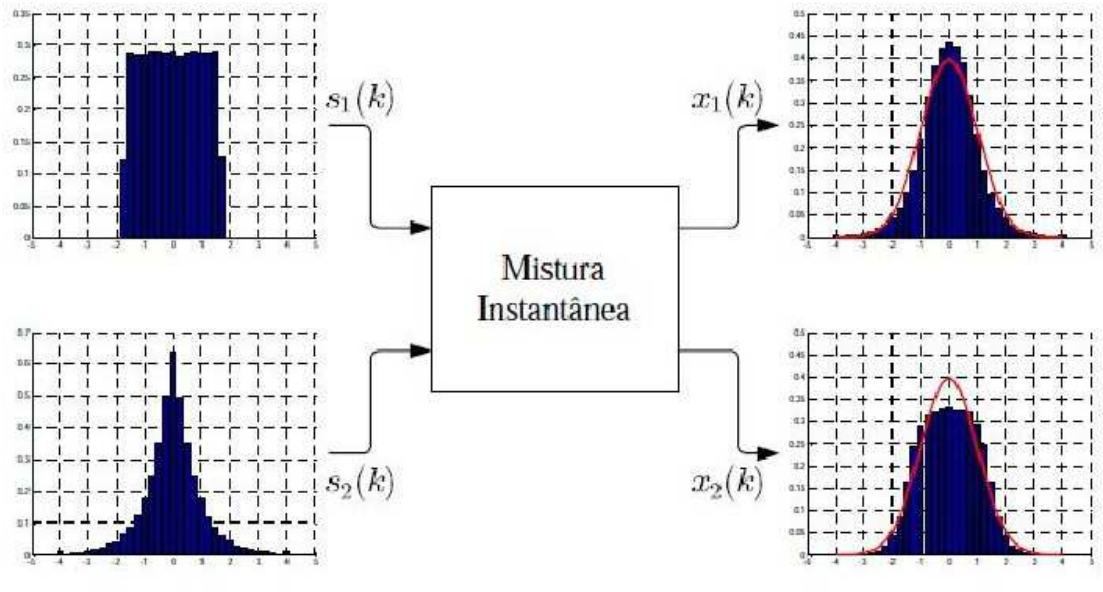

Figura 5 – Soma de variáveis aleatórias com distribuição de probabilidade não-gaussianas . Fonte:(SUYAMA, 2007).

Para estimar um elemento de **S**, considera-se a combinação linear de **X** representada pela equação 2.11

$$y = \mathbf{b}^T\mathbf{X}, \tag{2.11}$$

que pode ser escrita na forma

$$y = \sum_{i}^{n} b_i x_i. \tag{2.12}$$

Sendo b e b^T matrizes de coeficientes.

Usando a equação 2.2 pode-se escrever a equação 2.11 na forma

$$y = \mathbf{b}^T\mathbf{AS}. \tag{2.13}$$

A equação 2.12 mostra que y é uma combinação linear de **S**, com coeficientes dados por

$$\mathbf{q} = \mathbf{b}^T\mathbf{A} \tag{2.14}$$

Da equação 2.14 é possível inferir que se $\mathbf{b}^T$ corresponder a uma das linhas da inversa de **A**, **q** terá um elemento igual a 1 e todos outros iguais a zero, e portanto, y será igual a uma das componentes independentes de **S**. Assim, para encontrar uma componente independente, basta variar $\mathbf{b}^T$ na equação 2.13 e verificar a distribuição de probabilidade de y. Quanto mais distante sua distribuição estiver de uma distribuição gaussiana mais

próximo de uma das componentes de $\mathbf{S}$ estará y. Uma das formas de avaliar o quanto uma distribuição de probabilidade está próxima de uma distribuição gaussiana é medir sua negentropia.

2.3.3.1 Negentropia como medida de não gaussianidade

A negentropia é maior que zero para a maioria das variáveis aleatórias e zero para uma variável gaussiana. Matematicamente é definida através da equação 2.15:

$$J(y) = H(\mathbf{y}_{gaussiana}) - H(\mathbf{y}). \tag{2.15}$$

Sendo $J(y)$ a negentropia, $\mathbf{y}_{gaussiana}$ uma variável aleatória com distribuição gaussiana de mesma média e matriz de covariância que $\mathbf{y}$ e $\mathbf{H}(\mathbf{y})$ é a entropia de $\mathbf{y}$, dada pela equação 2.16.

$$H(\mathbf{y}) = -\sum f(\mathbf{y})log(\mathbf{y}), \tag{2.16}$$

na qual $f(y)$ é a distribuição de probabilidade da variável aleatória discreta y.

Na maioria das aplicações práticas são utilizadas medidas de Kurtose, que envolvem momentos de quarta ordem, ou aproximações por meio de funções não quadráticas no cálculo de $J(y)$ (HYVÄRINEN; KARHUNEN; OJA, 2001). Por Kurtose, a negentropia toma a forma aproximada

$$J(y) \approx \frac{1}{12}\mathbf{E}\{y^3\}^2 + \frac{1}{48}kurt(y)^2, \tag{2.17}$$

sendo $\mathbf{E}$ o operador esperança e $kurt(y)$ a kurtose de y, dada em termos de momentos de quarta ordem por:

$$kurt(y) = \mathbf{E}\{y^4\} - 3(\mathbf{E}\{y^2\})^2. \tag{2.18}$$

Contudo, $J(y)$ escrito na forma 2.17 não é considerado um estimador robusto, uma vez que depende de $\mathbf{E}\{y^4\}$ que é de fácil computação, mas tem o inconveniente de gerar valores elevados já que depende da quarta potência de y. Implementações de ICA que utilizam Kurtose tem o inconveniente de estarem sujeitos a *outliers*.

Utilizando funções G não quadráticas boas aproximações da negentropia são obtidas, entre elas:

$$J(y) = \sum_{i=1}^{N} k_i[\mathbf{E}(G_i(y)) - \mathbf{E}(G_i(y_{gaus}))]^2, \tag{2.19}$$

sendo cada k_i uma constante positiva, y_{gaus} variáveis gaussianas com variância unitária e média zero e os G_i são as funções não quadráticas. As funções G mais utilizadas para obtenção de aproximações da negentropia são (LEITE, 2013; HYVÄRINEN; KARHUNEN; OJA, 2001):

$$G_1(y) = \frac{1}{\beta}log(cosh(\beta y)). \tag{2.20}$$

Com a constante β assumindo valores no intervalo $1 \leq \beta \leq 2$.

$$G_2(y) = exp(-\frac{y}{2}) \text{ e} \tag{2.21}$$

$$G_3(y) = \frac{y^4}{4}. \tag{2.22}$$

Essas funções tem a vantagem de não crescerem muito rapido, o que diminui o erro de aproximação da negentropia. A figura 6 mostra a comparação das três funções com a negentropia, onde é possível observar a afirmação.

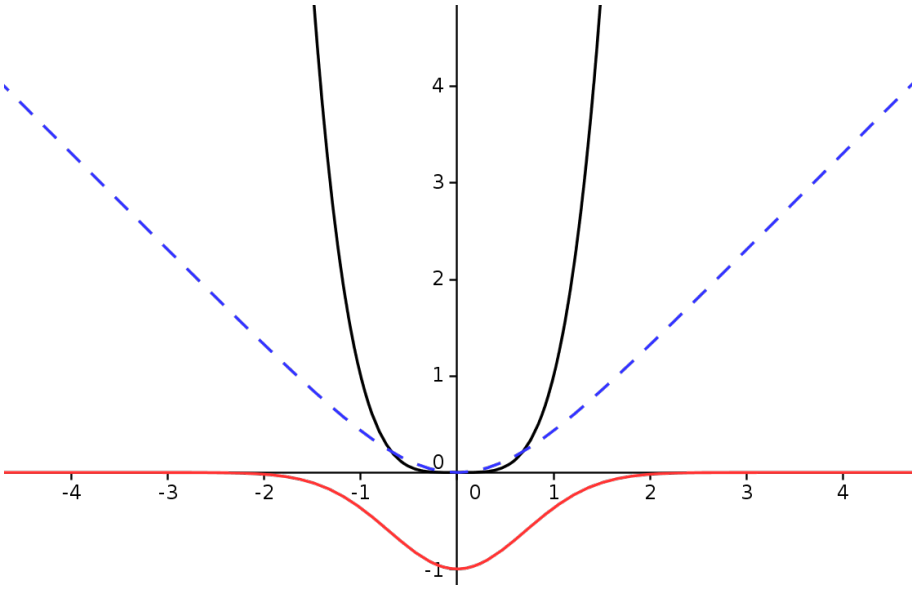

Figura 6 – Os gráficos nas cores azul e vermelho representam respectivamente as funções G_2 e G_3, enquanto o gráfico preto esboça a aproximação da curtose. Fonte:(LEITE, 2013).

2.4 Algoritmo de Ponto Fixo para a Estimação do Modelo ICA Usando Negentropia

A negentropia representada na equação 2.19 fornece uma função objetivo para a construção de um algoritmo de iteração de ponto fixo capaz de estimar o modelo ICA (LEITE, 2013). Como ela é mínima para variáveis gaussianas é de se esperar que sua maximização forneça uma estimativa das componentes independentes, já que estas são não gaussianas. Em outras palavras, uma componente independente é encontrada quando $J(y)$ atinge um valor máximo, sendo

$$J(W) = \sum_{i=1}^{N} k_i[E(G(y)) - E(G(y_{gaus}))]^2, \tag{2.23}$$

onde $\mathbf{W}$ é um vetor que deve ter a média $E[(G(\mathbf{WX})]^2 = ||\mathbf{W}||^2 = 1$.

Os valores ótimos de 2.23 ocorrem nos pontos onde o gradiente da lagrangiana

$$L(W) = E[(G(\mathbf{W} \cdot \mathbf{X})] + \beta \mathbf{W}, \tag{2.24}$$

é nulo, isto é,

$$E[\mathbf{W}(G'(\mathbf{W} \cdot \mathbf{X})] + \beta \mathbf{W} = 0. \tag{2.25}$$

Em que G' é a derivada de G e β é o multiplicador de Lagrange ligado a restrição $||\mathbf{W}||^2 = 1$. Sendo $\mathbf{W}_0$ uma solução ótima, a constante β é dada por $\beta = [\mathbf{W}_0 \mathbf{X} G'(\mathbf{W}_0 \cdot \mathbf{X})]$.

Chamando o lado esquerdo da equação 2.25 de F e derivando, tem-se que:

$$\frac{\partial F}{\partial W} = E[\mathbf{X}\mathbf{X}^T(G''(\mathbf{W} \cdot \mathbf{X})] + \beta \mathbf{I}. \tag{2.26}$$

Fazendo a aproximação $E[\mathbf{X}\mathbf{X}^T(G''(\mathbf{W} \cdot \mathbf{X})] + \beta \mathbf{I} \approx E[\mathbf{X}\mathbf{X}^T]E[(G''(\mathbf{W} \cdot \mathbf{X})] = E[(G(\mathbf{W} \cdot \mathbf{X})]\mathbf{I}$, a matriz F torna-se diagonal e facilmente inversível. Assim, se obtém a seguinte iteração do método de Newton:

$$\mathbf{W} \leftarrow \mathbf{W} - \frac{E[\mathbf{W}(G'(\mathbf{W} \cdot \mathbf{X})] + \beta \mathbf{W}}{E[(G''(\mathbf{W} \cdot \mathbf{X})] + \beta}, \tag{2.27}$$

que pode ser simplificada multiplicando ambos os lados de 2.27 por $E[(G''(\mathbf{W} \cdot \mathbf{X})] + \beta$ resultando em:

$$\mathbf{W} \leftarrow \mathbf{W} - E\{\mathbf{X}(G'(\mathbf{W} \cdot \mathbf{X})] + E[(G''(\mathbf{W} \cdot \mathbf{X})\mathbf{W}]\}. \tag{2.28}$$

O algoritmo descrito acima e resumido a seguir é denominado fastICA, pois é rápido e robusto para estimar as componentes independentes do modelo ICA. Seus passos de execução são (GUILHON, 2006):

Algoritmo: Algoritmo fastICA

1: Centraliza-se os dados subtraindo a média de cada dado;

2: Inicializa-se aleatoriamente uma estimativa para $\mathbf{W}$;

3: Ajusta-se $\mathbf{W}$

$$\mathbf{W}_{n+1} \longleftarrow E\{\mathbf{X}G(\mathbf{W}\mathbf{X}) - G'(\mathbf{W}\mathbf{X})\}\mathbf{W};$$

4: Normaliza-se $\mathbf{W}$

$$\mathbf{W}_{n+1} \longleftarrow \frac{\mathbf{W}_{n+1}}{\|\mathbf{W}_{n+1}\|};$$

5: Se não convergir repete-se a partir do passo 2

Implementações do fastICA nas linguagens R, C++, Python e MATLAB podem ser encontradas em (AAPO, 2015).

2.5 Seleção de Características

Tão importante quanto a extração de características e vital para o sistema de reconhecimento de padrões é a etapa de seleção de características, visto que o desempenho do classificador depende do número de características utilizadas e do quão discriminantes estas são. O uso de muitas características pode trazer problemas como dificultar a tarefa de classificação e superajuste, que é quando o classificador se especializa no conjunto de dados que foi usado no seu treinamento apresentando baixo desempenho na classificação de dados desconhecidos. Para evitar esse tipo de problema reduz-se o número de características utilizadas no processo de classificação (CATARINO, 2009; ARAUJO, 2014).

A seleção de características consiste na escolha de um subconjunto Z das características produzidas a partir dos sinais originais que conduza ao menor erro de classificação possível (CATARINO, 2009). Esse subconjunto, em geral, ajuda compreender melhor o conjunto original de dados, melhora a performance do classificador e permite reduzir custos computacionais, diminuindo o tempo da etapa de classificação (LEE, 2005).

Dado um conjunto de características Y e um rótulo c, um algoritmo de seleção deve encontrar dentro de Y um subconjunto Z menor, com as m características que melhor represente c. A tarefa de determinar o melhor Z pode ser feita de duas formas. A primeira, faz uso de uma função critério de seleção $J(Z)$ que deve indicar o melhor subconjunto Z de modo que a quantidade

$$J = 1 - P_e \tag{2.29}$$

tenha o maior valor possível, sendo p_e o menor erro de classificação. Todavia, o uso de P_e no critério de seleção faz o procedimento de escolha de atributos depender do classificador utilizado e do número de subconjuntos Z possíveis, que é dado pela combinação $\binom{d}{m} = \frac{d!}{m!(d-m)!}$, onde d é a cardinalidade de Y. Este procedimento requer o exame de todos $\binom{d}{m}$ subconjuntos de tamanho m para determinar o melhor. Tal processo, é considerado de alto custo computacional uma vez que o número de subconjuntos de Y cresce combinatorialmente. A outra forma de selecionar características, é ordenar os elementos de Y de acordo com alguma medida de importância e escolher os Q primeiros elementos para representar a classe c (LEE, 2005).

Em 2005 Ding e Peng (DING; PENG, 2005) propuseram um algoritmo de seleção de características que é baseado em medidas de dependência estatísticas e por isso não depende da combinação $\binom{d}{m}$, sendo computacionalmente mais eficiente. O algoritmo proposto organiza primeiramente as características do conjunto Y de acordo com sua medida de relevância e, em seguida, de acordo com a medida de redundância. Tal algoritmo é denominado Máxima Relevância e Mínima Redundância (mRMR).

2.5.1 Seleção de Características por Máxima Relevância e Mínima Redundância - mRMR

O algoritmo de Máxima Relevância e Mínima Redundância merece destaque por apresentar resultados superiores quando comparado com outros métodos de seleção de características existentes (DING; PENG, 2005). Esse algoritmo, como já falado, combina as medidas de relevância e redundância. Em outras palavras, ele maximiza a relevância D e minimiza a redundância R de uma determinada característica.

A máxima relevância seleciona do conjunto de características aquelas que mais se correlacionam com a variável de classe c. Matematicamente a máxima relevância é dada pela equação 2.30

$$max\ D(A,c),\ D = \frac{1}{\mid S \mid} \sum_{x_i \in A} I(x_i;c), \tag{2.30}$$

sendo D a medida de relevância, $I(x_i;c)$ a informação mútua entre a característica x_i e a classe c. Para duas variáveis discretas a e b, I é dado por:

$$I(a,b) = \sum_{i,j} p(a_i,b_j) log \frac{p(a_i,b_j)}{p(a_i)p(b_j)}, \tag{2.31}$$

onde $p(a_i)$ é a densidade de probabilidade de a, $p(b_i)$ é densidade de probabilidade de b e $p(a_i,b_j)$ é a densidade de probabilidade conjunta de a e b.

Como é possível que entre as características selecionadas via máxima relevância tenham informações redundantes (CATARINO, 2009; DING; PENG, 2005). Aplica-se o critério de mínima redundância, que é dado pela equação 2.32

$$min\ R(A),\ R = \frac{1}{A^2} \sum_{x_i,x_j \in A} I(x_i;x_j). \tag{2.32}$$

É importante ressaltar que as informações redundantes podem ser removidas do conjunto de características sem comprometê-lo devido a existência de uma outra característica semelhante.

Combinando as equações 2.30 e 2.32 chega-se a equação 2.33

$$max\ \Phi(D,R),\ \Phi(D,R) = D - R, \tag{2.33}$$

que fornece conjuntamente, após um processo de otimização, as características mais relevantes e menos redundantes. Essa equação foi utilizada por Ding e Peng (DING; PENG, 2005) para implementar o algoritmo de máxima relevância e mínima redundância. Tal algoritmo foi testado com varias bases de dados para selecionar características, em todas mostrou-se ser o mais eficiente (DING; PENG, 2005).

2.6 Máquina de Vetor de Suporte

A máquina de vetor de suporte (SVM-*Support Vector Machine*) é uma técnica de aprendizado de máquina criada por Vapnick em 1965 para resolver problemas de regressão e classificação (GUNN, 1998). Essa técnica estabelece princípios, baseados na Teoria do Aprendizado Estatístico (TAE), que permitem induzir um classificador para separar duas ou mais classes de forma que a distância das margens seja a máxima possível. A margem de uma classe é a distância entre o classificador e os dados mais próximos dele. A maximização da margem garante a SVM robustez na separação de dados com grandes dimensões, resistência a ruídos e boa capacidade de generalização (RODRIGUES et al., 2015). A generalização de um classificador refere-se a eficiência deste em classificar corretamente dados que não foram utilizados para indução.

Aplicações da técnica SVM podem ser encontradas em categorização de textos, análise de imagens e bioinformática (LORENA; CARVAHO, 2007).

Uma ideia simples de um classificador induzido por um algoritmo SVM é mostrado na figura 7. Este classificador está representado pela linha contínua mais forte, separando as classes quadrado e circulo ao mesmo tempo que maximiza a distância $2d$ entre as margens. Os pontos que estão sobre as linhas tracejadas são denominados vetores de suporte. Na técnica SVM apenas estes pontos são utilizados para encontrar o classificador de margem máxima (ótimo).

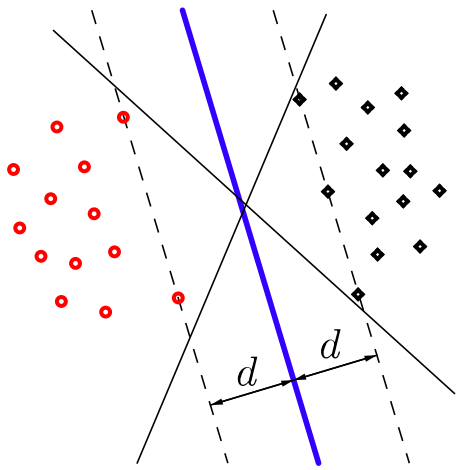

Figura 7 – Máquina de Vetor de Suporte.

Os classificadores induzidos a partir das SVMs são geralmente denominados de margens rígidas ou de margens suaves. Um classificador é de margem rígida se ele não permite que pelo menos um dado seja classificado de forma errada, por outo lado, a margem é dita suave se é permitido um certo limite de erro de classificação. Esses conceitos são importantes pois é a partir deles que se define o modelo matemático para encontrar o melhor classificador para uma determinada situação específica.

A seguir apresenta-se a determinação de classificadores de margens rígidas e de margens suaves a partir da técnica SVM.

2.6.1 Determinação do hiperplano ótimo para um conjunto linearmente separável: SVM de margem rígida

Um conjunto $\mathbf{X}$ é dito linearmente separável se é possível dividi-lo em duas classes com um hiperplano sem que existam erros de classificação. Para separar um conjunto desse tipo, a técnica SVM utiliza uma equação linear que deve ter seus parâmetros determinados a partir de um conjunto de treinamento T. O conjunto T é um subconjunto de $\mathbf{X}$ onde as entradas e as saídas desejadas são conhecidas (LORENA; CARVAHO, 2007; HAYKIN, 2007).

Seja $\mathbf{X} = \{(x_1, y_1), (x_2, y_2)...(x_i, y_i)\}^N$ um conjunto linearmente separável, onde x_i representa um certo padrão, y_i corresponde a classe desejada e T o conjunto de treinamento. A superfície de decisão que separa os elementos de $\mathbf{X}$ pode ser escrita na forma:

$$D(\mathbf{X}) = \langle \mathbf{W}, \mathbf{X} \rangle + b = 0. \tag{2.34}$$

Essa equação divide o espaço de características $\mathbf{X} = \{(x_1, y_1), (x_2, y_2)...(x_i, y_i)\}^N$ em duas regiões que podem ser rotuladas pelo conjunto $Y = \{-1, 1\}$. Sendo assim, uma função sinal $f : R^N \rightarrow \{-1, 1\}$ é capaz de decidir a que classe cada vetor x_i pertence. Por exemplo, um dado vetor x_i pertencerá a classe positiva $+1$, se $D(x_i) > 0$ ou pertencerá a classe negativa se $D(x_i) < 0$. Resumidamente, tem-se:

$$f(\mathbf{D}) = \begin{cases} D(x_i) > 0, & \text{se } y_i = +1 \\ D(x_i) < 0, & \text{se } y_i = -1 \end{cases} \tag{2.35}$$

O hiperplano $\mathbf{D}$ é determinado de modo que a distância entre os pontos mais próximos dele seja a maior possível. Isto é feito usando o subconjunto de treinamento T, cuja cardinalidade é menor do que a de $\mathbf{X}$. Para encontrar a equação que representa $\mathbf{D}$, a figura 8 é tomada como referência, nela observa-se que d é a distância entre os elementos de T mais próximos do hiperplano $\mathbf{D}$ e, por isso, atendem as seguintes restrições

$$\begin{cases} \mathbf{W}\mathbf{x}_i + b \geq +d & \text{se } y_i = +1 \\ \mathbf{W}\mathbf{x}_i + b \leq -d & \text{se } y_i = -1, \end{cases} \tag{2.36}$$

que podem ser escritas na forma 2.37

$$\begin{cases} \mathbf{W_0}\mathbf{x}_i + b_0 \geq +1 & \text{se } y_i = +1 \\ \mathbf{W_0}\mathbf{x}_i + b_0 \leq -1 & \text{se } y_i = -1 \end{cases} \tag{2.37}$$

e simplificada na equação 2.38

$$y_i(\mathbf{W_0}\mathbf{x}_i + b_0) - 1 \geq 0, \forall\,(\mathbf{x}_i, y_i) \in T. \tag{2.38}$$

Com $\mathbf{W}_0 = \frac{1}{d}\mathbf{W}$ e $\mathbf{b}_0 = \frac{1}{d}\mathbf{b}$.

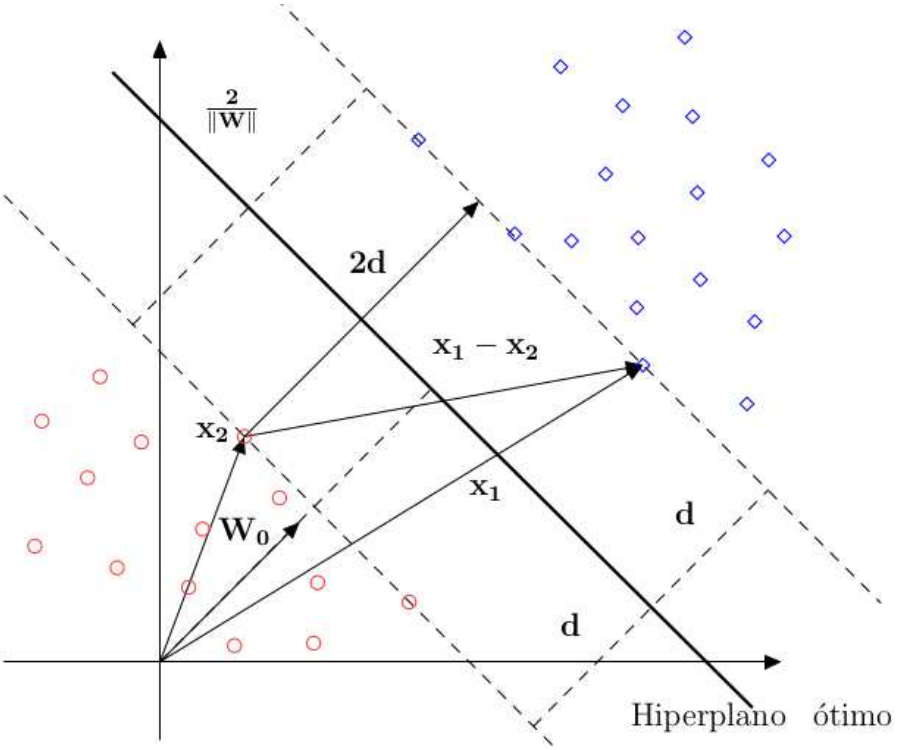

Figura 8 – Separação de duas classes pela máquina de vetor de suporte.

Da figura 8, percebe-se que o ponto $\mathbf{x}_1$ pertence a um hiperplano da forma $\mathbf{H}_1$: $\mathbf{W_0}\mathbf{x}_i + b_0 = +1$ e o ponto $\mathbf{x}_2$ a outro hiperplano $\mathbf{H}_2 : \mathbf{W_0}\mathbf{x}_i + b_0 = -1$. A distância d entre os hiperplanos $\mathbf{H}_1$ e $\mathbf{H}_2$ é dada pela projeção de $(\mathbf{X}_1 - \mathbf{X}_2)$ perpendicular ao hiperplano ótimo. De fato, o vetor $2\mathbf{d}$ é

$$2\mathbf{d} = proj_{\mathbf{W}_0}^{(\mathbf{x_1}-\mathbf{x_2})}. \tag{2.39}$$

A projeção de $(\mathbf{x_1} - \mathbf{x_2})$ sobre $\mathbf{W}_0$ é dada por

$$proj_{\mathbf{W}_0}^{(\mathbf{x_1}-\mathbf{x_2})} = \left[\frac{(\mathbf{x_1} - \mathbf{x_2}) \cdot \mathbf{W}_0}{|\mathbf{W}_0|^2}\right]\mathbf{W}_0 = \frac{(\mathbf{x_1}\mathbf{W}_0 + b - (\mathbf{x_2}\mathbf{W}_0 + b))}{|\mathbf{W}_0|^2}\mathbf{W}_0. \tag{2.40}$$

Utilizando as definições $\mathbf{H}_1$ e $\mathbf{H}_2$ na equação 2.40, encontra-se:

$$proj_{\mathbf{W}_0}^{(\mathbf{x_1}-\mathbf{x_2})} = \frac{2\mathbf{W}_0}{|\mathbf{W}_0|^2}. \tag{2.41}$$

A distância $2d$ entre as margens das classes é dada pela norma do vetor projeção na equação 2.41, isto é

$$2d = \left|proj_{\mathbf{W}_0}^{(\mathbf{x_1}-\mathbf{x_2})}\right| = \left|\frac{2\mathbf{W}_0}{|\mathbf{W}_0|^2}\right|. \tag{2.42}$$

Assim, o tamanho das margens de separação das classes é

$$2d = \frac{2}{|\mathbf{W}_0|}. \tag{2.43}$$

A equação 2.43 mostra que maximizar a margem de separação entre as duas classes consideradas é equivalente a minimizar a norma do vetor $\mathbf{W_0}$, uma vez que a margem $2d$ é inversamente proporcional a $|\mathbf{W}_0|$. Logo, para encontrar o hiperplano ótimo o problema de otimização com restrição definido na equação 2.44 deve ser resolvido

$$\begin{aligned} &\mathbf{W_0}, b_0 \min\ \phi(W_0) = \tfrac{1}{2}|\mathbf{W}_0|^2 \\ &\text{sujeito a}\ \ y_i(\mathbf{W_0}\mathbf{x}_i + b_0) - 1 \geq 0. \end{aligned} \tag{2.44}$$

A restrição é colocada para garantir que não existam dados do conjunto de treinamento entre a margem das classes.

Problemas de otimização com restrição, como o representado em 2.44, podem ser solucionados utilizando-se o método dos multiplicadores de Lagrange, que permite incorporar a função objetivo com a condição de restrição em apenas uma função chamada função de Lagrange, que deve ter o gradiente igual a zero em um mínimo (BERTSEKAS et al., 2003; FLETCHER, 2013). Para o problema em questão, a função de Lagrange é:

$$L(\mathbf{W_0}, b_0, \Lambda) = \frac{1}{2}\|\mathbf{W_0}\|^2 - \sum_{i=1}^{m} \lambda_i [y_i(\mathbf{W_0}\mathbf{x}_i + b_0) - 1], \tag{2.45}$$

sendo que os λ_i são os multiplicadores de Lagrange e Λ é o conjunto desses multiplicadores.

As soluções ótimas $\mathbf{W_0^*}$, b_0^* e λ_i^* devem satisfazer as condições do teorema de Kuhn-Tucker, isto é,

$$\frac{\partial L}{\partial \mathbf{W_0}}|_{W_0 = W_0^*} = 0 \tag{2.46}$$

$$\frac{\partial L}{\partial b_0}|_{b_0 = b_0^*} = 0. \tag{2.47}$$

Realizando essas derivadas parciais, equações 2.46 e 2.47, encontra-se:

$$\mathbf{W_0} = \sum_{i=1}^{m} \lambda_i y_i \mathbf{x}_i = 0 \qquad \lambda_i \geq 0 \tag{2.48}$$

e

$$\sum_{i=1}^{m} \lambda_i y_i = 0 \qquad \lambda_i \geq 0. \tag{2.49}$$

De acordo com o teorema de Kunh-Tucker, apenas os multiplicadores dos $(\mathbf{x}_i, y_i)$ que satisfazem a restrição de desigualdade 2.44, com igualdade dada pela equação 2.50, assumem valores diferentes de zero (FLETCHER, 2013; HAYKIN, 2007).

$$\lambda_i[y_i(\mathbf{W_0} \cdot \mathbf{x}_i + b_0) - 1)] = 0, \qquad i = 1, 2, 3, ...m. \tag{2.50}$$

Os valores de $\mathbf{x}_i$ para os quais $\lambda_i > 0$ são os vetores de suporte (JAKKULA, 2006; HAYKIN, 2007; THEODORIDIS; KOUTROUMBAS, 2010; RUFINO, 2011).

Um problema de otimização restrito com função objetivo convexa e restrição linear pode ser transformado em um outro problema de otimização chamado de problema dual, enquanto o primeiro é referido como primal. O problema dual tem as mesmas soluções do problema primal, porém com os multiplicadores de Lagrange fornecendo a solução ótima. Formula-se o dual do modelo representado na equação 2.44, substituindo as equações 2.49 e 2.48, que fornece:

$$\begin{aligned} &\max L(\Lambda) = \textstyle\sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j (\mathbf{x^T}_i \cdot \mathbf{x}_j) \\ &\text{sujeito a: } \lambda_i \geq 0; \\ &\qquad\qquad \textstyle\sum_{i=1}^{m} \lambda_i y_i = 0. \end{aligned} \tag{2.51}$$

Como pode ser visto na equação 2.51, o problema dual é formulado exclusivamente em termos do conjunto de treinamento e a função objetivo depende somente dos padrões de entrada na forma do produto escalar $\{\mathbf{x^T}_i \cdot \mathbf{x}_j\}$ (HAYKIN, 2007).

Após a determinação dos multiplicadores de Lagrange ótimos usando a formulação dual, encontra-se $\mathbf{W_0}^*$ substituindo os λ_i^* ótimos na equação 2.48. O resultado obtido é:

$$\mathbf{W_0}^* = \sum_{i=1}^{n} \lambda_i^* y_i \mathbf{x}_i. \tag{2.52}$$

O parâmetro b_0^* é determinado utilizando-se o $\mathbf{W}_0{}^*$ encontrado e a condição de que um vetor de suporte positivo $\mathbf{x}_s$ deve satisfazer a condição de igualdade na equação 2.37, de onde se obtém:

$$b^* = 1 - \mathbf{W_0}^* \mathbf{x}_s. \tag{2.53}$$

Um novo padrão x_k é classificado por meio da equação 2.54:

$$classe(x_k) = \begin{cases} +\mathbf{1} & \textbf{se } \mathbf{W_0 x_{ki}} + b_0 > 0 \\ -\mathbf{1} & \textbf{se } \mathbf{W_0 x_{ki}} + b_0 < 0 \end{cases} \tag{2.54}$$

Segue o algoritmo para determinar o hiperplano ótimo.

Algoritmo: Determinação do hiperplano ótimo para conjuntos linearmente separáveis

1: Para cada conjunto de treinamento linearmente separável $\{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$

2: Seja $\lambda^* = (\lambda_1^*, \lambda_2^*, ..., \lambda_n^*)$ a solução do seguinte problema de otimização com restrições:

3: Maximizar :$\sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j (\mathbf{x^T}_i \cdot \mathbf{x}_j)$

4: Sob as restrições: $\sum_{i=1}^{m} \lambda_i y_i = 0$, $\lambda_i \geq 0, \quad i = 1, 2, 3, ..., m.$

5: O par (W^*, b^*) apresentado a seguir define o hiperplano ótimo.

6: $\mathbf{W_0}^* = \sum_{i=1}^{n} \lambda_i^* y_i \mathbf{x}_i$

7: $\mathbf{b_0}^* = 1 - \mathbf{W_0}^* \mathbf{x}_s$

2.6.2 Hiperplano para padrões não linearmente separáveis: SVM de Margem Suave

Funções lineares são bons classificadores para padrões linearmente separáveis. Porém, na maioria das situações práticas os dados que se deseja classificar não são separáveis por funções lineares. Assim, não é possível separá-los com uma SVM de margem rígida. Contudo, é possível induzir um hiperplano que diminui a probabilidade de erros de classificação a partir do conceito de margem rígida.

Um erro de classificação ocorre quando determinado dado $(\mathbf{X}_i, y_i)$ viola a condição de desigualdade (2.35). Tal erro pode acontecer de duas formas (HAYKIN, 2007; RUFINO, 2011): O dado $(\mathbf{X}_i, y_i)$ localiza-se no lado correto, porém dentro da margem de separação, como mostra a figura 9(a), para duas características X_1 e X_2, ou ele se encontra no lado incorreto do hiperplano ótimo, como representado na figura 9(b).

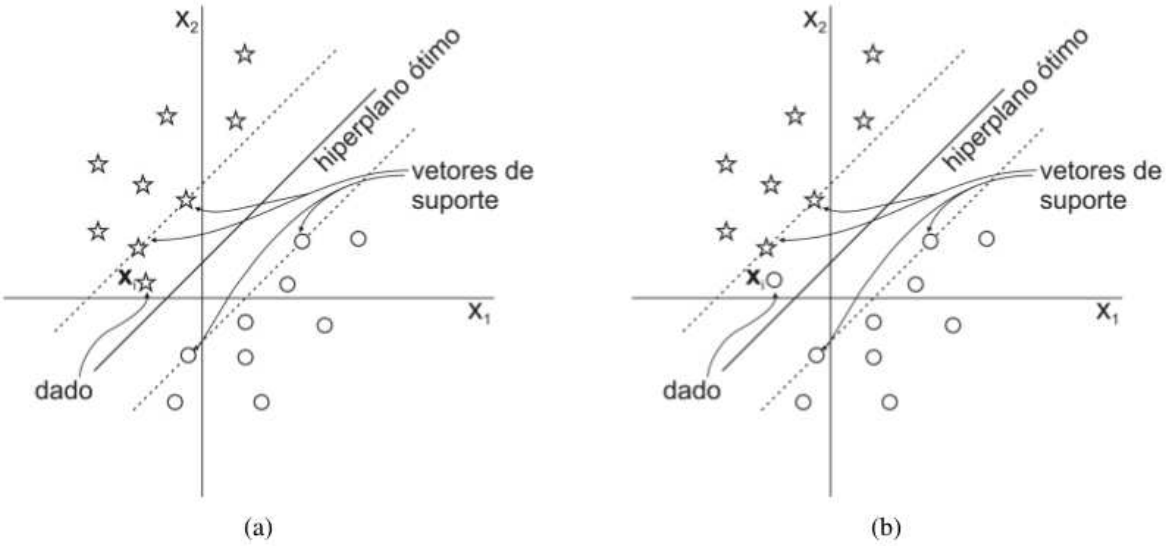

Figura 9 – Erros de classificação. Fonte: (RUFINO, 2011)

Para encontrar a superfície que diminui o erro de classificação de dados não

linearmente separáveis, introduz-se na definição de hiperplano de separação (equação 2.34) um conjunto de variáveis escalares $\xi = (\xi_1, \xi_2, ..., \xi_m)$ não negativas, que medem o desvio dos dados para a borda da margem(HAYKIN, 2007; LIMA, 2002).

Utilizando esse conjunto, o problema de otimização primal passa a ser agora:

$$\begin{aligned} &\min_{\xi, \mathbf{W_0}, b} \ \tfrac{1}{2} \parallel \mathbf{W_0} \parallel^2 + C \textstyle\sum_{i=1}^{m} \xi_i \\ &\text{sujeito a: } \ y_i(\mathbf{W_0} \cdot \mathbf{x_i} + b) \geq 1 - \xi; \\ &\qquad\qquad \xi_i \geq 1, \quad i = 1, 2, 3, ..., m. \end{aligned} \tag{2.55}$$

Sendo C um parâmetro a ser ajustado pelo usuário.

A equação 2.63 é também otimizada pelo método dos multiplicadores de Lagrange da mesma forma que a equação 2.45. Assim, tem-se para a equação 2.63 a lagrangiana L igual a :

$$L(\mathbf{W_0}, b, \Lambda, \xi) = \frac{1}{2} \parallel \mathbf{W_0} \parallel^2 + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \lambda_i [y_i(\mathbf{W_0} \cdot \mathbf{x}_i + b - 1 + \xi_i)] - \sum_{i=1}^{m} \mu_i \xi_i. \tag{2.56}$$

Sendo λ_i e μ_i os multiplicadores de Lagrange. Λ e M é o conjunto desses multiplicadores, respectivamente.

As condições de Karuch-Kuhn-Tucker (KKT) para esse problema são:

$$\frac{\partial L(\mathbf{W_0}, b, \Lambda, \xi)}{\partial \mathbf{W_0}} = \mathbf{W_0} - \sum_{i=1}^{m} \lambda_i y_i \mathbf{x_i} = 0, \tag{2.57}$$

$$\frac{\partial L(\mathbf{W_0}, b, \Lambda, \xi)}{\partial \mathbf{W_0}} = \sum_{i=1}^{m} \lambda_i y_i = 0 \tag{2.58}$$

$$\frac{\partial L(\mathbf{W_0}, b, \Lambda, \xi)}{\partial \xi_i} = C - \lambda_i - \mu_i = 0, \tag{2.59}$$

$$\lambda_i [y_i(\mathbf{W_0} \cdot \mathbf{x_i} + b) - 1 + \xi_i] = 0, \tag{2.60}$$

e

$$y_i(\mathbf{W_0} \cdot \mathbf{x_i} + b) - 1 + \xi_i \geq 0. \tag{2.61}$$

Com

$$\begin{aligned} \xi &\geq 0, \\ \lambda &\geq 0, \\ \mu &\geq 0, \\ \mu_i \xi_i &= 0 \end{aligned} \tag{2.62}$$

A forma dual da equação 2.56 é obtida fazendo a substituição das equações 2.57 e 2.58 na equação 2.56. O resultado após manipulações algébricas é:

$$\begin{aligned} &max \;\; L(\Lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j \mathbf{x}_i \mathbf{x}_j \\ &\text{sujeito a: } 0 \leq \lambda_i \leq C, \\ &\qquad \sum_{i=1}^{m} \lambda_i y_i = 0 \quad i = 1, 2, 3, ..., m. \end{aligned} \tag{2.63}$$

Essa formulação é semelhante a de SVMs de margem rígida, com apenas uma limitação nos valores assumidos pelos multiplicadores de Lagrange, estes devem estar entre 0 e C.

O vetor $\mathbf{W_0}$ ótimo é obtido resolvendo a equação 2.57. Já o b_0^* pode ser calculado usando a expressão 2.60, auxiliada pela equação 2.64 que foi encontrada substituindo a equação 2.59 na equação 2.62.

$$(C - \lambda_i^*)\xi_i^* = 0. \tag{2.64}$$

O algoritmo para classificação por SVM de margem suave é:

Algoritmo: Determinação do hiperplano ótimo para conjuntos linearmente separáveis

1: Para cada conjunto de treinamento $S = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$

2: Seja $\lambda^* = (\lambda_1^*, \lambda_2^*, ..., \lambda_n^*)$ a solução do seguinte problema de otimização com restrições:

3: Maximizar :$\sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j (\mathbf{x^T}_i \cdot \mathbf{x}_j)$

4: Sob as restrições: $0 \leq \lambda_i \leq C$, $\sum_{i=1}^{m} \lambda_i y_i = 0 \quad i = 1, 2, 3, ..., m.$

5: O par (W^*, b^*) apresentado a seguir define o hiperplano ótimo.

6: $\mathbf{W_0}^* = \sum_{i=1}^{n} \lambda_i^* y_i \mathbf{x}_i$

7: $\mathbf{b_0}^* = 1 - \mathbf{W_0}^* \mathbf{x}_s$

2.6.3 Mapeamento em alta dimensão e produto interno kernel

Uma outra forma de classificar dados não linearmente separáveis com a SVM e que apresenta resultados melhores que os descritos anteriormente, é fazer o mapeamento dos dados para um espaço de dimensão maior denominado espaço de características e classificá-los neste novo espaço. Segundo (HAYKIN, 2007; GUERRA, 2006), existe alta probabilidade que dados não linearmente separáveis numa dimensão X^n sejam linearmente separáveis na dimensão X^{n+m}, com $n, m = 1, 2, 3, ...$

O mapeamento é realizado por uma função $\Theta : \mathbf{X}^n \rightarrow \Im$, onde $\mathbf{X}$ é o espaço de entrada e $\Im$ é o espaço de características. A escolha adequada de Θ permite que o conjunto $\Im$ seja separado por um hiperplano.

Na figura 10, X^n é representado por pontos no espaço $\Re^2$, com superfície de separação dada por uma função não linear. Após o mapeamento para o espaço de características no $\Re^3$ estes mesmos dados são separados por uma SVM linear.

Segundo (LORENA; CARVAHO, 2007), a função de mapeamento Θ deve satisfazer duas condições: a primeira é que ela leve os dados de entrada a um espaço de dimensão suficientemente alta, a segunda é que Θ não seja linear.

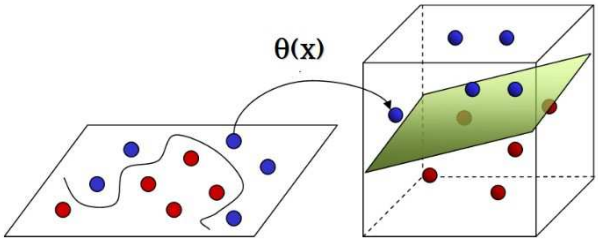

Figura 10 – Mapeamento para o espaço de características. Fonte: (LIAO, 2014).

Usando a mudança de espaço de representação na equação 2.63, o problema a ser resolvido agora é:

$$L(\Lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j [\theta(\mathbf{x_i}) \cdot \theta(\mathbf{x_j})] \tag{2.65}$$

No qual os dados do conjunto de entrada aparecem na forma de um produto escalar $\Theta(\mathbf{x}_i) \cdot \Theta(\mathbf{x}_j)$ denominado Kernel.

Um Kernel K é uma função que recebe dois pontos x_i e x_j do espaço de entrada e calcula o produto escalar no espaço de característica. Assim,

$$K(\mathbf{x_i}, \mathbf{x_j}) = \theta(\mathbf{x_i}) \cdot \theta(\mathbf{x_j}). \tag{2.66}$$

Os Kernels mais utilizados na prática são os Gaussianos ou R.B.F, os Polinomiais e os Sigmodais. A Tabela 2 mostra a forma matemática destes Kernels.

Tabela 2 – Kernel.

Função	**T**ipo de função
Função de base radial	$k(x_i, x_j) = e^{-\gamma \lvert x_i - x_j \rvert^2}$
Polinomial	$k(x_i, x_j) = (1 + x_i \cdot x_j)^n$
Sigmoid	$k(x_i, x_j) = tanh(ax_i \cdot x_j + b)$

Após a seleção do Kernel, a classe de um novo padrão x_k é definida por 2.67:

$$classe(x_k) = \begin{cases} +1 & se \ \ \mathbf{W}_0 \, \Theta(x_k) + b_0 > 0 \\ -1 & se \ \ \mathbf{W}_0 \, \Theta(x_k) + b_0 < 0 \end{cases} \tag{2.67}$$

O algoritmo para SVM não linear utilizando Kernel é:

Algoritmo: Determinação do hiperplano ótimo para conjuntos não linearmente separáveis

1: Para cada conjunto de treinamento $\Theta(S) = (\Theta(x_1), y_1), (\Theta(x_2), y_2), ..., (\Theta(x_m), y_m)$

2: Seja $\lambda^* = (\lambda_1^*, \lambda_2^*, ..., \lambda_n^*)$ a solução do seguinte problema de otimização com restrições:

3: Maximizar :$\sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j (\Theta(\mathbf{x}_i^T) \cdot \Theta(\mathbf{x}_j))$

4: Sob as restrições: $0 \leq \lambda_i \leq C,$
$\sum_{i=1}^{m} \lambda_i y_i = 0 \quad i = 1, 2, 3, ..., m.$

5: O par (W^*, b^*) apresentado a seguir define o hiperplano ótimo.

6: $\mathbf{W_0}^* = \sum_{i=1}^{n} \lambda_i^* y_i \Theta(\mathbf{x}_i)$

7: $\mathbf{b_0}^* = 1 - \mathbf{W_0}^* \Theta(\mathbf{x}_s)$

3 Material e Métodos

O método proposto consiste em extrair características dos sinais proteômicos utilizando a análise de componentes independentes (ICA), selecionar as melhores características com a técnica de máxima relevância e mínima redundância (mRMR) e classificá-las com a máquina de vetores de suporte (SVM). O resultado da classificação é a identificação dos sinais de portadores ou não portadores da Granulomatose de Wegener. A figura 11 mostra um diagrama do método proposto.

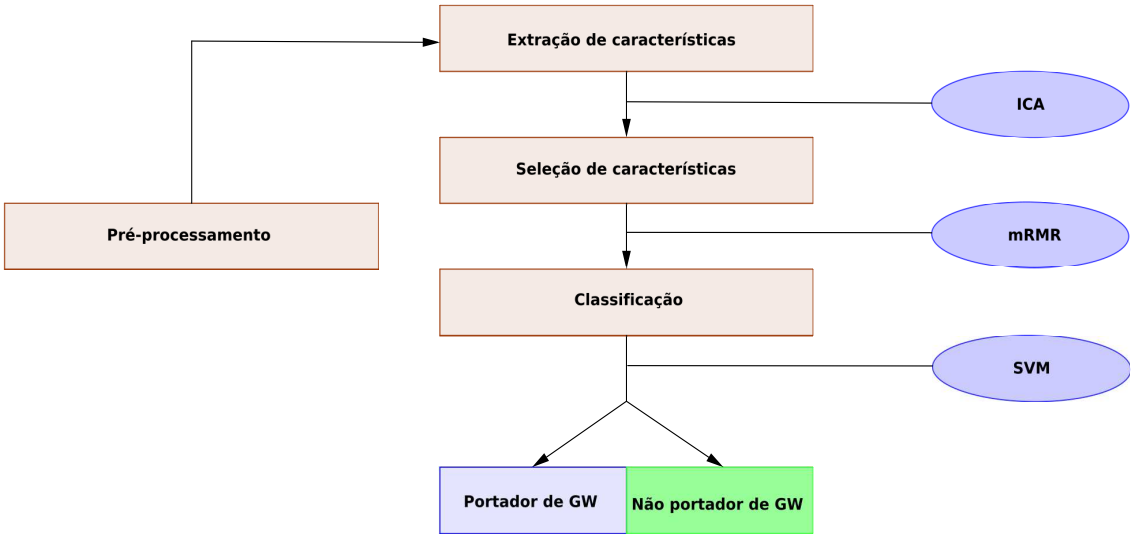

Figura 11 – Diagrama da metodologia proposta.

3.1 Base de Dados

Para testar a eficiência desse método, utilizou-se uma base de dados com 335 sinais proteômicos de alta resolução, que pode ser encontrada em (CLINICALPROTEOMICS-PROGRAM, 2015). 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 pontos. Foram utilizados o grupo ativo e o grupo controle que somam juntos 176 sinais.

A figura 12 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 abundância das moléculas que compõem esse espectro de massa.

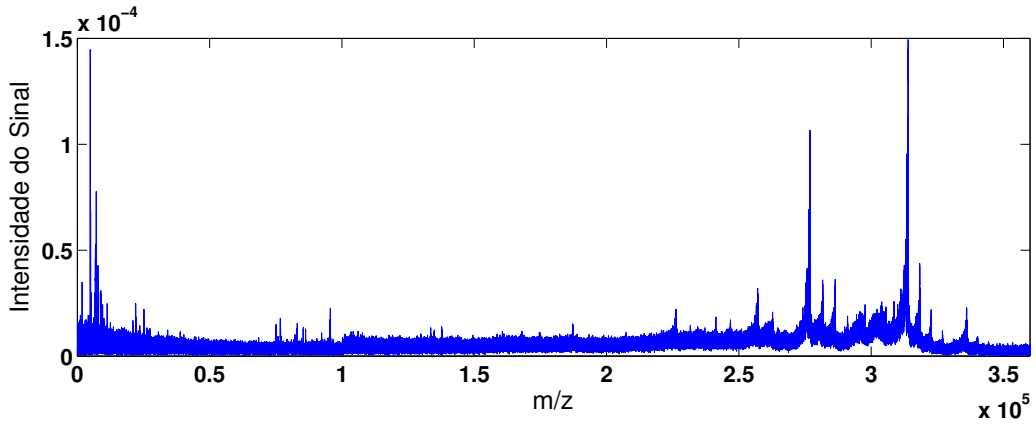

Figura 12 – Espectro de massa de um sinal proteômico retirado da base de dados de forma aleatória.

3.2 Pré-processamento

Como primeira etapa, antecedendo a extração de características, foi realizado o preprocessamento sobre o conjunto de sinais da base de dados, cujo o objetivo foi reduzir os ruídos verificados para evitar a degradação do desempenho do classificador SVM. Nesse 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 13 ilustra os resultados obtidos para dois sinais proteômicos com diagnósticos positivo e negativo, respectivamente, antes e depois do preprocessamento.

3.3 Aplicação de ICA na Extração de Características

A ideia principal que justifica a aplicação da análise de componentes independentes na extração de características dos sinais proteômicos, é considerar que todo sinal seja resultado da mistura de um conjunto de componentes independentes (sinais base) comuns aos sinais dos portadores e não portadores da GW. E cada componente contribui com alguma informação do sinal proteômico. Sendo assim, um sinal da base de dados pode ser representado pela combinação linear dessas componentes independentes. E a qualquer sinal corresponde-se um vetor cujos elementos são as coordenadas dos componentes independentes da mistura, como na figura 14. Essas coordenadas indicam a contribuição das componentes independentes no sinal, sendo, portanto, as características do sinal obtidas a partir do modelo ICA.

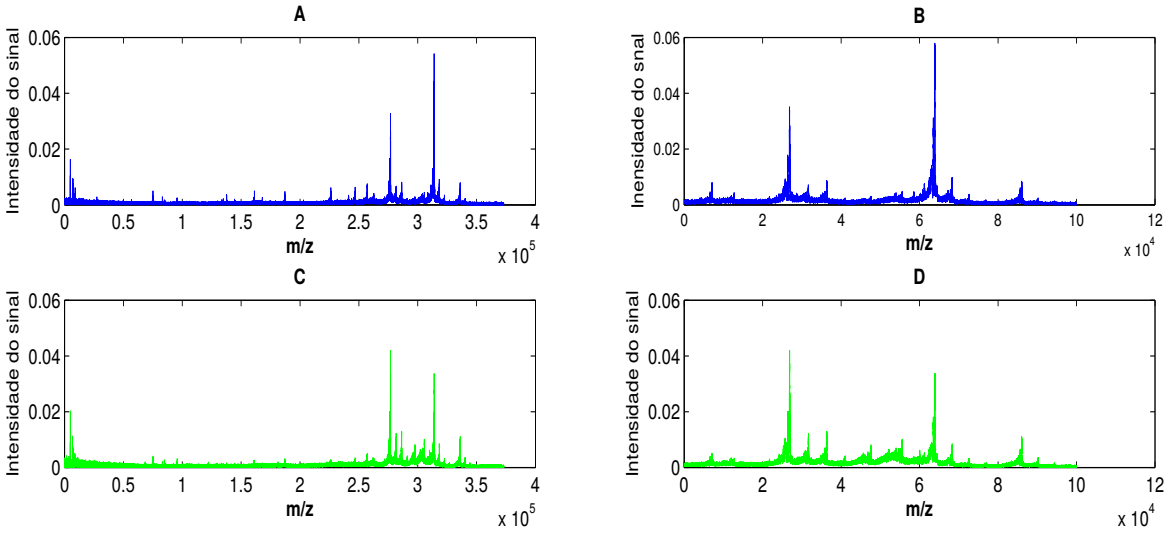

Figura 13 – Corte realizado nos sinais. A figura (A) corresponde a uma amostra completa (380000 pontos) da base de dados com diagnóstico negativo e a figura (B) mostra essa mesma amostra já reduzida para 100001 pontos. De forma semelhante, afigura (C) apresenta uma amostra com diagnóstico positivo e a figura (D) equivale a essa amostra com dimensão menor.

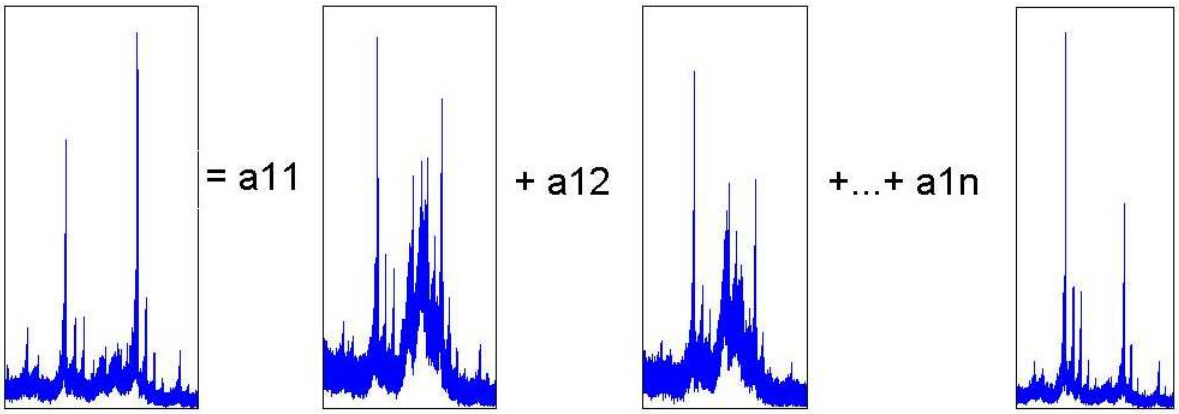

Figura 14 – Sinal proteômico representado como combinação linear de componentes independentes.

Para a extração das características dos sinais proteômicos foi utilizado o algoritmo fastICA implementado na linguagem *MATLAB* , onde se utilizou a biblioteca fastICA disponível em (COMPUTER; CENTRE, 2015). A entrada do algoritmo, nesse caso uma matriz **X** com os sinais proteômicos, é a única informação necessária na estimação do modelo. O fastICA oferece como resultado uma estimativa **W** da inversa de **A** e a matriz de componentes independentes **S**, como apresentado na equação 2.2.

A matriz **X** de entrada do modelo foi produzida unindo os sinais proteômicos dos casos ativos com os casos negativos pré-processados, com o intuito de reduzir ruídos. Os resultados obtidos são apresentados no capítulo seguinte.

3.4 Redução de Redundância com mRMR

Após a extração de características, foi utilizado o algoritmo de máxima relevância e mínima redundância com o objetivo identificar as melhores características para representar as classes portador e não portador de GW. Nessa etapa, cada vetor de característica da matriz **A** foi rotulado. Para tanto, adicionou-se uma coluna na matriz **A** com os valores numéricos $+1$ ou -1 , sendo 1 para os casos com diagnóstico positivo e -1 para os casos com diagnóstico negativo. Em seguida, essa nova matriz teve seus termos organizados pelo algoritmo mRMR que tem implementação baseada na equação 2.33.

3.5 Classificação pela Máquina de Vetores de Suporte

Ao realizar a extração de características, cada sinal passou a ser representado pelo vetor de pesos da combinação linear de componentes independentes. Assim, inicialmente todo sinal da base possuía 172 características, que foram organizadas posteriormente pelo algoritmo mRMR na ordem crescente de representatividade, do mais relevante para o menos redundante.

Para classificar as características por meio dos vetores de suporte, definiu-se um conjunto de treinamento, e por meio deste, o algoritmo SVM determinou o hiperplano capaz de identificar a classe de cada sinal proteômico da base de dados, cujas informações de entrada são as características extraídas por ICA e selecionadas por sua maior relevância e menor redundância em relação as duas classes (portador e não portador de GW) rotuladas por $+1$ e -1 respectivamente.

A partir deste conjunto de características foi criada uma função para variar de 5 em 5 a dimensão do vetor de características e testar cada conjunto formado no algoritmo SVM. Os resultados obtidos estão representados na tabela 3.

3.6 Medidas de Desempenho

Após a classificação com a SVM, foram realizadas medidas para certificar o método. 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 pela SVM. 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 utilizadas para calcular a sensibilidade, a especificidade e a acurácia foram, respectivamente (NEVES, 2012):

$$Acurácia = \frac{V_P + V_N}{V_P + V_N + F_P + F_N} \tag{3.1}$$

$$Sensibilidade = \frac{V_P}{V_P + F_N} \tag{3.2}$$

$$Especificidade = \frac{V_N}{V_N + F_P} \tag{3.3}$$

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.

Os melhores resultados obtidos podem ser encontrados na tabela 5 da seção seguinte.

4 Resultados e Discussão

Neste capítulo apresentam-se os resultados obtidos em cada uma das etapas da metodologia e faz-se as devidas discussões.

4.1 Extração de características

Na extração de características uniu-se os vetores de casos ativos com os vetores de casos negativos, já reduzidos, onde se obteve a matriz $\mathbf{X}$ de ordem 176×100001 que foi utilizada como entrada no modelo ICA. Cada linha da matriz obtidada representa um caso e cada coluna um nível de intensidade do sinal proteômico. Essa matriz foi utilizada no algoritmo fastICA para extrair as características dos sinais. De onde se obteve a matriz de características $\mathbf{A}$ de ordem 176×176.

As linhas da matriz $\mathbf{A}$ encontrada correspondem aos vetores de características dos sinais e permitem identificar cada uma das amostras entre presença ou ausência de Granulomatose de Wegener. Porém sua dimensão ainda é grande o que dificulta a determinação pelo algoritmo SVM de hiperplanos separadores. Além disso, entre o conjunto de características existem os dados pouco informativos e outros fortemente correlacionados representando praticamente a mesma característica, o que exigiu a redução de dimensionalidade selecionando as melhores características.

A tabela 3 apresenta parte da matriz $\mathbf{A}$ obtida do algoritmo fastICA. As cinco primeiras linhas dessa tabela correspondem a indivíduos com GW e as cinco últimas a não portadores.

Tabela 3 – Parte da matriz de características extraídas da base de dados pelo algoritmo FastICA.

Amostra	Características				
1	-0,16089878	-0,2084758	-1,07290435	0,57725916	0,15082361
2	-0,01673767	0,01269074	-0,99582612	0,11164136	0,07477289
3	-0,14102193	-0,75670779	-0,70301264	0,14420626	0,22699389
4	0,03671135	-0,49017676	-0,73873287	0,34567245	0,22532337
5	-0,01814825	0,18039134	-1,60282408	0,34921277	-0,03139734
6	-0,00619900	-0,02612083	-1,02161801	0,19369463	-0,10396357
7	-0,08221927	2,50648139	-2,76437211	0,41088437	-0,13837731
9	0,05849091	1,72781213	-2,03071892	0,38125612	-0,11457852
10	-0,04312896	6,07717516	-0,48123176	0,12472679	-0,05462446

No gráfico da figura 16 podem ser vistas as três primeiras características de cada

amosta da tabela 3, o que confirma a hipótese anterior, que existem dados correlacionados. Existem dois dados, um na cor azul e um na cor vermelha que estão praticamente sobrepostos, representando uma mesma característica, pode-se afirmar que elas são redundantes e sua retirada do conjunto de características não afeta o desempenho do sistema proposto. Da figura, é possível observar também que não dá para diferenciar entre os casos ativos e os casos negativos, pois as características que definem cada classe encontram-se misturadas. Isso acontece porque essas características não são adequadas para representar o conjunto **A** como um todo, sendo necessário assim, a etapa de seleção das características mais discriminantes com o algorítimo de Máxima Relevância e Mínima Redundância (mRMR).

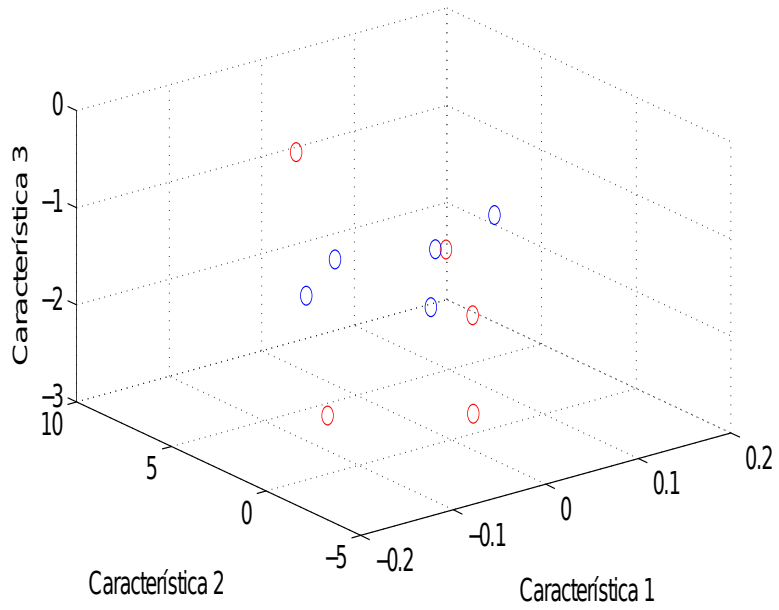

Figura 15 – Representação dos dados da tabela 3.

4.2 Seleção das Melhores Características

A identificação das melhores características para representar as duas classes 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 linhas da matriz $\mathbf{A_R}$ possuem os dados distribuídos em ordem decrescente de representatividade, o que possibilitou definir o número de características a serem utilizadas no classificador SVM para obter o seu melhor desempenho.

Na tabela 4 pode-se ver as cinco primeiras características de **A** organizadas segundo o critério mRMR, para dez amostras tomadas de forma randômica.

Tabela 4 – Características organizadas pelo algoritmo mRMR.

Amostra	Características				
1	0,64425629	-0,01362244	-0,08819685	-0,99958132	-0,19150331
2	2,02623471	0,19758404	-4,98695821	-0,79581367	0,21432692
3	1,21099043	0,210635806	-2,41868816	-1,06067648	0,03945212
4	0,03671135	-0,49017676	-0,73873287	0,34567245	0,22532312
5	-0,01814821	0,18039131	-1,60282401	0,34921276	-0,03139730
6	0,02799306	-0,05244586	-2,73747830	-0,22667632	-1,19399583
7	0,37128719	0,07693638	-4,33207933	-2,10614338	-2,32046923
8	1,63913034	-0,03844772	-4,43506604	-1,44283109	-0,72825764
9	0,52361205	0,02192241	-4,29302726	-1,74231196	-0,83667898
10	-0,67369629	-0,13357772	-2,66646357	-0,46705009	-1,56137035

A figura 16 traz as três primeiras características das 10 amostras da tabela $\mathbf{A_R}$. A partir dessa figura, nota-se que os dados de cada classe estão mais agrupados, sendo possível identificá-los facilmente. É possível determinar com mais facilidade um hiperplano para dividir as duas classes: portadores de GW e não portadores de GW, já que estas estão naturalmente divididas e bem próximas de um conjunto linearmente separável. No entanto, ainda existe um ponto (cor azul) que se encontra misturado aos vetores da outra classe (vermelho). Se uma função linear fosse traçada entre as duas classes, esse ponto seria um erro de classificação, significando que um doente foi identificado como saudável ou vice-versa. Para evitar erros dessa natureza utilizou-se os Kernels discutidos na seção 2.5.3.

4.3 Resultado da Classificação das Amostras e Avaliação do Método

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 (R.B.F) representado na tabela 2, com $\gamma = 0,5$.

A tabela 5 mostra as médias dos melhores resultados obtidos no processo de classificação pela SVM. Estes foram alcançados com vetores 5, 10, 15, 20 e 25 características, repetindo-se a execução do algoritmo trinta vezes para cada conjunto de características. Da observação dos dados da tabela 5 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 5). Para esse vetor, obteve-se 98, 24% de acurácia, 99, 73% de sensibilidade

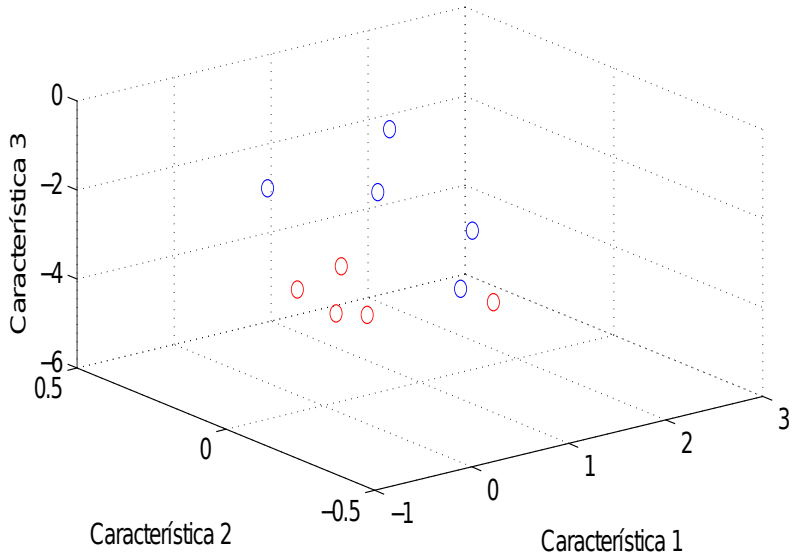

Figura 16 – Representação dos dados da tabela 4.

e 99, 50% de especificidade. 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 5 – Desempenho da SVM para 5, 10, 15, 20 e 25 características. A acurácia, a sensibilidade e a especificidade são apresentadas com seus respectivos desvios padrões.

Característica	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** ±1, 74)%	(**99,73** ±0, 35)%	(**99,50** ±0, 73)%
25	73	3	98	2	(96,22±2,92)%	(96,90±3,33)%	(96,33±2,40)%

5 Considerações Finais e Sugestões

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 permitiu selecionar as melhores características que identificam os portadores de GW, além de permitir a redução de custos computacionais. E a SVM implementada com um kernel gaussiano teve 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. Em comparação com a metodologia de diagnóstico clínico utilizado atualmente, proposta pelo *American College of Rheumatology*-ACR, para a GW, o método proposto foi capaz de superá-lo em cerca de 5%.

Apesar dos bons resultados obtidos a partir do sistema de reconhecimento de padrões proteômicos desenvolvido neste trabalho, 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 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.

5.1 Sugestões para Trabalhos Futuros

Como sugestão de trabalhos a serem desenvolvidos, propomos que a base de dados com sinais de GW seja utilizado de forma completa. É possível que dentre os sinais relacionados a fase de remissão da doença tenham informações que possam ajudar profissionais da área médica no acompanhamento de pacientes diagnosticados com GW durante a fase de remissão. Tal pesquisa seria importante do ponto de vista médico, porque junto ao tratamento da GW vem os malefícios das medicações, é essa fase que torna

o seu uso prolongado e como já foi falado, aplicação de imunossupressores durante um longo período pode causar hipertensão arterial, diabetes, doenças cardíacas, infertilidade e osteoporose.

Uma observação importante deve ser feita em relação a quantidade de características utilizadas no processo de classificação. Foi verificado que o sistema apresentado teve boa performance quando se utilizou apenas duas características no classificador SVM. Isso significa que estas duas informações são bem específicas da GW, pois dentro de um conjunto com 176 amostras as correspondentes a GW foram identificadas. Assim, essas duas características podem estar ligas a presença de biomarcadores específicos para a GW e precisam ser investigadas.

Referências

AAPO. *Independent Component Analysis (ICA) and Blind Source Separation (BSS)*. 2015. Disponível em: <http://research.ics.aalto.fi/ica/fastica/>. Acesso em: 2 mar. 2014. Citado na página 29.

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. Citado na página 21.

ANTUNES, T.; BARBAS, C. S. V. Granulomatose de wegener. *J Bras Pneumol*, SciELO Brasil, v. 31, n. 1, p. 21–6, 2005. Citado na página 19.

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. Citado 2 vezes nas páginas 15 e 59.

ARAUJO, W. B. D. 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. Citado 2 vezes nas páginas 20 e 30.

BERTSEKAS, D. P. et al. Convex analysis and optimization. Athena Scientific, 2003. Citado na página 35.

CATARINO, F. M. I. F. *Segmentação da íris em imagens com ruído*. Dissertação (Dissertação de Mestrado) — Universidade da Beira Interior, Covilhã, 2009. Citado 2 vezes nas páginas 30 e 31.

CIÊNCIAS, F. D. T. E. *Fundamentos de física e biofísica*. [S.l.]: Alagoinhas, 2009. Citado na página 17.

CLINICALPROTEOMICSPROGRAM. *Biomarker Profiling, Discovery and Identification*. 2015. Disponível em: <http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>. Acesso em: 2 mar. 2014. Citado na página 42.

COMPUTER, L. of; CENTRE, I. S. A. I. R. *fastICA*. 2015. Disponível em: <http://research.ics.aalto.fi/ica/fastica/code/dlcode.shtml>. Acesso em: 2 mar. 2014. Citado 2 vezes nas páginas 44 e 59.

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>. Citado 2 vezes nas páginas 30 e 31.

DOURADO, L. B. K. *Ativação endotelial na granulomatose com poliangeíte (granulomatose de Wegener)*. Tese (Faculdade de medicina da Universidade de São Paulo) — Universidade de São Paulo, 2015. Citado na página 18.

DRAUZIOVARELA. *Doenças e sintomas vasculite*. 2015. Disponível em: <http://drauziovarella.com.br/letras/v/vasculite/>. Acesso em: 2 jan. 2015. Citado 2 vezes nas páginas 17 e 18.

EDUCAÇÃO. *Sistema circulatório*. 2015. Disponível em: <http:$//$www.portaleducacao.com.br/medicina/artigos/38470/tipos-de-vasos-sanguineos-arterias-veias-e-capilares-sanguineos>. Acesso em: 2 mar. 2015. Citado na página 17.

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. Citado 2 vezes nas páginas 15 e 18.

FLETCHER, R. *Practical methods of optimization*. [S.l.]: John Wiley & Sons, 2013. Citado 2 vezes nas páginas 35 e 36.

GALDOS-RIVEROS, A. C. et al. Proteômica: novas fronteiras na pesquisa clínica. *Enciclopédia Biosfera*, v. 6, n. 11, p. 1–24, 2010. Citado na página 20.

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. Citado na página 15.

GRANULOMATOSEDEWEGENER. *Reumatismo e coisa séria*. 2016. Disponível em: <https://jmarcosrs.wordpress.com/2011/11/27/5257/>. Acesso em: 2 jan. 2015. Citado 2 vezes nas páginas 10 e 19.

GUERRA, F. A. *Análise de métodos de agrupamento para o treinamento de redes neurais de base radial aplicadas à identificação de sistemas*. Tese (Doutorado) — Pontifícia Universidade Católica do Paraná, 2006. Citado na página 39.

GUILHON, R. R. D. *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. Citado 2 vezes nas páginas 22 e 29.

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. Citado na página 32.

HAYKIN, S. (Ed.). *Redes neurais: princípios e prática*. Porto Alegre: Bookman, 2007. Citado 5 vezes nas páginas 33, 36, 37, 38 e 39.

HYVÄRINEN, A.; KARHUNEN, J.; OJA, E. *Independent component analysis*. [S.l.]: Wiley & Sons, 2001. Citado na página 27.

HYVÄRINEN, A.; KARHUNEN, J.; OJA, E. *Independent component analysis*. [S.l.]: John Wiley & Sons, 2004. Citado 2 vezes nas páginas 22 e 25.

JAKKULA, V. Tutorial on support vector machine (svm). *School of EECS, Washington State University*, 2006. Citado na página 36.

LEE, H. D. *Seleção de atributos importantes para a extração de conhecimento de bases de dados*. Tese (Doutorado em Ciências de Computação e Matemática Computacional) — Universidade de São Paulo, São Carlos, 2005. Citado na página 30.

LEITE, I. C. C. *Análise de componentes independentes aplicada a avaliação de imagem radiográfica de sementes*. Tese (Doutorado), 2013. Citado 6 vezes nas páginas 10, 22, 23, 25, 27 e 28.

LEITE, V. C. M. N. *Separaçao Cega de Sinais: Análise Comparativa entre Algoritmos*. Dissertação (Programa de Pós-Graduação em Engenharia de Eletrica) — UNIVERSIDADE FEDERAL DE ITAJUBÁ, São Luís, 2004. Citado na página 22.

LIAO, R. *SVM*. 2014. Disponível em: <https://www.google.com.br/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&ved=0ahUKEwi4rbu6uMbOAhVCF5AKHfTIAEoQjRwIBQ&url=http%3A%2F%2Fwww.cs.toronto.edu%2F~urtasun%2Fcourses%2FCSC411%2Ftutorial9.pdf&psig=AFQjCNEkKmmemoDBdTOv6Jb4wJNqtOmOCQ&ust=1471454262017237&cad=rjt>. Acesso em: 2 jan. 2015. Citado 2 vezes nas páginas 10 e 40.

LIBANES, S. *Medicina Avançada*. 2014. Disponível em: <hhttp://www.hospitalsiriolibanes.org.br/hospital/especialidades/re-umatologia/Paginas-/vasculite.aspx>. Acesso em: 2 jan. 2015. Citado 2 vezes nas páginas 17 e 18.

LIMA, A. R. G. *Máquinas de vetores suporte na classificaçao de impressões digitais*. Dissertação (Programa de Pós-Graduação em Engenharia de Computação) — Universidade Federal do Ceará, 2002. Citado na página 38.

LINHARES, J. do N. et al. Método computacional para o diagnóstico precoce da granulomatose de wegener. *Revista de Informática Teórica e Aplicada*, v. 23, n. 1, p. 277–292. Citado na página 68.

LORENA, A. C.; CARVAHO, A. C. P. L. F. de. 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>. Citado 3 vezes nas páginas 32, 33 e 40.

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. Citado 2 vezes nas páginas 15 e 16.

MATLABCENTAL. *minimum-redundancy maximum-relevance feature selection*. 2015. Disponível em: <http://www.mathworks.com/matlabcentral/fileexchange/14916-minimum-redundancy-maximum-relevance-feature-selection>. Acesso em: 2 mar. 2015. Citado na página 59.

MATWORKS. *Support Vector Machines (SVM)*. 2015. Disponível em: <http://www.mathworks.com/help/stats/support-vector-machines-svm.html?refresh=true>. Acesso em: 6 mar. 2015. Citado na página 59.

MORETO, F. A. d. L. *Análise de Componentes Independentes Aplicada à Separação de Sinais de Áudio*. Tese (Doutorado) — Universidade de São Paulo, 2008. Citado na página 25.

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. Citado na página 46.

PAPOULIS, A. (Ed.). *Probability, Random Variables and Stochastic Processes*. New York, USA: McGraw-Hill, 1991. Citado na página 25.

PEREIRA, I. C. et al. Granulomatose de wegener: relatos de casos. *Arq Bras Oftalmol*, v. 70, n. 6, p. 1010–5, 2007. Citado 2 vezes nas páginas 10 e 19.

PORTELA, L. *Análise estrutural de compostos orgânicos*. 2014. Disponível em: <https://www.google.com.br/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&ved=0ahUKEwi4rbu6uMbOAhVCF5AKHfTIAEoQjRwIBQ&url=http%3A%2F%2Fwww.cs.toronto.edu%2F~urtasun%2Fcourses%2FCSC411%2Ftutorial9.pdf&psig=AFQjCNEkKmmemoDBdTOv6Jb4wJNqtOmOCQ&ust=1471454262017237&cad=rjt>. Acesso em: 2 jan. 2015. Citado 2 vezes nas páginas 10 e 21.

PROGRAM, C. P. *Analise estrutural de compostos organicos*. 2015. Disponível em: <https://www.google.com.br/search?q=Espectrometro+de+massas&client>. Acesso em: 2 mar. 2015. Citado 2 vezes nas páginas 10 e 21.

PROGRAM, C. P. *Biomarker Profiling, Discovery and Identification*. 2015. Disponível em: <http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>. Acesso em: 2 mar. 2014. Citado na página 15.

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. Citado na página 15.

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. Citado na página 15.

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. Citado na página 15.

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. Citado na página 15.

RODRIGUES, T. A. d. 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_v14_n2_p43-67/3543>. Acesso em: 2 jul. 2015. Citado na página 32.

RUFINO, H. L. P. *Algoritmo de aprendizado supervisionado-baseado em máquinas de vetores de suporte-uma contribuição para o reconhecimento de dados desbalanceados*. Tese (Programa de Pós-Graduação em Engenharia Elétrica) — UNIVERSIDADE FEDERAL DE UBERLÂNDIA, Uberlândia, 2011. Citado 3 vezes nas páginas 10, 36 e 37.

SANTOS, I. de S.; SILVA, L. B. d. B. e; LOTUFO, P. A. *Clínica médica: diagnóstico e tratamento*. [S.l.]: Sarvier, 2008. Citado na página 18.

SANTOS, S. K. J. et al. Granulomatose de wegener: importância do diagnóstico precoce. relato de caso. *Revista Brasileira de clínica médica*, v. 69, n. 2, p. 427–433, 2009. ISSN 1809-4570. Disponível em: <http://www.scielo.br/pdf/rboto/v69n2/15634.pdf>. Acesso em: 2 mar. 2014. Citado na página 15.

SHINJO, S. K. et al. *Vasculites*. 2011. Disponível em: <https://jmarcosrs.wordpress.com/2011/11/27/5257/>. Acesso em: 2 jan. 2015. Citado 2 vezes nas páginas 10 e 19.

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>. Citado na página 15.

SUYAMA, R. *Proposta de métodos de separação cega de fontes para misturas convolutivas e não-lineares*. Tese (Doutorado em Engenharia Elétrica) — Universidade Estadual de Campinas, Campinas, 2007. Citado 2 vezes nas páginas 10 e 26.

THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition*. [S.l.]: Academic Press, 2010. Citado na página 36.

WEBCIENCIA. *Sistema circulatório*. 2015. Disponível em: <\http:www.webciencia.com/11-21circula.htm>. Acesso em: 2 mar. 2014. Citado na página 17.

WEGENER, G. D. *Patologia de Órgãos e Sistemas*. 2014. Disponível em: <http://patologiadeorgaosesistemas.blogspot.com.br/2010/10/granulomatose-de-wegener.html>. Acesso em: 2 jan. 2015. Citado 2 vezes nas páginas 10 e 19.

WILSON, K.; WALKER, J. *Principles and techniques of biochemistry and molecular biology*. [S.l.]: Cambridge university press, 2010. Citado na página 21.

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. Citado 2 vezes nas páginas 15 e 16.

Apêndices

APÊNDICE A – Algoritmos Utilizados

Nesta seção são apresentados os algoritmos usados para simular o sistema de reconhecimento de padrões proposto. Estes algoritmos foram implementados por (ARAUJO; CAMPOS; ALINE, 2014), na linguagem de programação *MATLAB*, com as bibliotecas *fastICA*, mRMR e SVM, disponíveis, respectivamente, em (COMPUTER; CENTRE, 2015), (MATLABCENTAL, 2015) e (MATWORKS, 2015).

A.1 Algoritmo Implementado

```
clear;
clc;
close all;
ds1_criado = 0;
gerou_X1 = 0;
gerou_fastica1 = 0;
gerou_mrmr1 = 0;
op = 0;
med = 0;
quantDS = 1;
while op < 3
op = menu('Escolha uma opção','Executar tudo','SAIR');
switch_expr = op;
switch switch_expr
case 1, % EXECUTAR TODAS AS ETAPAS
ds1_criado = 0;
gerou_X1 = 0;
gerou_fastica1 = 0;
gerou_mrmr1 = 0;
ds = switch_expr;
%chama a função/script para gerar a matriz com os dados de gw
%salva as informações no diretório database com o nome gw
[tamDsCa] = gera_db_gw(ds);
%chama a função/script para gerar a matriz com os dados de controle
%salva as informações no diretório database com o nome control
gera_db_control(ds);
ds1_criado = 1;
```

```
%inicia o tempo de execução do sistema
tstart = tic;
%chama a função/script para gerar a matriz consolidada com os dados gerais
[X,tamX] = carrega_dbs1(ds); %X = [gw;control];63
fprintf('salvando a matriz X........\n\n');
save('matrizes/X1','X');
gerou_X1 = 1;
fprintf('dataset control gerado com sucesso! \n');
fprintf('dataset cancer gerado com sucesso! \n');
fprintf('datasets carregados e gerada a matriz X!\n');
%%------------------ FastICA ---------------------%%
fprintf('Chama o FastICA\n\n');
[Y,W,A] = aapomagro(X,tamX);
save('matrizes/A1','A');
fprintf('FastICA concluído com sucesso\n');
gerou_fastica1 = 1;
%%------------------ mRMR ---------------------%%
fprintf('chamando o mRMR\n');
% GERA A MATRIZ DE RÓTULOS
f=ones(tamX,1);
for i=tamDsCa+1:tamX,f(i,1)=-1;end
save('matrizes/f1','f');
for k=5:5:175 % numero de características
[fea] = mrmr_mid_d(A,f,k);
gerou_mrmr1 = 1;
[fs] = feature_matrix(A,fea);
save('matrizes/fs1','fs');
%%------------------ SVM ---------------------%%
fprintf('Fazendo a Classificação SVM\n');
[acu,esp,sen,VP,FP,VN,FN] = fold_sens_espec_cross_validation(fs,f);
% tempo total gasto para executar todo o programa
tempototal = toc(tstart);
fprintf('Tempo Total gasto para o algoritmo: %0.2f \n \n',tempototal);
med = med + 1;
res1(med,1) = [k];
res1(med,2) = [acu];
res1(med,3) = [esp];
res1(med,4) = [sen];
res1(med,5) = [VP];
```

```
res1(med,6) = [FP];
res1(med,7) = [VN];
res1(med,8) = [FN];
save('resultados/res1','k','acu','esp','sen','VP','FP','VN','FN');64
save('resultados/res-consolidados1','res1');
fprintf('CLASSIFICAÇÃO para %d características CONCLUÍDA \n \n',k);
end
fprintf('RESULTADOS DATASET 1:\n');
res1
fprintf('MELHOR RESULTADO ALCANÇADO PARA O DATASET:\n');
fprintf('ACURÁCIA: %f\n',max(res1(:,2)));
fprintf('ESPECIFICIDADE: %f\n',max(res1(:,3)));
fprintf('SENSIBILIDADE: %f\n',max(res1(:,4)));
fprintf('VP: %d\n',max(res1(:,5)));
fprintf('FP: %d\n',min(res1(:,6)));
fprintf('VN: %d\n',max(res1(:,7)));
fprintf('FN: %d\n',min(res1(:,8)));
case 2,
break;
end
end
```

A.2 Algoritmo para Gerar a Base de Dados

```
% gera_db_gw.m
function [tamDsCa] = gera_db_gw(ds)
close all;
fprintf('gerando dataset %d de gw........\n',ds);
path = strcat('','ds',int2str(ds),'_gw/','*.csv','');
a = dir(path);
gw = [];
for i = 1:length(a)
temp = importdata(a(i).name);
gw = [gw, temp.data(:,2)];
end
gw=gw';
fprintf('salvando dataset %d gw no diretório databases........\n',ds);
save(strcat('databases/gw',int2str(ds)),'gw');
[tamDsCa] = size(gw,1);
```

```
end

========================================================================
% gera_db_control.m
function gera_db_control(ds)
close all;
fprintf('gerando dataset %d de controle........\n',ds);
path = strcat('','ds',int2str(ds),'_control/','*.csv','');
a = dir(path);
control = [];
for i = 1:length(a)
temp = importdata(a(i).name);
control = [control, temp.data(:,2)];
end
control=control';
fprintf('salvando dataset %d de controle no diretório databases........\n',ds);
save(strcat('databases/control',int2str(ds)),'control');
end

========================================================================
% carrega_dbs1.m
function [X,tamX] = carrega_dbs1(ds)
close all;
clc;
%fprintf('carregando dataset %d de gw........\n',ds);
load databases/cancer1;
%fprintf('carregando dataset %d de gw........\n',ds);
load databases/control1;
%fprintf('gerando a matriz X consolidada do dataset %d........\n',ds);
X=[gw;control];
%fprintf('Matriz X do dataset %d gerada com sucesso...\n\n',ds);
tamX = size(X,1);
end
```

A.3 Algoritmo FastICA

```
% aapomagro.m
function [Y,W,A]= aapomagro(x,pcomponents,no)
% FastICA - PIB
```

```
% Originally written by the Finish (Aapo) team;
% Modified by the PIB team in Aug. 21, 2007
fprintf ('Removing mean...\n');
%--------------------------------------------------------------------
% Meanize
% Removes the mean of X
%--------------------------------------------------------------------
[nn,M]=size(x);
if nn>M,
x=x';
[nn,M]=size(x);
end
X=double(x)-mean(x')'*ones([1,M]);
X1=X;
% Remove the mean.
%---- Meanize end ------
% Calculate the eigenvalues and eigenvectors of covariance matrix.
fprintf ('Calculating covariance...\n');
covarianceMatrix = X*X'/size(X,2);
[E, D] = eig(covarianceMatrix);
% Sort the eigenvalues and select subset, and whiten
%--------------------------------------------------------------------
%
PCA beggins
%--------------------------------------------------------------------
[dummy,order] = sort(diag(-D));
E = E(:,order(1:pcomponents));
d = diag(D);
d = real(d.^(-0.5));
D = diag(d(order(1:pcomponents)));
X = D*E'*X;
whiteningMatrix = D*E';
dewhiteningMatrix = E*D^(-1);
%--------------------------------------------------------------------
%
PCA end
%--------------------------------------------------------------------
N = size(X,2);69
B = randn(size(X,1),pcomponents);
```

```
B = B*real((B'*B)^(-0.5));
% orthogonalize
W1=randn(size(B' * whiteningMatrix));
W=rand(size(B' * whiteningMatrix));
iter=0;
while abs(norm(W)-norm(W1'))>1e-50,
iter = iter+1;
fprintf('(%d)',iter);
% This is tanh but faster than matlabs own version
hypTan = 1 - 2./(exp(2*(X'*B))+1);
% This is the fixed-point step
B = X*hypTan/N - ones(size(B,1),1)*mean(1-hypTan.^2).*B;
B = B*real((B'*B)^(-0.5));
W1=W;
W = B' * whiteningMatrix;
end
Y=W*X1;
A = dewhiteningMatrix * B;
fprintf(' Done!\n');
```

A.4 Algoritmo de Máxima Relevância e Mínima Redundância

```
function [fea] = mrmr_mid_d(d, f, K)
% function [fea] = mrmr_mid_d(d, f, K)
% MID scheme according to MRMR
% By Hanchuan Peng
% April 16, 2003
bdisp=0;
nd = size(d,2);
nc = size(d,1);
t1=cputime;
for i=1:nd,
t(i) = mutualinfo(d(:,i), f);
end;
fprintf('calculate the marginal dmi costs %5.1fs.\n', cputime-t1);
[tmp, idxs] = sort(-t);
fea_base = idxs(1:K);
fea(1) = idxs(1);
KMAX = min(1000,nd); %500
```

```
idxleft = idxs(2:KMAX);
k=1;
if bdisp==1,
fprintf('k=1 cost_time=(N/A) cur_fea=%d #left_cand=%d\n', ...
fea(k), length(idxleft));
end;
for k=2:K,
t1=cputime;
ncand = length(idxleft);
curlastfea = length(fea);
for i=1:ncand,
t_mi(i) = mutualinfo(d(:,idxleft(i)), f);
mi_array(idxleft(i),curlastfea) = getmultimi(d(:,fea(curlastfea)),
 d(:,idxleft(i)));
c_mi(i) = mean(mi_array(idxleft(i), :));
end;
[tmp, fea(k)] = max(t_mi(1:ncand) - c_mi(1:ncand));
tmpidx = fea(k); fea(k) = idxleft(tmpidx); idxleft(tmpidx) = [];
if bdisp==1,
fprintf('k=%d cost_time=%5.4f cur_fea=%d #left_cand=%d\n', ...
k, cputime-t1, fea(k), length(idxleft));
end;
end;
return;
%====================================
function c = getmultimi(da, dt)
for i=1:size(da,2),
c(i) = mutualinfo(da(:,i), dt);
end;
```

A.5 Algoritmo SVM e Crossvalidation

```
% fold_sens_espec_cross_validation.m
function
[acuracia,especificidade,sensibilidade,VP,FP,VN,FN]
fold_sens_espec_cross_validation(fs,f)
A_10 = fs;
At = f;
ker.ker=2;ker.gamma=0;
```

```
tp=[];
tn=[];
fp=[];
fn=[];
SENS=[];
ESPEC=[];
TP = 0;
TN = 0;
FP = 0;
FN = 0;
ACC=[];
TMP=[];
PRED=[];
=
% quantidade de divisões
fold = 10;
CVO = cvpartition(At,'k',fold);
err = zeros(CVO.NumTestSets,1);
for i = 1:CVO.NumTestSets
trIdx = CVO.training(i);
teIdx = CVO.test(i);
[ acc, predict_label ] = svmPrediction( A_10(trIdx,:), At(trIdx,:),
 A_10(teIdx,:) , At(teIdx,:), ker );
ACC=[ACC;acc];
temp = At(teIdx,:);
for m = 1:length(temp)
if (predict_label(m) == temp(m)) & (temp(m) == 1)
TP = TP + 1;
end
if (predict_label(m) == temp(m)) & (temp(m) == -1)
TN = TN + 1;
end
if (predict_label(m) ~= temp(m)) & (temp(m) == 1)
FN = FN + 1;
end
if (predict_label(m) ~= temp(m)) & (temp(m) ==-1)
FP = FP + 1;
end
end72
```

```
sens = 100*(TP / (TP+FN));
espec = 100*(TN / (FP+TN));
tp=[tp;TP];
tn=[tn;TN];
fp=[fp;FP];
fn=[fn;FN];
SENS=[SENS;sens];
ESPEC=[ESPEC;espec];
TMP=[TMP;temp];
PRED=[PRED;predict_label];
err(i) = sum(~strcmp(predict_label,At(teIdx)));
end
cvErr = sum(err)/sum(CVO.TestSize);
acuracia=sum(ACC)/fold;
sensibilidade=sum(SENS)/fold;
especificidade=sum(ESPEC)/fold;
VP=max(tp);
FP=max(fp);
VN=max(tn);
FN=max(fn);
end
```

APÊNDICE B – Artigos Publicados

Durante o desenvolvimento desta dissertação dois artigos foram desenvolvidos: Método Computacional para o Diagnóstico Precoce da Granulomatose de Wegener e Desenvolvimento de um Método Computacional para Detecção da Cardiotoxidade Utilizando Padrões Proteômicos, Análise de Componentes Independentes e Máquina de Vetores de Suporte em parceria com o grupo de processamento de sinais da Universidade Estadual do Maranhão. O primeiro artigo pode ser encontrado em: (LINHARES et al.,).

Abaixo seguem os artigos:

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,

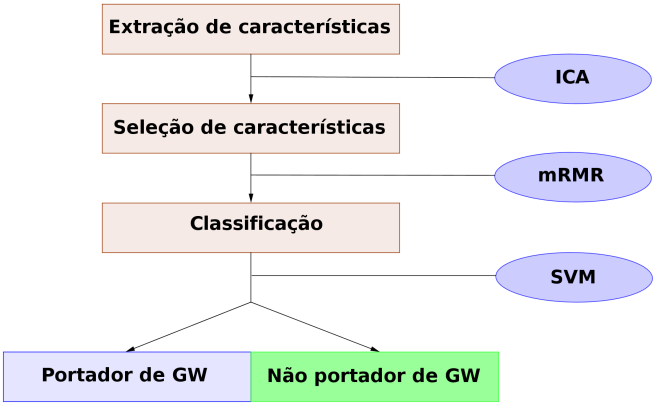

Figura 1. Diagrama da metodologia proposta.

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-

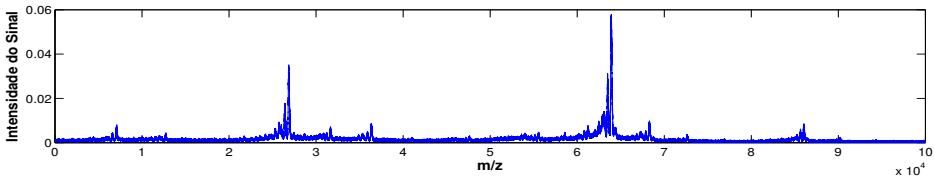

Figura 2. Espectro de massa obtido de um espectrômetro de massas.

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

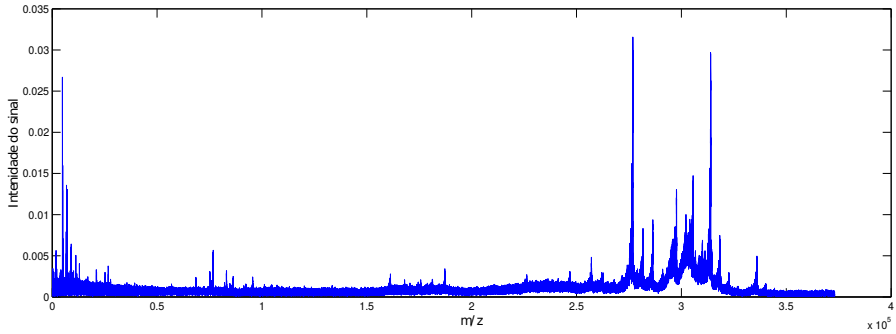

Figura 3. Espectro de massa de um sinal proteômico retirado da base de dados de forma aleatória.

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-

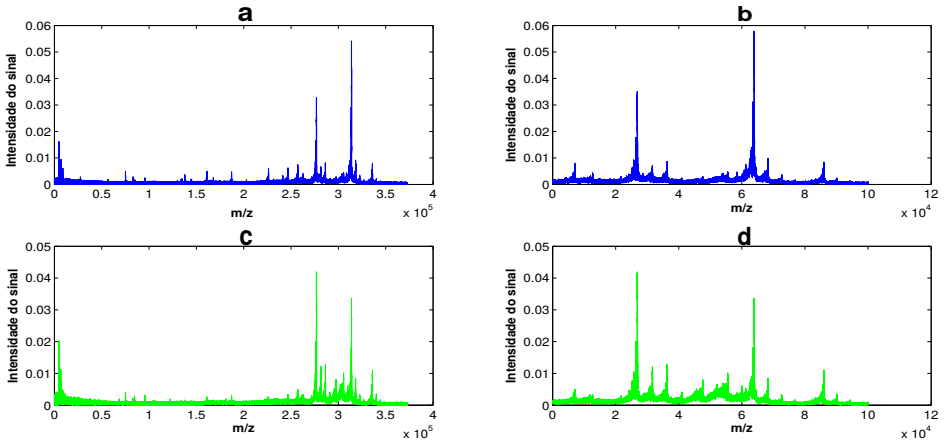

Figura 4. Espectros de massa. A figura (a) corresponde a uma amostra da base de dados com diagnóstico negativo de dimensão 380000 e a figura (b) mostra essa mesma amosta já reduzida para o intervalo $[250000; 350000]$. De forma semelhante, afigura (c) apresenta uma amostra com diagnóstico positivo e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

Método Computacional para Detecção da Cardiotoxidade Utilizando Padrões Proteômicos, Análise de Componentes Independentes e Máquina de Vetores de Suporte

Jardiel Nunes Almeida, Lúcio Flávio A. Campos, José do Nascimento Linhares, Flávia Larisse da Silva Fernandes

***Resumo*— Resumo: Em virtude da importância da detecção precoce da Cardiotoxidade, vários esquemas de diagnóstico auxiliados por computador estão sendo propostos com o objetivo de ajudar na identificação desta enfermidade. Isso porque o sucesso do tratamento no combate desta disfunção cardíaca depende de um diagnóstico rápido, pois quanto mais cedo iniciar-se o tratamento, maiores serão as chances de cura. Propomos um método de Diagnóstico Auxiliado por Computador (CAD) para diagnosticar pacientes com Cardiotoxidade, utilizando Análise de Componentes Independentes para extrair características de um sinal proteômico, depois fazendo uso da técnica de Máxima Relevância e Mínima Redundância para reduzir a dimensionalidade e com isso o custo computacional, e por fim a aplicação da Máquina de Vetores de Suporte para classificar as amostras entre presença ou ausência de Cardiotoxidade. O método foi testado com a base de dados de padrões proteômicos SELDI-TOF, cujo melhor desempenho obtido foi com um vetor de 20 características, resultando em uma acurácia de 88,718%, com 85% de especificidade e 97,26% de sensibilidade**

***Palavras-Chave*— Análise de Componentes Independentes, Cardiotoxicidade, Máxima Relevância e Mínima Redundância, Máquina de Vetores Suporte, Padrões Proteômico.**

***Abstract*— Because of the importance of early detection of cardiotoxicity, several schemes computer aided diagnosis has been being proposed in order to help the identification of this disease. Thats because the successful treatment to combat this cardiac dysfunction depends on rapid diagnosis because the earlier start the treatment, the greater the chances of healing. We propose a method of Diagnosis aided by computer (CAD) to diagnose patients with Cardiotoxicity, using Independent Component Analysis to extract characteristics of a proteomic signal. Then we use the technique of Maximum Relevance and Minimum Redundancy to reduce the dimensionality and thus the computational cost. And lastly the application of Support Vector Machine to classify the samples between the presence or absence of cardiotoxicity whose best performance was obtained with a vector of 20 features resulting in an accuracy of 88.718, with 85 of specificity and 97.26 of sensitivity.**

***Keywords*— Independent Component Analysis, cardiotoxicity, Maximum and Minimum Redundancy Relevance, Support Vector Machine.**

I. Introdução

Ao longo dos últimos anos, diferentes tratamentos para diversos tipos de câncer foram largamente desenvolvidos, levando a cura subsequente desta doença em alguns pacientes ou ao evidente aumento da sobrevida e qualidade de vida dos mesmos [1]. Porém, vários estudos comprovam que diferentes agentes antineoplásicos (antimetabólitos, antraciclinas e agentes biológicos, hormonais, alquilantes e antimicrotúbulos) utilizados no tratamento, têm potencial cardiotóxico[2]. Assim, eles podem causar o surgimento de uma nova enfermidade, conhecida como Cardiotoxicidade. Neste sentido, vários critérios de detecção e protocolos têm sido propostos para o tratamento e prevenção da mesma[2].

Sobre esta questão, vale ressaltar que a Cardiotoxicidade é definida pela situação na qual agentes externos (químicos ou físicos) interferem negativamente no coração, determinando alterações estruturais, elétricas e funcionais no miocárdio [3]. Este órgão torna-se mais fraco e não é tão eficiente em bombeamento, o que compromete a circulação do sangue [3]. Esta enfermidade pode ser causada por tratamentos de quimioterapia, complicações decorrentes da anorexia nervosa, efeitos adversos da ingestão de metais pesados, ou um medicamento administrado incorretamente como a bupivacaína [3].

Assim, torna-se necessária a prevenção desta patologia, que é realizada através de uma avaliação inicial dos pacientes oncológicos submetidos a quimioterapia cardiotóxica e esta avaliação tem como objetivos: excluir pacientes com evidências clínicas, laboratorial e radiológica de insuficiência cardíaca congestiva (IC) antes do início do tratamento quimioterápico, identificar pacientes com redução da fração de ejeção, associada a sintomas ou não, durante a quimioterapia [4]. E para fazer esta avaliação utiliza-se exames cardíacos, tais como: Eletrocardiografia, Eco-Dopplercardiografia, Cintilografia com Radionuclídeo, Teste Ergométrico, Biópsia Endomiocárdica, troponina T cardíaca, entre outros, menos utilizados.

O sucesso do tratamento no combate desta disfunção de eletro fisiologia do coração depende de um diagnóstico rápido, pois quanto mais cedo iniciar-se a assistência médica, maiores serão as chances de cura. Em virtude da importância da detecção precoce da ação cardiotóxica provocada por essas drogas, vários esquemas de diagnóstico auxiliados por computador estão sendo propostos com o objetivo de ajudar na identificação precoce desta enfermidade. Vale ressaltar que não existem métodos CAD associados a essa Patologia.

Neste trabalho propõe-se um método de Diagnóstico Aux-

Jardiel Nunes Almeida, Departamento de Engenharia da Computação, Universidade Estadual do Maranhão, São Luis-MA, Brasil, E-mail: jardieliguaiba@gmail.com .

iliado por Computador (CAD) para ajudar no reconhecimento precoce da Cardiotoxidade, utilizando dados ou sinais proteômicos. E para realizar a extração de característica destes sinais proteômicos será utilizada a técnica de Análise de Componentes Independentes (Independent Component Analysis-ICA), somada ao Algoritmo de Máxima Relevância Mínima Redundância (mRMR), para selecionar as características mais significantes e reduzir a dimensionalidade da matriz gerada. Após a seleção das características mais relevantes, estas serão classificadas utilizando-se a Máquina de Vetores Suporte (Support Vector Machinhe - SMV) entre duas classes, sendo que identificará o paciente com ou sem Cardiotoxidade.

II. Metodologia Proposta

A. Método Proposto

O método proposto é descrito pelo diagrama em blocos mostrado na Figura 1. Este consiste em: extrair as características significantes do sinal proteômico utilizando a Análise de Componentes Independentes (ICA), reduzir a dimensionalidade da matriz de característica gerada com o algoritmo de Máxima Relevância e Mínima Redundância (mRMR) e por fim, fazer a classificação com a de Máquina de Vetores de Suporte (SVM).

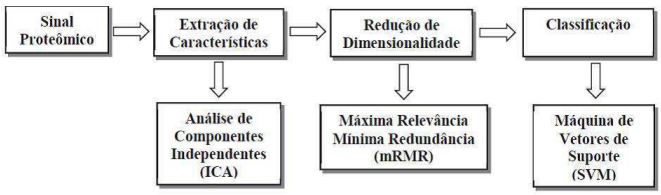

Fig. 1. Diagrama do método.

B. Espectrometria de massa e Sinal prteômico

Nos últimos anos diversos tipos de marcadores moleculares que auxiliam no diagnóstico precoce e no tratamento de várias doenças humanas, incluindo a Cardiotoxicidade como exposto em [5], vêm sendo analisados.

Para fazer a análise destes marcadores são utilizados conceitos como o da Espectrometria de Massa, uma técnica analítica utilizada para identificar compostos desconhecidos, modificar materiais conhecidos e elucidar as propriedades químicas e estruturais das moléculas. Nesta técnica, um composto é ionizado através de um método de ionização, os íons são separados na razão massa carga por meio de um método de separação, e o número de íons correspondentes a cada unidade de razão massa carga são registrados na forma de um espectro de massa. Para esse fim é necessário um espectrômetro de massa, um analisador que permite a determinação qualitativa e quantitativa dos compostos de uma amostra [6].

Neste sentido a Proteômica é entendida como sendo a análise em larga escala de um conjunto de proteínas, ou seja, a análise da expressão gênica de determinada célula, tecido ou organismo, sob determinadas condições ambientais, ou estágio de desenvolvimento que são responsáveis direta ou indiretamente pelo controle de todos ou quase todos os processos biológicos. Isto permite a identificação e caracterização de marcadores biológicos, ou seja, moléculas endógenas ou exógenas específicas de um determinado estado patológico[5]. E nesta perspectiva gerar listas de proteínas que aumentam ou diminuem em expressão como causa ou consequência de patologia [5]. A natureza desta informação pode nos levar a causa ou a uma consequência, de processos de doenças e de toxicidade. Além do mais, o recente progresso de metodologias nessa área tem aberto novas oportunidades para obtenção de informações relevantes sobre processos normais e anormais que ocorrem no organismo humano [5].

Os dados usados neste trabalho foram baseados em padrões proteômicos usando a técnica *SELDI-TOF*, que se mostrou um padrão de informação preciso para auxiliar no diagnóstico de pacientes com Cardiotoxidade [5]. As amostras utilizadas no trabalho foram adquiridas em [7].

A Figura 2 ilustra a amostra que foi extraída através de um espectrômetro de massa e foi posteriormente convertida em um sinal multinível através dos níveis de intensidade proteômicos encontrados no espectro de massa.

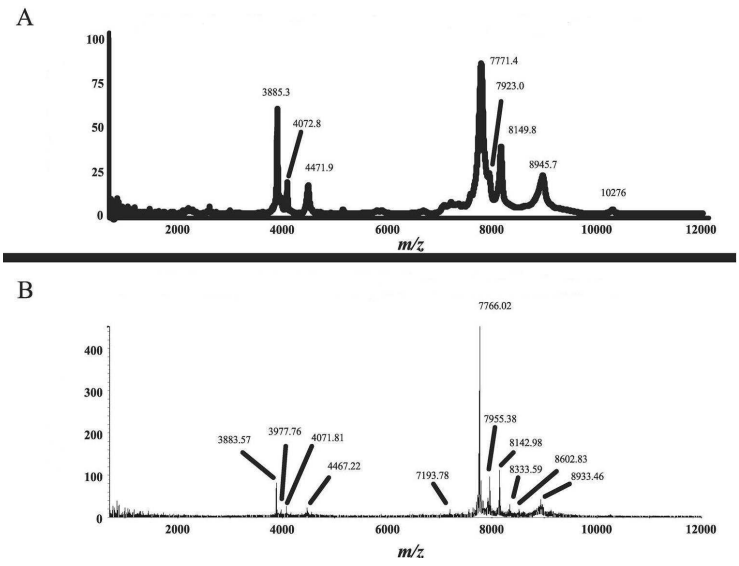

Fig. 2. Espectro de Massa. Comparação entre baixa resolução e alta resolução de um espectro de massa SELDI-TOF. Sendo, a figura 2.A corresponde a um espectro de massa de alta resolução e a figura 2.B a um de baixa resolução. O eixo vertical corresponde ao nível de intensidade do espectro de massa, enquanto o eixo horizontal corresponde à razão massa / carga. Fonte: [5]

C. Análise de Componentes Independentes

A análise de componentes independentes (ICA) é uma técnica estatística e computacional capaz de revelar componentes desconhecidos a um conjunto de variáveis aleatórias, medições, ou sinais observados multivariados [10]. Por isso ela é utilzada em: processamento de sinais Biomédicos, Telecomunicação e processamento de imagem. esste trabalho é baseado em processamento de sinais Biomédicos, que justifica a utilização da ICA. No modelo ICA, considera-se que dado sinal observado x_i pode ser representado como uma combinação linear de n variáveis aleatórias s_i que são estatisticamente independentes e não-gaussianas [8].

Desta forma pode-se escrever cada sinal x_i como:

$$x_i = a_{i1}s_1 + a_{i2}s_2 + ... + a_{in}s_n, \forall\, i, j = 1, 2, ..., n. \quad (1)$$

Onde os a_{ij} são os coeficientes de mistura (característica) e os s_i são os componentes independentes.

Este modelo pode ser expresso na forma matricial como:

$$\mathbf{X} = \mathbf{AS} \quad (2)$$

Sendo $\mathbf{A}$ a matriz de características e $\mathbf{S}$ as componentes independentes.

Verifica-se a partir da equação 2 que o problema da Análise de Componentes Independentes consiste em encontrar $\mathbf{A}$ e $\mathbf{S}$ conhecendo-se apenas $\mathbf{X}$. A solução deste problema pode ser obtida explorando-se a propriedade de independência ou não gaussianidade das componentes independentes [10]. O algoritmo *fastICA*, proposta por *Hyvärine*[10], é utilizado para estas matrizes. Este algoritmo tem rápida convergência, e, se comparado com algoritmos baseados em gradiente, é mais simples, pois não necessita de ajuste no passo de adaptação [10]. O *fastICA* usa como medida de não gaussianidade uma versão aproximada da negentropia dada pela equação 3

$$J(y) \propto [E(G(y)) - E(G(y_{gaus}))]^2. \quad (3)$$

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 [10], as funções G_1 e G_2, representadas nas equações 4 e 5, garantem boas aproximações da negentropia e melhoram a convergência do algoritmo fastICA.

$$G_1(y) = \frac{1}{\beta} log(cosh(\beta y)), \text{ com } 1 \leq \beta \leq 2 \quad (4)$$

$$G_2(y) = -exp\left(-\frac{y^2}{2}\right) . \quad (5)$$

D. Seleção das características mais significantes

Após a determinação da matriz de características $\mathbf{A}$, faz-se necessário reduzi-la selecionando as características mais discriminantes para melhorar a performance do classificador. Além disso, o uso de muitas características podem aumentar o erro de classificação e o custo computacional.

Neste trabalho a redução de características foi realizada através do algoritmo de Máxima Relevância e Mínima Redundância, que seleciona as características mais relevantes, através da equação 6, e retira as redundantes através da equação 7.

$$max\ M(\mathbf{a}, c), R = \frac{1}{\mid \mathbf{a} \mid} \sum_{a_i \in A} I(a_i, c) \quad (6)$$

$$min\ R(\mathbf{a}), R = \frac{1}{\mid \mathbf{a} \mid^2} \sum_{a_i a_j \in v} I(a_i, a_j) \quad (7)$$

Sendo **a** um vetor de características, c o vetor de classe, a_i e a_j duas características individuais e I a informação mútua. A informação mútua mede quanta informação uma variável aleatória (VA) possui sobre outra.[11]

O algoritmo mRMR combina M e R (equação 8) para obter simultaneamente as características mais relevantes e menos redundantes.

$$max\ \Phi(M, R),\ \Phi = M - R, \quad (8)$$

Com o vetor de características reduzido pela técnica de Máxima Relevância e Mínima Redundância, pode então ser feita a classificação das amostras. O que será mostrado na próxima Seção.

E. Classificação

Na classificação das amostras utilizou-se a Máquina de Vetores de Suporte. A classificação foi realizada a partir da análise do vetor de características já reduzido através da técnica mRMR, onde as amostras foram rotuladas em negativo (grupo controle) ou positivo(com Cardiotoxidade).

No entanto, para mensurar a aprendizagem da Máquina e assim aumentar a confiabilidade dos resultados, utilizou-se o *Cross-Validation* (validação cruzada) [14], uma técnica de particionamento de amostragem randômica, utilizada para estimar com maior precisão a acurácia (probabilidade de classificação correta de uma instancia selecionada estatisticamente) de um classificador. Existem basicamente três métodos distintos de validação cruzada, são eles: *Holdout*, *K-fold* e *Leave-one-out* [14]. Neste trabalho o método utilizado foi *K-fold*. Onde o conjunto de dados (exemplos) é aleatoriamente dividido em k partições mutuamente exclusivas (*folds*) e de tamanho aproximadamente igual a $\frac{n}{k}$ dados. As $(k-1)$ folds são utilizadas para para treinamento e o fold restante para testes. Este processo é repetido k vezes, e a cada vez é considerado um *fold* diferente para teste.

1) Máquinas de Vetores de Suporte: A Máquina de Vetores de Suporte (SVM) é um método de aprendizagem supervisionada, capaz de classificar a partir de n indivíduos observados pertencentes a diversos subgrupos, a que classe um indivíduo que deve ser classificado pertence [15].

As Máquinas de Vetores Suporte (SVMs), são algoritmos de aprendizagem bastante utilizados na área de aprendizagem de máquina. Elas constituem uma técnica embasada na Teoria de Aprendizado Estatístico [15] que vem recebendo grande atenção nos últimos anos [16]. As SVMs têm algumas características que tornam seu uso muito atrativo, tais como: Boa capacidade de generalização, robustez em grandes dimensões, Convexidade da função objetivo e Teoria bem definida. Além do mais a SVM pode ser utilizada para fazer a classificação e trabalha bem em espaço de alta dimensionalidade, atuando em problema de duas classes e assim podendo fazer a identificação dos pacientes com Cardiotoxidade e os sem Cardiotoxidade.

F. Métricas e Desempenho

Em processamento de sinais biomédicos e reconhecimento de padrões, a metodologia de desempenho usual é avaliada calculando-se algumas medidas estatísticas sobre o resultado dos testes [18]. Neste trabalho os resultados da classificação a partir da realização de testes são divididos em: Verdadeiro Positivo (VP), Falso Positivo (FP), Verdadeiro Negativo (VN) e Falso Negativo (FN).

Sendo VP e VN números de amostras que são corretamente identificadas como positivas e negativas, respectivamente pelo classificador. FP e FN representam o número de amostras correspondentes aos pacientes que são diagnosticados erroneamente como positivo (portador da doença) ou negativo (não portador da doença), respectivamente. Estes números são utilizados para gerar medidas capazes de quantificar o desempenho da metodologia, para avaliar a eficiência do nosso método e se os objetivos foram alcançados. As medidas de desempenho aqui utilizadas são: Acurácia, Especificidade, Sensibilidade.

III. Resultados e Discussões

A. Aquisição de Dados

A base de dados usada neste trabalho é composta por 62 sinais proteômicos divididos em: 28 amostras com diagnóstico positivo para a Cardiotoxidade e 34 amostras com diagnóstico normal (controle). Esses sinais proteômicos foram obtidos a parti da técnica *SELDI-TOF*, que se mostrou ser um padrão de informação preciso para auxiliar no diagnóstico de pacientes com Cardiotoxidade [5]. As amostras utilizadas no trabalho foram adquiridas em [7]. Cada amostra da base possui 373257 níveis de intensidade diferentes, no entanto para melhorar os resultados, reduzimos cada amostra para 100000 níveis de intensidade diferentes como ilustrado na figura 3

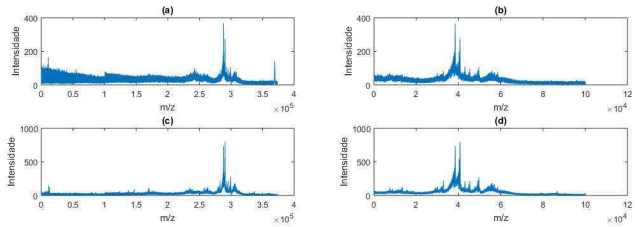

Fig. 3. Espectro de Massa. A figura (3.a) representa um sinal proteômico original de um paciente portador da Cardiotoxicidade contendo 373257 níveis de intensidade e a figura (3.b) mostra o sinal representado na figura (3.a) já reduzido para 100000 níveis de intensidades. Enquanto, a figura (3.c) ilustra um sinal proteômico de um paciente não portador da Cardiotoxicidade no tamanho original e a figura (3.d) mostra este sinal reduzido para 100000 níveis de intensidades.

B. Extração de Características

O objetivo desta etapa foi obter parâmetros a partir dos sinais proteômicos para discriminar os dois grupos (portador e não portador de Cardiotoxicidade) de informações capazes de identificar a Cardiotoxidade. As características selecionadas devem garantir que as amostras de um paciente sejam corretamente classificadas como portadora ou não da Cardiotoxidade.

Para extrair as características que distinguem os casos com Cardiotoxidade dos casos do grupo controle foi utilizada a Análise de Componentes Independentes (ICA), e para esse fim criou-se a matriz $\mathbf{X}$ de ordem 62 por 100000 que é a junção das matrizes com amostras de casos com Cardiotoxidade e grupo controle. A matriz $\mathbf{X}$ compõe o modelo ICA.

O algoritmo *FastICA* usou a matriz $\mathbf{X}$ para obter uma outra matriz que contém as característica de cada uma das amostras, e esta matriz é denominada matriz $\mathbf{A}$ de ordem 62 por 62. Sendo que cada linha desta matriz $\mathbf{A}$ corresponde a uma amostra, e cada coluna corresponde a uma característica. Desta maneira o classificador terá um parâmetro para distinguir pacientes com Cardioxicidade daqueles que não possuem a doença.

C. Seleção das Características mais Significantes

Nesta etapa foram obtidos os parâmetros que melhor representam as informações geradas a partir da extração de características. Pois, caso todos os parâmetros obtidos pela extração de características sirvam de entrada para o classificador, poderíamos ter resultados insatisfatórios, com baixa acurácia e grande esforço computacional. E para selecionar as características que melhor represente o banco de dados, foram realizados testes para a redução do vetor de características de cada amostra incrementando, de cinco em cinco, o número de características selecionadas através da técnica de Máxima Relevância e Mínima Redundância (mRMR) até 62, sendo que para encontrar o vetor de melhor desempenho, cada vetor gerado foi testado com a Máquina de Vetores de Suporte (SVM).

D. Métricas de Avaliação

Para entedermos melhor o que significa especificidade, sensibilidade e acurácia, pois são utilizadas neste trabalho, vamos definir estas variáveis que nos auxiliarão para melhor compreensão dos resultados obtidos.

Sendo Acurácia (A) a taxa de acerto do classificador durante a fase de teste, definida por:

$$A = (VP + VN)/(VP + VN + FP + FN) \qquad (9)$$

A Especificidade (E) é a proporção de verdadeiros negativos que são corretamente classificados pelo teste, definida por:

$$E = VN/(VN + FP) \qquad (10)$$

A Sensibilidade (S) é a proporção de verdadeiros positivos que são corretamente classificados pelo teste, definida por:

$$S = VP/(VP + FN) \qquad (11)$$

E. Classificação

Na última etapa, foi utilizada a SVM como classificador das amostras dos pacientes em controle e pacientes com Cardiotoxidade.

Utilizou-se para isto um classificador que tem núcleo baseado em RBF (*Radial-Basis Function*), com a configuração padrão dos parâmetros, sem otimização dos mesmos. As amostras estão contidas em um conjunto apenas, com o objetivo de realizar os testes de validação cruzada *10-fold cross-validation*.

Foram realizados vários testes para verificação da eficácia dos resultados demonstrados na tabela 1, contendo: especificidade, sensibilidade e acurácia. Os melhores resultados obtidos através do método *10-fold cross-validation* foram para os

vetores com 14, 16, 20, 53 e 49 características, eles tiveram bom desempenho durante o período de testes do classificador. Baseado nos resultados das tabelas, verifica-se que com 20 características das 62 possíveis, o método obteve $88,718\%$ de acurácia, $85,000\%$ de especificidade e $97,260\%$ de sensibilidade.

TABELA I
RESULTADOS

Carct	VP	FP	VN	FN	Esp(E)(%)	Sen(S)(%)	Acu(A)(%)
14	32	11	17	2	83,333	97,521	87,566
16	33	10	18	1	86,905	95,426	83,412
20	**33**	**10**	**18**	**1**	**85,000**	**97,260**	**88,718**
49	33	11	17	1	80,000	88,897	82,057
53	33	11	17	1	82,143	86,943	80,597

Considerando o vetor de 20 características, observou-se também, que das 34 amostras positivas para Cardiotoxidade, 33 foram classificadas corretamente (VP), logo dos 34 caso de Cardiotoxidade 1 foram classificados como normal (FN). Dos 28 casos com diagnóstico normal, somente em 10 casos (FP) houve erro de classificação, diagnosticando-os positivo para a Cardiotoxidade.

IV. CONCLUSÃO

Este artigo propõe um método computacional para fazer o diagnóstico precoce da Cardiotxicidade, através da classificação de padrões proteômicos, utilizando: Análise de Componentes Independentes (ICA), Algoritmo de Máxima Relevância e Mínima Redundância (mRMR), e Máquina de Vetores de Suporte.

Os resultados encontrados demonstraram que o conjunto de técnicas aplicadas é eficiente para diagnosticar a Cardiotxicidade, pois conseguiu identificar os indivíduos normais e os portadores desta enfermidade. Estes resultados são vistos nas métricas de desempenho encontradas: $88,718\%$ acurácia, 85% de especificidade e $97,26\%$ de sensibilidade, em um estudo que utilizou 62 amostras com baixa resolução. Das 62 amostras utilizadas, 51 foram classificadas corretamente (VP + VN) e 11 classificadas incorretamente (FP+FN). Apenas uma amostra foi classificada como negativa sendo positiva.

Para confirmar a eficácia do método, novos testes devem ser realizados em bases de dados maiores.

REFERÊNCIAS

[1] MORAES, A. d. J. P. Viabilidade do treinamento Físico Aeróbio por pacientes com Câncer Hematológico antes do transplante Autólogo de Células-Tronco Hematopoiéticas. Tese (Mestrado em Ciências do Movimento Humano)? Universidade de Estado de Santa Catarina, Florianópolis, 2014.

[2] Sá, M. P. B. de O. et al. Cardiotoxidade e Quimioterapia. 2009. Disponível em: ¡http://files.bvs.br/upload/ S/1679-1010/2009/v7n5/a010.pdf¿. Acesso em: 3 set 2014.

[3] I Diretriz Brasileira de Cardi o-Oncologia Pediatrica da Sociedade Brasileira de Cardiologia. Disponível em:

[4] R, A. et al. Cardiovascular toxicity caused by cancer treatment: strategies for early detection. 2009. Disponível em: ¡http://www.ncbi.nlm.nih.gov/pubmed/19341970¿. Acesso em: 2 abr. 2009.

[5] PETRICOIN, E. F. P. et al. Toxicoproteomics: Serum proteomic pattern diagnostics for early detection of drug induced cardiac toxicities and cardioprotection. Toxicological Pathology, Clinical Proteomics Program Website, v. 32, n. 1, p. 1?9, 2011. ISSN 0192-6233 print / 1533-160. Disponível em: ¡http://home.ccr.cancer.gov/ncifdaproteomics/pdf/ToxPath.pdf¿.

[6] BARBOSA, E. B. et al. Proteômica: Metodologias e aplicacões no estudo de doenças humanas. Revista da Associação Médica Brasileira, Revista da Associção Médica Brasileira, v. 58, n. 3, p. 366?375, 2012. ISSN 0104-4230. Disponível em: ¡http://dx.doi.org/10.1590/S0104-42302012000300019¿.

[7] PROGRAM, C. P. Toxicoproteomic analysis of anthracycline-induced cardiotoxicity. 2004. Disponível em: ¡http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp¿. Acesso em: 2 maio 2014.

[8] SILVA, A. P. O. d. Uma Implementacão da Análise de Componentes Independentes em Plataforma de Hardware Reconfigurável. Tese ((Mestrado em Automacão e Sistemas; Engenharia de Computacão; Telecomunicacões)?Universidade Federal do Rio Grande do norte, Natal, 2011.

[9] PROGRAM, C. P. Análise comparativa das abordagens de estimativa do modelo FestICA por máximação da negentropia verossimilhança. 2012. Disponível em: https://www.ime.usp.br/arquivos/4congresso/4120Henrique20$Morimitsu_N$.pdf. Acesso em: 3 set 2014.

[10] HYVARINEN J. KARHUNEN, E. O. A. (Ed.). Independent Component Analysis. USA: Pearson Prentice Hall, 2001.

[11] T. M. Cover and J. A Thomas, Elements of Information Theory. New York, USA: Wiley, 1991

[12] ARAUJO, W. B. D. Método de Detecção de Câncer de Ovário utilizando Padrões Proteômicos, Análise de Componentes Independentes e Máquina de Vetores de Suporte. Tese (Mestrado em Engenharia da Computacão e Sistema)? Universidade Estadual do Maranhão, São Luis, 2014.

[13] 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/2004JBCB feasel-04-06-15.pdf.

[14] KOHAVI, R. A. A study of cross-validation and bootstrap for accuracy estimation and model selection. Computational and Applied Mathematics, Digital Library, v. 2, n. 3, p. 1137 a 1143, 1995. ISSN 0101 − 205. Disponível em: http://dl.acm.org/citation.

[15] CHAPELLE, o.; VAPNIK, V. Model selection for support vector machines. In: SOLLA, S. A.; LEEN, T. K.; MÜLLER, K.-R. (Ed.). Advances in Neural Information Processing Systems 12. Cambridge, Mass: MIT Press, 2000. Disponível em: http://www.ens-lyon.fr/ ochapell/ms/nips99.ps.

[16] HEARST, M. A. et al. Trends and controversies - support vector machines. IEEE Intelligent Systems, v. 13, n. 4, p. 1828, 1998. Disponível em: http://computer.org/intelligent/ex1998/pdf/x4018.pdf.

[17] HAYKIN, S. O. (Ed.). Neural Networks A Comprehensive Foundation. USA: Pearson Prentice Hall, 1998.

[18] BUSCHBERG, J. T. et al. (Ed.). The Essential Physics of Medical Imaging. Philadelphia,PA, USA: Wolters Kluwer., 2012.