VITOR PRADO DE CARVALHO

DISCRIMINAÇÃO DE POPULAÇÃO POR MEIO DE INTELIGÊNCIA COMPUTACIONAL

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada a Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2016

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T

C896d
2016

Carvalho, Vitor Prado de, 1981-

Discriminação de população por meio de inteligência computacional / Vitor Prado de Carvalho. – Viçosa, MG, 2016.

xi, 50f. : il. (algumas color.) ; 29 cm.

Orientador: Moysés Nascimento.

Dissertação (mestrado) - Universidade Federal de Viçosa.

Inclui bibliografia.

1. Estatística. 2. Análise discriminatória. 3. Melhoramento genético. 4. Variabilidade genética. 5. Inteligência computacional. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.5

VITOR PRADO DE CARVALHO

DISCRIMINAÇÃO DE POPULAÇÃO POR MEIO DE INTELIGÊNCIA COMPUTACIONAL

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 25 de fevereiro de 2016.

Wagner Antônio Arbex
(Coorientador)

Cosme Damião Cruz
(Coorientador)

Leonardo Lopes Bhering

Moysés Nascimento
(Orientador)

AGRADECIMENTOS

Inicialmente agradeço muito a Deus por tornar esse sonho realidade, por ter encontrado pessoas maravilhosas que me incentivaram mesmo em meus momentos de tristeza para que eu chegasse ao final dessa jornada.

Não posso deixar de dar um agradecimento especial a minha amada esposa, mesmo não sabendo estatística, discutia e me ajudava nos estudos. Sonhou junto comigo e por meio de sua compreensão e carinho incentivou-me a alcançar este sonho sem jamais esmorecer, seu apoio é deveras importante e aumenta ainda mais nossa união.

Aos meus pais e irmão que sempre foram meus exemplos de integridade, amor e respeito, sonharam comigo nesta etapa e sempre estiveram ao meu lado com suas orações e apoio.

Ao meu orientador e amigo Moysés Nascimento pelos conselhos, auxilio paciência e disponibilidade sempre que eu precisava não somente para a execução deste trabalho quanto também para conversar em diversos aspectos. Agradeço a ele pela amizade, atenção e auxilio durante esses anos de convivência.

Aos professores e co-orientadores que aceitaram fazer parte desse trabalho e aos professores que aceitaram fazer parte da minha banca, agradeço pela disponibilidade e de antemão as sugestões para correção melhorando a execução desse trabalho.

Aos amigos de mestrado, pelos bons momentos e experiências que colecionamos juntos, em especial a Daiana Salles que sempre foi uma boa amiga desde a faculdade e me ajudou a alcançar o tão sonhado título de mestre.

Agradeço a Universidade Federal de Viçosa pela oportunidade em minha vida.

Agradeço também a todos que de alguma forma contribuíram e torceram para essa grande conquista!

BIOGRAFIA

VITOR PRADO DE CARVALHO, filho de Wadir Ribeiro de Carvalho e Amélia Prado Carvalho, nasceu em Vitória – ES, em 22 de maio de 1981. Graduou-se pela Universidade Federal do Espírito Santo como bacharel em estatística no ano de 2010 no período de graduação foi monitor no laboratório de Estatística (LESTAT) prestando conhecimentos estáticos a comunidade acadêmica. Soma-se a essa experiência a organização do evento da X semana de Estatística que trouxe como contribuição a reunião de estudantes, professores e profissionais de Estática e áreas correlatas.

No ano de 2012 atuou como professor substituto na Universidade Federal do Espírito Santo lecionando a disciplina de Estatística para diferentes cursos presente na instituição.

Em 2014 iniciou o curso de Mestrado no Programa de Pós-Graduação em Estatística Aplicada a Biometria na Universidade Federal de Viçosa, submetendo-se à defesa da dissertação em 25 de fevereiro de 2016.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

RESUMO

CARVALHO, Vitor Prado de, M.Sc., Universidade Federal de Viçosa, fevereiro de 2016. **Discriminação de população por meio de inteligência computacional**. Orientador: Moysés Nascimento. Coorientadores: Ana Carolina Campana Nascimento, Cosme Damião Cruz, Fabyano Fonseca e Silva e Wagner Antonio Arbex.

É importante para a preservação da variabilidade genética e da biodiversidade a correta classificação dos indivíduos. As técnicas de estatística multivariada comumente utilizada nessas situações são as funções discriminantes de Fisher e de Anderson, que permitem alocar um indivíduo inicialmente desconhecido em uma das g populações prováveis ou grupos pré-definidos. Entretanto, para o caso de populações não linearmente separáveis, esses métodos tem se mostrado pouco eficientes devido ao fato de não conseguir detectar a diferença entre as populações. Em alguns casos é preciso captar o máximo de informação possível e para tal outro método é necessário quando não for possível adquirir resultados pelos métodos multivariados. Portanto uma alternativa como possível solução para tal finalidade são as redes neurais artificiais, utilizadas em diversos problemas da Estatística, como agrupamento de indivíduos similares, previsão de séries temporais e em especial, os problemas de classificação. Outra técnica computacional que também vem adquirindo credibilidade e grande atenção nos últimos anos é conhecida como Máquina de Vetor Suporte (Support Vector Machines - SVMs). As SVMs vêm sendo utilizadas em diversas tarefas de reconhecimento de padrões, obtendo resultados superiores ou similares aos alcançados por técnicas similares em várias aplicações como em detecção de faces em imagens e na categorização de textos. Diante do exposto, o objetivo deste trabalho é avaliar a utilização da máquinas de vetores suporte em problemas de discriminação de populações com estruturas genéticas conhecidas. Além disso, os resultados obtidos pela técnica foram comparados com aqueles advindos de análises discriminante de Anderson e redes neurais. Cada população foi caracterizada por um conjunto de elementos mensurados por características de natureza contínua. Foram geradas considerados 50 locos independentes, cada qual com dois alelos. As relações de parentescos e a estruturação hierárquica foram estabelecidas considerando populações

genitoras geneticamente divergentes, híbrido F_1 e três gerações de retrocruzamentos em relação a cada um dos genitores, permitindo estabelecer parâmetros de eficácia das metodologias testadas. Os dados fenotípicos das populações foram utilizados para estabelecimento da função discriminante de Anderson e para o cálculo da taxa de erro aparente (TEA), que mede o número de classificações incorretas. As estimativas de TEA foram comparadas com as obtida por meio das Redes Neurais Artificiais e a Máquina de Vetor Suporte para verificação dos problemas de classificações, buscando minimizar o número de classificações incorretas em comparação aos obtidos pela função discriminante. De acordo com os resultados avaliados, a Rede Neural obteve resultados satisfatórios com TEA a 0% enquanto que o método SVM obteve TEA de 14,44% a 67,41% enquanto que a de Anderson manteve TEA entre 18,89% a 74,07%. No entanto são necessários mais estudos quanto a utilização da SVM com base em algoritmos de otimização de busca para o espaço de parâmetros para pôr fim tentar alcançar resultados mais satisfatórios.

ABSTRACT

CARVALHO, Vitor Prado de, M.Sc., Universidade Federal de Viçosa, February, 2016. **Discrimination of the population by means of computational intelligence.**Adviser: Moysés Nascimento. Co-Advisers: Ana Carolina Campana Nascimento, Cosme Damião Cruz, Fabyano Fonseca e Silva and Wagner Antonio Arbex.

It is important for the preservation of genetic variability and biodiversity the correct classification of the individuals. The techniques of multivariate statistics commonly used in these situations are the Fisher and Anderson discriminant functions, which allow you to allocate an individual initially unknown to one of g populations likely or groups pre-defined. However, for the case of populations that are not linearly separable, these methods have been shown little efficient due to the fact it's not able to detect the difference between the populations. In some cases, it is necessary capturing as much information as possible and for that other method is required when it is not possible to acquire the results frommultivariate methods. Therefore an alternative as a possible solution for this purpose is the artificial neural networks, used in various problems of Statistics, such as grouping of individuals with similar forecasting time series and in particular, the problems of classification. Another computational technique that has been acquiring credibility and great attention in recent years is known as the Support Vector Machines (SVM).The SVMs have been used in various tasks of pattern recognition, achieving higher results or similar to those achieved by similar techniques in various applications, such as detection of faces in images, and in the categorization of texts. The aim of this study is to evaluate the use of Support Vector Machines in problems of population's discrimination with a known genetic structure. In addition, the results obtained by the technique is compared with those resulting from analysis of Anderson discriminant function and neural networks. Each population was characterized by a set of elements measured by characteristics of continuous nature. Were generated considering 50 locos independent, each with two alleles. The relations of kinship and the hierarchical structuring were established considering populations genetically divergent, F1 hybrid and three generations of backcrossing in relation to each of the parents, allowing to establish parameters of effectiveness of the tested methodologies.

The phenotypic data of the populations were used to establish the discriminant function of Anderson and for the calculation of the error rate apparent (TEA), that measures the number of incorrect ratings. Estimates of TEA were compared with those obtained by means of Artificial Neural Networks and Support Vector Machine for verification of classification problems, seeking to minimize the number of incorrect ratings in comparison to discriminant function. According to the results, the neural network obtained satisfactory results with a TEA of 0%, while the SVM method obtained TEA between 14.44% and 67.41%, while the results of Anderson function have TEA between 18.89% and 74.07%. However, it is necessary more studies about the use of the SVM based on the optimization algorithms for the search of the space of parameters in order to try to achieve results that are more satisfactory.

INTRODUÇÃO GERAL

Análises da diversidade genética têm orientado a escolha de genitores apropriados em programas de melhoramento, levando à otimização dos ganhos seletivos, devido à variabilidade encontrada nos grupos divergentes. Além disso, as análises de diversidade genética têm permitido a quantificação da variabilidade existente e facilitado o gerenciamento dos bancos de germoplasma, poupando tempo e recursos (Sant'Anna, 2014).

Neste sentido, é de grande importância os estudos de diferenciação de populações por permitirem o entendimento das melhores estratégias para incrementar e preservar a diversidade das espécies, e ou, dos indivíduos dentro das espécies ou populações (Cruz et al., 2011).

Atualmente, existem na literatura muitas metodologias para quantificação e avaliação da diversidade em estudos populacionais, porém devido à grande variedade de informações a serem avaliadas e as particularidades de cada material biológico, a escolha correta da metodologia a ser aplicada é de grande importância para obtenção de resultados confiáveis.

Comumente, os métodos que se baseiam em análises estatísticas multivariadas tem sido uma alternativa eficaz nos estudos de diversidade, entre eles destacam-se as análises discriminantes, uma técnica da estatística multivariada cuja função é discriminar e classificar objetos separando-os de uma população em duas ou mais classes, ou seja, são utilizadas para classificar um indivíduo ou um grupo de indivíduos em diferentes populações conhecidas ou não (Cruz et al., 2011).

Entretanto, em muitas situações, mesmo se dispondo dos dados experimentais para realização das análises multivariadas, os resultados obtidos não são satisfatórios pela técnica não ser capaz de detectar as diferenças entre populações não linearmente separáveis (Martins et al., 2007).

Visando obter um solução para tal limitação Sant'Anna et al. (2014) propuseram o uso das redes neurais artificiais para a classificação de indivíduos, comparando diferentes funções discriminantes (Fisher e Anderson) e redes neurais artificiais em termos do número de classificações incorretas de indivíduos sabidamente pertencentes a diferentes populações simuladas de

retrocruzamentos, com crescentes níveis de similaridade e verificaram a eficiência das redes frentes as demais técnicas utilizadas.

Outra abordagem que pode ser utilizada para a solução de problemas de classificação é a Máquina de Vetor Suporte (SVM, do inglês *support vector machines*). Tal metodologia, a qual é fundamentada na Teoria da Aprendizagem Estatística (Vapnik, 1995), utiliza apenas algumas observações, denotadas por vetores de suporte, para a obtenção de uma regra de classificação, sendo então mais robusta a ruídos quando comparadas a técnicas tradicionais tais como de estatística multivariada e redes neurais (Lorena & Carvalho, 2003). Uma das principais diferenças quanto à capacidade da SVM com relação às redes neurais é a convergência local dos algoritmo de aprendizagem, ou seja, enquanto na RNA podem haver muitas soluções, convergindo para mínimos locais, o SVM converge para uma única solução ótima, o mínimo global (Rychetsky, 2001).

Apesar das SVM ser utilizada com sucesso em diversas áreas de conhecimento, como foi na comparação entre redes neurais artificias e máquinas de vetor suporte para reconhecimento de posturas manuais em tempo-real (Ticiano et al., 2007) e sistemas de detecção de vazamentos em dutos usando redes neurais e máquinas de vetor de suporte (Martins, 2007), a mesma ainda não foi utilizada e avaliada para estudos de diferenciação de populações.

Diante do exposto, este trabalho tem por objetivo propor e avaliar a utilização de SVM em estudos de diferenciação de populações. Especificamente, os objetivos podem ser discriminados em:

- Estudar a viabilidade da discriminação de populações com diferentes graus de dissimilaridade por redes neurais artificiais, máquina de vetor suporte e análise discriminante de Anderson.

- Comparar as técnicas de discriminação por máquina de vetor suporte, redes neurais e as técnicas de análise discriminante propostas por Anderson, por meio da taxa de erro aparente.

- Apresentar os aspectos teóricos envolvidos referentes as SVMs vinculados a teoria do aprendizado estatístico para estudos com classificação de populações.

1. REVISÃO DE LITERATURA

1.1. Diversidade Genética no Melhoramento

Os estudos envolvendo diversidade genética são muito importantes no processo de melhoramento de uma espécie. Apenas conhecer a diversidade genética não é suficiente para o sucesso de programas de melhoramento, é importante determinar a variabilidade existente em relação aos caracteres morfoagronômicos de interesse. Dessa forma, é necessário que se tenha informações fenotípicas confiáveis avaliadas nos genótipos existentes (Laviola 2010).

Os recursos genéticos devem ser devidamente caracterizados para permitir ganhos genéticos mais promissores no melhoramento, a mensuração e a preservação da biodiversidade se mostra crucial para a evolução e sobrevivência das espécies. Desse modo, a importância dos estudos sobre a divergência genética para o melhoramento reside no fato de cruzamentos envolvendo genitores geneticamente diferentes são os mais convenientes para produzir alto efeito heterótico (o valor genético da progênie seja igual à média do valor genético dos pais) e, também, maior variabilidade genética em gerações segregantes, para isso busca-se população base para seleção que alie ampla variabilidade genética com alta média para o caráter a ser selecionado (Cruz et al., 2011).

A diversidade genética pode ser avaliada de forma simultânea em relação às várias características, e para isso recomenda-se a utilização de medidas de dissimilaridade (Cruz et al., 2003). Uma forma prática e eficiente de se obter essas medidas é por meio da análise de agrupamentos (ou análise de cluster), a qual tem por finalidade reunir os indivíduos em grupos, de forma que exista a máxima homogeneidade dentro do grupo e a máxima heterogeneidade entre os grupos (Cruz et al., 2001). Dos métodos de agrupamento, os mais utilizados são os de otimização e os hierárquicos (Cruz et al., 2011).

Diversos estudos utilizam análises de agrupamento na visualização e interpretação da diversidade genética, com base em caracteres morfológicos e agronômicos em plantas como o cacau (Dias et al., 1997), o cupuaçu (Alves et al., 2003; Araujo et al., 2002) e o guaraná (Nascimento et al., 2001), a melancia

(Souza et al., 2005), o alface (Oliveira et al., 2004), a cana-de-açúcar (Silva et al., 2005), a aveia (Machioro et al.,2003), dentre outras.

1.2. Análise Discriminante de Anderson

A análise discriminante é uma técnica estatística utilizada para diferenciar populações e/ou classificar elementos de uma amostra ou população, ou seja, é uma técnica que estuda a separação de objetos em duas ou mais classes. Para tal, a discriminação (Separação) é a primeira etapa, necessitando de grupos previamente definidos para enfim procurar em sua fase exploratória características capazes de serem utilizadas para alocar novos objetos nos diferentes grupos, desse modo uma regra de decisão é incorporada a partir dos dados previamente utilizados (dados de treinamento) e utilizado para classificar novos elementos de maneira que a probabilidade de má classificação seja mínima (Johnson & Wichern, 2001; Cruz e Carneiro, 2006, Mingot, 2007).

A mais conhecida função discriminante conhecida é a análise discriminante de Fisher (Fisher, 1936) cuja única suposição para duas populações é de que suas matrizes de covariâncias das variáveis que caracterizam os elementos de ambos os grupos sejam homogêneas (Hair, 2007).

No entanto, para os casos em que se possuem mais do que dois grupos, ou seja, vamos supor $\pi_1, \pi_2, \ldots, \pi_g$ um grupo de g-populações, podemos então utilizar também outra técnica de discriminação proposta por Anderson (1958), a partir das informações dos indivíduos das diferentes populações, gera-se então funções de combinações lineares das características avaliadas de forma a obter uma melhor discriminação entre os indivíduos e assim podendo-se alocá-las em suas devidas populações. Para desenvolver esse tipo de classificador é necessário fazer algumas pressuposições, como segue: 1) As g-populações apresentam distribuição normal multivariada, 2) Definir as p_i probabilidades a priori de ocorrência das populações (geralmente são iguais e sua soma é igual a 1), 3) Apresentar os custos de má classificação (Geralmente as populações apresentam custos iguais de má classificação) (Pereira, 2009).

Quanto ao item 3, sobre custo de má classificação, esse é definido nas p_i probabilidades, recomenda-se antes de construir a regra de classificação um

estudo detalhado de quais p variáveis têm efeito significativo no fenômeno a ser estudado. É um ponto muito importante para a análise pois se dá pelo fato de termos de definir uma probabilidade a priori para as várias populações já que podem existir casos em que a probabilidade de um certo indivíduo com p-variáveis, pertencer a uma determinada população já ser conhecida pelo pesquisador por meio de sua experiência sobre a população pesquisada e do indivíduo a ser classificado (Pereira, 2009).

Com essas pressuposições estabelecidas e supondo homogeneidade das matrizes de variâncias e covariâncias, pode-se então estabelecer uma regra de classificação por meio da função discriminante de Anderson da seguinte forma:

$$D_i(X) = \ln(p_i) + \mu_i^T \Sigma^{-1} X - \frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i, \text{ para } i = 1, \dots, g \qquad (1.1)$$

E classifica-se X em $\pi_i$se:

$$d_i(X) = \max_j \left[d_j(X)\right], \text{ para } j = 1, \dots, g$$

Na prática, é bem difícil já se ter conhecimento a respeito dos parâmetros $\mu_i \text{e } \Sigma$, neste caso utiliza-se as estimativas ao invés dos parâmetros. No caso da matriz de variância e covariância tem-se:

$$\hat{\Sigma} = S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + \left(n_g - 1\right)S_g}{n_1 + n_2 + \cdots + n_g - g} \qquad (1.2)$$

Além do discutido, existem alguns casos que não é possível adquirir bons resultados utilizando-se desse método, podemos citar como exemplo: populações com poucas características, o método ser inadequado dependendo do problema e poucos dados para gerar uma regra de decisão (Nesse caso é necessário uma abordagem como ampliação dos dados).

1.3. Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) são conhecidas por funcionar conceitualmente de forma a simular a estrutura neural do cérebro humano, elas tem a capacidade de aprender por meio da experiência e assim resolver problemas de predição, reconhecimento de padrões e classificação afim de se fazer generalizações baseadas no seu conhecimento previamente acumulado (Mackay,1994).

Um grande desafio da computação é juntar a velocidade de processamento computacional com as características do cérebro humano e assim resolver vários problemas que vem surgindo.

Embora biologicamente inspiradas, elas encontraram aplicações em diferentes áreas científicas. Como por exemplo, Barroso et al. (2013) utilizaram as redes neurais e a análise discriminante para classificação de genótipos de alfafa quanto à adaptabilidade e estabilidade fenotípica utilizando como informação a priori o método de Eberhart e Russel. Já Nascimento et al.(2013) utilizaram redes neurais para avaliar uma metodologia de adaptabilidade e estabilidade fenotípica da alfafa considerando o método Eberhart e Russel.

Finalmente, podemos citar o estudo de Silva et al, (2014) no qual os autores utilizaram as redes neurais na predição de valores genéticos de dados simulados.

1.3.1. Inspiração Biológica da Redes Neurais Artificiais

As Redes Neurais são baseadas no funcionamento do cérebro humano, e este possui como unidade básica os neurônios, células especializadas pelo armazenamento e transmissão das informações.

Um neurônio mais simples apresenta três partes distintas que são os dentritos, o corpo celular e o axônio, como pode ser visto na Figura 1.

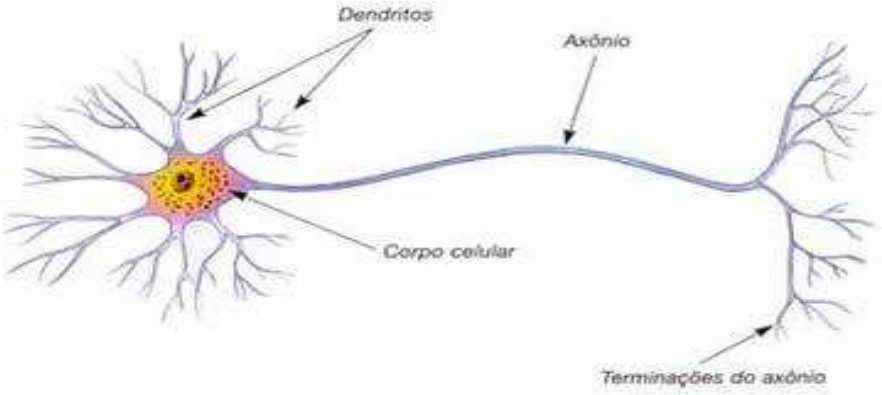

Figura 1-Modelo de Neurónio Biológico.
Fonte: (http://www.sogab.com.br/anatomia/neuronio.jpg)

Os dentritos tem a função de receber os estímulos e transmiti-los até o corpo celular. No corpo celular, esses estímulos serão processados e então

enviados até os axônios que por sua vez são responsáveis por transmitir os impulsos (estímulos) para a outra célula (Moreira, 2014).

O ponto de contato entre os dentritos e as terminações do axônio é chamado de sinapse, tem então a função de regular o fluxo dos estímulos e impulsos, de forma que, se estes impulsos atingem um certo limiar, então, o estímulo é transmitido e caso contrário não (Moreira, 2014).

Essas atividades, seja da mais simples a mais complexa, são reguladas em conjunto dos neurônios (Braga et al., 2000; Moreira, 2014).

1.3.2. Modelo Matemático dos Neurônios Biológicos

O primeiro modelo artificial de um neurônio biológico foi datado em 1943 como resultado de um trabalho feito pelo neuroanatomista e psiquiatra Warren McCulloch e do matemático Walter Pitts (McCulloch & Pitts, 1943). Alguns anos depois surge um modelo mais avançado denominado Perceptron Múltiplas Camadas (*Multi Layer Perceptron* – MLP) (Figura 2), com o algoritmo backpropagation as redes neurais artificiais se tornaram então uma técnica muito utilizada em várias áreas da ciência (Braga et al., 2007).

O princípio central da teoria de redes neurais está no fato de que fornecendo exemplos (população de treinamento) do relacionamento entre variáveis de entrada X e um alvo T, a rede neural irá capturar a relação entre as variáveis, podendo generalizar essas informações para novos casos (população de validação) (MACKAY, 1994). É esperado que uma RNA treinada tenha boa capacidade de generalização independente do aprendizado no treinamento ter sido supervisionado ou não.

A topologia de uma rede neural é definida pelo número de camadas (camada única ou múltiplas camadas), pelas conexões entre camadas, pelo número de neurônios em cada camada, pelo tipo de conexão entre eles (*feedforward* ou *feedback*) e pelo algoritmo de aprendizado (Haykin, 2001). Na Figura 2esta ilustrada o modelo básico de um neurônio artificial.

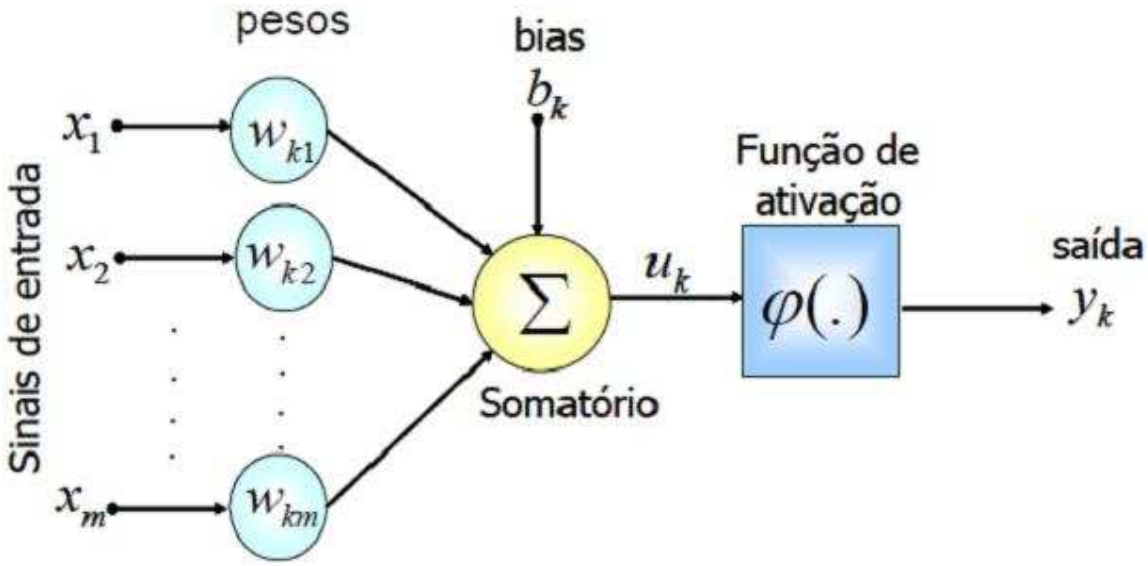

Figura 2- Modelo não linear de um neurônio artificial.
Fonte: Adaptado de Haykin, 2001.

Na Figura 2 $x_1, x_2, \ldots, x_m$ representam as entradas da rede; $W_{k1}, W_{k2}, \ldots, W_{km}$, os pesos sinápticos associados a cada entrada; b_k é o termo bias; u_k é a combinação linear dos sinais de entrada; φ(.) é a função de ativação e y_k é a saída do neurônio.

Portando em termos matemáticos podemos definir a saída de cada neurônio como sendo:

$$y_k = \varphi\left(\sum_{j=1}^{m} w_{kj}x_j + b_k\right)$$

Os pesos são os parâmetros ajustáveis que mudam e se adaptam à medida que o conjunto de treinamento é apresentado à rede. Assim, o processo de aprendizado supervisionado em uma RNA com pesos, resulta em sucessivos ajustes dos pesos sinápticos, de forma que a saída da rede seja a mais próxima possível da resposta desejada. O modelo neural também inclui um termo chamado de “bias”, aplicado externamente, simbolizado por b_k. O b_k tem o efeito do acréscimo ou decréscimo da função de ativação na entrada da rede, dependendo se é positiva ou negativa, respectivamente (Peixoto, 2013). Um neurônio biológico dispara quando a soma dos impulsos que ele recebe ultrapassa o seu limiar de excitação (*threshold*) ele é tomado fora do corpo do neurônio e conectado usando uma entrada adicional.

1.3.3. Funções de Ativação

Uma rede neural pode ser representada por funções de ativação $f(x, \alpha)$ referente a parte não linear de cada neurônio, onde f é a saída da rede, x é a variável de entrada e α os pesos (Mackay, 1994). Portanto as funções de ativação fornecem o valor da saída de um neurônio, elas limitam a saída do mesmo, geralmente nos intervalos de $[0,1]$ ou $[-1,1]$.

Estas funções são muito utilizadas já que são funções não lineares com um comportamento levemente linear, e assim é possível inserir a não linearidade para problemas de gravidade complexa e facilmente diferenciável, e assim obtendo estimadores de forma mais fácil (Haykin, 2001).

As principais funções de ativação são:

1.3.3.1. Função Limiar (Degrau)

Esta função foi utilizada no modelo de McCulloch e Pits e é expressa pela equação (1.3).

$$f(x) = \begin{cases} 1, se\ x \geq 0; \\ 0, se\ x < 0; \end{cases} \qquad (1.3)$$

Para esses neurônios, a saída de $f(x)$ será igual a zero se o valor de ativação x for negativo e 1 caso seja positivo (incluindo o zero).

1.3.3.2. Função “Sigmoidal” ou “Sigmóide” (logsig)

A função sigmóide é a função de ativação mais utilizada pelas Redes Neurais Artificiais (Haykin, 2001). Esta função pode assumir valores entre 0 e 1 e constitui-se de uma função monótona crescente com propriedades assintóticas e de suavidade. Um exemplo desse tipo de função pode ser visualizado pela equação (1.4).

$$f(x, \alpha) = \frac{1}{1 + e^{-\alpha x}} \qquad (1.4)$$

Em que α é o parâmetro da função sigmoide.

1.3.3.3. Função Tangente Hiperbólica

A função Tangente Hiperbólica (tansig) é uma função que se estende de -1 a +1 e assim assumindo uma forma não simétrica com relação a origem. Neste caso, utiliza-se uma forma correspondente à logsig que é definida pela equação (1.5).

$$\tanh\left(\frac{x}{2}\right) = \frac{(1 - e^{-x})}{(1 + e^{x})} \qquad (1.5)$$

1.3.4. Rede Perceptron Multicamadas

A rede neural mais utilizada para soluções de problemas de predição e classificação é conhecida como Perceptron Multicamadas (MLP). Essas redes possuem uma ou mais camadas de neurônios entre as camadas de entrada e saída conhecida como camada oculta (Sant'Anna, 2010).

A MLP também é conhecida como rede *feedforward* porque nessa rede o fluxo de informação se propaga pra frente, camada por camada, desde a camada de entrada até a camada de saída, como mostra a Figura 1Figura 3.

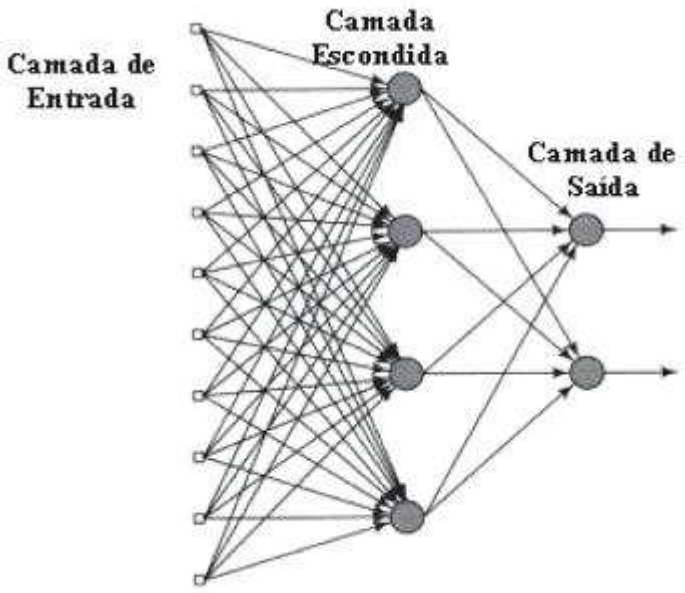

Figura 3- Exemplo de Rede Neural *Feedforward* Multicamadas.
Fonte: Santos et al., 2005.

Nesse modelo de rede, cuja maior importância se dá na resolução de problemas não linearmente separáveis, o aprendizado na fase de treinamento é o responsável pelo ajuste dos pesos. Geralmente é supervisionado e utiliza o

algoritmode retropropagação (*backpropagation*)em que o ajuste é feito por meio de um processo de correção do erro que é dado pela diferença entre a resposta da camada de saída e a desejada pela rede. (ZEIDENBERG, 1990).

1.4. Máquina de Vetor Suporte

1.4.1. Introdução

As maquinas de vetores suporte (MVS) são algoritmos de aprendizagem utilizados na área de aprendizagem de máquina fundamentados pela teoria do aprendizado estatístico (Mitchell, 2007). Tal abordagem foi proposta por Vapnik (1995) como um campo de pesquisa de Inteligência Computacional que estuda métodos para conseguir conhecimento por meio de amostras de dados. Segundo Lorena & Carvalho (2003), as técnicas de aprendizagem de máquinas são utilizadas de modo que, a partir de um conjunto de treinamento, pode ser capaz de prever uma classe qualquer do domínio em que foi treinado, logo é muito utilizado no reconhecimento de padrões. Dentro dessa área podemos dividi-la em duas partes: Aprendizado Supervisionado e não Supervisionado, neste trabalho estaremos lidando apenas com o aprendizado supervisionado em que tem-se a figura de um professor externo, o qual apresenta o conhecimento do ambiente por conjuntos de exemplos na forma: entrada e saída desejada (Lorena & Carvalho, 2003).

1.4.2. Teoria do Aprendizado Estatístico

A teoria do aprendizado estatístico visa estabelecer condições matemáticas que permitem escolher um classificador, com bom desempenho, para o conjunto de dados disponíveis para treinamento e teste (Lorena & Carvalho, 2003).

Uma escolha natural é definir, como função, aquela que produz o menor erro durante o treinamento, ou seja, aquela que possui uma maior capacidade de classificar corretamente os padrões os conjuntos de treinamento, denotado por S.

Logo, dado um classificador, devemos obter o menor erro durante o treinamento para melhor desempenho do mesmo, sendo esse erro mensurado

pelo número de predições incorretas de f. Sendo assim definimos o risco empírico $R_{emp}(f)$ como sendo a porcentagem de classificações incorretas obtidas em S. A equação (1.6) representa a definição do risco empírico.

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n} c(f(x_i), y_i) \qquad (1.6)$$

em que $c(\cdot)$ representa a função custo relacionada a previsão, $f(x_i)$, e com a saída desejada y_i (Muller et al., 2001). Um possível tipo de função custo é a conhecida como "função indicadora" definida pela equação (1.7).

$$c(f(x_i), y_i) = \begin{cases} 1, \text{se } y_i f(x_i) < 0 \\ 0, \text{caso contrário} \end{cases} \qquad (1.7)$$

O processo de busca por uma função, f', que apresente menor R_{emp} é denominado Minimização do Risco Empírico (Muller et al., 2001).

Pressupondo que os padrões de treinamento, (x_i, y_i), são gerados por uma distribuição de probabilidade, $P(x, y)$, independente e identicamente distribuídos (i.i.d.) em $\Re^n \times \{-1, +1\}$. O desempenho de generalização de um classificador f pode ser definido como a probabilidade de classificação incorreta do classificador f, logo denominamos Risco Funcional (ou Risco Esperado) $R(f) = P(f(X) \neq Y)$ a quantificação da capacidade de generalização de f[1] (Equação(1.8)) (Muller et al., 2001).

$$\mathrm{R}(f) = \int c(f(x), y) dP(x, y) \qquad (1.8)$$

No decorrer do processo de treinamento, o $R_{emp}(f)$ pode ser obtido de forma fácil, porém o mesmo não se pode dizer quanto ao $\mathrm{R}(f)$, pois em geral a distribuição de probabilidade $P(x, y)$ é desconhecida (Muller et al., 2001).

A ideia passa a ser minimizar o $R_{emp}(f)$ já que minimizar $\mathrm{R}(f)$ não é uma tarefa fácil devido a probabilidade muitas vezes desconhecida, espera-se com isso que ao minimizar o $R_{emp}(f)$ obtenha-se um menor $\mathrm{R}(f)$, o que nem sempre isso ocorre. Considere, por exemplo, um classificador que memoriza todos os dados de treinamento e gera classificadores aleatórios para outros exemplos, embora o risco empírico seja zero, o risco funcional será 0,5 (Smola, 1999).

[1] Se $\Re(f)$ for igual a 0, pode-se afirmar que f generaliza de forma "perfeita"

É possível no entanto encontrar uma f com pequeno risco empírico, porem existe o risco dos exemplos de treinamento se tornarem pouco informativos induzindo a um super ajustamento, nesse caso deve-se restringir a classe de funções da qual f^*é extraída, contudo, a teoria de aprendizado estatístico provê limites no risco esperado de uma função de classificação que podem ser empregados na escolha de um classificador (Smola, 2002), como será mostrado na seção (1.4.4.).

1.4.3. A dimensão Vapnik-Chervonenkis

A dimensão Vapnik-Chervonenkis (VC) é uma medida da capacidade ou poder expressivo da família de funções de classificação realizadas pela máquina de aprendizagem e pode ser definida para várias classes de funções. Em outras palavras a dimensão VC mede a complexidade intrínseca de uma classe de funções (Haykin, 1999).

A dimensão VC tem utilidade em teoria de aprendizagem estatístico, pois mede a complexidade provinda dos limites no risco funcional, relacionando o número de exemplos de treinamento, o risco empírico obtido na classificação dos dados e a complexidade do espaço de hipóteses. Está diretamente envolvido a problemas de classificação, geralmente considera-se h como sendo a dimensão VC do modelo de classificação, quanto maior o seu valor, mais complexas são as funções de classificação que podem ser induzidas a partir de G, sendo G o conjunto de todos os classificadores que um determinado algoritmo de aprendizado de máquinas pode gerar representado por funções sinais com fronteira linear, o que pode ser visto na Equação (1.11) (Haykin, 1999).

Na Figura 4, temos o caso de obtenção da dimensão VC para funções lineares no $\Re^2$, ou seja, retas. Nesse caso a dimensão VC é igual a 3, que é o número máximo de amostras que podem ser classificadas por uma reta para qualquer rotulação binária admitidas pelas amostras.

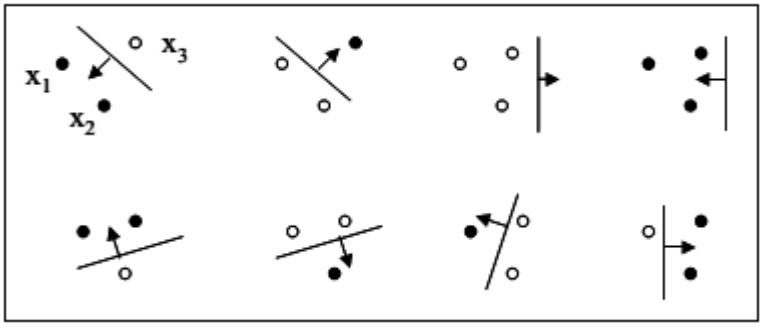

Figura 4-Rotulação de três amostras no $\boldsymbol{R^2}$.
Fonte: Lorena & Carvalho, 2003.

Para funções no R^n, $n \geq 2$, a dimensão VC é dada por $n + 1$, sendo n a dimensão de x. (Vapnik, 1995).

1.4.4. Minimização do Risco Empírico

Como visto anteriormente, existe uma busca pela minimização do erro empírico, $R_{emp}(f)$, para tal a lei dos grandes números garante sua convergência em probabilidade para $R(f)$, logo minimizar $R_{emp}(f)$ fornecera o mínimo de $R(f)$.

Porém, de acordo com a teoria de convergência uniforme em probabilidade, há um limitante para a variação do risco funcional $R(f)$ em função do risco empírico$R_{emp}(f)$. O limite sobre o erro de teste do modelo de classificação (Em dados de treinamento são independentes e cumprem uma distribuição aleatória da mesma distribuição), um limitante típico pode ser visto em (1.9) tal limitante é garantido com probabilidade $1 - \theta$, em que $\theta \in [0,1]$:

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{h\left(\ln\left(\frac{2n}{h}\right) + 1\right) - \ln\left(\frac{\theta}{4}\right)}{n}} \qquad (1.9)$$

Sendo o risco empírico, $R_{emp}(f)$, o erro devido ao treinamento, n é o tamanho do conjunto de treinamento e h a dimensão VC do modelo de classificação.

Sua maior importância se dá pelo fato de poder controlar a capacidade do conjunto de funções G do qual o classificador é extraído, apesar que na prática surgem alguns problemas como computar a dimensão h que pode não ser uma simples tarefa.

1.4.5. Maquinas de Vetor Suporte para Classificação

O problema de classificação pode ser tomado como um problema de duas classes, sem perda de generalidade (GUNN, 1998). O objetivo é fornecer um classificador que apresenta um bom desempenho para amostras não observadas durante o treinamento, isto é, com boa capacidade de generalização. Portanto a ideia principal da SVM é criar um hiperplano de separação como superfície de decisão de modo que a separação entre seus exemplos positivos e negativos sejam máxima (Haykin, 2001). Observando a Figura 5, para uma classe de funções lineares, diante dos possíveis hiperplanos de separação somente uma é escolhida como a melhor função que maximiza a margem de separação (distância do classificador e da amostra mais próxima de cada classe) representado como a reta de cor vermelha. Esse hiperplano com margem máxima é chamado de hiperplano de separação ótimo (ou hiperplano ótimo), que será o objeto de busca do classificador (Gunn, 1988).

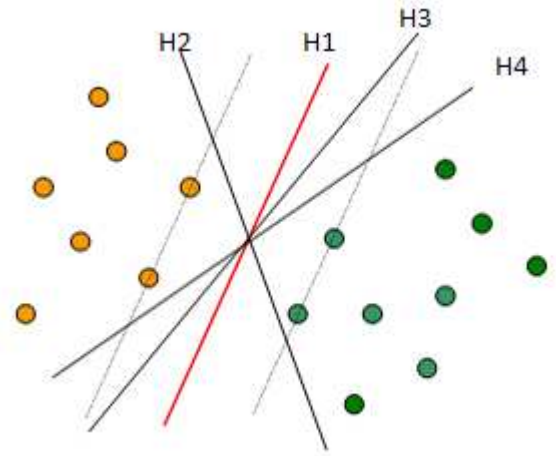

Figura 5- Possíveis hiperplanos de separação e um hiperplano ótimo (H1).

Fonte: Adaptado do trabalho de graduação de Gilson Medeiros de Oliveira Junior[2].

[2] Disponível em <http://www.cin.ufpe.br/~tg/2010-2/gmoj.pdf>
Acesso em fev. de 2016.

1.4.6. Hiperplano de Separação Ótimo

A ideia principal de uma máquina de vetor suporte é criar um hiperplano de separação como superfície de decisão. Em um espaço n-dimensional, um hiperplano é um subespaço desse plano de dimensão n-1.

Um conjunto de vetores, $X_1, X_2, \dots, X_n$, é dito ser separado otimamente pelo hiperplano se ele é separado sem erro e se a distância do hiperplano à(s) amostra(s) mais próxima(s) de uma classe, somada à distância do hiperplano à(s) amostra(s) mais próxima(s) da outra classe, denominada margem de separação é máxima (Lima, 2004). Essas amostras servem de base para construção da margem de separação e por isso são denominadas de vetores suporte.

A definição matemática para hiperplano é feita de modo simples, considere o problema de separação do conjunto de n vetores de treinamento pertencentes a duas classes linearmente separáveis:

$$(x_1, y_1), \dots, (x_n, y_n), com\ x \in \Re^n\ e\ y \in \{-1, +1\}.$$

Em que x é o vetor de entrada e y é a classificação desejada, portanto assumindo um hiperplano linearmente separável, o hiperplano de separação ótimo pode ser definido na forma:

$$w \cdot x + b = 0 \qquad (1.10)$$

Em que w é o vetor de peso ajustável e b é um viés.

A equação (1.10) divide o espaço dos dados X em duas regiões: $w \cdot x_i + b > 0$ e $w \cdot x_i + b < 0$. Uma função sinal $g(x) = sgn\big(f(x)\big) = sgn(w \cdot x + b)$ pode ser então aplicada sobre essa função, levando a classificação +1 se $f(x) > 0$ e -1 se $f(x) < 0$, como definido na equação (1.11) (Chang et al, 2011).

$$g(x) = sgn\big(f(x)\big) = \begin{cases} +1 \text{ se } w \cdot x_i + b > 0 \\ -1 \text{ se } w \cdot x_i + b < 0 \end{cases} \qquad (1.11)$$

Caso se possa classificar os conjuntos de treinamento de forma linearmente separável, então a SVM obtida é denominada de margens rígidas (Howlett, 2000).

Para facilitar ainda mais o entendimento, podemos reescrever a equação (1.11) da seguinte maneira:

$$(w \cdot x_i) + b \geq +1, \text{para} y_i = +1$$
$$(w \cdot x_i) + b \leq -1, \text{para} y_i = -1 \qquad (1.12)$$

Em que y_i a $sgn(f(x))$, dado $f(x_i)$=$(W \cdot x_i) + b$

A Figura 6 mostra um exemplo de como se comporta a função acima diante de um conjunto de dados.

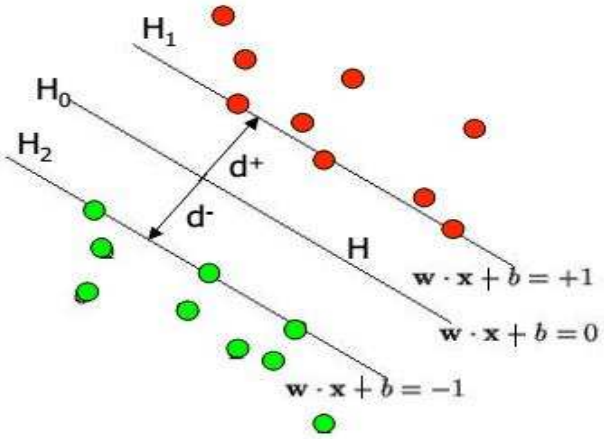

Figura 6- Conjunto de dados linearmente separável com o hiperplano e a margem de separação.
Fonte: Um guia para máquina de vetor suporte[3]

Em que d_+ é a menor distância do ponto positivo mais próximo e d_- é a menor distância do ponto negativo mais próximo (Vetores suportes).

Podemos combinar a equação (1.12) em um único conjunto de desigualdades, perceba que $(w \cdot x_i) + b \geq 0$ caso $y_i = +1$ e será negativo caso $y_i = -1$, logo sem erro de concordância podemos definir uma importante desigualdade como segue:

$$y_i[(w \cdot x_i) + b)] \geq 1, i = 1,2, \ldots, n \qquad (1.13)$$

A função da decisão é invariante sob uma alteração de escala positiva dentro da função sinal, conduzindo a uma ambiguidade em definir uma medida de distância e consequentemente a margem, assim nós definimos implicitamente uma escala para(w, b), assim da Figura 6 ajusta-se $w \cdot x + b = +1$ para os menores pontos representada pela cor vermelha e $w \cdot x + b = -1$ para os menores pontos representado pela cor verde. Os hyperplanos passando por $w \cdot$

[3] Disponível em: <http://www.svms.org/tutorials/Berwick2003.pdf>
Acesso em fev. de 2016

$x + b = +1$ e $w \cdot x + b = -1$ são chamados de hiperplanos canônicos e a região entre esses hiperplanos é chamada de margem de separação. Seja um ponto x_1 e x_2 dentro de cada hiperplano canônico (H_1 e H_2) como ilustrado na Figura 6, Projetando $x_1 - x_2$ na direção de w, perpendicular ao hiperplano separador $w \cdot x + b = 0$, é possível obter a distância entre os hiperplanos H_1 e H_2 (Kasabov, 2013; Lorena & Carvalho, 2003). Que pode ser visto pela equação (1.14):

$$(x_1 - x_2)\left(\frac{w}{||w||} \cdot \frac{(x_1 - x_2)}{||x_1 - x_2||}\right) \qquad (1.14)$$

Segue que, se $w \cdot x_1 + b = +1$ e $w \cdot x_2 + b = -1$ nos podemos deduzir pela diferença entre essas equações obtendo $w \cdot (x_1 - x_2) = 2$. Substituindo esse resultado pela equação (1.14) obtemos:

$$\left(\frac{2}{||w||} \cdot \frac{(x_1 - x_2)}{||x_1 - x_2||}\right) \qquad (1.15)$$

Logo, para obter o comprimento do vetor desejado, toma-se a norma da equação (1.15), e assim obtendo-se $2/||w||$ que é a distância d ilustrada na Figura 6. Portanto a metade da distância entre os dois hiperplanos canônicos pela margem é dado por $1/||w||$ (Kasabov, 2013; Lorena & Carvalho, 2003). Portanto para maximizar a margem de separação devemos minimizar $||w||$, isso decorre de um problema de otimização da forma:

$$\Phi(w) = \frac{1}{2}\|w\|^2 \qquad (1.16)$$

Sujeita as restrições impostas pela equação (1.13). O problema de minimização da equação (1.16) é um problema clássico de otimização, denominado programação quadrática (Hearst et al. 1998). Tal problema pode ser resolvido utilizando-se do método de multiplicadores de Lagrange (Lima, 2002).

Logo, por meio dos multiplicadores de Lagrange, podemos representar o problema (1.16), como:

$$\Phi(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i (y_i[w \cdot x_i + b] - 1) \qquad (1.17)$$

em que $i = 1, 2, \ldots, n$ e α_i são os multiplicadores de Lagrange. Logo, temos o problema de minimizar $\Phi(w, b, \alpha)$, com relação a w e b e a maximização dos

$\alpha_i \geq 0$ (Lorena & Carvalho, 2003). O problema dual equivalente assume a forma (BAZAARA et al., 1993):

$$\max_{\alpha} W(\alpha) = \max_{\alpha} \left(\min_{w,b} \Phi(w, b, \alpha) \right) \tag{1.18}$$

O método dos multiplicadores de Lagrange encontra os pontos ótimos igualando as deridadas parciais a zero que são obtidos por meio da resolução das seguintes igualdades apresentadas em (1.19):

$$\begin{cases} \dfrac{\partial \Phi}{\partial b} = 0 \\ \dfrac{\partial \Phi}{\partial w} = 0 \end{cases} \tag{1.19}$$

Sendo

$$\frac{\partial \Phi}{\partial w} = \left(\frac{\partial \Phi}{\partial w_1}, \frac{\partial \Phi}{\partial w_2}, \frac{\partial \Phi}{\partial w_3}, \dots, \frac{\partial \Phi}{\partial w_n} \right) \tag{1.20}$$

E a partir das equações em (1.19) obtém-se:

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \alpha_i y_i = 0 \\ w = \displaystyle\sum_{i=1}^{n} \alpha_i y_i x_i \end{cases} \tag{1.21}$$

Assim, das equações (1.17), (1.18) e (1.21) temos o seguinte problema dual:

$$\max_{\alpha} W(\alpha) = \max_{\alpha} \left(\sum_{l=1}^{n} \alpha_l - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j \right) \tag{1.22}$$

Logo, a solução do problema é dada por:

$$\alpha^* = \arg\min_{\alpha} \left(\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j - \sum_{l=1}^{n} \alpha_l \right) \tag{1.23}$$

Com as seguintes restrições:

$$\begin{cases} \alpha_i \geq 0, i = 1, \dots, n \\ \displaystyle\sum_{j=1}^{n} \alpha_j y_j = 0 \end{cases} \tag{1.24}$$

Resolvendo a equação (1.23) levando em consideração as restrições impostas na equação (1.24), a solução do multiplicador de lagrange e do hyperplano de separação ótimo é dado por:

$$w^* = \sum_{i=1}^{n} \alpha_i^* x_i y_i \qquad (1.25)$$

Portanto, o classificador assumira a seguinte formula:

$$f(x) = \text{sgn}(w^* \cdot x + b^*) \qquad (1.26)$$

Em que $\text{sgn}(\cdot)$ é a função sinal, definida por:

$$\text{sgn}(y) = \begin{cases} +1, \text{se } y \geq 0 \\ -1, \text{se } y < 0 \end{cases} \qquad (1.27)$$

Portanto a classificação é dada apenas pelo cálculo do produto interno entre o novo padrão e todos os vetores suporte, somente os pontos x_i que satisfazem $y_i(w^* \cdot x_i + b^*) = 1$ terão multiplicador de Lagrange diferente de zero e assim são denominados de vetores suporte (VS). O valor de b^* também pode ser obtido utilizando as equações de Karush-Kuhn-Tucker (KKT) (Dutra, 2011; Lima, 2004)

$$\alpha_i^*(y_i[(w \cdot x_i) + b] - 1) = 0, \qquad i = 1, \ldots, n \qquad (1.28)$$

Com isso tiramos a conclusão de que se as amostras de treinamento são linearmente separáveis, então os vetores suporte estarão a uma distância unitária do hiperplano ótimo, de modo que se tenha um número de vetores suportes muito pequenos. Logo conclui-se que o hiperplano ótimo é obtido pelos vetores suportes e é determinado por um pequeno subconjunto de amostras de treinamento e as demais amostras acabam por não participar da definição do hiperplano ótimo (Lorena & Carvalho, 2003).

Já o valor de b^* é calculado a partir dos vetores suportes e com as condições impostas por (1.28) (Smola et al, 2002; Lorena & Carvalho, 2003)

$$b^* = -\frac{1}{2}\left[\max_{\{i|y_j=-1\}}\left(w^* \cdot x_j\right) + \min_{\{i|y_i=+1\}}(w^* \cdot x_i)\right] \qquad (1.29)$$

Em que $w^* = \sum_{x_i \in SV} y_i \alpha_i^* x_i$

1.4.7. Máquina de Vetor Suporte para dados não lineares

Existem muitos casos em que não é possível dividir os dados de treinamento em um hiperplano, ou seja, não é possível criar um hiperplano de separação sem que aja erros de classificação.

Para se ter uma ideia, um conjunto de dados é dito ser não linearmente separável, caso não seja possível separar os dados com um hiperplano. A Figura 7 mostra um conjunto linearmente e outro não linearmente separável.

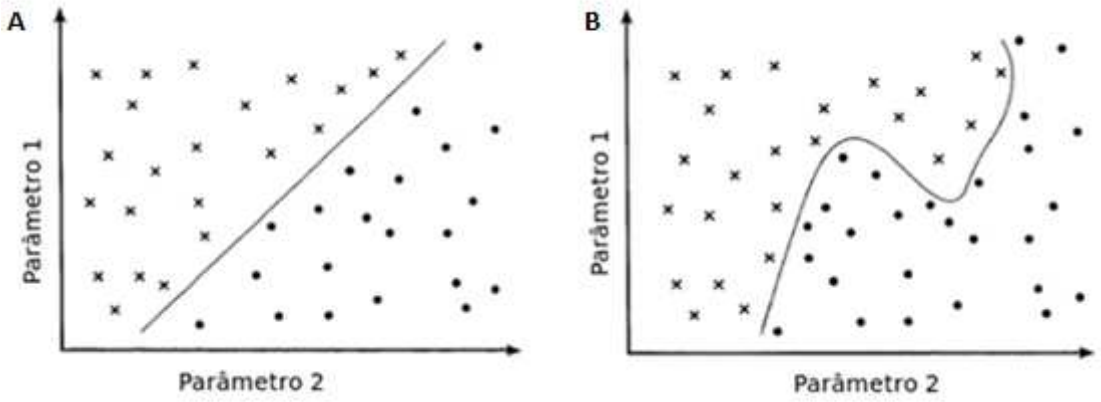

Figura 7- Exemplo de padrões linearmente (A) e não linearmente (B) separável respectivamente.
Fonte: Adaptado de Gonçalves, 2012.

Apesar desse entrave, é possível encontrar um hiperplano ótimo que minimize a probabilidade de erro de classificação, calculado sobre o conjunto de treinamento.

Assim, para um conjunto de dados não linearmente separáveis, é necessária a introdução de um novo conjunto de variáveis escalares não-negativas ε_i de forma a obter:

$$y_i(w \cdot x_i + b) \geq 1 - \varepsilon_i \qquad (1.30)$$

Em que $\varepsilon_i \geq 0$ e $i = 1, \dots, n$.

As ε_i são chamada de variaveis soltas ou de folga e medem o desvio de um ponto dado da condição ideal de separabilidade de padrões, ou seja, indica o erro de classificação associada a cada amostra. Esse procedimento suavisa as margens do classificador linear, permitindo que alguns dados permaneçam entre os hiperplanos de separação e assim acabando por ter alguns erros de

classificação. As SVMs obtidas nesse caso são conhecidas como SVMs de margens suaves (Bem-Hur et al., 2008).

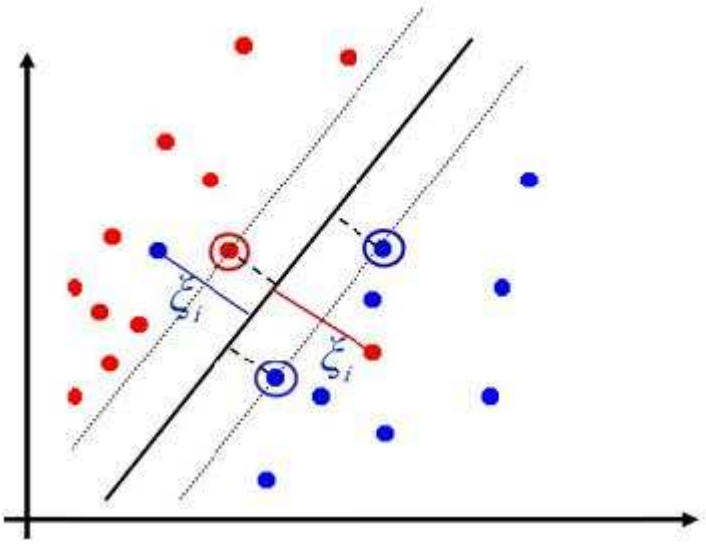

Figura 8- Exemplo ilustrativo de variáveis soltas.
Fonte: Tese de Danilo Althmann Maretto[4].

Observe na Figura 8 que para $\varepsilon_i > 1$, o hiperplano de separação não esta bem definido, ou seja, houve um erro no conjunto de treinamento e para $0 < \varepsilon_i < 1$, o hiperplano de separação estará bem definido. Para desencorajar o uso de variaveis de tranformação, já que a soma de ε_i significaria o limite do número de erros de treinamento, uma constante C em $\sum_{i=1}^{n} \varepsilon_i$ é introduzido na função para ser otimizada, logo:

$$\underset{w,b}{minimizar}\ \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n} \varepsilon_i\right) \qquad (1.31)$$

Em que C é um valor a ser arbitrado pelo pesquisador, alguns autores como Girosi (1997) e Smola & Schölkopf (1998) relacionam em algumas circunstâncias a constante C como uma função de regularização, que impõe um peso à minimização dos erros no conjunto de treinamento em relação à minimização da complexidade do modelo (Lorena & Carvalho, 2003). O efeito da escolha de C pode ser visto na Figura 9.

[4] Disponível em: <http://www.cprm.gov.br/publique/media/Tese_Danilo_Maretto.pdf> Acesso em fev de 2016.

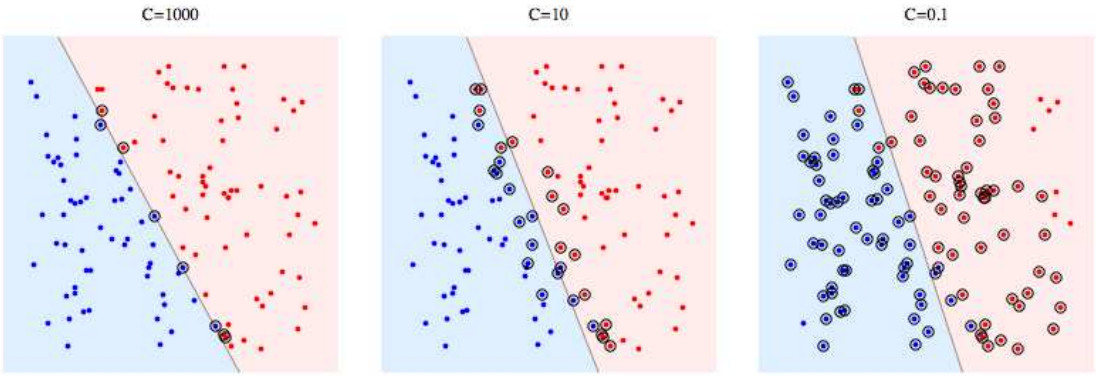

Figura 9- Efeito da margem suave da constante C.
Fonte: Pagina de programadores no stack overflow[5]

Da Figura 9 vemos que os pontos mais circulares são os vetores suportes, quando a constante C diminui isso faz com que os pontos se movam para dentro da margem, ou seja, ele sacrifica um pouco o separador linear para ganhar estabilidade mas quanto mais diminui-se a constante, mais o hiperplano sofre uma mudança de orientação e assim permitindoignorar ospontospróximos da fronteira, aumentandoa margem.

Logo, teremos um outro problema de otimização da mesma forma apresentada na equação (1.22) mas porém com uma nova restrição, ou seja:

$$\max_{\alpha}\left(\sum_{l=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i x_j\right) \tag{1.32}$$

Com as restrições:

$$\begin{cases} 0\le\alpha_i\le C, i=1,\dots,n \\ \sum_{j=1}^{n}\alpha_j y_j=0 \end{cases} \tag{1.33}$$

Tem-se como resultado final a mesma função encontrada em (1.26) com mesmo tipo de classificação, porém as variáveis α_i^* são determinadas pela solução da equação (1.32) com as restrições impostas em (1.33).

[5] Disponível em: <http://stackoverflow.com/questions/4629505/svm-hard-or-soft-margins> Acesso em fev. 2016.

1.4.8. Funções Kernels para Classificadores não Lineares

As SVMs lidam com os problemas não lineares mapeando o conjunto de treinamento de seu espaço original, para um espaço de mais alta dimensionalidade denominado de espaço de características (Smola et al., 1999b).

Neste caso, um problema não linear pode ser transformado com alta probabilidade em um problema linearmente separavel, em um espaço de maior dimensionalidade[6]. A SVM não linear realiza essa mudança de dimensionalidade por meio de funções conhecidas como Kernels e assim caindo em um problema de classificação linear e podendo achar um hiperplano ótimo.

A Figura 10 ilustra como é possível aumentar a dimensionalidade do espaço de entrada saindo de um dominio não linear para um espaço de dimensão linearmente separavel, onde é feito o mapeamente por meio de uma função Kernel (Gunn, 1998).

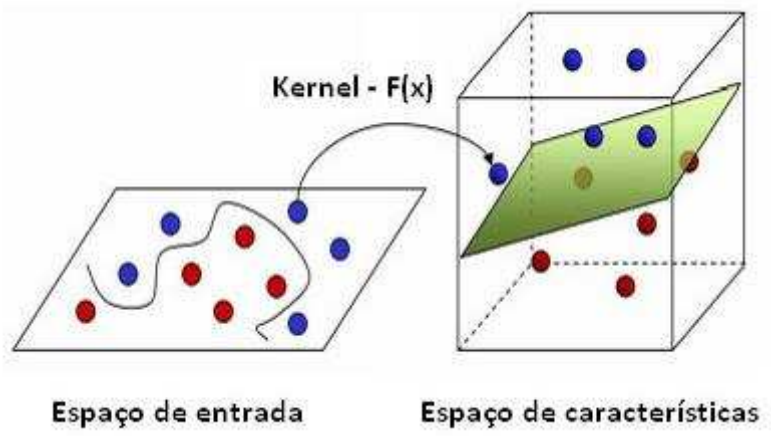

Figura 10- Transformação de um problema não linear em um problema linearmente separável.
Fonte: Trabalho de graduação de Gilson Medeiros de Oliveira Junior[7].

Transformando os dados de R^2 para R^3 como visto na Figura 10, o conjunto não linear se torna linearmente separavel. Seja $\varphi(x)$ representando uma função de kernel, é então possível encontrar um hiperplano capaz de separar esses dados, de tal forma que:

[6]Teorema de Cover
[7] Disponível em: <http://www.cin.ufpe.br/~tg/2010-2/gmoj.pdf>
Acessado em fev de 2016

$$\varphi(x) = \varphi(x_1, x_2) = \left(x_1^2, \sqrt{2x_1x_2}, x_2^2\right) \quad (1.34)$$

$$f(x) = w \cdot \varphi(x) + b = 0 \quad (1.35)$$

Logo, os dados são colocados em um espaço de maior dimensão utilizando o Kernel, $\varphi(x)$, e assim aplica-se a SVM linear sobre esse espaço. Neste caso o uso de uma SVM linear com margens suaves permite lidar com ruídos e *outliers* presentes nos dados, já que segundo Lorena & Carvalho (2003) um dos requisitos importante para as técnicas de aprendizagem de máquina é que sejam capazes de lidar com dados imperfeitos (ruídos) e minimizar a presença de outliers no processo de indução, e logo temos um outro problema de otimização com o mesmo formato visto na equação (1.32) mas porem aplicando-se $\varphi(x)$ ao problema de otimização representado na equação (1.33), como mostra a seguir:

$$\max_{\alpha}\left(\sum_{l=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left(\varphi(x_i)\cdot\varphi(x_j)\right)\right) \quad (1.36)$$

Sob as mesmas restrições impostas na equação (1.34), o classificador extraído assume a forma:

$$f(x) = sgn\left(\sum_{x_i \in SV} y_i\alpha_i^*\varphi(x_i)\cdot\varphi(x) + b^*\right) \quad (1.37)$$

Em que b^* é adaptado da equação (1.29) com as restrições impostas da equação (1.34) e assim se torna:

$$b^* = -\frac{1}{2}\left[\max_{\{i|y_j=-1\}}\left(w^*\cdot\varphi(x_j)\right) + \min_{\{i|y_i=+1\}}(w^*\cdot\varphi(x_i))\right] \quad (1.38)$$

Com $w^* = \sum_{x_i \in SV} y_i\alpha_i^*\varphi(x_i)$

Observe nas equações (1.36) e (1.37) o produto $\varphi(x_i)\cdot\varphi(x_j)$, para dois dados x_i e x_j, em conjunto. Esse produto é obtido por meio de funções denominadas Kernels. A função Kernel representa a relação entre o dado de entrada e a propriedade de saída a ser modelada (Howlett, 2000). Tem-se então:

$$K(x_i, x_j) = \varphi(x_i)\cdot\varphi(x_j) \quad (1.39)$$

No entanto, é necessário que a função $\varphi(\cdot)$ pertença a um domínio, onde seja possível o cálculo do produto interno e que também satisfaçam as condições do Teorema de Mercer[8] (Smola et al., 1999).

Segundo esse teorema, uma função é dita ser uma função Kernel, se a matriz K é positivamente definida, onde K é obtido por:

$$K = K_{ij} = K(x_i, x_j) \tag{1.40}$$

Uma matriz é dita ser positivamente definida, se seus autovalores são maiores que zero. As funções que satisfazem as condições do Teorema de Mercer, são chamadas de Kernels de Mercer (Smola et al., 1999)

Alguns dos Kernels mais utilizados na prática são os Polinomiais, os Gaussianos ou RBF (*Radial-Basis Function*) e os Sigmoidais (Tabela 1). Cada um deles contém parâmetros em que devem ser determinados previamente pelo pesquisador.

Tabela 1- Tipos de Kernels mais conhecidos.

Tipo de Kernel	Função $K(x_i, x_j)$	Tipo de classificador
Polinomial	$\gamma(x_i \cdot x_j + \alpha)^d$	Máquina de aprendizagem polinomial
Gaussiano (RBF)	$exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$	Rede RBF
Sigmoidal	$tanh(\beta_0(x_i \cdot x_j) + \beta_1)$	Perceptron de Duas Camadas

Porém algumas considerações devem ser tomadas sobre as funções Kernels, de acordo com Haykin (1999), na máquina de aprendizagem polinomial a potência d,γ e α são especificados a priori pelo usuário; Na Rede RBF a amplitude $\frac{1}{2\sigma^2}$, comum a todos os kernels, é especificada pelo usuário; No Perceptron de duas camadas o teorema de Mercer é satisfeito apenas para alguns valores de β_0 e β_1.

Os valores dos parâmetros a se considerar nas kernels são muito importantes, pois se ajustar um valor alto, o raio da área de influência dos vetores suporte incluirá apenas o próprio vetor do suporte e assim ocasionando um superajuste (*overfitting)*. Se for muito pequeno, o modelo fica muito limitado e não poderá captar a complexidade dos dados.

[8]Teorema de Mercer: Éa representaçãode uma função simétrica, positiva definida das somas de quadrados de uma sequência de produtos de funções convergêntes.

Logo, para obtenção de uma SVM, para o caso de modelos não lineares, devemos obter a função kernel, os parâmetros da função e a definição do algoritmo para determinação do hiperplano ótimo (Lorena & Carvalho, 2003).

1.4.9. Classificação via Multiclasses

As SVMs foram inicialmente utilizadas para problemas de classificação binária mas em muitos casos, quando lidamos com dados reais deparamos com problemas mais complexos que envolve classificação multiclasses, muitos métodos então foram propostos para se lidar com isso (Mayoraz, 1999).

Um método proposto para resolução de tal impasse foi a extensão do classificador binário para um de n classes. Para tal, temos que em um sistema multiclasses o conjunto de treinamento é composto por pares (x_i, y_i), tal que $y_i \in \{1, \cdots, k\}, \text{com } k > 2$, onde k é o número de classes.

Em problemas que envolvem mais de 2 classes, destacaremos um dos métodos mais comuns utilizados: Decomposição “Um-Contra-Todos”.

1.4.9.1. Decomposição Um-Contra-Todos

No método “Um-Contra-Todos” (1-c-t) a abordagem consiste na geração de k SVMs, ou seja, classificadores binários onde k é o número de classes (Lorena & Carvalho, 2003).

A ideia desse método é que para cada k SVM criada, uma classe é fixada como positiva e as restantes como negativas, independente do aprendizado utilizado no treinamento dos classificadores, e logo dado um novo padrão x a classe no qual este novo padrão pertencera será o que obtiver a saída com valor máximo entre os k classificadores.

É formalmente definido como:

$$f(x) = \underset{1 \le i \le k}{\operatorname{argmax}}(w_i \cdot \varphi(x) + b_i) \qquad (1.41)$$

Porém o método 1-c-t tem a desvantagem de não ser possível prever limites no erro de generalização através de seu uso e o tempo de treinamento é geralmente longo (Lorena & Carvalho, 2003).

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, T. W. An Introduction to Multivariate Statistical Analysis.**New York: John Wiley & Sons**, 345 p., 1958.

BAZARAA, M.S., SHERALI, H.D.,SHETTY, C.M. **Nonlinear Programming Theory and Algorithms**. 2nd Edn., Wiley, New York, 1993.

Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G.Support Vector Machines and Kernels for Computational Biology. **Editor: Fran Lewitter, Whitehead Institute**, United States of America. 2008.

BRAGA, A. P.; CARVALHO, A. C. P. L. F.; LUDERMIR, T. B. **Redes neurais artificiais: teoria e aplicações**. Rio de Janeiro, TCL – Livros Técnicos e Científicos, 262 p., 2000.

BRAGA, A. P.; FERREIRA, A. C. P. L.; LUDERMIR, T. B. **Redes neurais artificiais: teoria e aplicações**. LTC Editora, 2007.

BRAGATTO, T. A. C. ; RUAS, G. I. S. ; LAMAR, M. V. . Uma comparação entre redes neurais artificiais e máquinas de vetores de suporte para reconhecimento de posturas manuais em tempo-real. In: **VIII Congresso Brasileiro de Redes Neurais**, Florianópolis. Anais do VIII CBRN, 2007.

BURGES, C. J. C. A tutorial on support vector machines for pattern recognition. **Knowledge Discovery and Data Mining**, 2 (2): 1-43. 1998.

C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines. **ACM Transactions on Intelligent Systems and Technology**, 2:27:1--27:27, 2011.

HOWLETT, R. J.; JAIN, L. C. Radial Basis Function Networks: Design and Applications. **Physica-Verlag, Wurzbury**, 2000.

CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. 2011. A**.** Biometria aplicada ao estudo da diversidade genética. **Suprema, Visconde do Rio Branco**, 620p. 2011.

FERREIRA, D. F. **Estatística Multivariada**, 2ª ed. Editora UFLA, 2011.

GUNN, S. Support vector machine for classification and regression, Technical Report ISIS-1-98, **Image Speech & Intelligent Systems Group**, University of Southampton, 1998.

GONALVES, A. R. (2012). **Máquina de vetores suporte**. <www.dca.fee.unicamp.br/~andreric/arquivos/pdfs/svm.pdf>. Ultimo acesso: Fevereiro de 2016.

HAYKIN, S; NETWORK, Nl. A comprehensive foundation. **Neural Networks**, v. 2, n. 2004, 2004.

HAYKIN, S. **Redes Neurais, Princípios e prática.**2. Ed. [S. l.]: Bookman, 1999.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. Support vector machines. **IEEE Intelligent Systems and their Applications**, v. 13, n. 4, p. 18-28, 1998.

Kijsirikul, B. and Ussivakul, N. Multiclass support vector machines using adaptive directed acycle graph. In. **Proceedings of International Joint Conference on Neural Networks (IJCNN 2002)**, pages 980-985, 2002.

LIMA, A. R. G. Máquinas de Vetores Suporte na Classificação de Impressões Digitais. **Dissertação (Mestrado) — Universidade Federal do Ceará**, Fortaleza, Ceará, 2002.

LIMA, C. A. M. Comitê de Máquinas: Uma Abordagem Unificada Empregando Máquinas de Vetores-Suporte. **Tese (Doutorado) — Universidade Estadual de Campinas**, São Paulo, 2004.

LORENA, A. C. ; CARVALHO, A. C. P. L. F.; Introdução às Maquinas de Vetores Suporte. Relatórios técnicos do icmc**.** Nº 192. **Instituto de Ciências Matemáticas e de Computação**. ISSN - 0103-2569. São Carlos – SP, Ano 2003.

MACKAY, D. J. C. Bayesian nom-linear modelling for the prediction competition. In: **ASHRAE Transaction**, ASHRAE, Atlanta Georgia. Vol. 100, pp. 1053-1062, 1994.

MARTINS, R. S.; DUARTE, V. J. L.; MAITELLI, A. L.; SALAZAR, A. O. e DÓRIA NETO, A. D. Sistemas de Detecção de Vazamentos em Dutos Usando Redes Neurais e Máquinas de Vetor de Suporte. **Anais do VIII Congresso Brasileiro de Redes Neurais**, pp. 1-6, Florianópolis-SC, outubro 2007.

MAYORAZ, E.; ALPAYDIN, E.; Support vector machines for multi-class classification. In: **International Work-Conference on Artificial Neural Networks**. Springer Berlin Heidelberg, p. 833-842.1999.

Muller, K. R., Mika, S., Ratsch, G., Tsuda, K., and Scholkopf, B. An introduction to kernel-based learning algorithms. **IEEE Transactions on Neural Networks**, 12(2):181–201, 2001.

PLATT, John C. 12 fast training of support vector machines using sequential minimal optimization. **Advances in kernel methods**, p. 185-208, 1999.

RYCHETSKY, Matthias. **Algorithms and architectures for machine learning based on regularized neural networks and support vector approaches**. Shaker, 2001.

SANT'ANNA, I. C.; Tomaz, Rafael Simões; SILVA, G. N.; BHERING, L. L.; NASCIMENTO, M.; CRUZ, C. D.Artificial neural networks in genetic classification. **Genetics and Molecular Research**, 2014.

Smola, A. J. and Schölkopf, B. Learning with Kernels. **The MIT Press**, Cambrige, MA, 2002.

Smola, A. J., Barlett, P., Schölkopf, B., and Schuurmans, D. Introduction to Large Margin Classifiers, **The MIT Press Cambrige**, Massachusetts London, England. 1999.

VAPNIK, Vladimir. **The nature of statistical learning theory**. Springer Science & Business Media, 2013.

VAPNIK, V; CHERVONENKIS A. On the uniform convergence of relative frequencies of events to their probabilities. **Theory of Probability and its Applications**, 16(2):264--280, 1971.

ZEIDENBERG, M.;Neural Networks in Artificial Inteligence. **Ellis Horwood Series in Artificial Intelligence**, 268p. 1990.

CAPÍTULO 1

Máquina de vetor suporte na discriminação de populações genéticas com diferentes graus de similaridade.

Resumo: É importante para a preservação da variabilidade genética e da biodiversidade a correta classificação dos indivíduos. Técnicas tradicionalmente utilizadas nessas situações são as funções discriminantes de Fisher e de Anderson que permitem a alocação de um indivíduo dentro de uma população conhecida com características semelhantes. No entanto, há situações em que tais métodos não são tão eficazes em detectar diferenças entre populações como é o caso de populações com altos níveis de similaridade. Métodos computacionais, como as Redes Neurais Artificiais (RNAs) e a Máquina de Vetor Suporte (*Support Vector Machines - SVMs*), vêm ganhando notável destaque na solução de problemas mais complexos. Em especial, este último, vem adquirindo grande atenção e credibilidade por seus resultados em diversas tarefas envolvendo o reconhecimento de padrões, contudo seu uso no que tange a classificação de populações com auto nível de similaridade ainda não foi estudado. O objetivo deste trabalho foi utilizar aSVM na obtenção de uma solução para um problema de discriminação de populações com autos níveis de similaridade e compará-la com as RNAs e com a análise discriminante de Anderson por meio da taxa de erro aparente. Os resultados obtidos por meio da SVM foram equivalentes aos obtidos pela análise discriminante de Anderson e não superaram a eficiência obtida pelas RNAs. Contudo, são necessários mais estudos a respeito da técnica de SVM, tal como a ampliação do algoritmo de busca com a finalidade de otimizar os parâmetros do modelo almejado, para confirmar sua viabilidade para a classificação de populações altamente similares.

1. INTRODUÇÃO

As análises da diversidade genética têm orientado na escolha de genitores apropriados em programas de melhoramento, levando à otimização dos ganhos seletivos, devido à variabilidade encontrada nos grupos divergentes. A diversidade genética também tem sua importância na quantificação da variabilidade existente, a fim de diminuir tempo e recursos, facilitando o gerenciamento dos bancos de germoplasma (Sant'Anna, 2014).

Uma ferramenta bastante utilizada em estudos de diversidade provem de métodos baseado na área de estatística multivariada, dentre estes, podem-se destacar as análises discriminantes, bastante utilizadas em estudos de discriminação de objetos com o objetivo de separá-los em duas ou mais classes e assim utiliza-las para classificar um indivíduo ou grupo de indivíduos em diferentes populações, sendo elas conhecidas ou não (Cruz et al., 2011).

No entanto, é de se esperar que existam situações nas quais os métodos convencionais se tornem pouco satisfatórios para a resolução do problema, já que nem sempre a técnica convencional é capaz de detectar as diferenças entre populações não linearmente separáveis, mesmo dispondo dos dados experimentais (Martins et al., 2007).

Visando obter um solução para tal limitação Sant'Anna et al. (2014) propuseram o uso das RNAs para a classificação de indivíduos. Os autores compararam diferentes funções discriminantes (Fisher e Anderson) e redes neurais artificiais em termos do número de classificações incorretas de indivíduos sabidamente pertencentes a diferentes populações simuladas de retrocruzamentos, com crescentes níveis de similaridade e verificaram a eficiência das redes frentes as demais técnicas utilizadas.

Outra abordagem que pode ser utilizada para a solução de problemas de classificação é a Máquina de Vetores de Suporte (SVM, do inglês *support vector machines*). Tal metodologia, a qual é fundamentada na Teoria da Aprendizagem Estatística (Vapnik, 1995), utiliza apenas algumas observações, denotadas por vetores de suporte, para a obtenção de uma regra de classificação, sendo então mais robusta a ruídos quando comparadas a técnicas tradicionais tais como de estatística multivariada e redes neurais.

As SVMs foram utilizadas em diversas áreas de conhecimento com sucesso, como exemplo Ticiano et al. (2007) utilizaram redes neurais artificias e SVM para reconhecimento de posturas manuais em tempo-real. Martins(2007) usou redes neurais e máquinas de vetor suporte para criar um sistemas de detecção de vazamentos. Entretanto, não se encontra na literatura a utilização de máquina de vetor suporte para estudos de diferenciação de populações.

Diante do exposto, este trabalho tem por objetivo a utilização da técnica de SVM para um problema de discriminação de populações com autos níveis de similaridade e em seguida e compará-la com as RNAs e com a análise discriminante de Anderson por meio da taxa de erro aparente (TEA).

2. MATERIAL E MÉTODOS

2.1. O métodos de simulação dos dados:

A simulação dos dados genotípicos de populações estruturadas no delineamento genético de retrocruzamentos foram obtidos pelo uso do programa computacional Genes (Cruz, 2013), desenvolvido pelo laboratório de Bioinformática da Universidade Federal de Viçosa, localizado no instituto de Biotecnologia aplicada a Agropecuária (BIOAGRO).

Inicialmente, foram simulados 10 populações em equilíbrio de Hardy-Weinberg (Cruz, 2013) para os dados genotípicos com 100 indivíduos cada. Foram geradas então para o cálculo da medida de dissimilaridade fenotípica de Nei (Nei, 1972) informações relativas a 50 locos manifestando dois alelos codominantes. Foi considerada para cada simulação da população que suas matrizes de variância e covariância seriam iguais, uma vez que sem essa pressuposição haveria perda da linearidade das funções discriminantes.

Das 10 populações simuladas, foram escolhidas duas mais divergentes para gerar o híbrido F_1e três gerações de retrocruzamentos, para tal, foram considerados previamente conhecidos o sistema de parentesco e nível de hierarquia, desse modo com um par de genitores (P_1 e P_2) divergentes foram geradas 7 outras populações considerando para cada conjunto de dados 8 características com herdabilidade de 55% até 90%.

O sistema hierárquico do retrocruzamentos pode ser visto na Figura 11.

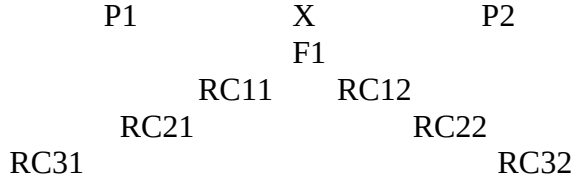

Figura 11- Cruzamentos entre os genitores $\boldsymbol{P_1}$ e $\boldsymbol{P_2}$ e seus respectivos RCij que representa o i-ésimo retrocruzamento referente ao j-ésimo genitor recorrente em que i={1,..., 3} e j={1, 2}.

2.2. Simulação dos Fenótipos

Para simulação dos fenótipos, foram utilizadas 10 populações constituídas de 100 indivíduos cada, de tal forma que P_i representa P_1 para$i = 1$; P_2 para $i = 2$; F_1 para $i = 3$; Para i=4, 5, 6 representam respectivamente RC11, RC21, RC31; Para i= 7, 8, 9 representam respectivamente RC12, Rc22, Rc32. Tais populações foram mensurados com base em 8 características quantitativas cada representando um nível de herdabilidade (mede o nível da correspondência entre o fenótipo e o valor genético), para tal, estabeleceu-se previamente uma média e herdabilidade conhecida. Os valores da herdabilidade assumidas foram respectivamente 55, 60, 65, 70 ,75, 80 ,85, 90, tais valores são representados em percentual. Por meio da ação de alelos de 20 locos randomizados nas quais tais características foram estabelecidas, tomados ao acaso entre os 50 previamente genotipados, com efeito aditivo diferencial determinados pelos pesos dado por uma distribuição binomial, está representa a importância do locos na variabilidade genotípica total das características quantitativas com grau médio de dominância nulo.

Deve-se então para cada variável estabelecer um modelo estatístico para os valores de média e herdabilidade, usamos o seguinte modelo:

$$Y_{ij} = \mu + G_i + \varepsilon_{ij} \qquad (2.1)$$

Em que:

Y_{ij}: observação simulada de uma dada característica;

μ: média geral da caraterística;

G_i: efeito associado ao i-ésimo indivíduo;

ε_{ij}: erro aleatório, sendo $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Neste trabalho, o modelo foi empregado de tal forma que o valor genotípico de cada indivíduo foi gerado considerando o efeito ambiental admitindo-se proveniente de um modelo normal de média zero e variância σ^2. Segundo Cruz et al. (2012), esse modelo é o mais empregado em programas de melhoramento além de ser o mais simples.

2.3. Simulação dos cenários

As características simuladas da população fenotípica para analise discriminante foram aplicadas em quatro diferentes cenários, para cada cenário estudado a proporção do grau de similaridade dos indivíduos nas populações de retrocruzamentos com relação a P_1 (1º genitor recorrente) e P_2 (2º genitor recorrente) é aumentado dificultando o processo de classificação, sendo para cada cenário 1, 2, 3 e 4 respectivamente 50%, 75%, 87,5% e 93,75% de grau de similaridade (Tabela 2), desse modo será comparado técnicas de analise discriminante de Anderson, Redes Neurais e Máquina de Vetor Suporte.

Segundo Sant'Anna et al.(2014), conseguir diferenciar as populações de retrocruzamentos têm sua importância devido ao grau de dificuldade exigido para diferenciá-los, já que cada geração vem recuperando a genética do seu parente (Em nosso estudo, chamamos de P_1 e P_2) chegando a ser quase indistinguível depois da 4ª a 5ª geração de retrocruzamentos, superando essa dificuldade acarretaria uma significativa vantagem da técnica para poder ser utilisada em outros problemas de semelhante complexidade como analise genômica.

O método de treinamento e validação dos dados consiste em separar os dados em dois grupos, um chamado grupo de treinamento e outro de grupo de teste, neste trabalho optou-se em usar 70% dos dados para o grupo de treinamento e 30% para o grupo de teste.

Tabela 2: Definição dos cenários para comparação das técnicas de analise discriminante de Anderson, Redes Neurais e Máquina de Vetor Suporte para as características de alta herdabilidade.

Cenários de distinguibilidade	Delineamento Genético	Similaridade máxima	Observações
1	P1, P2, F1	50%	300
2	P1, P2, F1, RC1a, RC1b	75%	500
3	P1, P2, F1, RC1a, RC1b, RC2a, RC2b	87,5%	700
4	P1, P2, F1, RC1a, RC1b, RC2a, RC2b, RC3a, RC3b	93,75%	900

2.4. Funções Discriminantes

A análise discriminante é uma técnica estatística no qual para uma população já conhecida e um conjunto de informações que as definem (variáveis explicativas), é possível então estudar diferenças entre dois ou mais grupos (Ferreira, 2011). O objetivo dessa análise é a separação de objetos em duas ou mais classes para então construir uma regra de decisão afim de que se possam alocar novos indivíduos segundo uma regra de decisão com menor erro possível.

2.4.1. Análise Discriminante

As funções discriminantes de Anderson e de Fisher são as mais conhecidas em termos de classificação de indivíduos, porem para o caso de p>2 Anderson (1958) propôs outro procedimento para a regra de decisão levando em consideração uma probabilidade a “priori” para as várias populações já que pode ocorrer um indivíduo que seja muito distinto com relação a uma determinada população acarretando em uma chance menor de pertencer a uma determinada população com relação à outra.

Considere π_i as populações a serem comparadas, dado μ_i e Σ_i, o vetor de médias e a matriz de covariâncias destas populações, respectivamente, com i={1, 2,..., n} com n variando de 3, 5, 7 e 9 já que temos para cenário populações constituídas de três a nove grupos. Como na simulação foram consideradas populações de mesma variância, temos então $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_n = \Sigma_p$. Sendo Σ_p a matriz de covariância comum dada pela equação (1.2) e função discriminante é dada pela expressão:

$D_i(X) = \ln(p_i) + \mu_i^T \Sigma^{-1} X - \frac{1}{2}\mu_i^T \Sigma^{-1} \mu_i$, para $i = 1, \dots, n$ (Ferreira, 2011). Em que p_i é a probabilidade a priori do grupo i pertencer a população π_i. De posse dessas funções o classificador de um indivíduo k a fim de classificá-lo a uma determinada população é dado por $D_i(X_k) = \max(D_1(X_k), \dots, D_n(X_k))$.

Para os cenários definidos na Tabela 2 as probabilidades foram admitidas como sendo iguais para todos os indivíduos e também admitimos as médias e matriz de variância covariância de cada população como sendo homogêneas.

2.4.2. Construção da Rede Neural

A arquitetura da Rede Neural utilizada neste trabalho foi a *Multilayer Perceptron* (MLP), para tal foi utilizada uma camada de entrada, três camadas intermediárias e uma camada de saída. A Rede foi processada no *software Mathworks Matlab* R2011a V. 7.12.0.635 integrado ao aplicativo computacional GENES (Cruz, 2013).

Foram utilizadas todas as funções de ativação implementadas no aplicativo GENES tal como a linear (purelin), para a camada de saída e, para as camadas ocultas, foram utilizadas a tangente hiperbólica (tansig) e a logarítimica (Logsig). Para o treinamento da Rede, foi escolhido o Trainbr (Bayesian Regulation backpropagation) e o número máximo de épocas foi fixado em 1500 para que não se torne excessivo, de acordo com Sant'Annaet al. (2014), uma interação muito alta pode levar a perda do poder de generalização.

Para as camadas intermediárias, utilizou-se o mesmo procedimento que resultou em um melhor resultado encontrado por Sant'Anna et al. (2014), para tal variou-se de 6 a 15 neurônios na primeira camada, de 10 a 40 na segunda camada e 10 a 40 na terceira camada. Para a camada de saída, a composição se deu por um neurônio e a saída representada por um vetor contendo em seus elementos o número da população, esse valor era conhecido no treinamenteo mas não na validação. De acordo com Sant'Anna et al. (2014), a rede que apresentar uma acurácia média superior, calcula-se todas as possíbilidades dentre o número de neurônios em cada camada e as funções de ativação possíveis $(15 \times 40 \times 40 \times 3 \times 3 \times 3)$, será a que possui a melhor arquitetura da rede.

Dessa forma, para cada cenário (A e B), adotamos como críterio a menor taxa de erro aparente, tal procedimento da rede pode ser melhor visualizado pela Figura 12.

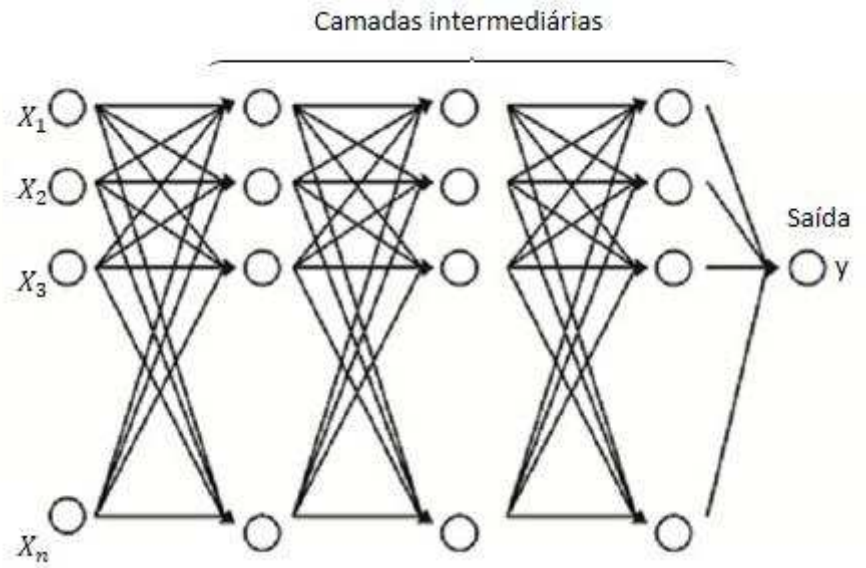

Figura 12- Representação das camadas existentes em um modelo de redes neurais Perceptron Múltiplas Camadas utilizado para classificação.
Fonte: Adaptado de Sant'Anna et al, 2014

2.4.3. Construção da Máquina de Vetor Suporte

Para a construção da Máquina de Vetor Suporte utilizada neste trabalho, foi utilizada uma SVM não linear com a decomposição um-contra-todos como meio de classificação das populações. A SVM foi processada no software livre R[9] V. 3.3.2 para verificação de resultados.

Foram utilizadas três funções kernels implementadas no aplicativo R com diferentes parâmetros de entradas, tais como a PolyKernel (Função Polinomial), NormalizedPolyKernel (Função de base radial gaussiana) e RBFPolyKernel (Função de base radial exponencial) (Tabela 3). Para os dados de treinamento, foi utilizado um método de normalização a fim de agilizar o andamento da SVM.

Tabela 3- Tipos de funções kernels e respectivos parâmetros

Tipo de Kernel	Função $K(x_i, x_j)$	Tipo de classificador
Polinomial	$\gamma(x_i \cdot x_j + \alpha)^d$	Máquina de aprendizagem polinomial

[9] Disponível em: <https://cran.r-project.org/>

Gaussiano (RBF)	$exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right)$	Rede RBF
Sigmoidal	$tanh\left(\beta_0\left(x_i \cdot x_j\right) + \beta_1\right)$	Perceptron de Duas Camadas

Para lidar com a não linearidade dos dados, utilizou-se uma constante de suavização (C), quanto maior esse valor mais rígido é à margem de separação, e os parâmetros dos respectivos kernels.

Realizou-se uma procura pelos valores que proporcionariam melhores parâmetros para o modelo, para tal, algumas medidas foram adotadas, como definir um espaço de busca, um método procura e uma validação cruzada.

O espaço de busca foi definido de forma variada para cada modelo kernel exceto para os valores da constante C em que o espaço de busca permaneceu fixado entre 1 até 100 variando de um em um para cada kernel escolhido. Para os parâmetros da kernel, definimos de forma como se encontra na Tabela 4.

Tabela 4: Escolha dos parâmetros nos diferentes kernels.

Kernel	Intervalo de valores
Gaussiano (RBF)	$1 \leq C \leq 10$ $0{,}001 \leq \frac{1}{2\sigma^2} \leq 0{,}5$
Polinomial	$1 \leq C \leq 10$ $0{,}001 \leq \gamma \leq 0{,}2$ $-1 \leq \alpha \leq 1$ $1 \leq d \leq 10$
Sigmoidal	$1 \leq C \leq 10$ $0{,}001 \leq \beta_0 \leq 0{,}5$ $0{,}001 \leq \beta_1 \leq 0{,}5$

Como método de procura, utilizou-se o *grid search* (grade de valores), este método é o mais recomendado em conjunto com N *fold* da validação cruzada (Howlett, 2000; Gaspar et al., 2012),tal método requer que a máquina de vetor suporte seja treinada múltiplas vezes para se obter uma performance ótima por meio dos valores de parâmetros definidos no espaço de busca. Para validar os resultados, alterou-se a semente aleatória mudando a partição dos dados e repetindo o processo 10 vezes gerando assim uma abordagem próxima a validação cruzada com 10 *fold* com o objetivo de avaliar se a SVM mantém um resultado próximo para todas as partições geradas.

Por fim, o método um-contra-todos foi escolhido para a classificação das populações de retrocruzamentos, o problema é decomposto em N classificadores binários como mostra a Figura 13 para $N = 3$.

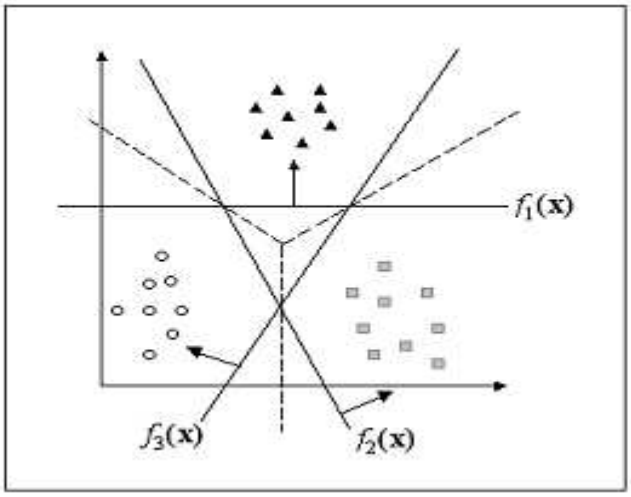

Figura 13- Ilustração das regiões com fronteiras estendidas pela abordagem um-contra-todos.
Fonte: Lorena & Carvalho, 2003.

2.4.4. Taxa de Erro Aparente

Conhecida então a função a ser utilizada, utilizou-se a taxa de erro aparente (TEA) para avaliar a eficácia do método quanto à alocação de novos indivíduos para as populações previamente definidas. Tal técnica também foi utilizada para o uso de Redes Neurais e Máquina de Vetor Suporte.

A taxa de erro aparente é dada de forma simples:

$$TEA = \frac{1}{N}\sum_{j=1}^{k} m_j \qquad (2.2)$$

em que m_j é o número de observações retiradas de uma população, que por meio da técnica avaliada, foi classificada em outra população, N é o número total de observações avaliadas e k é o número de populações consideradas.

3. RESULTADOS E DISCUSSÃO

3.1. Análise Discriminante de Anderson

A análise discriminante de Anderson foi ineficaz para distinguir as populações de retrocruzamentos em todos os cenários estabelecidos (Tabela 2,

Tabela 5, Tabela 6). Especificamente, os valores de TEA variaram entre 18,89% a 78,89% de acordo com cada cenário estudado.

Tais resultados podem ser explicados devido a proporção esperada de similaridade entre os três retrocruzamentos que é aproximadamente 93,75% (Tabela 5) (Sannt'Anna et al., 2014), ou seja, existe grande sobreposição entre as populações e esse fato acarreta em um aumento do nível de dificuldade a cada cenário de distinguibilidade.

No estudo de Sant'Anna (2014), foi utilizado redes neurais artificiais para classificar 5 populações de retrocruzamentos. Os resultados encontrados com relação a taxa de erro aparente (TEA) em percentual variaram de 22,67% à 80,01%, essa classificação foi bem próxima da obtida nesse trabalho, confirmando assim com os resultados encontrados.

De acordo com a Tabela 5, que representa o grupo de variáveis quantitativas, observa-se que os resultados foram poucos satisfatórios já que a taxa de erro aparente a partir do cenário 2 fica acima de 50%. Para o cenário 1 que representa a discriminação dos indivíduos das populações P_1, P_2 e F_1 tem-se uma TEA de aproximadamente 19%. Considera-se essa alta para os objetivos desse trabalho.

A partir do segundo cenário a dificuldade encontrada já era de se esperar devido às características semelhantes na população de retrocruzamentos. A similaridade do RC3 e do seu genitor recorrente é de aproximadamente 93,75%, esse grau de similaridade faz com que ocorra uma grande dificuldade de discriminar essas diferentes populações de retrocruzamentos e seus parentes recorrentes.

Tabela 5: Taxa de Erro Aparente (TEA) aplicada a análise discriminante de Anderson aplicadas aos 4 cenários.

Cenários	Tamanho Populacional	Similaridade máxima	TEA (%) Anderson
1	300	50%	18,89
2	500	75%	52,00
3	700	87,5%	67,14

4	900	93,75%	74,07

De acordo com a Tabela 6, representando a classificação incorreta das classes para o cenário 4, é possível notar que não houve classificação incorreta quanto a P_1 ser classificado como sendo P_2 e o mesmo de forma inversa. A população F_1 possuiu o maior percentual de acerto tanto na população de alta quanto baixa herdabilidade, tendo respectivamente 70% e 47% de acertos na classificação correta dos indivíduos.

Tabela 6: Resumo da classificação incorreta pela análise discriminante de Anderson para o cenário 4 aplicada ao grupo com alta e baixa herdabilidade (%).

População[10]	P1	P2	F1	RC1a	RC1b	RC2a	RC2b	RC3a	RC3b
P1	30	0	7	10	3	27	3	20	0
P2	0	47	3	0	13	0	10	0	27
F1	3	0	70	17	3	3	3	0	0
RC1a	27	0	23	13	7	13	0	17	0
RC1b	3	13	27	0	30	0	13	0	13
RC2a	20	0	7	27	13	13	3	13	3
RC2b	3	27	0	0	17	0	13	0	40
RC3a	17	0	3	37	3	20	3	17	0
RC3b	0	33	7	3	20	0	23	0	13

3.2. Aplicação das Redes Neurais Artificiais

De acordo com a Tabela 7, a taxa de erro aparente para os dados de treinamento foi nula para todos os cenários pesquisados confirmando os resultados obtidos por Sant'Anna et al. (2014). Desse modo a rede neural mostra se bastante capaz de diferenciar perfeitamente as populações de três gerações de retrocruzamentos e assim superando aqueles obtidos pela Análise Discriminante de Anderson.

Tabela 7: Taxa de Erro Aparente (TEA) com relação a fase de treinamento da rede neural artificial nos 4 cenários.

Cenários	Tamanho Populacional	Similaridade máxima	TEA (%) RNA
1	300	50%	0
2	500	75%	0

[10] Híbrido F1 provindo de P1, P2 e três gerações de retrocruzamentos obtidas em relação a cada um dos genitores a e b respectivamente.

3	700	87,5%	0
4	900	93,75%	0

Para o conjunto de dados de validação, o resultado encontrado equivale ao obtido por Sant'Anna et al. (2014), com TEA em 0% como pode ser visto na Tabela 8. Segundo Sant'Anna et al. (2014), isso se dá pelo fato das RNAs terem melhor capacidade de extrair características dos dados, não se baseando apenas em parâmetros estáticos como média e variância.

Tabela 8: Taxa de Erro Aparente (TEA) com relação a fase de validação da rede neural artificial nos 4 cenários.

Cenários	Tamanho Populacional	Similaridade (*)	TEA (%) RNA
1	300	50%	0
2	500	75%	0
3	700	87,5%	0
4	900	93,75%	0

Nas configurações utilizadas para a construção da rede, foram utilizadas três camadas ocultas com variações dos números de neurônios de 6 a 15 na primeira camada, 15 a 40 na segunda e 15 a 40 na terceira camada e por fim combinações das funções de ativação logsig e tansig. Ambas mostraram-se adequadas a solução do problema apresentado. Foi notado que nas camadas um e dois, para o cenário 4, houve predomínio da função logsig enquanto que para a terceira houve predominância da logsig o que pode ser visto na Tabela 9.

Tabela 9: Topologia das redes neurais. Números de neurônios por camada (C_1, C_2, C_3) e as funções de ativação (F_1, F_2, F_3) para cada camada nos 4 cenários estudados.

Cenário	Números de neurônios			Função de ativação		
	C_1	C_2	C_3	F_1	F_2	F_3
1	13	35	35	logsig	logsig	logsig
2	14	40	40	tansig	tansig	tansig
3	15	35	40	logsig	tansig	logsig
4	15	40	40	logsig	logsig	tansig

Baseando-se no estudo de Sant'Anna et al. (2014), foram fixadas três camadas intermediárias para o uso da rede, o quê conduziu para o resultado

obtido. As configurações da rede também foram satisfatórias ocasionando TEA nulo para todas as populações de retrocruzamentos estudados. De maneira geral, os resultados indicaram semelhanças alcançadas por Sant'Anna (2014) e mostraram-se superiores quando comparado a análise discriminante de Anderson, tal superioridade já era esperada, pois, segundo Braga et al. (2011), as redes neurais destacam-se em relação a outros métodos convencionais mais comumente utilizados para discriminação de populações.

Vale ressaltar que houve um gasto de tempo longo para o processamento dos dados, em torno de 1 mês de processamento, isso se dá devido a quantidade de interações, camadas ocultas, quantidade de neurônios em cada camada e o conjunto de funções de ativação. Tal processo era necessário devido à complexidade dos dados provenientes da similaridade das populações de retrocruzamentos. Encontra-se na literatura outro trabalho semelhante como visto em Pereira (2009), que também usou populações de retrocruzamentos comparando-se as técnicas de Rede Neurais com Analise Discriminante de Fisher e Anderson para três diferentes cenários, os resultados por eles encontrados também se constatou eficiente com TEA entre 0,38 à 13,23, superando os resultados encontrados pela análise discriminante de Anderson com TEA entre 11,52 à 26,67.

No trabalho de Sant'Anna et al. (2014), em que a população foi ampliada foram encontrados resultados semelhantes ao do nosso estudo. A ampliação deu-se para que a população fosse separada em treinamento e os dados originais na validação, essa geralmente é utilizado quando se tem poucos dados para subdividi-los. Sendo assim a estrutura genética da população inicial se manteve para checar quais parâmetros utilizar nas camadas ocultas e treinar a rede, porém nesse trabalho mesmo sem a utilização de dados ampliados a classificação foi satisfatória, alcançando o resultado desejado.

3.3. Aplicação da Máquina de Vetor Suporte

As análises discriminantes de populações de retrocruzamentos aplicados aos diferentes cenários por meio da máquina de vetor suporte se mostraram ineficazes comparados às redes neurais, porém obteve melhores resultados com relação à análise discriminante de Anderson quando utilizado a função kernel

sigmoidal, com TEA variando entre 14,44% à 67,41%, os resultados encontrados para diferentes tipos de kernels e quais parâmetros melhores se ajustaram ao modelo podem ser vistos na Tabela 10 e Tabela 11. Apesar das SVMs também terem a capacidade de aprendizagem como o perceptron (Lorena, 2003; Martins et al., 2007), uma rede neural de uma camada também chegaria em tal resultado devido a complexidade do modelo e o superajustamento da rede, o SVM também possui tal dificuldade necessitando de uma abordagem para estruturação dos parâmetros de busca mais amplos. Como a SVM possui uma alta complexidade computacional na busca de soluções, caso o número de dados de treinamento seja muito alto, a complexidade da solução se agrava (Lorena, 2003).

Tabela 10: Máquina de Vetor Suporte nos 4 cenários para diferentes valores da constante C e parâmetros das funções ótimos.

Cenários	RBF		Polinomial				Sigmoidal		
	C	$1/2\sigma^2$	C	d	γ	α	C	β_0	β_1
1	1	0.0018	1	5	0.04	-0.4	11	0.102	0.028
2	1	0.387	51	7	0.003	-1	1	0.05	0.258
3	11	0.0883	3	6	0.029	1	19	0.127	0.434
4	2	0.037	10	3	0.005	-0.9	4	0.055	0.02

Tabela 11: Taxa de Erro Aparente (TEA%) com relação à fase de validação da Máquina de Vetor Suporte nos 4 cenários.

Cenários	TEA		
	RBF	Polinomial	Sigmoidal
1	22.22%	17.78%	14.44%
2	54.67%	50,00%	51.33%
3	61.61%	61.61%	62.09%
4	70.74%	69.26%	67.41%

O método “um-contra-todos”, segundo Almeida (2010), possui a desvantagem de não ser possível prever limites no erro de generalização através de seu uso, além de tomar um tempo geralmente longo. Alternativas podem ser encontradas na literatura, como por exemplo, Inoue e Abe (2001 e 2002) encontraram outra forma de combinação da SVM a qual denominaram de *Fuzzy Suport Vector Machines* (FSVMs) a fim de introduzir funções para determinar o grau com que um padrão pertence a uma determinada classe. Outra maneira de obter melhor desempenho da técnica seria pelo uso do método todos-contra-todos (et al, 2003) para diferenciação das classes. Na literatura existem vários outros métodos que não serão abordados aqui, mas que podem ser vistos pelo leitor em Lorena & Carvalho. (2003).

Na procura de parâmetros ótimos para o modelo, se definimos C^* como valor ótimo encontrado, então $C < C^*$ ira fazer com que sejam criadads muitas superfícies de separação e caso seja $C > C^*$, isso pode deixar a generalização comprometida informando que as amostras de dados são equivalentes aos pontos de vetor suporte (Martins et al., 2007).

Desse modo não foi possível encontrar um resultado que se coincidisse com as Redes Neurais Artificiais necessitando assim de um estudo quanto ao uso da técnica SVM para populações com certo grau de similaridade já que não se encontra na literatura tal método para esse tipo de problema.

Contudo, são necessários mais estudos a respeito da técnica de SVM, tais como: a ampliação do algoritmo de busca com a finalidade de aperfeiçoar os parâmetros do modelo em um espaço de busca maior; considerar outros tipos de kernels e utilizar a validação cruzada para encontrar os valores ótimos de C e parâmetros da kernels com intuito de usá-los para treinar todo o conjunto de treinamento (o melhor parâmetro pode ser afetado pelo tamanho do conjunto de dados, mas na prática, os dados obtidos porvalidação cruzada demonstram-se adequados para a totalidade do conjunto de treinamento).

4. CONCLUSÕES

A análise discriminante de Anderson não foi capaz de diferenciar as populações de retrocruzamentos demonstrando resultados ineficientes para cada cenário estudado.

A abordagem por meio de redes neurais com estrutura da rede de três camadas com 6 a 15 neurônios na primeira camada, 15 a 40 na segunda camada e 15 a 40 na terceira camada mostrou-se eficaz para classificar populações derivadas de retrocruzamentos sem a necessidade do uso de ampliação dos dados.

Já os resultados obtidos por meio de máquina de vetor suporte foram semelhantes à técnica multivariada de análise discriminante proposta por Anderson exceto para os cenários 1 e 4 (similaridades de 50% e 93,75%, respectivamente) em que o kernel sigmoidal apresentou uma taxa de erro aparente menor.

REFERÊNCIAS BIBLIOGRÁFICAS

ABE, S., & INOUE, T. (2002). Fuzzy support vector machines for multiclass problems. ESANN 2002. In:**Proceedings of European symposium on artificial neural networks**, pages 113–118,

ALMEIDA, E. D. Algoritmos de Classificação Com a Opção de Rejeição, Faculdade de engenharia da universidade do porto - **Relatório do Estado da Arte**. Mestrado Integrado em Engenharia Informática e Computação. 2010.

ANDERSON T. W. An Introduction to Multivariate Statistical Analysis. **New York: John Wiley & Sons**, 1958, 345 p.

BOSER, B.; GUYON, I.; VAPNIK, V.An training algorithm for optimal margin classifiers**. Fifth Annual Workshop on Computational Learning Theory**, pages 144-52, Pittsburgh, ACM. 1992.

BRAGA, A. P.; CARVALHO, A. P. L. F.; LUDERMIR, T. B. **Redes Neurais Artificiais – Teoria e aplicações** – 2. Ed. Rio de Janeiro: LTV, 2011. 226p.

CAMPBELL, Colin. An introduction to kernel methods. **Studies in Fuzziness and Soft Computing**, v. 66, p. 155-192, 2001.

HOWLETT, R. J.; JAIN, L. C. Radial Basis Function Networks: Design and Applications. **Physica-Verlag, Wurzbury**, 2000.

CRISTIANINI,N. andSHAWE–TAYLOR, J. An introduction to support vector machines. **Cambridge University Press**. 1999.

CRUZ, C. D. GENES – a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum**.**Agronomy**, V.35, n.3, p. 271-276, 2013.

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos Biométricos Aplicados ao Melhoramento Genético** – Vol 2. 2.ed. Viçosa: UFV, 585p.2006.

FERREIRA, D. F. **Estatística Multivariada**, 2ª ed. Editora UFLA, 2011.

GASPAR, P., CARBONELL, J., OLIVEIRA, J. L., On the parameter optimizationof Support Vector Machines for binary classification, **Journal of Integrative Bioinformatics**, 9(3):201, 2012.

INOUE, Takuya; ABE, Shigeo. Fuzzy support vector machines for pattern classification. In: **Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on**. IEEE, p. 1449-1454, 2001.

KAVZOĜLU, Taşkin. An investigation of the design and use of feed-forward artificial neural networks in the classification of remotely sensed images. **Tese (Doutorado) - University of Nottingham**.2001.

KAVZOGLU, T.; MATHER, P. The use of backpropagating artificial neural networks in land cover classification. **International Jorurnal of Remote Sensing**, v. 24, n. 23, p. 4907-4938, 2003.

KAVZOGLU, T. Increasing the accuracy of neural network classification using refined training data. **Environmental Modelling & Software**, v. 24, n. 7, p. 850-858, 2009.

LIMA, C. A. M. Comitê de Máquinas: Uma Abordagem Unificada Empregando Máquinas de Vetores Suporte. **Tese (Doutorado) — Universidade Estadual de Campinas**, São Paulo, 2004.

LORENA, A. C. ; CARVALHO, A. C. P. L. F.; Introdução às Maquinas de Vetores Suporte. **Relatórios técnicos do icmc. Instituto de Ciências Matemáticas e de Computação**. ISSN - 0103-2569. Ano 2003.

MARTINS, R. S.; DUARTE, V. J. L.; MAITELLI, A. L.; SALAZAR, A. O. e DÓRIA NETO, A. D. Sistemas de Detecção de Vazamentos em Dutos Usando Redes Neurais e Máquinas de Vetor de Suporte. **Anais do VIII Congresso Brasileiro de Redes Neurais**, pp. 1-6, Florianópolis-SC, outubro 2007.

NEI, Masatoshi. Genetic distance between populations. **American naturalist**, p. 283-292, 1972.

PEREIRA, T. M. **Discriminações de populações com diferentes graus de similaridade por redes neurais artificiais**.Viçosa: UFV, 2009. 73 p.

SANT'ANNA, I. C.; Tomaz, Rafael Simões; SILVA, G. N.; BHERING, L. L.; NASCIMENTO, M.; CRUZ, C. D.Artificial neural networks in genetic classification. **Genetics and Molecular Research**, 2014.

SANTOS, E. M. Teoria e Aplicação de Support Vector Machines à Aprendizagem e Reconhecimento de Objetos Baseado na Aparência. 121 p. **Tese (Mestrado) – Universidade Federal da Paraíba**, Paraíba, 2002.

CONSIDERAÇÕES FINAIS

Este trabalho teve como finalidade abordar a utilização dos diferentes métodos de classificações comumente utilizados, avaliando-os conjuntamente com a técnica de máquina de vetor suporte (SVM) objetivando assim o estudo de populações.

No Capítulo inicial, foram apresentados os aspectos teóricos das metodologias envolvidas, porém com ênfase na construção das SMVs.

No Capítulo 1 foi realizado um estudo envolvendo populações de retrocruzamentos com disparidade graus de similaridades com o alvo de comparar as diferentes técnicas de análises discriminantes a fim de diferenciar tais populações. Para esse afazer foi utilizado como medida de comparação a TEA (taxa de erro aparente).

De acordo com os resultados obtidos, constatou-se que não foi possível, pelos métodos apresentados, encontrar resultados positivos com o uso da SVM mostrando se inicialmente ineficiente para discernir as populações com auto grau de similaridade. No no entanto, a técnica das Redes Neurais mostrou –se novamente eficiente corroborando com os resultados encontrados por Sant'anna et al. (2014).

Diante do exposto os autores sugerem mais estudos quanto ao uso da técnica SVM em populações com alto grau de similaridade, cita-se como exemplo: um espaço de busca maior no algoritmo para otimização dos parâmetros e considerar outros tipos de kernels utilizados na literatura.