VITOR PRADO DE CARVALHO

APRENDIZADO DE MÁQUINA E ESTATÍSTICO NA DISCRIMINAÇÃO DE POPULAÇÕES NA PRESENÇA DE MATRIZES DE COVARIÂNCIAS HETEROGÊNEAS E VETORES ALEATÓRIOS NÃO NORMAIS MULTIVARIADOS

Tese 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 *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2019

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

T
C331a
2019

Carvalho, Vitor Prado, 1981-
Aprendizado de máquina e estatístico na discriminação de populações na presença de matrizes de covariâncias heterogêneas e vetores aleatórios não normais multivariados / Vitor Prado Carvalho. – Viçosa, MG, 2019.
x, 47 f. : il. (algumas color.) ; 29 cm.

Orientador: Moyses Nascimento.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 45-47.

1. Análise discriminatória. 2. Análise multivariada. 3. Métodos de simulação. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 515.252

VITOR PRADO DE CARVALHO

APRENDIZADO DE MÁQUINA E ESTATÍSTICO NA DISCRIMINAÇÃO DE POPULAÇÕES NA PRESENÇA DE MATRIZES DE COVARIÂNCIAS HETEROGÊNEAS E VETORES ALEATÓRIOS NÃO NORMAIS MULTIVARIADOS

Tese 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 *Doctor Scientiae*.

APROVADA: 22 de julho de 2019.

Assentimento:

Vitor Prado de Carvalho
Autor

Moysés Nascimento
Orientador

AGRADECIMENTOS

Inicialmente agradeço muito a Deus pela determinação de poder tornar mais esse sonho em realidade.

Não posso deixar de agradecer primeiramente a minha querida esposa que desde o mestrado continua sonhando junto comigo para alcançar um futuro melhor, sua compreensão e carinho me incentiva a jamais desistir.

A meus filhos Maria Clara e Lucas, razão da minha vida, eles me mostram o quanto ainda tenho que caminhar para levar um futuro melhor para nossa família.

Aos meus pais, sogros 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, disponibilidade e muita paciência, agradeço muito pelas orientações sempre visando a melhoria dos trabalhos realizados, agradeço também pela compreensão, amizade e auxilio durante esses anos de convivência que não esquecerei.

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 para melhora na entrega 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 e doutor.

Agradeço a Universidade Federal de Viçosa pela oportunidade em minha
vida e a CAPES pela bolsa de estudo sem a qual eu não poderia estar concluindo este trabalho.

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

RESUMO

CARVALHO, Vitor Prado de, D.Sc., Universidade Federal de Viçosa, julho de 2019. **Aprendizado de máquina e estatístico na discriminação de populações na presença de matrizes de covariâncias heterogêneas e vetores aleatórios não normais multivariados**. Orientador: Moysés Nascimento. Coorientadores: Mauro César Martins Campos e Isabela de Castro Sant'anna.

Na análise discriminante, é avaliado a diversidade ou classificação dos indivíduos nas populações, para tal um grande número de metodologias está disponível, dentre as quais destacam-se os métodos multivariados de análise discriminante que têm sido utilizados em estudos preditivos da diversidade genética. Tal metodologia visa identificar as populações nas quais um indivíduo deva pertencer, admitindo previamente, que este indivíduo pertença a uma das populações avaliadas, no entanto esta análise pressupõe que as populações sejam provenientes de uma distribuição normal multivariada. Dentre as diversas metodologias de análise discriminante destaca-se a função discriminante linear de Fisher que possui para sua utilização a pressuposição de que as matrizes de covariância entre as populações sejam homogêneas, e na quebra desse pressuposto outras abordagens são necessárias como a análise discriminante quadrática ou auxilio de métodos computacionais como os de aprendizado de máquina. Desse modo o presente trabalho visa avaliar a robustez da função discriminante linear de Fisher na presença de matrizes de covariâncias heterogêneas e vetores aleatórios não normais multivariados, já que na literatura não exemplifica o critério de escolha quanto ao uso de tal função. Os dados foram gerados por meio de simulação com cenários caracterizados por matrizes de covariâncias heterogêneas e vetores aleatórios não normais multivariados e seus resultados foram comparados com outras metodologias de mesmo proposito, tais como a Análise Discriminante Quadrática, Redes Neurais Artificiais, Máquina de Vetor Suporte e Árvore de Classificação. De acordo com os resultados foi possível observar que as técnicas para classificação de indivíduos devem ser utilizadas seguindo suas pressuposições. Especificamente, para situações em que os dados apresentam

normalidade multivariada e heterocedasticidade de matrizes de covariâncias, a função discriminante Quadrática apresentou melhores resultados quanto ao valor de Taxa de Erro Aparente (TEA). Para situações em que os dados apresentaram distribuição Poisson multivariada e homogeneidade de matrizes de covariância, a Função Discriminante de Fisher apresentou menores valores de TEA. As demais metodologias, Redes Neurais Artificiais, Máquina de Vetor Suporte, Árvores de Decisão e seus refinamentos (Poda, *Bagging* e *Random Forest*) e *Boosting* apresentaram valores razoáveis de TEA e se apresentam como técnicas alternativas para situações em que os pressupostos necessários para aplicação das técnicas da Função Discriminante de Fisher e da Função Discriminante Quadrática não são atendidos.

ABSTRACT

CARVALHO, Vitor Prado de, D.Sc., Universidade Federal de Viçosa, July, 2019. **Machine and statistical learning in discrimination of the population in the presence of heterogeneous covariance matrices and multivariate non-normal random vectors.** Adviser: Moysés Nascimento. Co-Advisers: Mauro César Martins Campos and Isabela de Castro Sant'anna.

In discriminant analysis, is evaluated the diversity or classification of individuals in populations, for that a large number of methodologies are available, among which stand out the multivariate methods of discriminant analysis that have been widely used in predictive studies of genetic diversity. This methodology aims to identify the populations in which an individual should belong, previously admitting that this individual belongs to one of the evaluated populations, however this analysis assumes that the populations come from a normal multivariate distribution. Among the various discriminant analysis methodologies, stands out the Fisher's linear discriminant function, which has for its use the assumption that the covariance matrices between populations are homogeneous, and in breaking this assumption other approaches are necessary such as quadratic discriminant analysis or the aid of computational methods such as machine learning. Thus, the present work aims to evaluate the robustness of Fisher's linear discriminant function in the presence of heterogeneous covariance matrices and multivariate non-normal random vectors since in the literature it does not exemplify the criterion of choice regarding the use of such function. The data were generated by simulation with scenarios characterized by heterogeneous covariance matrices and multivariate non-normal random vectors and their results were compared with other methodologies of the same purpose, such as Quadratic Discriminant Analysis, Artificial Neural Networks, Support Vector Machine and Decision Tree. According to the results it was observed that the techniques for classification of individuals should be used following their assumptions. Specifically, for situations in which data present multivariate normality and heteroscedasticity of covariance matrices, the Quadratic discriminant function presented better results regarding the Apparent Error Rate (AER) value. For situations in which the data presented multivariate Poisson distribution and homogeneity of

covariance matrices, Fisher Discriminant Function presented lower AER values. The other methodologies such as, Artificial Neural Networks, Support Vector Machine, Decision Trees and their refinements (Pruning, Bagging and Random Forest) and Boosting presented reasonable values of AER and are presented as alternative techniques for situations where the necessary assumptions for the application of Fisher Discriminant Function and Quadratic Discriminant Function techniques are not met.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO GERAL

Há um grande interesse na conservação dos recursos genéticos de espécies vegetais e animais, pois a escolha correta dos genitores abrem a possibilidade de obter híbridos com maior heterose e populações com maior variabilidade, visto que cruzamentos mais heteróticos estão associados a uma maior divergência entre os genitores, dessa forma os programas de melhoramento vem acumulando conhecimento da quantidade de variação presente nas espécies (Cruz et al., 2011; Sant'Anna, 2014).

Pode-se observar diversas metodologias referentes a avaliação da diversidade genética, temos como exemplo, os métodos baseados em agrupamento (Cruz et al., 2011; Rodrigues et al., 2017; Santos et al., 2017b), e a analise discriminante (Cruz et al., 2011; Santos et al., 2017a). No entanto é comum estas metodologias estarem associadas a algum tipo de pressuposição. A quebra de determinada pressuposição leva a busca de outras formas de abordagem do problema, como por exemplo, o uso de métodos de agrupamento que geralmente não requerem nenhuma pressuposição acerca da estrutura dos dados.

Logo, pode-se notar que a escolha do método mais apropriado não é tarefa simples já que diferentes técnicas utilizam diferentes conceitos estatísticos que devem ser analisados dentro de cada problema a ser abordado, a fim de viabilizar seus resultados. Entretanto, em alguns casos, a escolha dentre os vários métodos disponíveis para a análise da diversidade genética é realizada sem a preocupação com os pressupostos de cada técnica, sendo escolhida por exemplo, por meio de estatísticas tais como o coeficiente de correlação cofenética (Sokal & Rohlf, 1962).

No que diz respeito ao estudo da diversidade genética, consideram-se diversas metodologias de modo a analisar a potencialidade de cada uma, dentre as opções destaca-se a função discriminante linear de Fisher como sendo uma reconhecida técnica de agrupamento e classificação (Fisher, 1936), porém uma pressuposição quanto ao seu uso requer que as matrizes de covariância entre as populações sejam homogêneas (Ferreira, 2008) e, em alguns casos que o vetor aleatório seja proveniente de uma função distribuição de densidade normal multivariado.

Dentre as aplicações dessa técnica na diversidade, cita-se a aplicação na diversidade fenotípica entre cruzamentos de ovinos Dorper com raças locais, em

que auxiliou a redução dos dados de um espaço p dimensional a um espaço unidimensional (Carneiro et al., 2007), a estimação da diversidade fenotípica em búfalas (Bubalus bubalis) das raças Jafarabadi, Murrah e Mediterrâneo (Rezende et al., 2017) e na avaliação do efeito de diferentes doses de esterco bovino curtido adicionado aos substratos vermiculita e substrato comercial para a formação de mudas de café em tubetes, por meio da técnica da análise discriminante de Fisher os dados puderam receber um tratamento multivariado (Cogo et al., 2019).

Caso haja rejeição da hipótese de igualdade quanto as matrizes de covariâncias serem homogêneas, existem outras formas de obtenção dos resultados como por exemplo a análise discriminante quadrática em que não existe pressuposição acerca dos grupos possuírem matrizes de covariância iguais, porém é mais sensível à suposição da normalidade. Caso não haja normalidade dos dados, há outras formas de abordagem, por exemplo o uso de estratégias como transformações de dados é uma opção (Ferreira, 2008; Azevedo et al., 2015).

Porém, vem se destacando os métodos de aprendizado de máquina para resolução de diversos problemas inclusive envolvendo discriminação de populações (Sousa & Salame, 2018; Nicoletti, 2018), é apresentado um algoritmo que especifica passo a passo de como o problema deve ser resolvido e há pouco ou nenhum pressuposto antes da sua realização. Esses métodos utilizam as informações apresentadas pelo pesquisador para melhorar sua base de conhecimentos e desse modo o processo vai melhorando seu desempenho até alcançar a resposta desejada (Haykin, 2007). Os métodos de aprendizagem de máquinas mais conhecidas e que vem se destacando são as Redes Neuras Artificiais, Máquina de Vetor Suporte e Árvore de Classificação\Regressão (James et al., 2013; Cruz & Nascimento, 2018).

Apesar destas indicações, não há na literatura estudos que avaliem a robustez das técnicas quanto a quebra dos pressupostos e geralmente nos trabalhos que fazem uso de tais funções lineares não há uma explicação quanto ao critério na escolha de tal metodologia.

Logo, este trabalho tem por objetivo avaliar, por meio de simulação de dados, a robustez da função discriminante linear quanto a falta de homogeneidade de matrizes de covariância e a presença de vetores não normais multivariados. Tais avaliações visam nortear pesquisadores quanto a escolha adequada do método a ser utilizado em estudos de diversidade genética. Os resultados serão comparados

por meio de outras metodologias com o mesmo proposito tais como a Análise Discriminante Quadrática, Redes Neurais Artificiais, Máquina de Vetor Suporte e Árvore de Classificação.

2. REVISÃO DE LITERATURA

2.1. Diversidade Genética no melhoramento

A diversidade genética pode ser definida como sendo a variação biológica hereditária acumulada durante todo o processo evolutivo e adaptativo, existindo uma gama de variações entre as espécies e entre indivíduos de uma mesma espécie, desta forma possui grande importância na área do melhoramento genético sendo possível indicar qual cruzamento entre as espécies acarretaria na descoberta de genótipos com melhor resistência a pragas e doenças, aliadas a um elevado potencial produtivo, dessa forma respondendo de forma positiva às mudanças climáticas e a todos os tipos de estresses bióticos e abióticos (Cruz et. al., 2011).

Uma das estratégias para garantir a conservação da diversidade genética das espécies são os bancos de germoplasma, viabilizando o estudo de um grande número de indivíduos de uma mesma de espécies ou de espécies parentes (Sant'anna, 2014).

A importância do estudo da diversidade genética em um programa de melhoramento está em quantificar a variabilidade existente, bem como a distribuição entre ou dentro de grupos que podem ser representados por populações, cultivares, linhagens e acessos de bancos de germoplasta (Bold, 2011, Cruz; et al., 2011).

Dessa forma esses estudos auxiliam em problemas envolvendo hibridações artificiais em que o melhorista deve selecionar genitores que darão origem a populações segregantes estimando a divergência genética entre as progênies (Benin et al., 2002). Nessa etapa os pesquisadores podem ter interesse em agrupar genótipos semelhantes para que as diferenças ocorram entre os grupos formados, dessa forma as técnicas multivariadas vêm sendo bastante aplicadas nesse tipo de estudo como exemplo, a análise discriminante, componentes principais, análise de coordenadas e de agrupamento (Cruz et al., 2011; Deniz Filho, 2000; Santos et al., 2018). Outras formas também predominam nos estudos de divergência genética, dentre os quais, citam-se as análises de variáveis canônicas e os métodos aglomerativos. A escolha do método a ser utilizado tem sido determinado pelo pesquisador dependendo dos seus objetivos, a facilidade da análise e pela forma como os dados foram obtidos (Miranda et al.,1988; Cruz, 1990; Cruz et al., 1994).

2.2. Análise Discriminante Linear de Fisher

As análises discriminantes são metodologias que buscam diferenciar e/ou classificar os elementos provenientes de uma amostra (ou população) com base no conhecimento preestabelecido da população estudada. Por meio das características dos grupos são elaboradas as regras de discriminação e após constatada sua eficácia serão utilizadas para alocar novos indivíduos dos quais se desconhecem a origem (Cruz et al., 2011; Mingot, 2007).

Para discriminação das classes, é necessário primeiramente procurar características dentro das amostras que possam ser utilizadas para compor as regras que ajudarão a alocar novos objetos em diferentes grupos de tal maneira que a probabilidade de má classificação seja mínima, desse modo estabelecem-se funções das variáveis observadas que sejam responsáveis ou possam explicar as diferenças entre populações (Johnson & Wichern, 2001; Cruz et al., 2011; Ferreira, 2011).

O modelo matemático proposto para discriminação de classes foi inicialmente estabelecido por Fisher (1936), que propôs um critério para separação de três populações de plantas por meio das medidas das suas folhas. A ideia era transformar observações multivariadas em observações univariadas por meio de combinações lineares das variáveis estudadas de tal forma que era possível um critério de separação entre as populações. Tal método é conhecido como análise discriminante de Fisher e a pressuposição quanto a seu uso é que as duas populações em estudo tenham matrizes de covariâncias homogêneas, ou seja, as variáveis que caracterizam os elementos dos grupos possuem covariâncias estatisticamente iguais (Cruz et al., 2011; Ferreira, 2011).

Considere que haja diversas variáveis para caracterizar as populações (π_i, com $i = 1, 2, \cdots, g$), as quais apresentam matriz de variâncias e covariâncias homogêneas, denotada por $\sum$. Com isso dito, admita então que em cada população o vetor de observações $\tilde{X}$ tem distribuição $N_v(\mu_i, \sum)$. Assim, $\tilde{X}$, em uma determinada população π_i, apresenta a seguinte função densidade de probabilidade:

$$f_i(\tilde{X}) = |2\pi\sum|^{-1/2} e^{-\frac{1}{2}\left[(\tilde{X}-\mu_j)'\Sigma^{-1}(\tilde{X}-\mu_j)\right]} \qquad (1)$$

Tomando um par de populações π_1 e π_2, aloca-se um indivíduo, com vetor de observações $\tilde{X}$, em π_1 se:

$$\frac{f_1(\tilde{X})}{f_2(\tilde{X})} > 1 \qquad (2)$$

O que, por analogia ao descrito anteriormente, significa que:

$$(\tilde{X} - \mu_1)' \sum^{-1} (\tilde{X} - \mu_1) < (\tilde{X} - \mu_2)' \sum^{-1} (\tilde{X} - \mu_2) \qquad (3)$$

Dessa forma que a equação (3) possa ser reescrita como:

$$(\tilde{X} - \mu_1)' \sum^{-1} (\tilde{X} - \mu_1) - (\tilde{X} - \mu_2)' \sum^{-1} (\tilde{X} - \mu_2) < 0 \qquad (4)$$

Expandindo-a tem-se:

$$\begin{matrix} (\tilde{X}' \sum^{-1} \tilde{X} - \tilde{X}' \sum^{-1} \mu_1 - \mu_1' \sum^{-1} \tilde{X} + \mu_1' \sum^{-1} \mu_1) \\ - (\tilde{X}' \sum^{-1} \tilde{X} - \tilde{X}' \sum^{-1} \mu_2 - \mu_2' \sum^{-1} \tilde{X} + \mu_2' \sum^{-1} \mu_2) \end{matrix} \qquad (5)$$

Tendo-se para os escalares, a seguinte igualdade:

$$\tilde{X}' \sum^{-1} \mu_1 = \mu_1' \sum^{-1} \tilde{X}$$

$$\tilde{X}' \sum^{-1} \mu_2 = \mu_2' \sum^{-1} \tilde{X}$$

Logo:

$$-2\tilde{X}' \sum^{-1} \mu_1 + \mu_1' \sum^{-1} \mu_1 + 2\tilde{X}' \sum^{-1} \mu_2 + \mu_2' \sum^{-1} \mu_2 < 0 \qquad (6)$$

Arranjando os elementos, somando e subtraindo os escalares $\mu_1' \sum^{-1} \mu_2 = \mu_2' \sum^{-1} \mu_1$ na expressão anterior, chegamos a:

$$-2\tilde{X}' \sum^{-1} (\mu_1 - \mu_2) + (\mu_1' \sum^{-1} \mu_1 + \mu_1' \sum^{-1} \mu_2) - (\mu_2' \sum^{-1} \mu_2 + \mu_2' \sum^{-1} \mu_1) < 0 \qquad (7)$$

$$-2\tilde{X}' \sum^{-1} (\mu_1 - \mu_2) + \mu_1' \sum^{-1} (\mu_1 + \mu_2) - \mu_2' \sum^{-1} (\mu_2 + \mu_1) < 0 \qquad (8)$$

$$-2\tilde{X}' \sum^{-1} (\mu_1 - \mu_2) + (\mu_1 - \mu_2)' \sum^{-1} (\mu_1 + \mu_2) < 0 \qquad (9)$$

Dessa forma tem-se:

$$\tilde{X}' \sum^{-1} (\mu_1 - \mu_2) - (\mu_1 - \mu_2)' \sum^{-1} \frac{1}{2} (\mu_1 + \mu_2) > 0 \qquad (10)$$

Segundo Cruz et al. (2011), com a função discriminante estimada, adota-se a regra de classificação de modo que, dado um vetor de observações:

a) Aloca-se $\tilde{X}$ em π_1 se:

$$\tilde{X}' \sum^{-1} (\mu_1 - \mu_2) \geq (\mu_1 - \mu_2)' \sum^{-1} \frac{1}{2} (\mu_1 + \mu_2) \qquad (11)$$

b) Aloca-se $\tilde{X}$ em π_2 se:

$$\tilde{X}'\sum^{-1}(\mu_1 - \mu_2) < (\mu_1 - \mu_2)'\sum^{-1}\frac{1}{2}(\mu_1 + \mu_2) \qquad (12)$$

2.3. Análise Discriminante Quadrática

Quando o pressuposto da igualdade das matrizes de covariância (∑) não é atendido, ou seja, as p matrizes de variância são heterogêneas, recorre-se então à análise discriminante quadrática.

Considere, como exemplo, o caso em que se tem p > 1 variáveis medidas em cada elemento amostral de cada população e provenientes de distribuições normais p-variadas. Suponha $\tilde{X}$ como um vetor das características da população proveniente de uma distribuição normal, dado duas populações A e B com vetor de médias $\boldsymbol{\mu}_A$e $\boldsymbol{\mu}_B$ e matrizes de covariâncias $\boldsymbol{\Sigma}_A$ e $\boldsymbol{\Sigma}_B$, sendo que $\boldsymbol{\Sigma}_A \neq \boldsymbol{\Sigma}_B$. Para um vetor de observações fixo $\boldsymbol{x}^T = [x_1 x_2 \dots x_p]$, a razão entre as funções densidade de probabilidade das duas populações, em termos de logaritmo neperiano, será:

$$-2\ln\big(\lambda(x)\big) = -2ln\left\{\frac{(2\pi)^{\frac{p}{2}}\left(|\Sigma_A|^{\frac{1}{2}}\right)^{-1}}{(2\pi)^{\frac{p}{2}}\left(|\Sigma_B|^{\frac{1}{2}}\right)^{-1}}\left[\frac{exp\left\{-\frac{1}{2}(x-\mu_A)'{\Sigma_A}'(x-\mu_A)\right\}}{exp\left\{-\frac{1}{2}(x-\mu_B)'{\Sigma_B}'(x-\mu_B)\right\}}\right]\right\} \qquad (13)$$

Assim, um novo indivíduo com vetor de observações $\boldsymbol{x}$ será classificado como pertencente à população A, se $-2\ln\big(\lambda(\boldsymbol{x})\big)$ for maior ou igual a zero e será classificado como sendo da população B, se for menor que zero, tal função passa a ser denominada como função discriminante quadrática (Mingoti, 2007).

2.4. Pressupostos da Análise Discriminante

De acordo com Ferreira (2011) a Análise Discriminante linear de Fisher não pressupõe que as populações sejam provindas de distribuições normais multivariadas, mas assume-se homogeneidade das matrizes de covariâncias. Se este pressuposto não for satisfeito é indicado o uso da função discriminante quadrática em que o seu pressuposto assume que as populações sejam provindas de distribuições normais multivariadas. Alguns trabalhos fazem uma comparação entre o desempenho das duas funções para classificação e a alocação correta dos indivíduos assumindo populações normais multivariadas.

No trabalho de Gilbert (1969), a função quadrática se sobressai para quando as matrizes de covariâncias não são iguais, mas o autor comenta que apesar disso a

função linear foi adequada para a classificação. Já Dunn e Marks (1974) observaram que para grandes amostras e autovalores grandes, a função quadrática se sobressai com relação a linear, mas para autovalores pequenos a diferença é mínima, no entanto para pequenas amostras a função linear tem um desempenho muito melhor do que a quadrática para autovalores pequenos se tornando mais claro a medida que aumenta o número de parâmetros. Kronmal e Wahl (1977) afirmam que mesmo com a heterogeneidade das matrizes de covariâncias, a função linear é a mais comum devido à sua simplicidade e a indicam para uso quando o tamanho da amostra é pequeno. Krzanowski (1977) já apresenta alguns problemas que surgem com o desempenho da função linear sob condições não ótima, como a perda gradativa da performance, mas destaca que a simplicidade da técnica a torna bastante utilizável para comparar e avaliar a classificação.

Para a verificação quanto a população ser provinda de uma distribuição normal multivariada, geralmente a primeira inspeção dos dados se faz graficamente, mas para sua comprovação utiliza-se um teste de hipótese, como exemplo, podemos citar o teste de Mardia que é baseado nas extensões multivariadas para assimetria e curtose (Shipley, 2016).

No que tange a homocedasticidade das matrizes de covariância, é necessário um teste que compare a variação em duas ou mais amostras multivariadas, desta forma o Teste M de Box (Box M-test) pode ser aplicado desde que a premissa de normalidade seja satisfeita (Manly & Navarro Alberto, 2017).

2.5. Redes Neurais Artificiais

As Redes Neurais Artificiais (RNA) funcionam de forma a simular o comportamento do cérebro humano, ou seja, sua estrutura neural. Dessa forma a rede adquire a capacidade de aprender por meio da experiência e desse modo resolvendo problemas de maiores complexidades para o cérebro humano como por exemplo, os problemas de predição, reconhecimento de padrões e classificação. Com base nos resultados é possível se fazer generalizações baseadas no seu conhecimento previamente acumulado (Mackay,1994; Cruz & Nascimento, 2018).

Dessa forma, juntando-se a velocidade de processamento computacional com o procedimento humano de aprendizagem torna-se possível resolver vários problemas que vem surgindo. Embora biologicamente inspiradas, as Redes Neurais

encontram aplicações em diferentes áreas científicas. Cita-se como exemplo, Barroso et al. (2013) que utilizaram as redes neurais juntamente com a análise discriminante de forma a comparar seus resultados com os obtidos pela metodologia de Eberhart e Russel para análise da adaptabilidade e estabilidade fenotípica de genótipos da alfafa. 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, (2016) no qual os autores propuseram as redes neurais como alternativa na predição de valores genéticos utilizando para tal uma base de dados simulados.

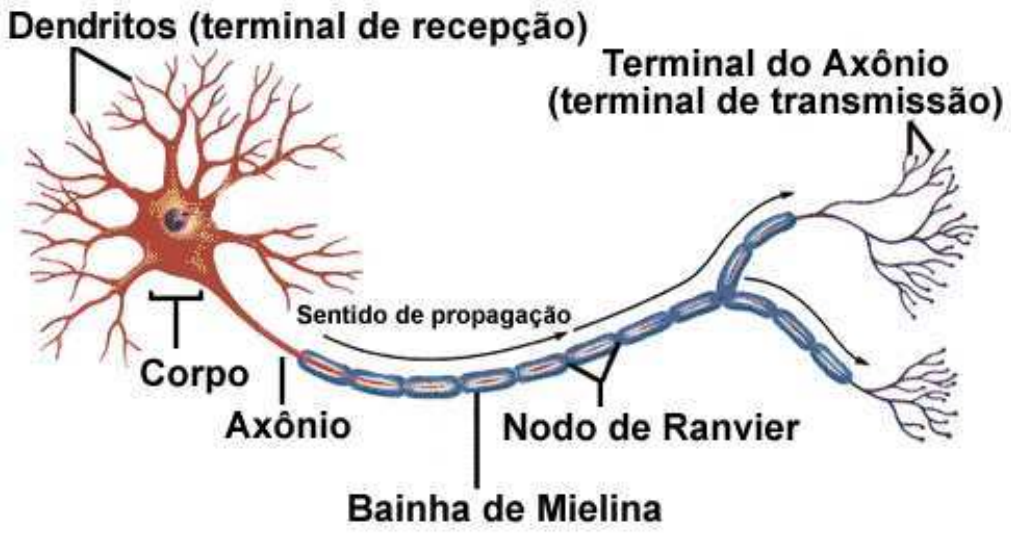
2.5.1. Inspiração Biológica da Redes Neurais Artificias

Como dito anteriormente, as Redes Neurais foram inspiradas baseando-se no comportamento do cérebro humano, que possui como unidade básica os neurônios, que são os responsáveis pela transmissão das informações de um neurônio para outro.

A

Figura 1 apresenta o modelo de um neurônio biológico, o qual é constituído por corpo celular, axônio e dentritos.

Figura 1 - Modelo de Neurónio Biológico.



Fonte: Inteligência Artificial[1]

[1] Disponível em: <https://guildadocodigo.atelie.software/intelig%C3%AAncia-artificial-parte-i-ebd62adbc10> Acesso em: 11 jul. 2019.

Os dendritos tem a função de receber os estímulos e enviar os sinais das extremidades para o corpo celular. No corpo celular, esses estímulos serão processados e então enviados até os axônios, as extremidades do axônio de um neurônio são conectados com os dentritos de outros neurônios por meio das sinapses, que por sua vez controlam a transmissão de impulsos, formando assim as redes (Braga et al., 2000; Cruz & Nascimento, 2018).

2.5.2. Modelo Matemático dos Neurônios Biológicos

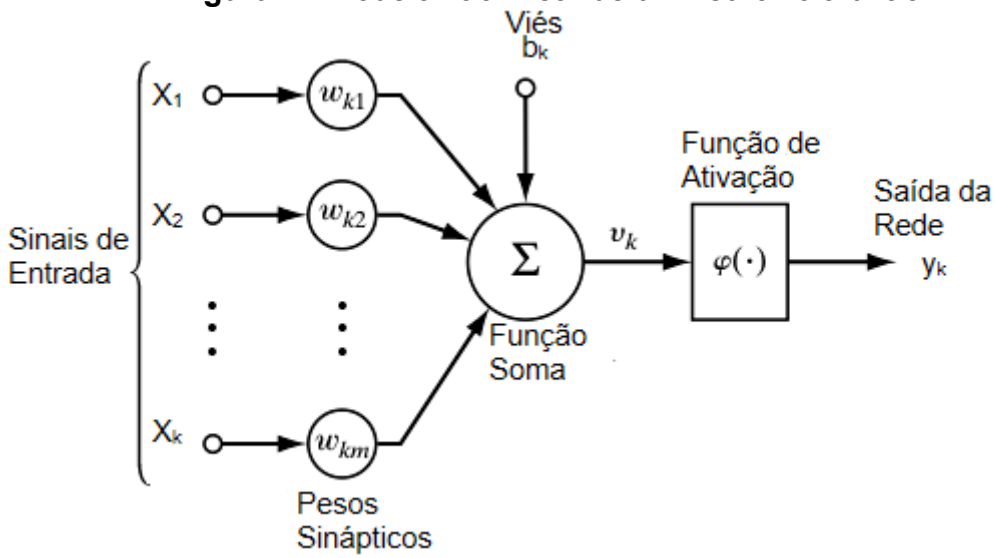
O surgimento de um modelo artificial de neurônio biológico começou por volta de 1943 graças ao trabalho feito pelo neuroanatomista e psiquiatra Warren McCulloch e do matemático Walter Pitts (McCulloch & Pitts, 1943), estas publicações introduziram o primeiro modelo de redes neurais por simulação.

Porém anos mais tarde surgem novos modelos mais avançados para resolução de problemas mais complexos, sendo o mais conhecido e ainda aplicado o modelo de Perceptron Múltiplas Camadas (Multi Layer Perceptron – MLP) (Figura 2), com o algoritmo backpropagation que torna o treinamento da rede muito mais rápida e eficiente, as redes neurais artificiais vem se tornando bem recebida em várias áreas da ciência (Braga et all., 2000).

Logo, a utilização de uma RNA envolve um conjunto de exemplos para que a rede possa aprender, dessa forma o treinamento ocorre de forma a ajustar os pesos das sinápticos, esses estão sujeitos a sucessivos ajustes, para que a saída da rede seja a mais próxima possível da resposta desejada (Cruz et al, 2018).

A arquitetura de uma rede neural pode ser definida como o 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, 2007). Na Figura 2 é ilustrado um modelo básico de um neurônio artificial.

Figura 2 - Modelo não linear de um neurônio artificial



Fonte: Adaptado de Haykin, 2001.

Nesta figura $x_1, x_2, \dots, x_m$ representam as entradas da rede; $W_{k1}, W_{k2}, \dots, W_{km}$, 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:

$$y_k = \varphi\left(\sum_{j=1}^{m} w_{kj} x_j + b_k\right) \qquad (14)$$

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 tal 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.

2.5.3. Funções de Ativação

A função de ativação tem grande importância no uso das redes neurais, com ela gera-se a saída do neurônio a partir das somas ponderadas recebidas pelo neurônio. Dessa forma a saída de um neurônio é limitada geralmente nos intervalos de $[0,1]$ ou $[-1,1]$ (Haykin, 2001).

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).

Diversos trabalhos utilizam as funções de ativação na Rede Neural, seu uso é dependente do comportamento dos dados ou da própria experiência do pesquisador, dentre os trabalhos mais recentes citam-se:

Junior et al. (2019) utilizaram uma Rede Neural do tipo *feed-forward* com uma camada escondida utilizando função ativação tipo sigmoide para classificação de imagens radiografadas de sementes de girassol quanto ao seu nível de dano.

Barreto et al. (2017) desenvolveram uma Rede Neural para auxiliar no diagnóstico de Câncer Cervical por meio da biópsia em um paciente, os dados consistiam nos fatores de risco do câncer cervical, e para o treinamento das Redes foram escolhidas duas opções para as funções de ativação: *Exponential Linear Unit* e a Sigmoid.

Gongorra et al. (2016) propuseram um modelo alternativo baseado nas Redes Neurais Artificiais para classificação e detecção de falhas de rolamentos em indução trifásica motores conectados diretamente à rede elétrica, dentre as etapas quanto a estrutura da rede Neural, contou-se com apenas dois neurônios de camada de saída tendo como função de ativação a tangente hiperbólica.

Dessa forma, as funções de ativação desempenham um papel importante para a arquitetura da Rede Neural, descrevemos abaixo as principais funções de ativação geralmente utilizadas:

2.5.3.1. Função Limiar (Degrau)

Esta função foi utilizada no modelo de McCulloch e Pits e é expressa pela equação (15).

$$f(x) = \begin{cases} 1, se\ v \geq 0; \\ 0, se\ v \leq 0; \end{cases} \qquad (15)$$

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.

2.5.3.2. Função "Sigmoidal" ou "Sigmóide"

A função sigmoide 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 (16).

$$f(x,\alpha) = \frac{1}{1+e^{-\alpha x}} \qquad (16)$$

Em que α é o parâmetro da função sigmoide.

2.5.3.3. 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 (17)

$$\tanh\left(\frac{x}{2}\right) = \frac{(1-e^{-x})}{(1+e^{x})} \qquad (17)$$

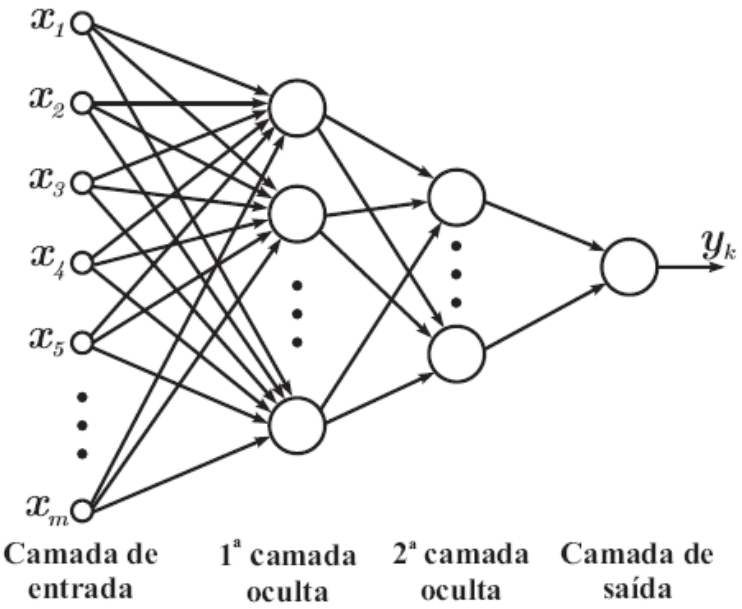
2.5.4. Rede Perceptron Multicamadas

A rede neural mais utilizada para soluções de problemas de predição e classificação é conhecida como Perceptron Multicamadas (MLP) (Nakai et al, 2015; Santos, 2016; Aquino, 2016). Tais redes podem possuir uma ou mais camadas de neurônios entre as camadas de entrada e saída, estas camadas centrais são conhecidas como camada oculta (Sant'anna, 2014).

O fluxo da informação da MLP se propaga para frente camada por camada desde a entrada até a saída, por isso ela também é conhecida como rede Feed-Forward, o exemplo do seu funcionamento pode ser visto na

Figura 3.

Figura 3 - Exemplo de Rede Neural *Feedforward* Multicamadas



Fonte: (Neto et al., 2009)

Os modelos de redes propostos por McCulloch e Pitts, apresentavam apenas camadas de entrada e saída, dessa forma o sistema da rede não conseguia aprender problemas não linearmente separados, dessa forma, anos mais tarde surgiu-se o modelo Perceptron Multicamadas (Cruz, 2018).

Este modelo é baseado no algoritmo de retro propagação (*back-propagation*), baseando-se na retropropagação dos erros, os ajustes dos pesos das camadas intermediárias são obtidos reajustando os pesos de cada entrada que é apresentada na rede, por meio do aprendizado supervisionado resolveu-se assim muitos problemas não linearmente separáveis. Dado um número grande de neurônios (Zeidenberg, 1990)

2.6. As Árvores de Decisão

A Árvore de Decisão é um processo utilizado para o desenvolvimento de modelos que auxiliam na tomada de decisão, são caracterizadas como sendo estruturas gráficas hierárquicas de fácil entendimento e aplicação, sua importância se dá por sua capacidade preditiva, ou seja, o modelo tem a capacidade de predizer o valor de uma variável com um certo nível de certeza (Cervantes et al., 2015; Ramya et al., 2015).

Uma árvore de decisão é composta por uma cadeia nós interconectados por ramificações estendendo-se desde o nó raiz até os nós folhas (Last et al., 2016; Larose, 2014). Segundo Kliemann Neto (2011), existem quatro tipos de nós: o nó raiz que marca o início da árvore, nós de probabilidade, nós de decisão e nós de término. O nó de probabilidade mostra as probabilidades de certos resultados, o nó de decisão mostra uma decisão a ser tomada dada as alternativas, e um nó de término mostra o resultado final de um caminho de decisão.

Sua principal vantagem está em demonstrar claramente as decisões que devem ser tomadas, a ordem de cada decisão deverá eventualmente ser tomada, as consequências relacionadas de cada decisão e o grau de incerteza do mesmo. Porém as árvores de decisão podem se tornar imensas dependendo do número de alternativas que possam ser tomadas, ou seja, muitos dos nós da árvore podem possuir ruídos de classificação acarretando um problema de super ajustamento (*Overfitting)* impossibilitando ao modelo generalizar seus resultados.

Para contornar essa situação, são utilizados os métodos de poda, cujo objetivo é melhorar a taxa de acerto do modelo para se adequar aos novos exemplos que foram utilizados no conjunto de treinamento (Han, 2001).

Dentre eles tem-se os métodos de *Bagging*, *Random Forest* e *Boosting*.

2.6.1. Arvores de Classificação

Quando a variável resposta é uma variável qualitativa dizemos então que a Análise Discriminante é classificada como árvore de classificação.

Para construção dessa árvore de classificação deve-se estabelecer regras para atribuição das classes, desse forma o objetivo passa a ser obter regiões do tipo $R_1, R_2, R_{3,} \cdots, R_M$ em que minimizam o erro esperado de classificação. Dentre os métodos propostos para esta finalidade Hastie et al. (2009) apresentam três deles:

Taxa de Erro Aparente:

$$TEA = 1 - \max_{k}(\hat{p}_{mk}) \quad (18)$$

Índice de Gini:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad (19)$$

Entropia Cruzada:

$$EC = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \quad (20)$$

Sendo que $\hat{p}_{mk}$ é a proporção de observações na m-ésima região e pertencente a k-ésima classe.

Todos os três métodos são semelhantes, mas a Entropia Cruzada e o índice de Gini são diferenciáveis e, portanto, mais receptivo quanto a otimização numérica e também tende a serem mais sensíveis a mudanças nas probabilidades do nó do que a taxa de erro aparente. Por essa razão, o índice de gini ou a entropia cruzada são os melhores quanto ao crescimento da árvore (James et al., 2013).

2.6.2. Bagging

Bagging (*Bootstrap* Aggregation) é uma técnica proposta por Breiman (1996a, 1996b). Ele pode ser utilizado para melhorar o poder de predição e estabilidade das árvores de classificação por estas possuírem uma alta variabilidade, ou seja, se o número de níveis for muito alto o modelo tende a super ajustar. Para contornar esse problema a melhor alternativa seria construir várias árvores e em seguida obter a média/moda dos valores preditos. Mas como na prática nem sempre essa alternativa é possível, Breiman, (1996a; 1996b) propôs uma nova abordagem por meio da técnica de *bootstrap*. Tal técnica envolve uma amostragem aleatória de pequenos subconjuntos dos dados originais com reposição e com mesma probabilidade, visando diminuir a variabilidade dos algoritmos de treinamento como por exemplo as árvores de decisão.

Sua aplicação envolve uma amostra de tamanho B com reposição do tipo $\hat{f}^1(x), \hat{f}^2(x), \cdots, \hat{f}^B(x)$, treinar modelo para cada amostra e com os novos dados calcular uma predição baseado na amostragem de cada modelo como segue:

$$\hat{f}_{médio}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x) \qquad (21)$$

Diminuindo assim a alta variabilidade gerada pelas árvores de decisão (Han, 2001).

O único parâmetro a ser considerado quanto a utilização do *bagging* é o número de amostras e, portanto, o número de árvores que devem ser incluídas. Isso pode ser escolhido aumentando gradativamente as amostras até que a precisão comece a parar de mostrar melhoria. Um grande número de modelos pode levar muito tempo para ser preparado, mas não sobrecarregará os dados de treinamento sendo que nem todos são utilizados, sendo estes chamados de OOB *(out-of-bag*).

2.6.3. Random Forest

Ao utilizar o *Bagging*, todas as variáveis são escolhidas para cada partição das amostras, dessa forma as predições de uma árvore de decisão acabam por serem altamente correlacionadas devido a semelhanças de cada resultado. Isto influência diretamente na variância já que uma média correlacionada não resulta em menor variabilidade. Dessa forma para contornar essa situação foi proposto uma modificação dentro do *bagging* criando assim o *Random Forest* (Breiman, 2001).

O *Random Forest* adiciona uma aleatoriedade extra ao modelo feito pelo *bagging*, como vimos o *bagging* constrói cada árvore usando uma diferente amostra *bootstrap* dos dados, em adição a isso o *Random Forest* mudam como a árvore de classificação é gerada. Nas árvores feitas de forma padrão, cada nó é dividido usando a melhor divisão entre todas as variáveis, já na *Random Forest* cada nó é dividido usando o melhor entre um subconjunto de preditores escolhidos aleatoriamente nesse nó. Isto acaba por gerar um desempenho muito melhor em comparação a muitos outros classificadores, incluindo análises discriminantes, máquinas de vetores de suporte e redes neurais, e é robusta contra super ajustamento (Breiman, 2001). É necessário, no entanto estipular o número de

variáveis preditoras (p) que serão utilizados em cada partição, sendo $m < p$. Para árvores de classificação sugere-se utilizar $m = \sqrt{p}$ (Hastie, 2009).

2.6.4. Boosting

Considerada uma poderosa ferramenta para predição (Hastie, 2009), foi proposto inicialmente por Breiman (1996a, 1996b) que encontrou vários ganhos na precisão pela combinação dos classificadores. A partir do seu desenvolvimento muitas versões do *boosting* surgiram como o Real *AdaBoost* (Freund and Schapire 1996; Schapire and Singer 1999), *LogitBoost* (Friedman, Hastie, and Tibshirani 2000), e o *gradient boosting* (Friedman 2001), no entanto é considerado um método mais complexo de aplicação do que relação ao *Bagging*.

Porém de modo geral a maior diferença com relação ao *Bagging* está no esquema de amostragem, enquanto ao *Bagging* usa uma média dos resultados obtidos dos vários modelos amostrados, o *Boosting* possui um peso nas observações, e a cada passo do treinamento esses pesos são modificados a fim de que os melhores classificadores sejam incorporados ao novo conjunto amostrado melhorando assim o resultado final na predição (Sutton, 2005).

Porém possui uma aprendizagem lenta, necessitando que a quantidade amostrada seja grande, isso, no entanto pode causar um super ajustamento do modelo, dessa forma recomenda-se utilizar a validação cruzada para escolher o melhor número de árvores a ser construída.

2.7. Máquina de Vetor Suporte para Classificação

As Máquinas de Vetores Suporte (MVS) são algoritmos de aprendizagem de máquina supervisionado, ou seja, dado um conjunto de dados de treinamento é possível descobrir relações entre as variáveis independentes (Variável de entrada) e a variável dependente (Variável de Saída) (Huang & Learned-Miller, 2014). Foi proposta inicialmente por Cortes e Vapnik (1995) como um campo de pesquisa de Inteligência Computacional possuindo propriedade que permitiram uma boa generalização dos resultados dos dados (Smola & Schölkop, 2004).

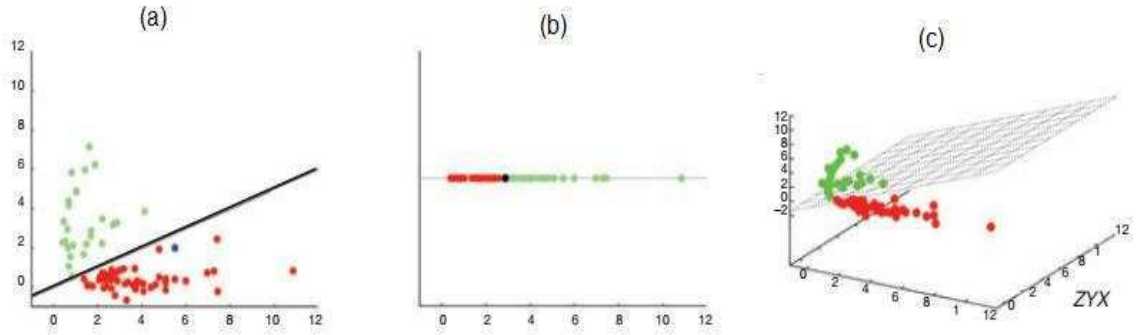
Para melhor compreensão do algoritmo é importante detalhar alguns conceitos como: hiperplano de separação ótimo, vetores suporte, margem máxima, hiperplano de separação ótimo, margem suave e Funções Kernel.

2.7.1. Hiperplano de Separação Ótimo

O hiperplano e separação constitui uma função capaz de separar os dados de treinamento de forma que a separação entre seus exemplos positivos e negativos seja máxima, ou seja, a função maximiza a distância em relação aos dados do conjunto de treinamento (Haykin, 2007; Lorena & Carvalho, 2003). 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.

A função que separa os dados de treinamento pode ser um simples ponto, ou de classificação unidimensional (Figura 4a), uma reta (Figura 4b) ou um plano (Figura 4c) (Noble, 2006)

Figura 4 - Hiperplano de separação com dimensão $\boldsymbol{R^2, R^1}$ *e* $\boldsymbol{R^3}$ respectivamente.



Fonte: Noble (2006)

O hiperplano pode ser descrito como uma função da forma:

$$f(x) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n = 0 \qquad (22)$$

Que pode ser reescrita como (Vapnik, 1995; Lorena & Carvalho, 2003):

$$w \cdot x + b = 0 \qquad (23)$$

Onde w_0 é uma constante descrita como b (Vapnik, 1995), esta equação poder também ser reescrita de forma matricialmente por $w^Tx + b = 0$.

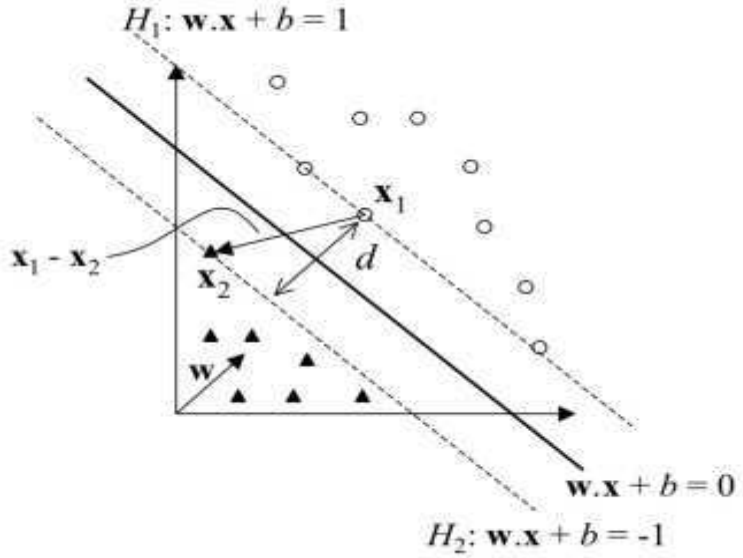
Em um problema de classificação binário, a equação (23) divide o espaço dos dados X em duas regiões: $w \cdot x_i + b \geq 0$ e $w \cdot x_i + b < 0$. Uma função sinal é aplicada sobre essa equação resultando na classificação +1 se $f(x) \geq 0$ e -1 se $f(x) < 0$, esses hiperplanos que se ajustam para $w \cdot x_i + b = +1$ e $w \cdot x_i + b = -1$ são chamados de margem de separação (Lorena & Carvalho, 2003).

Para gerar então um hiperplano de separação ótimo, ou seja, que possui essa margem máxima, conta-se com os chamados vetores de suporte (*support*

vectors), que são os pontos de treinamento mais próximos ao hiperplano e que coincidem com a margem de separação, a função da MVS então passa a ser orientar o hiperplano de modo que ele fique o mais distante possível dos vetores suporte de ambas as classes. Dessa forma o algoritmo trabalha de forma a maximizar a distância entre o hiperplano e os vetores suporte (Fletcher, 2009).

A Figura 5 mostra um hiperplano de separação de classificação binária e os vetores suportes ligado a ele (H_1 e H_2).

Figura 5 - Hiperplano de separação e as margens do hiperplano que passam pelos vetores suportes



Fonte: Lorena & Carvalho (2003).

Sendo:

- w é um vetor de parâmetros normal ao hiperplano;
- d é a distância entre os hiperplanos H_1 e H_2, paralelos ao hiperplano separador.

A distância d forma o comprimento da margem da MVS, o algoritmo trabalha de forma a maximizar a distância entre o hiperplano e os pontos mais próximos a ele, através da maximização do cálculo da norma dos vetores de forma a possuir margem máxima calculada com a maximização das normas dos vetores, a esse hiperplano damos o nome de hiperplano ótimo (Fletcher, 2009; Lorena & Carvalho, 2003).

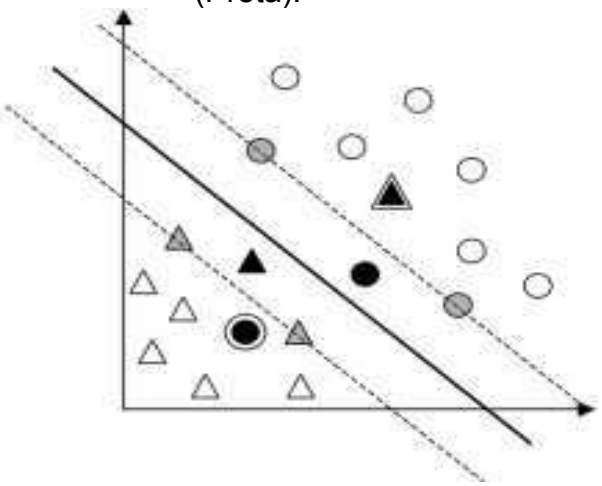
2.7.2. Margens Suave

Porém existem muitos casos em que a classificação não é possível de se separar por uma linha reta, devido à natureza do fenômeno estudado ou por apresentar ruídos (*outliers*). Isso acarreta em erros de classificação pois podem existir pontos fora da região de classificação de uma determinada classe por não terem sido captados adequadamente (Lorena & Carvalho, 2003).

Porém ainda é possível encontrar um hiperplano ótimo que minimize a probabilidade de erro de classificação, calculado sobre o conjunto de treinamento. Para o conjunto de dados não linearmente separáveis é necessária a introdução de uma variável de folga não negativa ε que indica o erro de classificação associada a cada amostra. Dessa forma esse parâmetro ira controlar a quantidade de exemplos que poderão violar a função classificadora permitindo alguns erros de classificação, isto faz com que outliers sejam ignorados e possam estar em regiões de classificação diferentes sem alterar o resultado final (Lorena & Carvalho, 2003; Noble, 2006; Bem-Hur et al., 2008).

As Máquinas de Vetores Suporte (MVS) obtidas nesse caso são conhecidas como MVS de margens suaves. Nela, pontos incorretamente classificados, ou seja, que estão no lado incorreto do hiperplano são penalizados, a Figura 6 mostra um exemplo de variáveis soltas.

Figura 6 - Exemplo ilustrativo de vetores suporte (cinza) e variáveis de folga (Preta).



Fonte: Lorena & Carvalho (2003)

O uso dessa ideia parte de que alguns erros são esperados, dessa forma a margem de aceite desses erros pode ser minimizada. Seja ε_i a margem de folga, temos então que:

$$\text{Minimizar: } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}\varepsilon_i \qquad (24)$$

$$\text{Sujeito a } y_i(x_i \cdot w + b) - 1 + \varepsilon_i \leq 0\ \forall \qquad (25)$$

Em que C é o parâmetro de regularização do algoritmo de aprendizagem de máquina, isso impõe um peso à minimização dos erros no conjunto de treinamento em relação à minimização da complexidade do modelo evitando assim o super ajustamento (overffiting) (Girosi, 1997; Smola & Schölkopf, 1998; Lorena & Carvalho, 2003).

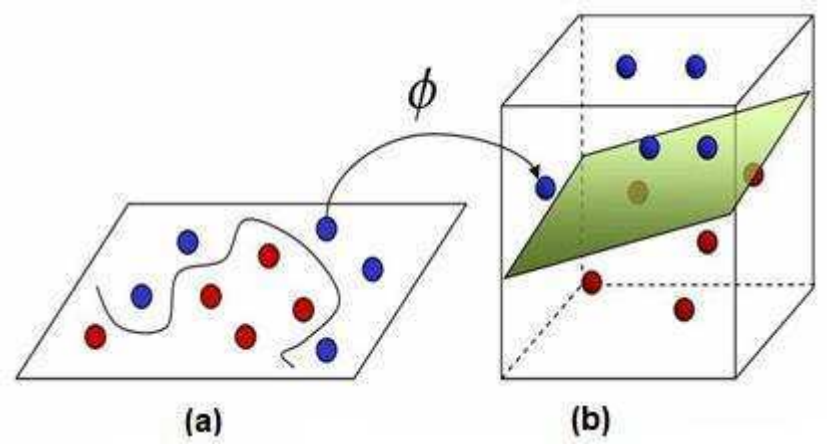
2.7.3. Funções Kernel

Em casos de classificações ditas não lineares, a utilização de um hiperplano pode se tornar mais complexa devido aos dados de entrada, exigindo que o separador também seja não-linear. Dessa forma a MVS (Máquina de Vetor Suporte) lida com esse problema não linear transformado com alta probabilidade o problema em algo linearmente separável, criando um espaço de maior dimensionalidade.

A MVS 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 (Smola & Schölkopf, 2004)

As funções kernels baseiam-se no produto interno dos dados de entrada e os reformula em um espaço de dimensão maior através de uma função não linear capaz de mapear $x \rightarrow \Phi(\mathrm{x})$ (Figura 7)

Figura 7 - Transformação de um problema não linear em um problema linearmente separável.



Fonte: Estratégia computacional para avaliação de propriedades mecânicas de concreto de agregado leve - *Scientific Figure on ResearchGate*[2]

A função kernel pode ser descrita por $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$, onde K é representado pelo produto interno dos dados de entrada mapeados em um espaço de características maior transformados por Φ (Fletcher, 2009; Lorena & Carvalho, 2003). A Tabela 1 apresenta as principais funções kernels usadas pela SVM.

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

Fonte: Lorena & Carvalho (2003)

Porém a função RBF (*Radial Basis Function*) é a opção mais amplamente utilizada para as MVS, devido as respostas finitas e localizadas por toda a variação do eixo x real (Lorena & Carvalho, 2003).

2.8. A importância da Simulação Computacional em programas de melhoramento genético.

Os programas de melhoramento genético têm por objetivo desenvolver indivíduos com características desejáveis. Para tal finalidade vários métodos de

[2] Adaptado de <https://www.researchgate.net/figure/Figura-215-Classicacao-perfeita-pelo-hiperplano-otimo-do-SVM-com-kernel-nao-linear_fig1_318598388> Acesso em: 11 jul. 2019.

seleção podem ser utilizados, como por exemplo a seleção fenotípica e a seleção genômica. Porém a maioria desses métodos de seleção requer a obtenção de um banco de dados, o que exige tempo e recurso (Silva, 2014).

Os avanços computacionais trouxeram grande contribuição para a área da pesquisa e melhoramento, isso se deve ao fato de um fenômeno poder ser estudado simulando uma situação complexa em que os parâmetros a serem estudados possam ser controlados pelo pesquisador observando o estudo de seus efeitos. Simulação pode ser definida como um caminho para imitar o comportamento de um sistema real para estudar seu funcionamento sob algumas condições e envolvendo uma certa lógica levando a imitar o máximo possível o modelo real (Cruz, 2013). Uma grande vantagem da simulação é a possibilidade de replicação do experimento já que nem sempre é fácil fazê-lo ao se tratar de problemas reais.

Para os estudos de melhoramento genéticos, a simulação é de grande importância por proporcionar o estudo de populações, indivíduos ou do próprio genoma. Porém é necessário o desenvolvimento do modelo biológico apropriado no qual represente o fenômeno de interesse dado pelo pesquisador e os parâmetros cuja influência no modelo será estudada (Cruz, 2013).

REFERÊNCIAS

AZEVEDO, A. M.; DE ANDRADE JÚNIOR, V. C.; FERNANDES, J. S. C. Transformação Box-Cox na homecedasticidade e normalidade uni e multivariada em experimentos de batata-doce. **Horticultura Brasileira**, v. 34, n. 1, 2015.

AQUINO, C. F.; SALOMÃO, L. C. C.; AZEVEDO, A. M. Fenotipagem de alta eficiência para vitamina A em banana utilizando redes neurais artificiais e dados colorimétricos. **Bragantia**, v. 75, n. 3, p. 268-274, 2016.

BARRETO, R. G.; MARINHO, G. M. G. A.; BARRETO, G. F. M.; BARRETO, R. G.; AVERSARI, L. O. C.; DANTAS, B. L. Utilizando redes neurais artificiais para o diagnóstico de câncer cervical. **Revista Saúde & Ciência Online**, v. 7, n. 2, p. 59-67, 2018.

BARROSO, L. M. A.; NASCIMENTO, M.; NASCIMENTO, A. C. C.; Silva, F. F.; FERREIRA, R. D. P. Uso do método de EBERHART e RUSSELL como informação a priori para aplicação de redes neurais artificiais e análise discriminante visando a classificação de genótipos de alfafa quanto à adaptabilidade e estabilidade fenotípica. **Embrapa Pecuária Sudeste-Artigo em periódico indexado (ALICE)**, 2013.

BEN-HUR, A.; Ong, C. S.; SONNENBURG, S.; SCHÖLKOPF, B.; RÄTSCH, G. Support vector machines and kernels for computational biology. **PLoS computational biology**, v. 4, n. 10, p. e1000173, 2008.

BENIN, Giovani et al. Identificação da dissimilaridade genética entre genótipos de feijoeiro comum (Phaseolus vulgaris L.) do grupo preto. **Current Agricultural Science and Technology**, v. 8, n. 3, 2002.

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.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, nº 2, pag. 123-140, 1996a.

BREIMAN, L. Heuristics of instability and stabilization in model selection. **The annals of statistics**, v. 24, nº 2, pag. 2350-2383, 1996b.

BREIMAN, L. Random forests. **Machine learning**, v. 45, nº 1, pag. 5-32, 2001.

CARNEIRO, P. L. S.; MALHADO, C. H. M.; DE SOUZA JÚNIOR, A. A. O.; DA SILVA, A. G. S.; DOS SANTOS, F. N.; SANTOS, P. F.; PAIVA, S. R. Desenvolvimento ponderal e diversidade fenotípica entre cruzamentos de ovinos Dorper com raças locais. **Pesquisa Agropecuária Brasileira**, v. 42, n. 7, p. 991-998, 2007.

SANTOS, J. N., GUSMÃO, M. S., & COELHO, C. J. Reconhecimento de Congestionamento de Veículos em Semáforos Empregando Análise de Componentes Principais. **Revista Arithmós-Revista da Escola de Ciências Exatas e da Computação**, v. 1, n. 1, p. 54-60, 2019.

SOUSA, A. D. L.; SALAME, M Tecnologia para identificação de cultivares de guaranazeiro. In: **Embrapa Amazônia Ocidental-Artigo em anais de congresso (ALICE)**. In: JORNADA DE INICIAÇÃO CIENTÍFICA DA EMBRAPA AMAZÔNIA OCIDENTAL, 14., 2017, Manaus. Anais... Brasília, DF: Embrapa, 2018., 2018.

CERVANTES, J.; LAMONT, F. G.; LÓPEZ-CHAU, A.; MAZAHUA, L. R; RUÍZ, J. S. Data selection based on decision tree for svm classication on large data sets, **Applied Soft Computing** 37: 787–798, 2015.

COGO, F. D.; DE ASSIS, T. L.; DE ALMEIDA, S. L. S.; CAMPOS, K. A. Produção sustentável de mudas de cafeeiro: substituição do substrato comercial por esterco bovino. **Intercursos Revista Científica**, 2019.
CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, v. 20, n. 3, p. 273-297, 1995.
CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. A. Biometria aplicada ao estudo da diversidade genética. **Visconde do Rio Branco: Suprema**, v. 620, 2011.
CRUZ, C. D. Genes: a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum. Agronomy**, *35*(3), 271-276, 2013.
CRUZ, C. D.; Nascimento, M. **Inteligência Computacional Aplicada ao Melhoramento Genético.** Viçosa: Editora UFV, 2018.
DOS SANTOS, Michelli de Souza et al. Resistance to water deficit during the formation of sugarcane seedlings mediated by interaction with Bacillus sp. **Científica**, v. 45, n. 4, p. 414-421, 2017.
DUNN, O. J.; MARKS, S. Discriminant functions when covariance matrices are unequal. **Journal of the American Statistical Association**, Alexandria, v. 69, n. 346, p. 555-559, 1974.
FAGUNDES, F. L.; BORGES, C. C. H.; NETO, R. F. Aprendizado de Métrica Utilizando uma Função de Distância Parametrizada e o Algoritmo K-means. **XIII Encontro Nacional de Inteligência Artificial e Computacional, Recife, Brasil**, 2016.
FERREIRA, D. F. **Estatística multivariada**. Editora UFLA, 2008.
FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of human genetics**, 7(2), 179-188, 1936.
FLETCHER, T. Support vector machines explained. **Tutorial paper**, 2009.
FREUND, Y.; SCHAPIRE, R. E. Experiments with a new boosting algorithm. In **Proceedings of the 13th International Conference on Machine Learning**, Vol. 96, pp. 148-156, 1996.
FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). **The annals of statistics**, v. 28, n. 2, p. 337-407, 2000.
FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, p. 1189-1232, 2001.
GILBERT, E. S. The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. **Biometrics, Arlington**, v. 25, p. 505-515, 1969.
GONGORA, W. S.; GOEDTEL, A.; DA SILVA, S. A. O.; GRACIOLA, C. L. Neural Approach to Fault Detection in Three-phase Induction Motors. **IEEE Latin America Transactions**, v. 14, n. 3, p. 1279-1288, 2016.
HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. Elsevier, 2011.
HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The elements of statistical learning: data mining, inference, and prediction, **Springer Series in Statistics**, 2009.
HAYKIN, S. **Redes neurais: princípios e prática**. Bookman Editora, 2007.
HUANG, G. B.; LEARNED-MILLER, E. Labeled faces in the wild: Updates and new reporting procedures. **Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep**, 14-003, 2014.
JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning with Applications in R**. 1st ed. New York, NY: Springer, 2013. 426p.

JOHNSON, R.; WICHERN, D. **Applied Multivariate Statistical Analysis**. 5th Edition, Prentice Hall, Upper Saddle River, 761, 2001.
KLIEMANN NETO, J. F et al. **A Gestão de Riscos como Ferramenta para Aumento da Competitividade das Empresas**. In: OLIVEIRA, V. F.; CAVENAGHI, V.; MÁSCULO, F. S. (Org.). **Tópicos Emergentes e Desafios Metodológicos em Engenharia de Produção: Casos, Experiências e Proposições**. Rio de Janeiro: ABEPRO, 2011. p. 151-210.
KRONMAL, R. A.; WAHL, P. W. Discriminant functions when covariances are unequal and sample sizes are moderate. **Biometrics**, Arlington, v. 33, n. 3, p. 479-484, 1977.
KRZANOWSKI, W. J. The performance of Fisher's linear discriminant function under nonoptimal conditions. **Technometrics, Washington**, v. 19, n. 2, p. 191-200, 1977.
LAROSE, D. T.; LAROSE, C. D. **Discovering knowledge in data: an introduction to data mining**. John Wiley & Sons, 2014.
LAST, M.; TOSAS, O.; CASSARINO, T. G.; KOZLAKIDIS, Z.; EDGEWORTH, J. Evolving classification of intensive care patients from event data. **Artificial intelligence in medicine**, v. 69, p. 22-32, 2016.
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.
MANLY, B. F. J.; ALBERTO J. A. N. **Multivariate Statistical Methods: A Primer**. Fourth edition. Boca Raton: CRC Press. 2017.
MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2007.
NAKAI, M.; JUNIOR, H. G.; AGUIAR, P.; BIANCHI, E.; SPATTI, D. Neural tool condition estimation in the grinding of advanced ceramics. **IEEE Latin America Transactions**, v. 13, n. 1, p. 62-68, 2015.
NASCIMENTO, M.; PETERNELLI, L. A.; CRUZ, C. D.; NASCIMENTO, A. C. C.; FERREIRA, R. D. P.; BHERING, L. L.; SALGADO, C. C. Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. **Crop Breeding and Applied Biotechnology**, v. 13, n. 2, p. 152-156, 2013.
NETO, J. A. O.; VIEIRA, G. M.; FELIX, L. B.; CASTRO, M. A. A. Classificação e Localização de Faltas em um Sistema de Distribuição Industrial Contendo Harmônicos. In: **The 8th Latin-American Congress on Electricity Generation and Transmission-Clagtee**. 2009.
NICOLETTI, M. C. **Tópicos de Aprendizado de Máquina e Técnicas Subjacentes**, Ed. 1ª, Editora CRV, 2018.
NOBLE, W. S. What is a support vector machine?. **Nature biotechnology**, v. 24, n. 12, p. 1565, 2006.
PEIXOTO, L. A. **Redes neurais artificiais na predição do valor genético,** Dissertação (Mestrado em Genética e Melhoramento) – Universidade Federal de Viçosa, 97p, 2013.
RAMYA, M.; LOKESH, V.; MANJUNATH, T.; HEGADI, R. S. A predictive model construction for mulberry crop productivity, **Procedia Computer Science** 45: 156–165, 2015.

REZENDE, M. P. G. D., FERRAZ, P. C., CARNEIRO, P. L. S., & MALHADO, C. H. M. Phenotypic diversity in buffalo cows of the Jafarabadi, Murrah, and Mediterranean breeds. **Pesquisa Agropecuária Brasileira**, v. 52, n. 8, p. 663-669, 2017.
RODRIGUES, Daniele Lima et al. Contribuição de variáveis de produção e de semente para a divergência genética em maracujazeiro-azedo sob diferentes disponibilidades de nutrientes. **Pesquisa Agropecuária Brasileira**, v. 52, n. 8, p. 607-614, 2017.
SANT'ANNA, I. C.; Tomaz, R. S.; SILVA, G. N.; BHERING, L. L.; NASCIMENTO, M.; CRUZ, C. D. Artificial neural networks in genetic classification. **Genetics and Molecular Research**, 2014.
SANTOS, D. S.; ARCE, A. I. C.; PIZA, L. V.; SILVA, A. S.; COSTA, E. J. X.; TECH, A. R. B. Redes bluetooth associadas a redes neurais artificiais para monitoramento de suínos. **Archivos de zootecnia**, v. 65, n. 252, p. 557-563, 2016.
SCHAPIRE, R. E.; SINGER, Y. Improved boosting algorithms using confidence-rated predictions. **Machine learning**, v. 37, n. 3, p. 297-336, 1999.
SHIPLEY, Bill. **Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R**. Cambridge University Press, 2016.
SILVA, G. N. **Redes neurais artificiais: novo paradigma para a predição de valores genéticos**. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, 105p, 2014.
SILVA, G. N.; TOMAZ, R. S.; SANT'ANNA, I. C.; CARNEIRO, V. Q.; CRUZ, C. D.; NASCIMENTO, M. Evaluation of the efficiency of artificial neural networks for genetic value prediction. **Genetic Molecular Research**, v. 15, p. 1-11, 2016.
SMOLA, A. J.; SCHÖLKOPF, B. A Tutorial on Support Vector Regression. **Statistics and Com puting**, v. 14, n. 3, p. 199–222, 2004.
SOKAL, R. R.; ROHLF, F. J. The comparison of dendrograms by objective methods. **Taxon**, 11(2), 33-40, 1962.
SOUSA, A. D. L.; Salame, M. F. A. Uma abordagem comparativa de algoritmos de aprendizado supervisionado para classificação dos cultivares da planta Paullinia cupana. In: **Embrapa Amazônia Ocidental-Artigo em anais de congresso (ALICE)**. In: ENCONTRO REGIONAL DE COMPUTAÇÃO E SISTEMAS DE INFORMAÇÃO, 6., 2017, Manaus. Resumos e artigos completos. Manaus: Fucapi, 2017. p. 121-129., 2017.
SUTTON, C. D. Classification and regression trees, bagging, and boosting. **Handbook of statistics**, *24*, 303-329, 2005.
ZEIDENBERG, M. **Neural networks in artificial intelligence**. Ellis Horwood, 1990.

CAPÍTULO 1

Artigo Cientifico

Discriminação de populações na presença de heterogeneidade de matrizes de covariâncias e vetores aleatórios não normais em estudos de diversidade genética

Discrimination of populations in the presence of heterogeneity of non-normal covariance matrices and random vectors in genetic diversity studies

Resumo: O objetivo desse trabalho foi avaliar, por meio de simulação de dados, a robustez da função discriminante de Fisher na presença de matrizes de covariâncias heterogêneas e vetores aleatórios não normais multivariados. Além disso, os resultados obtidos foram comparados com aqueles provenientes de outras metodologias comumente utilizadas para esse fim, tais como análise discriminante quadrática, Redes Neurais Artificiais, Máquina de Vetor Suporte e Árvore de Classificação. Para tanto, foram simulados cenários caracterizados por matrizes de covariâncias heterogêneas e vetores aleatórios não normais multivariados. Os percentuais de acerto utilizando-se do método da discriminante quadrática apresentaram melhores resultados entre os cenários estudados. Para as situações em que os dados apresentam distribuição Poisson multivariada e homogeneidade de matrizes de covariância a Discriminante de Fisher se destaca com melhores resultados. Outras metodologias apresentam-se como técnicas alternativas apresentando resultados razoáveis quanto ao percentual de acerto.
Palavras-chave adicionais: Função discriminante quadrática; análise multivariada, simulação.

Abstract: The aim of this study was to evaluate, through simulation data, the robustness of the discriminant function of Fisher in the presence of heterogeneous covariance matrices and non - normal multivariate random vectors. In addition, the results were compared with those form other methodologies commonly used for this purpose, such as Quadratic Discriminant Analysis, Artificial Neural Networks, Support Vector Machine and Classification Tree. For this, scenarios characterized by matrices

of heterogeneous covariance and non-normal multivariate random vectors were simulated. The percentages using the quadratic discriminant method presented better results among the scenarios, for the situations in which the data present a multivariate Poisson distribution and homogeneity of covariance matrices. Fisher Discriminant stands out with better results, other methodologies present as alternative techniques presenting reasonable results as to the percentage of correctness.
Additional keywords: Quadratic discriminant function; multivariate analysis, simulation.

1. Introdução

Estudos de diversidade genética têm orientado a escolha de genitores apropriados em programas de melhoramento, levando à obtenção de híbridos com maior heterose e populações segregantes com maior variabilidade. Além disso, tais análises permitem a quantificação da variabilidade existente facilitando o gerenciamento dos bancos de germoplasma (Sant'Anna, 2014).

A literatura apresenta diversas metodologias para quantificação e avaliação da diversidade genética em estudos populacionais (Cruz et al, 2011; Cruz & Nascimento, 2018). Comumente, métodos provenientes da Estatística Multivariada têm sido uma alternativa eficaz nos estudos de diversidade genética, como por exemplo, os métodos baseados em análises de agrupamento (Rodrigues et al., 2017; Santos et al., 2017b) e análise discriminante (Santos et al., 2017a). Entretanto, a obtenção de resultados confiáveis está associada ao atendimento das pressuposições da metodologia a ser aplicada, a qual deve ser escolhida de acordo com as características do conjunto de informações disponíveis e com os objetivos estabelecidos pelo pesquisador.

De maneira geral, o uso de métodos de agrupamento não requer pressuposições quanto a estrutura dos dados e geralmente envolvem procedimentos hierárquicos formados por meio de uma matriz simétrica de distâncias, análise de otimização ou dispersão gráfica. A escolha do método mais apropriado não é tarefa simples e, mesmo dentro de uma classe de análise como, por exemplo, agrupamento hierárquico deve-se ter escolha cuidadosa tendo em vista que diferentes técnicas utilizam diferentes conceitos estatísticos e biológicos. Entretanto,

em alguns casos, a escolha de um, dentre estes vários métodos, é norteada, de maneira simplista, por meio de estatísticas tais como o coeficiente de correlação cofenética (Sokal & Rohlf, 1962).

Em situações de estudo da diversidade genética, não existem métodos únicos para se atingir aos objetivos do pesquisador, o melhorista deve avaliar cada situação de forma a alcançar seus objetivos dentro da melhor relação custo-benefício (Borém e Miranda, 2005). O auxílio de técnicas multivariadas tem sua função também reconhecida em estudos de diversidade genética, como exemplo, técnicas como análise por componentes principais, método de otimização Tocher (Rao, 1952) e as de análise discriminante.

Para a técnica da análise discriminante, o objetivo é classificar elementos da população ou de uma amostra por meio de uma técnica estatística, é utilizado para classificar novos elementos nos grupos já existentes. Dentre estas técnicas, merecem destaque as funções discriminantes lineares de Anderson e de Fisher. Entretanto, a utilização de funções discriminantes lineares (Fisher, 1936) requer que as matrizes de covariância entre as populações sejam homogêneas (Ferreira, 2008) e, em alguns casos que a distribuição do vetor aleatório seja normal multivariado. Caso a hipótese de igualdade seja rejeitada, ou seja, as matrizes de covariâncias sejam homogêneas, a função discriminante quadrática é recomendada (Mingoti, 2007), porém tal função exige que a distribuição do vetor aleatório seja normal multivariado, dito isso, caso a normalidade não seja alcançada, estratégias como transformações de dados são sugeridas. Apesar destas indicações, a literatura apresenta poucos estudos que avaliem a robustez desta técnica quanto a quebra de tais pressupostos e, sem análise aprofundada.

Diante do exposto, este estudo teve por objetivo avaliar, por meio de simulação de dados, a robustez da função discriminante linear quanto a falta de homogeneidade de matrizes de covariâncias e na presença de vetores aleatórios não normais multivariados. Tais avaliações visam nortear pesquisadores quanto a escolha adequada do método a ser utilizado em estudos de diversidade genética. Os resultados obtidos serão comparados com aqueles provenientes de outras metodologias comumente utilizadas para esse fim, tais como análise discriminante quadrática, Redes Neurais Artificiais, Máquina de Vetor Suporte e Árvore de Classificação.

2. Material e Métodos

Neste estudo realizou-se a análise discriminante de indivíduos pertencentes a duas populações simuladas com matrizes de covariâncias particularizadas em, em alguns casos, não normais. Confrontou-se resultados obtidos pelas funções discriminantes lineares com aqueles advindos de outras metodologias, como análise discriminante quadrática, Redes Neurais Artificiais, Máquina de Vetor Suporte e Árvore de Classificação.

3. Conjunto de dados simulados

Para avaliar a robustez da função discriminante quanto à heterogeneidade entre as matrizes de covariâncias e a presença de vetores aleatórios não normais, foram simulados conjuntos de dados com diferentes estruturas de covariâncias e distribuição de probabilidades multivariadas. Avaliou-se o desempenho das metodologias considerando duas populações (A e B), com tamanho da amostra n = 100, o número de variáveis (p) foi estabelecido como p = 5 e as estruturas das matrizes de covariâncias ($\boldsymbol{\Sigma}$) foram definidas como:

$$\boldsymbol{\Sigma}_A = \begin{bmatrix} 1 & \cdots & 0{,}9 \\ \vdots & \ddots & \vdots \\ 0{,}9 & \cdots & 1 \end{bmatrix} \text{ e } \boldsymbol{\Sigma}_B = \begin{bmatrix} 1 & \cdots & 0{,}1 \\ \vdots & \ddots & \vdots \\ 0{,}1 & \cdots & 1 \end{bmatrix}; \quad (1)$$

$$\boldsymbol{\Sigma}_A = \begin{bmatrix} 1 & \cdots & 0{,}9 \\ \vdots & \ddots & \vdots \\ 0{,}9 & \cdots & 1 \end{bmatrix} \quad \text{e} \quad \boldsymbol{\Sigma}_B = \begin{bmatrix} 1 & \cdots & 0{,}5 \\ \vdots & \ddots & \vdots \\ 0{,}5 & \cdots & 1 \end{bmatrix}; \quad (2)$$

$$\boldsymbol{\Sigma}_A = \begin{bmatrix} 1 & \cdots & 0{,}9 \\ \vdots & \ddots & \vdots \\ 0{,}9 & \cdots & 1 \end{bmatrix} \text{ e } \boldsymbol{\Sigma}_B = \begin{bmatrix} 1 & \cdots & 0{,}9 \\ \vdots & \ddots & \vdots \\ 0{,}9 & \cdots & 1 \end{bmatrix}; \quad (3)$$

$$\boldsymbol{\Sigma}_A = \begin{bmatrix} 1 & \cdots & 0{,}1 \\ \vdots & \ddots & \vdots \\ 0{,}1 & \cdots & 1 \end{bmatrix} \text{ e } \boldsymbol{\Sigma}_B = \begin{bmatrix} 1 & \cdots & 0{,}1 \\ \vdots & \ddots & \vdots \\ 0{,}1 & \cdots & 1 \end{bmatrix}. \quad (4)$$

Para a distribuição normal multivariada, os valores paramétricos dos vetores de médias foram considerados iguais à $\boldsymbol{\mu}_A = [0 \quad \cdots \quad 0]^T$ e $\boldsymbol{\mu}_B = [i \quad \cdots \quad i]^T$, em que i = 1, 2 e 3. Por outro lado, para os vetores paramétricos da distribuição Poisson

multivariada considerou-se $\boldsymbol{\lambda}_A = [0 \quad \cdots \quad 0]^T$ e $\boldsymbol{\lambda}_B = [j \quad \cdots \quad j]^T$ em que a média varia para j = 2, 3 e 4. A diferença entre os vetores de médias visa representar diferentes níveis de discriminação considerando 1, 2 e 3 desvios padrão. A combinação entre as diferentes estruturas de covariâncias e distribuição de probabilidade resulta em 24 cenários distintos (Tabela 2).

Tabela 2 - Cenário avaliados para avaliação da robustez da função discriminante linear quanto a falta de homogeneidade de matrizes de covariâncias e na presença de vetores aleatórios não normais multivariados. *Scenario evaluated for the robustness of the linear discriminant function regarding the lack of homogeneity of covariance matrices and the presence of random vectors not multivariate normal.*

Cenário	Distribuição Multivariada	Estrutura de Covariância	Diferenças entre vetores de médias em desvios padrão
1	Normal	(1)	1
2		(1)	2
3		(1)	3
4		(2)	1
5		(2)	2
6		(2)	3
7		(3)	1
8		(3)	2
9		(3)	3
10		(4)	1
11		(4)	2
12		(4)	3
13	Poisson	(1)	1
14		(1)	2
15		(1)	3
16		(2)	1
17		(2)	2
18		(2)	3
19		(3)	1
20		(3)	2
21		(3)	3

22	(4)	1
23	(4)	2
24	(4)	3

Especificamente, cenários simulados considerando as matrizes de covariâncias definidas em (1) e (2) representam situações nas quais as matrizes de covariâncias possuem maior e menor heterogeneidade respectivamente. Já cenários simulados considerando as estruturas (3) e (4) representam situações em que as matrizes de covariância possuem maior e menor homogeneidade respectivamente. A combinação entre essas estruturas de covariância com as diferentes distribuições de probabilidade multivariadas e graus de discriminação de acordo com os vetores paramétricos compõem todo o conjunto de cenários simulados. Para garantir heterogeneidade entre matrizes de covariância a hipótese $H_0: \Sigma_A = \Sigma_B$ foi avaliada por meio da estatística de Box's M (Morison, 1976) derivada dos testes de razão de verossimilhanças (Vuong, 1989). Todo o processo de simulação foi repetido 25 vezes já que acima desse valor não se observou muita diferença nos resultados.

4. Análise Discriminante Linear e Quadrática

Considere o caso em que se tem p > 1 variáveis medidas em cada elemento amostral de cada população e provenientes de distribuições normais p-variadas. Suponha que, para a população A, o vetor X seja normal com vetor de médias $\boldsymbol{\mu}_A$ e matriz de covariâncias $\boldsymbol{\Sigma}_A$, e que para a população B, $\boldsymbol{X}$ seja normal com vetor de médias $\boldsymbol{\mu}_B$ e matriz de covariâncias $\boldsymbol{\Sigma}_B$. Para um vetor de observações fixo $\boldsymbol{x}^T = [x_1 x_2 \dots x_p]$, a razão entre as funções densidade de probabilidade das duas populações, em termos de logaritmo neperiano, será:

$$-2\ln\big(\lambda(x)\big) = -2ln\left\{\frac{(2\pi)^{\frac{p}{2}}\left(|\Sigma_A|^{\frac{1}{2}}\right)^{-1}}{(2\pi)^{\frac{p}{2}}\left(|\Sigma_B|^{\frac{1}{2}}\right)^{-1}}\left[\frac{exp\left\{-\frac{1}{2}(x-\mu_A)'\Sigma_A{}'(x-\mu_A)\right\}}{exp\left\{-\frac{1}{2}(x-\mu_B)'\Sigma_B{}'(x-\mu_B)\right\}}\right]\right\}.$$

Assim, um elemento amostral com vetor de observações $\boldsymbol{x}$ será classificado como pertencente à população 1, se $-2\ln\big(\lambda(\boldsymbol{x})\big)$ for maior que zero e será classificado como sendo da população 2, se menor que zero. Casos $-2\ln\big(\lambda(\boldsymbol{x})\big) = 0$ o elemento amostral poderá ser classificado em qualquer uma das duas populações.

Se $\boldsymbol{\Sigma_A \neq \Sigma_B}$., essa função é denominada como função discriminante quadrática (Mingoti, 2007).

Quando as matrizes Σ_A e Σ_B são homogêneas, a função se torna equivalente a "função discriminante linear de Fisher" (Fisher, 1936) que é expressa como:

$$f(x) = (\mu_A - \mu_B)'\Sigma^{-1}x - \frac{1}{2}(\mu_A - \mu_B)'\Sigma^{-1}(\mu_A + \mu_B)$$

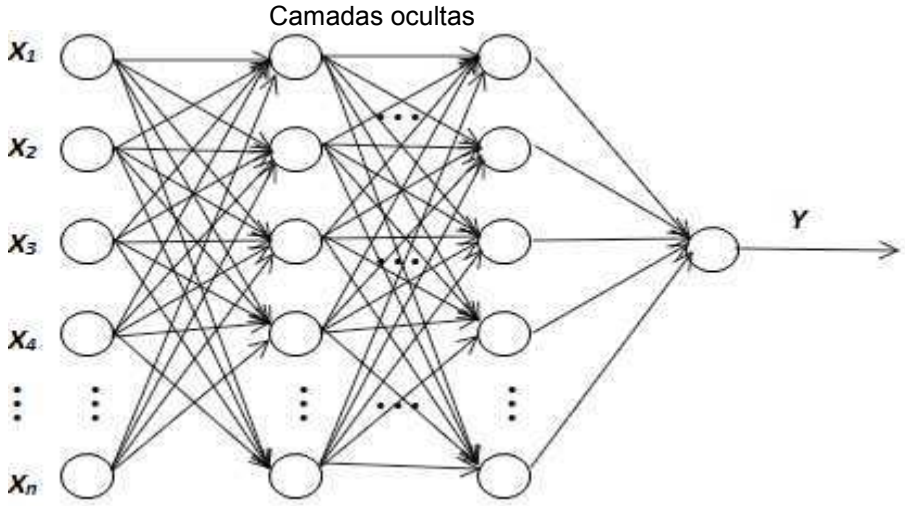
Sendo Σ^{-1}a inversa da matriz de covariâncias das duas populações que é estimada por $S_p = \left[\frac{n_A-1}{(n_A-1)+(n_B-1)}\right]S_1 + \left[\frac{n_B-1}{(n_A-1)+(n_B-1)}\right]S_2$.

Neste caso, um indivíduo será classificado como pertencente à população 1, se $f(x)$ for maior que zero e será classificado como sendo da população 2, se $f(x)$ for menor que zero. Caso $f(x)$= 0, o elemento amostral poderá ser classificado em qualquer uma das duas populações.

5. Redes Neurais Artificiais

Uma Rede Neural Artificial (RNA) é formada pela combinação de diversos neurônios artificiais, os quais são uma estrutura lógica que procuram simular o comportamento e as funções de um neurônio biológico. De maneira geral, uma RNA pode ser decomposta em três camadas, a de entrada, as intermediárias e a de saída (Figura 8).

Figura 8 - Representação das camadas existentes em um modelo de Redes Neurais Artificiais (variáveis $\boldsymbol{X_i}$ em que i=1, 2, 3, 4,⋯, n), camadas ocultas e uma saída ($\boldsymbol{Y}$). *Representation of existing layers in a model of Artificial Neural Networks (variables* $\boldsymbol{X_i}$ *in which i=1, 2, 3, 4, 5), hidden layers and one output* ($\boldsymbol{Y}$)*.*



Fonte: Do autor (2019)

Nesse estudo foi utilizada uma rede *feed-forward* (alimentada a frente) a qual assume que a saída de qualquer camada não é afetada na mesma camada, ou seja, não existe *feedback* (retro alimentação). A camada de entrada é alimentada pelos valores fenotípicos (simulados conforme os cenários 1 a 24, já descritos) que serão utilizados para a análise de diversidade, ou seja, um conjunto de dados composto de n = 100 indivíduos (acessos, genótipos, etc..) mensurados em p =5 caracteres. A camada de entrada é conectada à camada oculta composta de T neurônios (T variando de 1 a 40), os quais são conectados a camada de saída composta de um único neurônio. Essas conexões são direcionadas por meio de pesos estimados que medem a influência das variáveis preditoras na resposta variável. Adicionalmente aos pesos, o viés (b_t), também denotado por intercepto é estimado (Glória et al., 2016).

Matematicamente, na i-ésima camada intermediária o j-ésimo neurônio é formado pelo vetor de pesos, w_{ij}^T, somado ao intercepto. A combinação linear resultante é então transformada por meio de uma função de ativação, f(.), gerando a saída do referido neurônio, $a_i^T = \mathrm{f}\left(\sum_{j=1}^{5} w_{ij}^T x_{ij}\right) + b_t$. A função de ativação pode ser linear ou não linear. Entretanto, de acordo com Bishop (2006), para problemas complexos as funções de ativação não lineares fornecem melhores resultados quando comparados com aqueles provenientes de funções lineares. Na última camada, considerando sem perda de generalidade, uma RNA com apenas uma camada oculta, todas saídas dos neurônios que compõem as camadas intermediárias são entradas em uma nova combinação linear, a qual é novamente transformada por uma função de ativação g(.). Assim, a saída da RNA, y_i, depende de um novo vetor de pesos e um escalar (viés):

$$y_i = \mathrm{g}\left[\sum_{j=1}^{T} w_{2t}\, \mathrm{f}\left(\sum_{j=1}^{5} w_{1j}^T x_{ij} + b_t\right) + b\right].$$

Nesse estudo foi considerada uma RNA com apenas uma camada oculta com funções de ativação Tangente Sigmoidal Hiperbólica ou a Logarítmica Sigmoide. O número de neurônios (T) variou de 1 a 40 e o número de interações foi fixado em 100000.

6. Máquina de Vetor Suporte

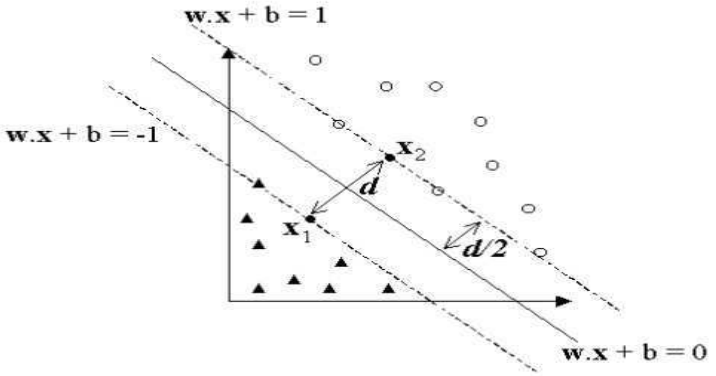
A Máquina de Vetor Suporte (MVS) (Lorena & Carvalho, 2007) baseia-se na teoria do aprendizado estatístico, que visa estabelecer condições matemáticas que permitem escolher um classificador com bom desempenho para o conjunto de dados disponíveis. A ideia principal da MVS é criar um hiperplano de separação como superfície de decisão, de modo que a separação entre seus exemplos positivos e negativos seja máxima (Campbell, 2000; Lorena & Carvalho, 2007).

Matematicamente, um hiperplano pode ser escrito da seguinte maneira:

$$w \cdot x + b = 0,$$

em que w é o vetor de pesos ajustável e b, da mesma forma que na RNA, é o termo de viés. A partir dessa equação divide-se o espaço das observações X em duas regiões: $w \cdot x_i + b > 0$ e $w \cdot x_i + b < 0$ tal que $g(x) = sgn(w \cdot x + b)$, de tal modo que, a classificação será +1 se $f(x) > 0$ e -1 se $f(x) < 0$ (Lorena & Carvalho, 2007) (**Figura 9**).

Figura 9 - Ilustração de um conjunto de dados linearmente separável e a distância d entre os hiperplanos **w** · **x**1 + **b** = -1 e **w** · **x**2 + **b** = +1. *Illustration of a data set linearly separable and the distance d between hyperplanes* ***w*** · ***x1*** + **b** = *-1 and* ***w*** · ***x2*** + **b** = *+1*.



Fonte: Lorena & Carvalho (2007)

Para lidar com situações em que não seja possível dividir satisfatoriamente os dados por um hiperplano linear mapeia-se os conjuntos de treinamento para um novo espaço denominado espaço de características, dessa forma pode-se mapear o conjunto de treinamento para um espaço de mais alta dimensionalidade visando com o aumento do espaço reduzir o problema para o caso de uma classificação linear. É necessário porém a escolha de uma função de mapeamento ϕ apropriada bastando aplicar a função de mapeamento a cada padrão na equação f$(x) = w \cdot$

$\phi(x) + b$, por meio desse procedimento, a informação necessária para o mapeamento da função é definida pelo produto interno $\phi(x_i) \cdot \phi(x_j)$. Isto é obtido com a introdução do conceito Kernels (Lorena et al., 2007), que são funções que recebem dois pontos x_i e x_j do espaço de entradas e computam o produto escalar $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ no espaço de características (Haykin, 1999). Os kernels mais utilizados na prática são os Polinomiais, RBF (*Radial-Basis Funcrion*) e os Sigmoidais.

Outra forma de lidar com a não linearidade dos dados é implementando uma constante de suavização ("C"), que determina a rigidez da margem de separação (Lorena & Carvalho, 2007). Neste estudo, utilizou-se a função Kernel RBF (Shanthini et al, 2017) dado por:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

Ao qual depende somente de um parâmetro, sigma (σ) usualmente definido como o desvio padrão de uma distribuição Gausiana. O espaço de busca do parâmetro sigma foi definido por $0{,}001 \leq \frac{1}{2\sigma^2} \leq 2$ e a constante de suavização ("C"), teve seu espaço de busca de 10^i, com i=0, 1, 2 e 3.

7. Árvore de Decisão e seus refinamentos

Para a construção da árvore de classificação, objetiva-se obter regiões R_1, R_2, ..., R_M que minimizam o índice Gine dado por (James et al., 2013):

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

em que $\hat{p}_{mk}$ representa a proporção de observações na m-ésima região pertencentes a k-ésima classe. O índice Gini diminui de acordo com o crescimento da árvore que ocorre por meio da divisão binária recursiva. Para evitar o super ajustamento do modelo aos dados indica-se a poda da mesma por meio o *custo complexidade* (Hastie et al., 2009). Além disso, sugere-se que nenhuma região contenha mais que 5 indivíduos.

A construção de uma única árvore não é uma estratégia indicada, visto que tal abordagem apresenta modelos com grande variabilidade. Para contornar o problema a literatura sugere o uso do *bootstrap aggregation (Bagging) (*Friedman et al., 2001*)*.

O *Bagging* consiste em obter B amostras com reposição (tamanho igual a N) do conjunto de dados, ajustando assim B modelos [$\hat{f}^1(x)$, $\hat{f}^2(x)$,..., $\hat{f}^B(x)$] que serão utilizados como classificadores individuais. Um novo indivíduo será classificado na classe mais comum dentre as predições dos B classificadores individuais. Outra abordagem que visa aumentar a acurácia na classificação dos indivíduos é o *Random Forest* (RF). Nesse procedimento da mesma forma que no *Bagging*, B amostras são extraídas da população. Entretanto, o número de variáveis preditoras utilizadas em cada partição é inferior ao número total de variáveis disponíveis (m<p). Segundo James et al. (2013) o RF resulta em um processo de descorrelacionar as árvores geradas melhorando ainda mais a acurácia das predições.

Outro refinamento utilizado para melhorar o resultado da árvore de classificação é o *Boosting*. Ao contrário do *Bagging* que cria múltiplas árvores independentes, o *Boosting* cria árvores sequencialmente utilizando informações das árvores anteriores. O classificador do *Boosting* possui a forma $\boldsymbol{H(x)} = \sum_t \alpha_t \boldsymbol{h_t(x)}$ que busca minimizar uma função de perda L através da otimização do escalar α_t (importância atribuída a $\boldsymbol{h_t(x)}$) e do classificador individual $\boldsymbol{h_t(x)}$ (árvore de decisão individual) a cada iteração t (Freund & Schapire, 1999) . Os classificadores individuais $\boldsymbol{h_t(x)}$ possuem um poder classificatório baixo, porém quando utilizados conjuntamente $\boldsymbol{H(x)}$, apresentam bons resultados (Appel, et al., 2013)

.

8. Comparação entre as metodologias

Para acessar a habilidade preditiva dos métodos avaliados, a Taxa de Erro Aparente (TEA) foi calculada de acordo com a seguinte expressão:

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

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 2 é o número de populações consideradas. Esses valores foram obtidos utilizando um procedimento considerando 80% do conjunto de dados para o ajuste/treinamento e o restante 20% para validação. O valor final da TEA é dado pela média obtida nas 25 repetições.

9. Aspectos Computacionais

Todo o processo de simulação das populações e ajuste/treinamento dos modelos foi conduzido por meio software R (R Development Core Team, 2017). Amostras das distribuições Normal e Poisson multivariada foram obtidas por meio, respectivamente das funções mvrnorm e gen.PoisBinOrd dos pacotes MASS (Venables & Ripley, 2002) e PoisBinOrd (Inan & Demitras, 2016). As funções discriminantes linear e quadrática, Redes Neurais, Maquina de Vetor Suporte, Árvore e seus refinamentos e Boosting foram ajustados por meio das funções lda, qda, neuralnet, ksvm, tree, randomFores e gbm dos pacotes MASS, kernlab (Karatzoglou et al., 2004), neuralnet (Fritsch & Guenther, 2016), tree (Ripley, 2016), ..., respectivamente.

10. Resultados e Discussão

A homogeneidade de variâncias foi testada para todos os cenários e repetições. E, como resultados, rejeitou-se a hipótese de homogeneidade ($P \leq 0,01$) nos casos em que as matrizes foram simuladas heterogêneas e não rejeitou nos demais casos ($P > 0,01$). Estes resultados indicam que os dados estão de acordo com os valores dos cenários simulados.

Os valores obtidos da TEA para todas as metodologias avaliadas variaram de 0,00 até 0,40 (Tabela 2). De maneira geral, para os cenários que contemplam a heterogeneidade de matrizes de covariâncias e normalidade multivariada (Cenários 1, 2, 3, 4, 5 e 6) a Função Discriminante Quadrática (FDQ) apresentou menores valores de TEA comparados com aqueles obtidos por todos as outras abordagens avaliadas. Vale ressaltar que tais canários representam situações ideais para a aplicação desta metodologia, visto que a mesma requer que as matrizes de covariâncias sejam heterogêneas e que o vetor aleatório tenha distribuição normal multivariada (Ferreira, 2008). Outra técnica que apresentou destaque nessas situações foi a MVS a qual apresentou um valor médio de TEA igual a 0,06 nestes cenários (**Tabela 3** - Taxa de Erro Aparente (TEA) obtidas para 24 diferentes cenários por meio de diferentes técnicas de classificação. *Apparent error rate (AER) obtained for 24 different scenarios by means of different techniques of classification.*Tabela 3). Diferentemente da FDQ as MVS não possuem pressuposição quanto a distribuição dos dados e matrizes de covariâncias. A MVS

tem como princípio particionar os pontos buscando alocar tantos pontos quanto possível em uma classe pré-definida (Briges et al., 2011). As demais abordagens, que também não possuem pressuposições quanto a distribuição dos vetores aleatórios, apresentaram valores médios variando de 0,09 (*Random Forest*) até 0,15 (Árvores de Decisão e Árvore de Decisão com Poda). Por outro lado, para os cenários que englobam situações nas quais o conjunto da dados foi simulado considerando normalidade multivariada do vetor aleatório e homocedasticidade das matrizes de covariâncias (Cenários 7, 8, 9, 10, 11 e 12) a MVS apresentou melhores resultados. Os valores de TEA variaram de 0,00 até 0,26, enquanto que para as demais metodologias tais valores variaram de 0,00 até 0,40. Os resultados indicam que suposições quanto as matrizes de covariâncias têm maior efeito na classificação dos indivíduos. Esses resultados são esperados visto que as FDQ levam a limiares de decisão não linear (Zhang et al., 2000). Assim, quanto maior a heterogeneidade de matrizes de covariâncias, mais não lineares serão os limiares de classificação e melhor será a performance de metodologias que modelam essa estrutura. Um ponto a se ressaltar é que a RNA utilizada nesse estudo possui apenas uma camada oculta. Assim, visto que as camadas ocultam proporcionam maior complexidade ao modelo e não-linearidade (Nascimento et al., 2013), o uso de um maior número de camadas pode aumentar a eficiência da rede.

Tabela 3 - Taxa de Erro Aparente (TEA) obtidas para 24 diferentes cenários por meio de diferentes técnicas de classificação. *Apparent error rate (AER) obtained for 24 different scenarios by means of different techniques of classification.*

Dist	Cen	Fis	Quad	RNA	SVM	Arv	Poda	Bagg	RanFor	Boost
Normal	1	0,32	0,06	0,13	0,09	0,21	0,21	0,13	0,12	0,25
	2	0,14	0,03	0,08	0,05	0,12	0,11	0,07	0,06	0,10
	3	0,06	0,01	0,04	0,01	0,05	0,06	0,03	0,02	0,03
	4	0,31	0,12	0,19	0,14	0,30	0,29	0,22	0,20	0,29
	5	0,15	0,06	0,11	0,06	0,16	0,15	0,11	0,10	0,13
	6	0,06	0,03	0,06	0,02	0,06	0,06	0,04	0,04	0,05
	7	0,32	0,32	0,29	0,26	0,40	0,35	0,35	0,34	0,30
	8	0,16	0,16	0,16	0,13	0,20	0,18	0,18	0,17	0,14
	9	0,06	0,06	0,08	0,05	0,08	0,08	0,06	0,06	0,06
	10	0,19	0,19	0,17	0,14	0,25	0,25	0,20	0,19	0,19
	11	0,03	0,03	0,04	0,03	0,10	0,11	0,06	0,04	0,04
	12	0,00	0,00	0,01	0,00	0,06	0,06	0,02	0,01	0,01
Poisson	13	0,24	0,10	0,14	0,09	0,25	0,26	0,17	0,15	0,29
	14	0,15	0,04	0,08	0,05	0,15	0,15	0,10	0,09	0,15
	15	0,07	0,02	0,04	0,01	0,11	0,11	0,06	0,05	0,07
	16	0,24	0,17	0,21	0,14	0,31	0,31	0,25	0,24	0,32
	17	0,14	0,09	0,13	0,06	0,20	0,20	0,15	0,13	0,21
	18	0,07	0,05	0,07	0,02	0,14	0,14	0,10	0,09	0,12
	19	0,27	0,37	0,34	0,26	0,37	0,37	0,38	0,37	0,34
	20	0,15	0,23	0,21	0,13	0,26	0,25	0,28	0,27	0,24
	21	0,08	0,15	0,15	0,05	0,17	0,17	0,17	0,16	0,15
	22	0,18	0,26	0,21	0,14	0,31	0,32	0,26	0,25	0,25
	23	0,05	0,10	0,09	0,03	0,16	0,17	0,11	0,10	0,11
	24	0,01	0,03	0,04	0,00	0,09	0,09	0,05	0,04	0,04

Dist = Distribuição Multivariada; Cen = Cenário; Fis = Análise discriminante de Fisher; Quad = Análise discriminante Quadrática; RNA = Redes Neurais Artificiais;

SVM = Máquina de Vetor Suporte; Arv = Árvore de Decisão; Poda = Àrvore de Decisão Podada; Bagg = Bagging; RanFor = Random Forest; Boost = Boosting.

Para os cenários simulados considerando a distribuição Poisson multivariada e heterocedasticidade das matrizes de covariâncias (Cenários 13, 14, 15, 16, 17 e 18) a FDQ apresentou maior TEA média (0,26). Um resultado interessante é a TEA obtida por meio da aplicação da Função Discriminante de Fisher (FDF) o qual teve um desempenho superior quando comparado com aqueles obtidos nos quais os cenários consideraram a normalidade multivariada do vetor aleatório. Esse resultado enfatiza que para a obtenção da FDF não é necessário o pressuposto de normalidade multivariada. A derivação do método considera apenas que as matrizes de covariâncias sejam homogêneas (Ferreira, 2008). Dentre todos os métodos avaliados, aquele que apresentou melhores resultados quanto ao valor da TEA foi a MVS (0,06). Considerando os resultados para o caso de homocedasticiadade de matrizes de covariâncias e vetores aleatórios Poisson multivariados, a FDF apresentou os melhores resultados em termos da TEA média (0,05). Deve ressaltar que tais cenários (Cenários 19, 20, 21, 22, 23 e 24) são situações ideais para a aplicação da técnicas visto que o único pressuposto para a aplicação da mesma é a homogeneidade de matrizes de covariâncias (Ferreira, 2008). Novamente, a MVS apresentou bom desempenho com valor de TEA média igual à 0,10. Já as demais metodologias apresentaram valores de TEA razoáveis com médias variando de 0,11 (RNA) até 0,23 (*Bagging*).

Os resultados apresentados indicam que as técnicas para classificação de indivíduos devem ser utilizadas seguindo suas pressuposições visto que as FDQ e FDF apresentaram melhores desempenhos nos cenários que atendiam os seus pressupostos de utilização. Outras metodologias tais como as RNA, MVS, Árvores e seus desenvolvimentos posteriores apresentaram performance satisfatórias e podem ser alternativas interessantes em estudos nos quais os pressupostos não são atendidos. Entretanto, deve-se ressaltar que a escolha da metodologia depende de diversas características do conjunto de observações em estudo. Neste trabalho, foram avaliadas situações englobando diferentes distribuições de probabilidade multivariadas e presença ou não de homocedasticidade de variâncias. Outras características devem ser levadas em consideração para a escolha da metodologia,

como por exemplo, tipos de variáveis, presença de *outliers*, habilidade para lidar com valores perdidos e habilidade para extrair padrões lineares dos dados. Para todas essas situações, exceto na habilidade para extrair padrões lineares dos dados a literatura indica o uso de Árvores de Decisão e seus refinamentos (Hastie et al. 2008). Já para extrair padrões lineares as RNA e MVS são indicadas (Hastie et al. 2008).

Estudos de diversidade genética apresentam dados com diferentes tipos de caracteres, como por exemplo o estudo de Vargas et al. (2015) em que os autores avaliaram a diversidade genética de acessos de tomates heirloom da coleção do Departamento de Fitotecnia da UFRRJ por meio descritores quantitativos (por exemplo, comprimento e largura do fruto) e qualitativos (por exemplo, forma do fruto e presença de pedicelo com joelho). Já em relação a estrutura dos dados, no que diz respeito a heterocedasticidade de matrizes de covariância, não é comum na literatura apresentar trabalhos de diversidade genética em que os autores se atentem para esse fato, como por exemplo, o estudo de Nogueira et al. (2008) e Santos et al. (2017) nos quais os autores não avaliaram tal hipótese. Desta forma, visto os resultados acima, a avaliação do pressuposto de homogeneidade de matrizes de covariância é importante diante do possível decréscimo do desempenho da técnica aplicada.

11. Conclusões

- As técnicas para classificação de indivíduos devem ser utilizadas seguindo suas pressuposições.
- Para situações em que os dados apresentam normalidade multivariada e heterocedasticidade de matrizes de covariâncias a função discriminante Quadrática apresentou melhores resultados quanto ao valor de Taxa de Erro Aparente (TEA).
- Para situações em que os dados apresentaram distribuição Poisson multivariada e homogeneidade de matrizes de covariância, a Função Discriminante de Fisher apresentou menores valores de TEA.
- As demais metodologias, Redes Neurais Artificiais, Máquina de Vetor Suporte, Árvores de Decisão seus refinamento (Poda, *Bagging* e *Random Forest*) e *Boosting* apresentaram valores razoáveis de TEA e se apresentam

como técnicas alternativas para situações em que os pressupostos necessários para aplicação das técnicas FDQ e FDF não são atendidos.

REFERÊNCIAS

APPEL, R.; FUCHS, T.; DOLLÁR, P.; PERONA, P. Quickly Boosting Decision Trees – Pruning underachieving features early -. In: Journal of Machine Learning Research: Workshop and Conference Proceedings, 30. **Atlanta: international Conference on Machine Learning**, V. 28, p. 594-602, 2013.

BISHOP, C. **Pattern Recognition and Machine Learning (Information Science and Statistics)**, 1st edn. 2006. corr. 2nd printing edn. Springer, New York, 2007.

BORÉM, A., MIRANDA, G. V., & FRITSCHE-NETO, R. **Melhoramento de plantas**. Viçosa: Editora Universidade Federal de Viçosa, 2017.

BRIDGES, M.; HERON, E. A.; O'DUSHLAINE, C.; SEGURADO, R.; MORRIS, D.; CORVIN, A.; ... & International Schizophrenia Consortium. Genetic classification of populations using supervised learning. **PloS one**, 6(5), e14802, 2011.

CAMPBELL, C. An introduction to kernel methods. In **Howlett, R. J. and Jain, L. C., editors**, Radial Basis Function Networks: Design and Applications, pages 155-192. Springer Verlag, 2000.

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

FERREIRA, D. F. **Estatística multivariada**. Editora UFLA, 2008.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of human genetics**, 7(2), 179-188, 1936.

FREUND, Y.; SCHAPIRE, R. E. A Short Introduction to Boosting. **Journal of Japanese Society for Artificial Intelligence**, v.14(5), p.771-780, 1999.

FRIEDMAN, J., HASTIE, T., & TIBSHIRANI, R. **The elements of statistical learning**. New York: Springer series in statistics, 2001.

FRITSCH S.; GUENTHER F. **Neuralnet: Training of Neural Networks.** R package version 1.33, 2016. Disponível em <https://CRAN.R-project.org/package=neuralnet>. Acesso em 26 dez. 2017.

GLÓRIA, L. S.; CRUZ, C. D.; VIEIRA, R. A. M.; DE RESENDE, M. D. V.; LOPES, P. S.; DE SIQUEIRA, O. H. D.; SILVA, F. F. Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. **Livestock Science**, 191, 91-96, 2016.

HASTIE, T.; TIBISHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. **New York: Springer**. 745 p., 2009.

HAYKIN, S. **Neural Networks: A Comprehensive Foundation**. Prentice Hall, 1999.

INAN, G.; DEMIRTAS, H. **PoisBinOrd: Data Generation with Poisson, Binary and Ordinal Components**, R package version 1.2, 2016. Disponível em <https://CRAN.R-project.org/package=PoisBinOrd>. Acesso em 26 dez. 2017.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to Statistical Learning with Applications in R**. New York: Springer. 426 p, 2013.

KARATZOGLOU, A.; SMOLA, A.; HORNIK, K.; ZEILEIS, A. kernlab - An S4 Package for Kernel Methods in R. **Journal of Statistical Software**, 11(9), 1-20, 2004. Disponível em <http://www.jstatsoft.org/v11/i09/>. Acesso em 26 dez. 2017

LORENA, A. C.; DE CARVALHO, A. C. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, 14(2), 43-67, 2007.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2007.
MORRISON, D.F. **Multivariate Statistical Methods**. New York: Editora McGraw-Hill. 415 p, 1976.
NASCIMENTO, M.; PETERNELLI, L. A.; CRUZ, C. D.; NASCIMENTO, A. C. C.; FERREIRA, R. D. P.; BHERING, L. L.; SALGADO, C. C. Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. **Crop Breeding and Applied Biotechnology**, 13(2), 152-156, 2013.
NOGUEIRA, A. P. O.; SEDIYAMA, T.; CRUZ, C. D.; REIS, M. S.; PEREIRA, D. G.; JANGARELLI, M. Novas características para diferenciação de cultivares de soja pela análise discriminante. **Ciência Rural**, 38(9), 2427-2433, 2008.
R Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2017. Disponível em <https://www.R-project.org/>. Acesso em 26 dez. 2017.
RAO, C. R. **Advanced statistical methods in biometric research**, 1952.
RIPLEY B. **Tree: Classification and Regression Trees**. R package version 1.0-37, 2016. Disponível em <https://CRAN.R-project.org/package=tree>. Acesso em 26 dez. 2017.
RODRIGUES, D. L.; VIANA, A. P.; VIEIRA, H. D.; SANTOS, E. A.; de LIMA, F. H.; SANTOS, C. L. Contribuição de variáveis de produção e de semente para a divergência genética em maracujazeiro-azedo sob diferentes disponibilidades de nutrientes. **Pesquisa Agropecuária Brasileira**, 52(8), 607-614, 2017.
SANT'ANNA, I. C.; Tomaz, R. S.; SILVA, G. N.; BHERING, L. L.; NASCIMENTO, M.; CRUZ, C. D. Artificial neural networks in genetic classification. **Genetics and Molecular Research**, 2014.
SANT'ANNA, I. C.; TOMAZ, R. S.; SILVA, G. N.; NASCIMENTO, M.; BHERING, L. L.; CRUZ, C. D. Superiority of artificial neural networks for a genetic classification procedure. **Genet Mol Res**, 14(3), 9898-9906, 2015.
SANTOS, B. W. C.; FERREIRA, F. M.; DE SOUZA, V. F.; CLEMENT, C. R.; ROCHA, R. B. Análise discriminante das características físicas e químicas de frutos de pupunha (Bactris gasipaes Kunth) do alto Rio Madeira, Rondônia, Brasil. **Científica**, 45(2), 154-161, 2017a.
SANTOS, M. D. S.; STANCATTE, R. S.; FERREIRA, T. C.; DORIGHELLO, D. V.; PAZIANOTTO, R. A. A.; de MELO, I. S.; ... & RAMOS, N. P. Resistance to water deficit during the formation of sugarcane seedlings mediated by interaction with Bacillus sp. **Científica**, 45(4), 414-421, 2017b.
SHANTHINI, D.; SHANTHI, M.; BHUVANESWARI, M. C. A Comparative Study of SVM Kernel Functions Based on Polynomial Coefficients and V-Transform Coefficients. **International Journal Of Engineering And Computer Science (IJECS)**, Volume 6 Issue 3, 20765-20769, 2017.
SOKAL, R. R.; ROHLF, F. J. The comparison of dendrograms by objective methods. **Taxon**, 11(2), 33-40, 1962.
TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. 2nd ed. New York, NY: Springer, 745p, 2001.
TIBSHIRANI, R.; JAMES, G.; WITTEN, D.; HASTIE, T. **An introduction to statistical learning-with applications** in R. 1st ed. New York, NY: Springer, 426p, 2013.
VARGAS, T. O.; ALVES, E. P.; ABBOUD, A. C. S.; CARMO, M.; LEAL, M. D. A. Diversidade genética em acessos de tomateiro heirloom. **Embrapa Agrobiologia-Artigo em periódico indexado (ALICE)**, 2015.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S**. Fourth Edition. Springer, New York, 498 p, 2002.
VUONG, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. **Econometrica: Journal of the Econometric Society**, 307-333, 1989.
ZHANG, M. Q. Discriminant analysis and its application in DNA sequence motif recognition. **Briefings in Bioinformatics**, 1(4), 331-342, 2000.