Universidade Federal de Juiz de Fora
Instituto de Ciência Exatas
Bacharelado em Ciência da Computação

Estudo da utilização de métodos de seleção de características aplicados ao problema de seleção de marcadores genômicos

Amanda Ferreira de Castro

JUIZ DE FORA
MARÇO, 2016

Estudo da utilização de métodos de seleção de características aplicados ao problema de seleção de marcadores genômicos

Amanda Ferreira de Castro

Universidade Federal de Juiz de Fora

Instituto de Ciência Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Saulo Moraes Villela

Coorientador: Wagner Antonio Arbex

JUIZ DE FORA

MARÇO, 2016

Estudo da utilização de métodos de seleção de características aplicados ao problema de seleção de marcadores genômicos

Amanda Ferreira de Castro

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIA EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Saulo Moraes Villela
D.Sc. em Engenharia de Sistemas e Computação

Wagner Antonio Arbex
D.Sc. em Engenharia de Sistemas e Computação

Saul de Castro Leite
D.Sc. em Modelagem Computacional

Carlos Cristiano Hasenclever Borges
D.Sc. em Engenharia Civil

JUIZ DE FORA

02 DE MARÇO, 2016

Este trabalho é dedicado ao meu Amor, Eder.

Resumo

Com os avanços nas tecnologias de sequenciamento e genotipagem, a quantidade e a qualidade de dados de marcadores genômicos aumentou consideravelmente. Com isso, estudos que identificam marcadores associados ao desenvolvimento de doenças, distúrbios ou características fenotípicas têm recebido grande atenção nos últimos anos. Além da complexidade das interações entre os marcadores genômicos, há o desafio em analisar dados de alta dimensionalidade. Assim, torna-se interessante o uso de métodos e procedimentos computacionais capazes de selecionar, dentre todos os marcadores genômicos de um conjunto, apenas aqueles que estejam relacionados à determinado fenótipo observado. Neste trabalho, são apresentados os resultados produzidos pela aplicação de métodos de seleção de características a uma base de dados de marcadores genômicos de bovinos relacionados ao fenótipo para produção de leite.

Palavras-chave: Seleção de características, Polimorfismo de um único nucleotídeo, Classificadores de larga margem.

Abstract

With the advances in sequencing and genotyping technologies, the quantity and quality of data from genomic markers increased considerably. With that, studies that identify markers associated with the development of diseases, disorders or phenotypic features have received great attention in recent years. In addition to the complexity of interactions among genomic markers, there is the challenge in analyzing high-dimensional data. Therefore, it becomes interesting the use of computational methods and procedures able to select, from among all the genomic markers of a set, only those that are related to a disease or phenotype that has been observed. In this work, it's presented the results produced by applying features selection methods on a database of bovine genomic markers related to a milk production trait.

Keywords: Feature selection, Single Nucleotide Polymorphisms, Large margin classifiers.

Agradecimentos

Ao meu esposo Eder, pelo encorajamento, dedicação, carinho e apoio.

Aos professores Saulo Moraes Villela e Wagner Antonio Arbex pela oportunidade, orientação e paciência.

Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e a todos que de alguma forma contribuíram para meu enriquecimento pessoal e profissional.

"Decidi há muito tempo não caminhar à sombra de alguém. Se eu fracassar ou obtiver sucesso, terei vivido acreditando em mim".

Whitney Houston

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Contexto e motivação

O DNA é uma molécula que contém a maior parte das informações necessárias para construção e desenvolvimento dos organismos. Nos últimos anos foram realizados muitos avanços nas tecnologias que envolvem sequenciamento e genotipagem de DNA. Esses avanços possibilitaram avanços no estudo de características fenotípicas, doenças e distúrbios em animais, seres humanos e plantas. Em bovinos é possível realizar a seleção de fenótipos de interesse econômico com mais rapidez e precisão utilizando marcadores genômicos que são identificados durante o processo de genotipagem. As tecnologias de leitura de marcadores genômicos são capazes de realizar identificação de milhares de nucleotídeos situados ao longo do genoma dos bovinos, e, após esse processo, são mapeados milhares de marcadores dentre os quais apenas uma parte possui relação com determinada característica fenotípica a ser analisada. Além de serem redundantes ou irrelevantes, dependendo da característica que está sendo observada, os marcadores genômicos também podem não estar diretamente relacionados ao fenótipo, mas "marcar" uma região do genoma responsável pelo mesmo.

Devido ao grande volume de marcadores e à complexidade das interações existentes entre os mesmos, um dos desafios é selecionar, dentre o conjunto de todos os marcadores genômicos, aqueles que possuem efeito em relação à característica que está sendo estudada. Após a escolha desse subconjunto, outro aspecto de estudo refere-se à representatividade desse sobconjunto em relação ao fenótipo observado.

A seleção de um subconjunto de marcadores que seja suficiente para representar um fenótipo pode diminuir os custos na produção de *chips* de genotipagem personalizados os quais possuem um número menor de marcadores relacionados somente à característica desejada.

Existem várias metodologias que são utilizadas para realizar seleção de marcadores genômicos. Alguns métodos são tradicionalmente utilizados, como regressão linear, por

exemplo; outros são uma alternativa aos métodos tradicionais e utilizam, principalmente, aprendizagem de máquinas e mineração de dados.

Vários algoritmos de aprendizagem de máquinas são utilizados para separar, ou classificar, elementos de um conjunto em classes. Esses algoritmos de classificação baseiam-se em características desses elementos, também chamados de padrões, para efetuar a classificação. Assim, uma das maneiras de se melhorar o desempenho de classificadores é utilizar apenas características relevantes ao processo de classificação.

A tarefa de selecionar atributos ou características relevantes em um determinado conjunto de dados tem recebido grande atenção na área de Inteligência Computacional nas últimas décadas. O problema de seleção de características (*feature selection* ou *feature subset selection*) ou seleção de atributos consiste em selecionar um subconjunto de características que consiga reter ou representar a mesma (ou quase a mesma) informação do conjunto original de características.

Como resultado da aplicação do processo de seleção de características tem-se a redução da dimensionalidade do espaço representativo do problema e a remoção dos atributos redundantes e/ou irrelevantes. Dessa forma, pode-se destacar as seguintes vantagens ao se empregar a seleção de características: melhoria na qualidade dos dados, melhor compreensão dos dados e a melhoria no tempo de computação dos algoritmos que utilizarão o conjunto filtrado de dados.

1.2 Objetivos

O principal objetivo deste trabalho é o estudo dos resultados produzidos pela aplicação de métodos de seleção de características em uma base de dados de marcadores genômicos de bovinos que, por sua vez, estão associados ao fenótipo para produção de leite.

Este trabalho analisa também os resultados produzidos pelos métodos de seleção de características após a partição da base de dados em relação a valores altos do fenótipo. As partições criadas buscam utilizar os animais que se destacam positivamente e os animais que se destacam negativamente em relação ao fenótipo analisado, eliminando-se assim os demais animais que tendem a apresentar baixo poder discriminatório durante o processo de classificação.

1.3 Organização

O restante deste texto encontra-se organizado em outros quatro capítulos. Além deste capítulo introdutório, o Capítulo 2 apresenta conceitos e aplicações relacionados aos marcadores genômicos. O Capítulo 3 trata das definições relacionadas ao processo de classificação e também descreve o algoritmo de classificação IMA_p. O capítulo 4 mostra diferentes métodos de seleção de características e apresenta os algoritmos RFE e AOS. O Capítulo 5 é destinado aos experimentos e resultados. Por fim, o Capítulo 6 apresenta uma análise dos resultados produzidos e a conclusão do trabalho.

2 *Single Nucleotide Polymorphisms*

2.1 Definição

Com o sequenciamento do genoma de animais, plantas e, principalmente, o sequenciamento do genoma humano em 2003, tornou-se ainda mais evidente a necessidade do estudo das variações das sequências do DNA. Essas variações genéticas explicam a diversidade de características dos indivíduos, como por exemplo, cor dos olhos, cor da pele e grupo sanguíneo nos seres humanos. Essas diferenças no DNA também podem estar associadas com a predisposição para o surgimento de doenças ou distúrbios.

Muitas dessas variações já são conhecidas, como é o caso do projeto *1000 Genomes* que busca catalogar as variações do genoma humano (CONSORTIUM et al., 2015). Sabe-se também que 90% dessas variações são do tipo SNP (do inglês *Single Nucleotide Polymorphism*) ou polimorfismo de um único nucleotídeo (BROOKES, 1999).

O DNA é composto de nucleotídeos que, por sua vez, são formados pela associação de três moléculas: uma base nitrogenada, um grupamento fosfato e uma desoxirribose (açúcar). As bases nitrogenadas são quatro: Adenina (A), Citosina (C), Guanina(G) e Timina (T). Essas bases ligam-se em pares na estrutura do DNA, a Timina se liga à Adenina e Guanina se liga à Citosina conforme a Figura 2.1. No exterior dessa estrutura, encontra-se a cadeia fosfato-desoxirribose.

SNPs são variações de pares de bases nitrogenadas em uma única posição em uma dada sequência de DNA entre indivíduos, Figura 2.2. Considerando as quatro bases, essas variações para uma dada sequência de DNA, chamados alelos, poderiam ocorrer em até quatro formas. No entanto, a forma bialélica é a mais presente na reprodução das populações e por isso os SNPs, algumas vezes, são simplesmente chamados de "marcadores bialélicos". Além disso, para ser considerado um SNP, o alelo menos frequente deve ocorrer em no mínimo 1% da população (BROOKES, 1999).

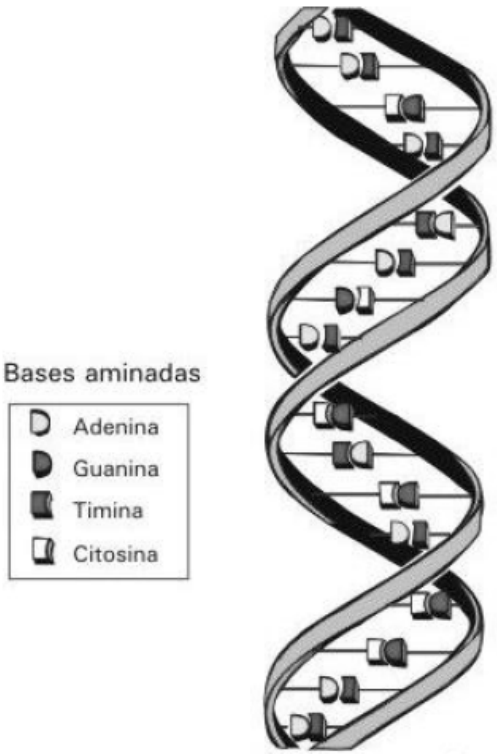

Figura 2.1: Estrutura do DNA (ARIAS, 2004)

2.2 Aplicações dos SNPs

Diversos estudos têm identificado SNPs relacionados ao desenvolvimento de doenças, características fenotípicas, distúrbios e outras interações. Em Kang at al. (2015) é examinada a relação entre SNPs e o desenvolvimento de câncer colorretal. Distúrbios do sono são analisados com relação à genética em Parsons (2015) e em Li et al. (2014) são identificados genes e cromossomos que possuem grandes efeitos sobre a composição de gordura do leite em vacas *Chinese Holstein*. Esses estudos de associação são denominados GWAS, do inglês *Genome-wide Association Studies*, que pode ser traduzido como estudos de associação em genoma amplo.

Em bovinos, torna-se economicamente interessante identificar quais regiões do DNA estão relacionadas à certas características fenotípicas, como por exemplo, alta produção de leite, maciez da carne, maior percentual de gordura no leite e resistência à doenças ou à infecções por ecto ou endo parasitas com o intuito de realizar melhoramento genético dos animais. Essas regiões do genoma animal relacionadas à características quantitativas de interesse são denominadas QTL, do inglês *Quantitative Trait Locus*.

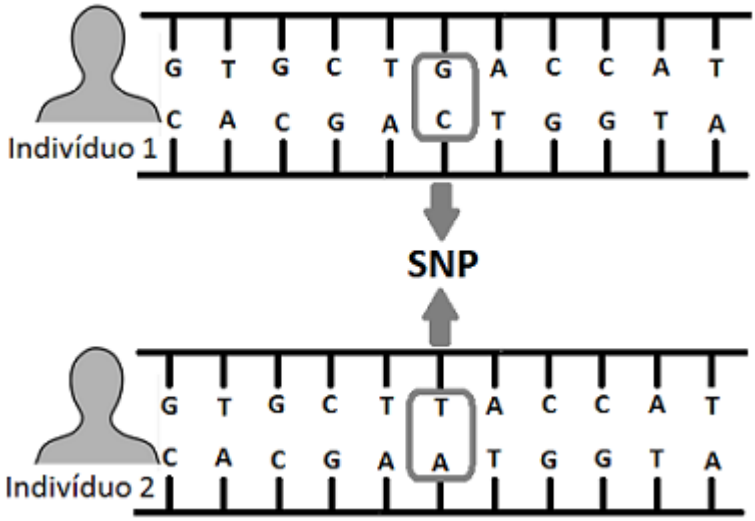

Figura 2.2: Representação de um SNP

2.3 Métodos para realização de GWAS

Os avanços na tecnologia de genotipagem trouxeram mais rapidez ao processo de mapeamento de SNPs. Por exemplo, em 2009, foi apresentado o Illumina BovineSNP50 Beadchip, um *chip* capaz de identificar milhares de marcadores genômicos em bovinos em um único ensaio. Com isso, tem-se uma grande quantidade de dados a serem analisados.

Para realizar os estudos de associação podem ser utilizados diferentes métodos como estatítica multivariada, mineração de dados e apredizagem de máquinas (MOORE; ASSELBERGS; WILLIAMS, 2010). No entanto, além do desafio proporcionado pela alta dimensionalidade dos dados, algumas interações devem ser consideradas:

- o ambiente também pode contribuir para a manifestação de um determinado fenótipo;
- múltiplos SNPs contribuem para a manifestação de um determinado fenótipo;
- um SNP pode não estar associado diretamente a um fenótipo, mas pode estar associado a uma região do genoma responsável pelo mesmo.

Por serem capazes de lidar com a alta dimensionalidade dos dados, o efeito de múltiplos SNPs e não dependerem de nenhuma informação genética subjacente, as técnicas de aprendizagem de máquinas e mineração de dados têm sido crescentemente utilizadas no estudo das relações entre genótipos e fenótipos (ARBEX; MARTINS; MARTINS, 2014).

3 Classificação

O processo de classificação consiste em atribuir uma classe a um determinado elemento de um conjunto. Para realizar esta classificação, os algoritmos classificadores baseam-se em características desses elementos (também chamados padrões) e um conjunto de pares do tipo (padrão, classe) é denominado conjunto de treinamento. Quando determinado elemento é rotulado como pertencente uma determinada classe sendo que ele percence à outra classe, diz-se que ocorreu um erro de classificação. A menor distância entre os padrões do conjunto de treinamento e um hiperplano que os separa é denominada margem de um classificador. Vários estudos são realizados com o intuito de maximizar o valor da margem e obter classificadores mais genéricos, ou seja, que consigam classificar corretamente entradas que não fizeram parte do treinamento. É o caso, por exemplo, do algoritmo SVM (*Support Vector Machine*) que busca melhorar a generalização do classificador encontrando um hiperplano separador com a margem maximizada.

3.1 Classificação binária

Dado um conjunto de treinamento Z de tamanho m, composto de pontos $x_i \in \mathbb{R}^d$ e suas respectivas classes $y_i \in \{+1, -1\}$, um problema de classificação linear consiste em determinar um hiperplano, dado pelo vetor normal $w \in \mathbb{R}^d$ e por uma constante b (bias) $\in \mathbb{R}$, tal que:

$$y_i \left(w \cdot x_i + b\right) \geq 0, \forall (x_i, y_i) \in Z. \tag{3.1}$$

Se esse hiperplano existir, o conjunto de treinamento é dito linearmente separável, caso contrário o conjunto Z é denominado não linearmente separável.

Rosenblatt (1958) propôs a primeira Rede Neural Artificial utilizada para reconhecimento de padrões na qual o algoritmo, em sua forma mais simples, é capaz de determinar o vetor w em um número finito de iterações e classificar corretamente os dados caso eles sejam linearmente separáveis. Nesse procedimento, denominado Perceptron, a atualização dos pesos é baseada na comparação do valor da saída com o valor desejado.

3.2 Algoritmo de Margem Incremental

Leite e Fonseca Neto (2008) apresentaram dois algoritmos que constituem uma nova solução para o problema de maximização da margem. O primeiro algoritmo é uma extensão do algoritmo Perceptron denominada FMP (*Fixed Margin Perceptron*) que encontra um hiperplano classificador respeitando uma margem fixa γ_f para a norma euclidiana conforme expressão 3.2. Dessa forma, os exemplos, além de serem classificados corretamente, precisam estar a uma distância mínima do hiperplano para não serem considerados como erro.

$$y_i \left(w \cdot x_i + b\right) \geq \gamma_f \left\|w\right\|_2, \forall (x_i, y_i) \in Z \tag{3.2}$$

O segundo algoritmo, denominado IMA (*Incremental Margin Algorithm*), utiliza sucessivas soluções do algoritmo FMP com valores crescentes de margem fixa para obter uma solução aproximada para a máxima margem.

Em Villela et al. (2016) o problema de classificação linear e os algoritmos FMP e IMA foram adaptados para uma norma p permitindo o uso de diferentes valores para a mesma. Nesse sentido, considera-se o problema de encontrar o vetor normal w e a constante b tal que:

$$y_i \left(w \cdot x_i + b\right) \geq \gamma_f \left\|w\right\|_q, \forall (x_i, y_i) \in Z \tag{3.3}$$

onde $1/p + 1/q = 1$.

A adaptação do algoritmo FMP deu origem ao Perceptron de Margem Fixa com norma p (*Fixed p-Margin Perceptron* - FMP$_p$) e é exibido no Algoritmo 1.

O algoritmo inicia-se com um valor inicial (w^0, b^0) e a cada iteração t um par $z_i = (x_i, y_i)$ é escolhido e verificado se constitui um erro, ou seja, se $y_i \left(w^t \cdot x_i + b^t\right) < \gamma_f \left\|w^t\right\|_q$. Caso tenha ocorrido um erro, o vetor normal w^{t+1} e a constante b^{t+1} são atualizados conforme a regra de correção a seguir:

$$w^{t+1} = w^t - \eta \left(\gamma_f \left\|w\right\|_q^{1-q} \left|w^t\right|^{q-1} sinal(w^t) + y_i x_i\right) \tag{3.4}$$

$$b^{t+1} = b^t + \eta y_i, \tag{3.5}$$

onde $\eta \in (0, 1]$ é a taxa de aprendizado, $|w| := (|w_i|, ..., |w_d|)'$ e $sinal(w) := (sinal(w_i), ..., sinal(w_d))'$.

Algoritmo 1: Perceptron de Margem Fixa com norma p

Entrada: conjunto $Z = \{(x_i, y_i)\}, i \in \{1, \cdots, m\}$;
margem geométrica fixa γ_f;
limite superior no número de iterações max;
Saída: Vetor de pesos w e bias b;
início
  inicializar (w^0, b^0);
  $j \leftarrow 0$;
  $t \leftarrow 0$;
  $stop \leftarrow$ falso;
  enquanto $j \leq max$ **e** $\neg stop$ **faça**
    $erro \leftarrow$ falso;
    para i **de** 1 **até** m **faça**
      se $y_i \left(w^t \cdot x_i + b^t\right) < \gamma_f \left\|w^t\right\|_q$ **então**
        $w^{t+1} \leftarrow w^t - \eta \left(\gamma_f \left\|w\right\|_q^{1-q} |w^t|^{q-1} sinal(w^t) + y_i x_i\right)$;
        $b^{t+1} \leftarrow b^t + \eta y_i$;
        $t \leftarrow t + 1$;
        $erro \leftarrow$ verdadeiro;
      fim se
    fim para
    se $\neg erro$ **então**
      $stop \leftarrow$ verdadeiro;
    fim se
    $j \leftarrow j + 1$;
  fim enquanto
fim

A adaptação do algoritmo IMA deu origem ao Algoritmo de Margem Incremental com norma p (*Incremental p-Margin Algorithm* - IMA$_p$) e é exibido no Algoritmo 2. A formulação para o problema de maximização da margem é desenvolvida a partir da constatação de que, na obtenção de máxima margem, os pontos ou vetores suporte das classes contrárias se encontram à mesma distância do hiperplano separador.

A margem geométrica fixa γ_f inicia-se com zero e tem seu valor incrementado a cada chamada do algoritomo FMP$_p$. O procedimento é repetido até que a convergência não seja atingida em um número max de iterações. Se os valores de margem positiva e negativa forem diferentes, a solução encontrada não é a ótima, então, atualiza-se o valor da margem fixa conforme expressão a seguir:

$$\gamma_f^{t+1} = \frac{\gamma^+ + \gamma^-}{2}.$$

Caso os valores de margem positivas e negativas sejam iguais, realiza-se um in-

cremento $(1+\Delta), \Delta \in (0,1)$, no valor da margem fixa a fim de verificar se não se trata apenas de um ótimo local.

Algoritmo 2: Algoritmo de Margem Incremental

Entrada: conjunto $Z = \{(x_i, y_i)\}, i \in \{1, \cdots, m\}$;
limite superior no número de iterações max;
Saída: Vetor de pesos w e bias b;
margem de parada γ^w;
início
  $\gamma_f \leftarrow 0$;
  repita
    $(w, b) \leftarrow \text{FMP}_p(Z, \gamma_f, max)$;
    $\gamma^+ \leftarrow Min_{i^+}(w \cdot x_i + b)/\left\|w\right\|_q$;
    $\gamma^- \leftarrow Min_{i^-}(w \cdot x_i + b)/\left\|w\right\|_q$;
    $\gamma^w \leftarrow Min(\gamma^+, \gamma^-)$;
    se $\gamma^+ \neq \gamma^-$ **então**
      $\gamma_f \leftarrow (\gamma^+ + \gamma^-)/2$;
    senão
      $\gamma_f \leftarrow (1+\Delta)\gamma_f$;
    fim se
  até *que a convergência do FMP$_p$ em p max iterações não seja atingida*;
fim

3.3 Formulação L_∞

Neste trabalho será utilizada a formulação L_∞ do IMA$_p$. Essa variante é indicada para o processo de seleção de características onde muitas vezes é necessário soluções em que as componentes do vetor w são esparsas (Villela et al., 2016).

Para obter um hiperplano classificador com margem L_∞ é necessário minimizar a norma L_1 do vetor w. Nesse caso, a margem é calculada de forma a maximizar o valor da maior componente do vetor w. Assim, tem-se o seguinte problema de otimização:

$$\text{Max}_{w,b}\text{Min}_{(x_i,y_i)\in Z}\left\{\frac{y_i\,(w\cdot x_i+b)}{\left\|w\right\|_1}\right\}$$

No entanto considerando uma margem funcional de valor unitário:

$$\text{Min}_{x_i\in X}\left\{y_i\,(w\cdot x_i+b)\right\} = 1,$$

tem-se o seguinte problema equivalente:

$$\text{Min}\left\|w\right\|_1$$

$$\text{Sujeito a}$$

$$y_i\,(w\cdot x_i+b) \geq 1.$$

Adotando $\|w\|_1 = \sum_j |w_j|$ e substituindo, tem-se:

$$\text{Min} \sum_j |w_j|$$

$$\text{Sujeito a}$$

$$y_i \left(w \cdot x_i + b\right) \geq 1.$$

Tomando w como $w = w_+ - w_-$ e $|w_j| = w_j^+ + w_j^-$, Kecman e Hadzic (2000) propõem uma solução para o problema na forma:

$$\text{Min} \sum_j (w_j^+ + w_j^-)$$

$$\text{Sujeito a}$$

$$y_i \left((w_j^+ + w_j^-) \cdot x_i + b\right) \geq 1, w_j^+ \geq 0 \text{ e } w_j^- \geq 0.$$

4 Seleção de características

As características são utilizadas para realizar a classificação de dados de um determinado conjunto conforme mencionado anteriormente. O processo de seleção de características (*feature selection* ou *feature subset selection*), ou selecão de atributos, consiste em selecionar apenas as características necessárias para representar a informação contida no conjunto eliminando-se os atributos irrelevantes e/ou redundantes.

Um ponto de estudo no problema de seleção de características é a maneira (metodologia) de se realizar a seleção de características. Para esta tarefa, diversos métodos de seleção de características (CHANDRASHEKAR; SAHIN, 2014) são encontrados na literatura. No entanto, esses métodos podem ser representados por uma busca heurística na qual alguns aspectos determinam sua natureza (BLUM; LANGLEY, 1997). Esse aspectos são o ponto de partida, a organização da busca, o critério de avaliação e o critério de parada. A seguir são apresentadas três técnicas de seleção de características: métodos de seleção em filtro, métodos de seleção embutidos e métodos de seleção *wrapper*.

4.1 Métodos de seleção em filtro

Métodos de seleção em filtro são aplicados antes do algoritmo de classificação e objetivam filtrar as características menos relevantes do conjunto de acordo com algum critério. Eles são considerados métodos computacionalmente leves e podem produzir bons resultados dependendo do conjunto de dados e do classificador. Como exemplo desse método de seleção, pode-se citar o Golub (GOLUB at al., 1999).

4.2 Métodos de seleção embutidos

Os métodos de seleção embutidos implementam a seleção de características como parte do processo de treinamento do classificador e, em geral, são específicos para um determinado algoritmo de classificação. Uma função objetivo é utilizada para maximizar uma medida

de desempenho do classificador. Um exemplo de método de seleção embutido é o uso do algoritmo IMA$_p$ na sua formulação L_∞. Essa formulação minimiza a norma L_1 do vetor w selecionando as características pelo valor da maior componente do vetor. Nesse sentido, outro exemplo de método embutido é a utilização de uma solução de programação linear na formulação L_∞.

4.3 Métodos de seleção *wrapper*

Nessa abordagem, algoritmos de busca geram subconjuntos de atributos como candidatos. Esses métodos buscam o subconjunto ótimo de atributos tendo um classificador como referência para a construção de uma função objetivo. Métodos de seleção *wrapper* são dependentes do algoritmo de aprendizado e podem apresentar um custo computacional caro pelo fato de ser executado para cada subconjunto de características gerado. Como exemplo deste método de seleção, tem-se o RFE e o AOS.

4.3.1 *Recursive Feature Elimination*

O método de eliminação recursiva de características elimina, recursivamente, a menor componente do vetor w por ela não exercer grande influência sobre a posição do hiperplano. A cada passo do procedimento, um número fixo de características é elinimado e o classificador é executado novamente. O RFE, neste trabalho, executa a retirada de uma características por vez e considera como medida de avaliação os pesos gerados pelo algoritmo IMA$_p$ na sua formulação L_∞.

4.3.2 *Admissible Ordered Search*

Em Villela at al. (2011) é apresentado um algoritmo de seleção de características denominado *Admissible Ordered Search* (AOS). O algoritmo executa uma busca ordenada admissível e possui a capacidade de encontrar, em cada dimensão do problema, o classificador de maior margem. O AOS gera hipóteses contendo cada qual um conjunto único de características selecionadas. As hipóteses são inseridas em uma fila ordenada por valores de margem obtidos da solução do problema de classificação. Para não exigir, a cada

hipótese gerada, a solução de um problema de maximização da margem para encontrar o valor da margem geométrica real, utiliza-se um valor de margem projetada. A margem projetada é uma estimativa otimista da margem geométrica real e corresponde a um limite máximo para a mesma.

A cada iteração do algoritmo, a hipótese que possui a maior margem é escolhida para ser expandida e gerar novas hipóteses com dimensão menor. Dessa forma, duas situações podem ocorrer:

- se o valor da margem para a hipótese escolhida corresponde ao valor projetado, calcula-se o valor da margem geométrica real usando SVM e compara-o com o maior valor da fila de prioridade. Caso o valor da margem projetada ainda seja maior, fecha-se esse estado e geram-se as suas hipóteses. Caso contrário, substitui-se o valor da margem projetada dessa hipótese pelo valor real e reinsere o mesmo na fila;
- se o valor da margem da hipótese escolhida já corresponde a valor real, fecha-se esse estado e geram-se as suas hipóteses.

Para controlar a quantidade de combinações possíveis, o algoritmo AOS realiza dois esquemas de podas adaptativas. O primeiro esquema é baseada em um limite de margem inferior que é atualizado toda vez que uma hipótese escolhida é a primeira a alcançar uma nova dimensão. Para tanto, é utilizada uma estratégia de seleção míope (RFE) que avalia os valores de margem até uma dimensão inferior escolhida eliminando os estados de dimensão maior que tenham valores de margem inferiores a esse limite. A segunda poda, que também ocorre toda vez que uma hipótese escolhida é a primeira a alcançar uma nova dimensão, é baseada em um valor de corte que elimina todas as hipóteses da fila que tenham dimensão superior à hipótese em questão mais o valor do corte. Isso elimina os candidatos que tendem a ter valores de margem relativamente baixos.

O Algoritmo 3 descreve o pseudocódigo do algoritmo AOS.

Algoritmo 3: Admissible Ordered Search

Entrada: conjunto de treinamento Z;
conjunto de caracteríticas $F = 1, 2, 3, \cdots d$;
fator de ramificação b;
profundidade da poda p;
profundidade do corte c;
nível de parada s;
Dados: nível alcançado *nivel*;
limite inferior usado para podar a árvore de busca *limite*;
Saída: último estado aberto;
início
    inicializar heap H e a tabela hash HT;
    computar a solução SVM para o estado inicial $S_{inicial}$ com o conjunto F;
    $nivel \leftarrow d$;
    inserir $S_{inicial}$ em H;
    enquanto *nivel* $> s$ **e** H *não vazio* **faça**
        selecionar a melhor hipótese S de H;
        se S *possuir margem projetada* **então**
            computar a margem real de S usando SVM;
            se *solução SVM convergiu* **então**
                reiserir S em H;
            fim se
        fim se
        senão
            se *dimensão do estado S for igual a nivel* **então**
                usar RFE até $nivel - p$ e encontrar o novo valor para *limite*;
                eliminar de H todos os estados com margem menor que *limite*;
                eliminar de H todos os estados com nível maior que $nivel + c$;
                $nivel \leftarrow nivel - 1$;
            fim se
            ordenar F em S pelo critério de ramificação;
            para i **de** *1* **até** b **faça**
                criar um novo estado S' com o conjunto $F' = F - \{f_i\}$;
                computar γ_{pj} para S; `// onde` γ_{pj} `é a margem projetada`
                se $\gamma_{pj} > limite$ **e** F' *não esta em* HT **então**
                    inserir F' em HT;
                    inserir S' em H;
                fim se
            fim para
        fim se
    fim enquanto
    retorna *último estado aberto;*
fim

5 Experimentos e resultados

5.1 Base de dados QTL-MAS 2012

Para avaliar a capacidade do algoritmos em selecionar marcadores que delimitam regiões relacionadas a características de interesse econômico foi utilizada a base de dados simulados do *workshop* QTL-MAS do ano de 2012 (QTL-MAS, 2012). Nesta base, 1.020 indivíduos (20 machos e 1.000 fêmeas) não relacionados foram gerados em 5 cromossomos constituindo a geração base (G0). Cada cromossomo possui 2.000 SNPs igualmente distribuídos. Cada uma das quatros gerações seguintes (G1-G4) é composta de 20 machos e 1.000 fêmeas que foram geradas por meio do acasalamento aleatório de cada macho com 51 fêmeas. As gerações não se sobrepõem. Três características quantitativas de produção de leite foram simuladas nas fêmeas da geração G1 a G3. Neste evento, o desafio consistiu em encontrar os QTLs e suas possíveis ações pleiotrópicas [1]. Também foi disponibilizado o valor genético verdadeiro (*true breeding value* - TBV) da base de dados simulada.

Neste trabalho é utilizado apenas o primeiro fenótipo. Os valores do fenótipo são contínuos e podem ser positivos ou negativos conforme Tabela 5.1.

Tabela 5.1: Resumo base de dados QTL-MAS 2012

	Amostras			Fenótipo			
Atributos	Pos	Neg	Total	Mínimo	Mediana	Média	Máximo
20.000	1.491	1.509	3.000	-584,99365	-1,71149	-0,000003	587,18972

Cada SNP pode assumir os valores 1 1, 1 2, 2 1 ou 2 2 que é uma codificação que se refere às possibilidades de variação bialélicas. Dessa forma, os 10.000 SNPs (2.000 SNPs por cromossomo) deram origem a 20.000 características que foram utilizadas pelos algoritmos de classificação e seleção descritos neste trabalho.

[1] pleiotropia é o fenômeno em que um único gene afeta múltiplas características (PAVLICEV; WAGNER, 2012)

5.2 Experimentos

Para utilizar os dados da base QTL-MAS 2012 em um problema de classificação binária, as amostras foram divididas em duas classes de acordo com os valores positivos e negativos do fenótipo. Assim, obteve-se 1.491 exemplos com fenótipos de valores positivos e 1.509 exemplos com fenótipos de valores negativos. Essa divisão que contempla todos os animais foi chamada de Partição 100%. Após a divisão das amostras em duas classes, optou-se por manter 50% dos animais que possuem os maiores valores absolutos de cada classe eliminado-se os demais. Essa divisão obteve 745 exemplos com fenótipos de valores positivos e 754 exemplos com fenótipos de valores negativos e foi chamada de Partição 50%. Por último, foi criada uma terceira partição, no mesmo sentido da segunda, que mantém os 25% animais com os maiores valores absolutos de cada classe. Essa partição foi denominada Partição 25% e obteve-se 372 animais com fenótipos de valores positivos e 377 animais com fenótipos de valores negativos. A Tabela 5.2 resume as partições realizadas.

Tabela 5.2: Partições realizadas na base QTL-MAS 2012

	Amostras								
	Partição 100%			Partição 50%			Partição 25%		
Atributos	Pos	Neg	Total	Pos	Neg	Total	Pos	Neg	Total
20.000	1.491	1.509	3.000	745	754	1.499	372	377	749

Após gerar as partições da base de dados, foi realizada uma etapa de pré-processamento. Essa etapa consistiu em obter a solução exata usando a formulação L_∞ de programação linear em cada uma das partições. Os problemas de programação linear foram resolvidos utilizando uma versão acadêmica do software IBM CPLEX (IBM CPLEX, 2016). A solução de programação linear proporcionou uma redução na dimensão em cada uma das partições porque apresentou muitas componentes do vetor w com valor zero. As componentes do vetor w com valor zero foram eliminadas por não exercerem influência na posição hiperplano durante o processo de classificação.

Para a Partição 100% a dimensão foi reduzida de 20.000 para 1.952. Na Partição 50% o número de atributos foi reduzido de 20.000 para 984 e na Partição 25% a dimensão foi reduzida de 20.000 para 525 após a solução de programação linear.

Mesmo com redução dimensional proporcionada pela solução de programação

linear, a seleção de características, por meio do desenvolvimento de uma árvore de possibilidades como no AOS, ainda é totalmente inviável. Desse modo, foi utilizado o algoritmo RFE em cada uma das partições a partir da respectiva dimensão atingida pela solução de PL. Foi realizada a remoção de uma característica por vez e as medidas de avaliação utilizadas foram são os valores das componentes do vetor w encontrado pelo algoritmo IMA_∞. A característica relacionada à menor componente do vetor w é eliminada a cada passo do algoritmo RFE. Após a aplicação do RFE em cada uma das partições novas dimensões foram alcançadas. A Partição 100% atingiu a dimensão 410, a Partição 50% alcançou a dimensão 140 e a Partição 25% a dimensão 62.

Para efeito de comparação entre os algoritmos AOS e RFE, optou-se por executar o algoritmo AOS, a partir da dimensão alcançada pela solução de PL, para encontrar subconjuntos com a mesma cardinalidade dos encontrados pelo RFE. Para tanto, devido à possível explosão combinatória, foram feitas pausas no AOS (a cada 20 dimensões) eliminando todas as hipóteses da fila. Foi utilizado o fator de ramificação 3, profundidade da poda 2 e profundidade do corte 2. Para o cálculo da margem real aproximada foi utilizado o algoritmo IMA_∞.

Por último, aplicou-se o AOS a partir da dimensão alcançada pelo RFE, em cada partição, com o intuito de avaliar quais seriam as novas dimensões atingidas pelo algoritmo AOS.

A Tabela 5.3 exibe as dimensões alcançadas após a solução de PL e também após a execução dos algoritmos RFE e AOS.

Tabela 5.3: Quantidade de atributos após aplicação dos algoritmos

	Atributos		
Algoritmo	Partição 100%	Partição 50%	Partição 25%
PL	1952	984	525
RFE	410	140	62
AOS	400	135	58

5.3 Resultados

A Tabela 5.4 exibe os 10 melhores SNPs ranqueados pelos métodos utilizados neste trabalho. A leitura ordenada dos marcadores apresentados deve ser feita da esquerda para

direita e de cima para baixo em cada um dos grupos de 10 SNPs. Os SNPs destacados em negrito correspondem a SNPs encontrados na listagem do valor genético verdadeiro (TBV) disponibilizada pelo *workshop* QTL-MAS 2012.

Nota-se que nas três partições, os métodos de seleção de características encontraram o SNP mais representativo em relação fenótipo para produção de leite, isto é, o SNP 6499 que delimita a região 6499-6500 da listagem TBV. Nota-se também que, quando comparada uma mesma dimensão, os algoritmos apresentaram resultados melhores na Partição 25%. O SNP 6499 aparece em todos subconjuntos da Partição 25% e em todos eles na primeira posição.

Pode-se observar que a similaridade entre os SNPs foi maior na segunda e terceira partição. Dos 70 SNPs apresentados em cada partição, 26 são distintos (não se repetem) na Partição 100%, 16 são distintos na Partição 50% e 17 são distintos na Partição 25% conforme apresentado na Tabela 5.5.

A Tabela 5.6 mostra as dimensões alcançadas pelo algoritmo AOS e os 10 melhores SNPs de cada partição.

A Tabela 5.7 mostra os valores de margem produzidos pelo RFE e o AOS. Os valores de margem do AOS foram maiores.

Tabela 5.4: TOP 10 SNPs

	Partição 100%			Partição 50%			Partição 25%		
Dim	PL	RFE	AOS	PL	RFE	AOS	PL	RFE	AOS
1952	4358,**6499**, 2685,6189, 5290,2460, 8690,372, 5991,3816	-	-	-	-	-	-	-	-
984	-	**6499**,6498, 3519,8098, 4696,6249, 3162,5631, 2460,2694	6189,4955, 1161,2460, 6249,8690, **6499**,807, 5813,8098	**6499**,7441, 6498,9776, 7284,8721, 9277,9610, 6584,**4738**	-	-	-	-	-
525	-	2685,6615, **6499**,6079, 1161,1338, 3234,8690, 9086,3496	8690,2685, 3234,3544, 1161,3628, 5396,6780, 1208,9504	-	**6499**,7441, 6498,8721, 9277,9776, 7677,707, 4574,2071	**6499**,7441, 9776,6498, 7284,8721, 4308,1661, 661,2071	**6499**,4675, 1661,**9314**, 7543,4193 **293**,3976, 1615,3094	-	-
410	-	2685,2279, 6615,**6499**, 6079,494, 9585,4358, 1338,3496	2685,6531, 9917,7587, 6780,5396, 494,714, 1188,3191	-	**6499**,7441, 7284,8721, 9776,6498, 7677,1085, 6738,1661	**6499**,7441, 6498,4308, 8721,7525, 9776,6293, 7284,707	-	**6499**,4675, 1661,4193, **9314**,7543, **293**,1615, 3094,3976	**6499**,4675, **9314**,1661, 4193,**293**, 7543,3094, 1615,8422
140	-	-	-	-	7441,**6499**, 7284,6584, 9776,6254, 373,4445, 1943,4257	7284,**6499**, 7441,4445, 3595,296, 7701,884, 7498,9791	-	**6499**,1661, 4193,9497, 1683,1116, 2221,4675, 3764,5612	**6499**,1683, 4193,439, **9314**,1661, 5326,3764, 3720,9818
62	-	-	-	-	-	-	-	**6499**,4468, 3764,7391, 7432,6055, 4193,252, 3721,3094	**6499**,1683, 3611,9528, 1661,843, 5326,3720, 4789,6903

Tabela 5.5: Quantidade de SNPs distintos por partição

	Partição 100%	Partição 50%	Partição 25%
Quantidade	26	16	17

Tabela 5.6: TOP 10 SNPs AOS

Dim	Partição 100%	Partição 50%	Partição 25%
400	2685,6531, 9917,3191, 5396,7284, 1188,7587, 6780,714	-	
135		7284,**6499**, 7441,884, 9791,9776, 296,4445, 3048,3595	
58	-	-	**6499**, 1683 3611,1661, 9497,7456, 5326,896, 4193,4789

Tabela 5.7: Comparativo valores de margens do RFE e AOS

	Partição 100%		Partição 50%		Partição 25%	
Dim	RFE	AOS	RFE	AOS	RFE	AOS
410	0,000489	0,000567	0,016941	0,017667	0,037353	0,037347
140			0,000338	0,001207	0,027427	0,027850
62					0,001278	0,003908

6 Conclusão

O problema de seleção de marcadores genômicos não é trivial, pois, além da alta dimensionalidade dos dados, são complexas as interações existentes entres os marcadores. Neste trabalho, métodos de seleção de características foram aplicados a esse tipo de problema. Os algoritmos AOS e RFE, associados ao classificador IMA_∞, e a solução de programação linear foram capazes de encontrar SNPs que são bastante expressivos com relação ao fenótipo para produção de leite observado sem nenhuma informação genética subjacente. No entanto, isso não dispensa uma validação sob a ótica da área de conhecimento do problema, e assim, a interpretação dos resultados no domínio da genômica. Os resultados mostraram também que a retirada de exemplos do conjunto de treinamento de baixo poder discriminatório facilitou o problema de seleção e apresentou melhores resultados no que se refere a seleção de marcadores genômicos que contribuem para fenótipo analisado.

Neste trabalho, foi realizada a divisão de classes baseada nos valores positivos e negativos do fenótipo observado. No entanto, em trabalhos futuros, outros tipos de divisão em classes podem ser analisados como a divisão baseada em porcentagens ou em um valor específico do fenótipo. O tamanho das partições utilizadas neste trabalho com o intuito de manter os maiores e os menores valores do fenótipo é uma sugestão; diversos tamanhos podem ser explorados. Para justificar a relevância das partições, experimentos que realizam o processo contrário, ou seja, removem as amostras com valores extremos do fenótipo, também podem ser realizados. Além disso, outros métodos também podem ser avaliados como máquina de vetores de suporte com regressão por exemplo.

Referências Bibliográficas

Arbex, W.; Martins, N. F. ; Martins, M. F. **Talking About Computing and Genomics - TACG**, volume I. 1. ed., Brasília, DF: Embrapa Gado de Leite, 2014.

Arias, G. Em 1953 foi descoberta a estrutura do DNA: etapas de um grande avanço científico. **Embrapa Trigo**, 2004.

Blum, A. L.; Langley, P. Selection of relevant features and examples in machine learning. **Artificial intelligence**, v.97, n.1, p. 245–271, 1997.

Brookes, A. J. The essence of SNPs. **Gene**, v.234, n.2, p. 177–186, 1999.

Chandrashekar, G.; Sahin, F. A survey on feature selection methods. **Computers & Electrical Engineering**, v.40, n.1, p. 16–28, 2014.

Consortium, T. . G. P. A global reference for human genetic variation. **Nature**, v.526, n.7571, p. 68–74, 2015.

Golub, T. R.; Slonim, D. K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J. P.; Coller, H.; Loh, M. L.; Downing, J. R.; Caligiuri, M. A.; Bloomfield, C. D. ; Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. **Science**, v.286, n.5439, p. 531–537, 1999.

CPLEX, I. **IBM academic initiative**. https://developer.ibm.com/academic/, 2016. Site, acessado em 04 mar. 2016.

Kang, B. W.; Jeon, H.-S.; Chae, Y. S.; Lee, S. J.; Park, J. Y.; Choi, J. E.; Park, J. S.; Choi, G. S. ; Kim, J. G. Association between GWAS-identified genetic variations and disease prognosis for patients with colorectal cancer. **PLoS One**, v.10, n.3, p. e0119649, 2015.

Kecman, V.; Hadzic, I. **Support vectors selection by linear programming**. In: Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on, volume 5, p. 193–198. IEEE, 2000.

Leite, S. C.; Neto, R. F. Incremental margin algorithm for large margin classifiers. **Neurocomputing**, v.71, p. 1550–1560, 2008.

Li, C.; Sun, D.; Zhang, S.; Wang, S.; Wu, X.; Zhang, Q.; Liu, L.; Li, Y. ; Qiao, L. Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in chinese holstein. **Plos One**, v.9, n.5, 2014.

Moore, J. H.; Asselbergs, F. W. ; Williams, S. M. Bioinformatics challenges for genome-wide association studies. **Bioinformatics**, v.26, n.4, p. 445–455, 2010.

Pavlicev, M.; Wagner, G. P. A model of developmental evolution: selection, pleiotropy and compensation. **Trends in Ecology & Evolution**, v.27, n.6, p. 316 – 322, 2012.

Parsons, M. J. On the genetics of sleep disorders: genome-wide association studies and beyond. **Advances in Genomics and Genetics**, v.2015, p. 293–303, 2015.

QTL-MAS. **16th QTL-MAS Workshop**. http://qtl-mas-2012.kassiopeagroup.com/en/index.php, 2012. Site, acessado em 21 jan. 2016.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, v.65, n.6, p. 386, 1958.

Villela, S. M.; Fonseca Neto, R.; Leite, S. C. ; Xavier, A. E. **Seleção de características utilizando busca ordenada e um classificador de larga margem**. In: X CBIC - Congresso Brasileiro de Inteligência Computacional, Fortaleza, CE, 2011.

Villela, S. M.; de Castro Leite, S. ; Neto, R. F. Incremental p-margin algorithm for classification with arbitrary norm. **Pattern Recognition**, 2016.