SELEÇÃO CLONAL DE CARACTERÍSTICAS RANKEADAS POR FILTROS UNIVARIADOS PARA CLASSIFICAÇÃO DE TIPOS DE LEUCEMIA AGUDA

Honovan P. Rocha, Antônio P. Braga

Programa de Pós-Graduação em Engenharia Elétrica - Universidade Federal de Minas Gerais - Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brasil
E-mails: honovan@gmail.com, apbraga@ufmg.br

Resumo – Com o crescimento dos estudos quanto ao genoma humano a utilização de expressões gênicas para identificação de tipos de câncer tem se tornado uma boa alternativa para auxiliar o direcionamento do tratamento. A dificuldade na utilização destes dados se encontra na grande quantidade de variáveis a serem analisadas. Neste trabalho apresenta-se uma abordagem utilizando um algoritmo inspirado no sistema imunológico biológico baseado no princípio da seleção clonal para busca do menor conjunto de variáveis com eficácia na classificação entre duas variações de câncer relacionadas à leucemia. Efetuou-se uma pré-seleção das variáveis utilizando-se filtros uni-variados antes da submissão ao algoritmo, o que caracterizou a abordagem utilizada como um método híbrido para seleção de características. Com a utilização de alguns classificadores geralmente utilizados na literatura encontrou-se boas taxas de classificação correta em relação às variáveis selecionadas pelo algoritmo.

Palavras-chave – Seleção de características, seleção clonal, classificação, filtros uni-variados, expressão gênica.

Abstract – With the growth of studies on the human genome, the use of gene expressions to identify types of cancer has become a good alternative to assist the planning of the treatment. The difficulty in using these data is the large number of variables to be analyzed. This paper presents an approach using an algorithm inspired by biological immune system based on clonal selection principle to search the smaller set of variables efficiently in the classification between two variations of cancer-related leukemia. We conducted a pre-selection of variables using uni-varying filters prior to submission to the algorithm, which characterized the approach as a hybrid method for feature selection. Which the use of some commonly used classifiers in the literature we found good rates of correct classification in relation to the variables selected by the algorithm.

Keywords – Feature selection, clonal selection, classification, uni-varying filters, gene expression.

1 Introdução

O tratamento contra o câncer teve melhorias significativas nos últimos anos, conseguindo cada vez mais eficácia no tratamento e menores danos à saúde dos pacientes. A classificação entre tipos de câncer tem ganhado grande importância devido à especificidade dos tratamentos para cada uma das variações da doença.

Os avanços recentes nas ciências biológicas em relação ao genoma humano têm trazido muita informação útil a ser explorada em relação ao tratamento dos mais diversos tipos de câncer, com isso, a utilização de informações provenientes de expressões gênicas tem se mostrado uma boa alternativa para análise de variações cancerígenas, ao invés de se analisar apenas a aparência morfológica do tumor, como acontece na abordagem clássica a este problema.

As informações contidas em expressões gênicas possuem milhares de variáveis, o que torna difícil sua utilização em tarefas como a classificação de padrões. Nem toda essa enorme quantidade de informações é necessária numa tarefa específica como, por exemplo, a classificação de tipos de câncer e, além disso, existe grande quantidade de redundância nestes dados. Com isso observa-se a necessidade de se efetuar uma seleção das características mais relevantes ao problema que se quer resolver, conduzindo ao uso de métodos de seleção de características.

Neste contexto, propõe-se neste trabalho a utilização de uma abordagem que utiliza o algoritmo de seleção clonal para selecionar características relevantes ao problema de classificação de tipos de leucemia aguda. O método tem o objetivo de buscar o menor conjunto de características que maximize a taxa de classificações corretas. O método utilizado é denominado método híbrido de seleção de características por utilizar filtros uni-variados para pré-seleção e um método multivariado, a seleção clonal, para seleção final das características. Após a realização da seleção utilizou-se dois classificadores muito explorados na literatura para se avaliar os resultados da abordagem, verificando-se bons resultados com a utilização do método proposto.

O trabalho está organizado da seguinte forma: a seção 2 aborda o problema da classificação de tipos de Leucemia Aguda, a seção 3 apresenta conceitos relativos à seleção de características, a seção 4 mostra conceitos de classificação de padrões, na seção 5 são vistos alguns conceitos de sistemas imunológicos artificiais, a seção 6 apresenta o método proposto, na seção 7 é descrita a metodologia utilizada, a seção 8 expõe-se os resultados obtidos e, por fim, a seção 9 expressa as conclusões do trabalho.

2 O Problema de Classificação de Tipos de Leucemia Aguda

As primeiras bases para classificação de leucemia aguda datam da década de 60, onde através de análises histoquímicas baseadas em enzimas obteve-se a categorização de leucemias decorrentes de precursores linfoides (em inglês, *acute lymphoblastic leukemia - ALL*) e precursores mieloides (em inglês, *acute myeloid leukemia - AML*). Na década de 70 esta categorização tornou-se ainda mais estável com o desenvolvimento de anticorpos que reconhecem a superfície celular de moléculas linfoide ou mieloide [1].

Mesmo estando bem definida a distinção entre *ALL* e *AML* devem ser feitos inúmeros testes em laboratórios altamente especializados para se obter uma classificação, muitas vezes ainda imperfeita, com ocorrência de erros. A correta categorização entre estes tipos de leucemia é crítica devido às diferentes substâncias utilizadas no tratamento de cada tipo. Embora o tratamento de um tipo utilizando terapia para o outro tipo tenha efeitos positivos na diminuição da doença, tem-se uma diminuição das taxas de cura.

A utilização de expressões gênicas tem-se mostrado uma alternativa para a realização de analise e categorização de variações de câncer. O monitoramento deste tipo de informação, que possui milhares de variáveis, tem sido facilitado através da utilização de tecnologias de arranjo, que extrai informação durante a diferenciação e resposta celular. O conjunto de dados de expressões gênicas obtidos por estas tecnologias são representados por grandes sequências numéricas, apresentando a necessidade de utilização de ferramentas para análise e consequente obtenção de informação útil [2].

2.1 Obtenção dos Dados

Os dados utilizados neste trabalho são oriundos do trabalho apresentado em [1]. A base de dados utilizada consiste de expressões gênicas de 7129 sondas referentes a 6817 genes humanos e 72 amostras de dados referentes à pacientes com *ALL* e *AML*. Estas amostras foram divididas em dois conjuntos, onde 38 destas (27 *ALL*, 11 *AML*), provenientes de medula óssea, foram definidas como conjunto de treinamento. As outras 34 amostras (20 *ALL*, 14 *AML*) foram definidas como conjunto independente, sendo que 24 destas foram obtidas da medula óssea e as outras 10 de sangue periférico.

3 Seleção de Características

Em problemas com grande dimensionalidade, como é o caso dos problemas com expressões gênicas, existem muitos atributos irrelevantes e um número reduzido de amostras, o que ocasiona em aumento de complexidade computacional e perda de exatidão na tarefa de classificação. Nestes casos torna-se necessário a remoção de características irrelevantes e a definição de um subconjunto reduzido de características discriminativas para melhorias na classificação [3].

Uma desvantagem da seleção de características é o aumento de uma camada de complexidade no processo devido ao custo de se obter um subconjunto adequado à resolução do problema num espaço de busca relativamente grande.

No contexto de classificação as técnicas de seleção de características se diferem quanto à forma utilizada para incorporar a busca no espaço adicional dos subconjuntos de características à escolha do modelo, dividindo-se em três categorias: métodos de filtro, métodos *wrapper* e métodos embarcados [4]. Os métodos de filtro e *wrapper* se diferem na forma de avaliação dos subconjuntos de características. Os filtros utilizam critérios baseados em informações intrínsecas aos dados sem utilização de nenhuma técnica de aprendizagem de máquina enquanto que *wrappers* utilizam o desempenho de uma máquina de aprendizagem treinada utilizando um subconjunto específico de características. Os métodos de filtros são também conhecidos como métodos de rankeamento de características, pois na maioria dos casos realizam o cálculo de um índice de relevância das características em relação à discriminação obtida em relação às categorias encontradas nos dados. Filtros e *wrappers* também podem ser combinados formando técnicas híbridas onde se utiliza os filtros para criação do rank e, em seguida, utiliza-se uma abordagem *wrapper* levando em consideração as características mais relevantes. Estas duas técnicas utilizam estratégias de busca para explorar o espaço de subconjuntos de características devido à inviabilidade de se efetuar uma busca exaustiva num espaço com muitas dimensões. Nos métodos embarcados a busca por um subconjunto ótimo de características é realizada dentro do processo de construção da máquina de aprendizagem [5]. Neste trabalho utilizou-se um método híbrido para seleção de características. Os filtros utilizados são baseados em análise uni variada, onde se realiza uma análise relativa à relevância individual de cada uma das características considerando-se independência entre elas. Estes filtros são vistos a seguir enquanto a estratégia de busca utilizada e a abordagem *wrapper* serão vistos nas seções posteriores.

3.1 F-Score

O filtro *F-Score* (*Fisher score*) é um critério simples e eficiente que, através de características estatísticas dos dados, mede a relevância das características para discriminação entre classes [6]. Considerando-se o um problema de classificação binário com as classes C_1 e C_2, ele é definido por:

$$f(i) = \frac{(\mu_i^{c_1} - \mu_i) + (\mu_i^{c_2} - \mu_i)}{\sigma_i^{c_1} + \sigma_i^{c_2}} \quad (1)$$

onde i corresponde ao índice da i-ésima características e, μ_i^c e σ_i^c são média e desvio padrão para a classe C em relação à característica i.

3.2 Pearson Correlation Coeficient

O coeficiente de correlação de Pearson é outro método geralmente utilizado para rankear características em relação ao seu poder discriminativo para as classes em problemas de classificação binários [5], sendo definido por:

$$f(j) = \frac{\sum_{i=1}^{p}(x_{ij}-\bar{x})(y_i-\bar{y})}{\sigma_{xj}\sigma_y} \quad (2)$$

onde j corresponde à j-ésima característica, i é um padrão de entrada e *p* é o número total de amostras. O vetor x_j contem todos os valores da característica j para todas as amostras de treinamento e y é o vetor contendo todos os valores alvos representando a classe referente a cada amostra.

4 Classificação de Padrões

A classificação de padrões é a tarefa em que se atribui um determinado objeto (padrão) a uma categoria (classe), dado um conjunto de características (também chamado conjunto de variáveis ou atributos) que representam este objeto. Nesta tarefa, de forma geral, determina-se a probabilidade de um objeto pertencer a uma determinada categoria, sendo geralmente impossível uma classificação ótima [7]. O classificador de *Bayes* e o classificador baseado na regra dos k vizinhos mais próximos (em inglês, *k-nearest-neighbor – K-NN*), são técnicas geralmente utilizadas na tarefa de classificação de padrões.

4.1 Classificador *de Bayes*

O Classificador de *Bayes* baseia-se na suposição de que o problema de decisão é visto de uma forma probabilística onde se conhece todos os valores de probabilidades relevantes. A classificação de um objeto a uma determinada classe é feita de acordo com a probabilidade de o objeto pertencer à classe [7]. Um classificador de *Bayes* simples (também denominado classificador de *Bayes* ingênuo) supõe independência entre as variáveis, o que não ocorre na maioria dos problemas de classificação, mas ainda assim obtém resultados competitivos com a maioria dos classificadores além de possuir menor complexidade computacional devido à facilitação nos cálculos utilizados obtida pela suposição de independência. A fórmula geral utilizada pelo classificador de *Bayes* é dada por:

$$P(C_j|X) = \frac{p(X|C_j)P(C_j)}{p(X)} \quad (3)$$

onde $P(C_j \mid X)$ é o termo definido como probabilidade à posteriori que indica a probabilidade da classe ser C_j dado que o padrão X foi mensurado. O termo $p(X \mid C_j)$ é uma probabilidade condicional denominada verossimilhança que representa a probabilidade de X dado que a classe C_j foi apresentada e $P(C_j)$ é a probabilidade a priori, sendo a informação que reflete o conhecimento prévio que se tem sobre os dados em relação à predição de determinado objeto pertencer à classe C_j levando em consideração apenas as quantidades de objetos amostrados em cada classe. O termo p(X) é definido como evidência e pode ser visto como um mero fator de escala que garante que a soma das probabilidades à posteriori é igual a um.

4.2 *K-NN*

O K-NN é um classificador também muito utilizado na literatura [2], pertencente à categoria dos algoritmos de aprendizagem baseados em memória. Neste classificador os dados de treinamento são utilizados para formação de uma memória de exemplos com padrões de entrada e suas respectivas saídas corretas. Neste contexto, a classificação de um padrão ainda desconhecido ocorre através da análise dos padrões armazenados na memória, onde se atribui o rótulo de determinada classe a este padrão de acordo com a classe dos k padrões mais similares a ele, levando-se em consideração alguma métrica de distância para avaliação da similaridade [8]. Uma medida geralmente utilizada para se avaliar a similaridade entre padrões é a distância euclidiana.

5 Sistemas Imunológicos Artificiais

O sistema imunológico é responsável pela principal forma de proteção do hospedeiro contra agentes infecciosos. Podem ser geradas duas formas de resposta a estes invasores, uma rápida e efetiva efetuada pelo sistema imune inato e outra mais lenta e duradoura oriunda do sistema imune adaptativo [9].

As células do sistema imune inato constituem uma resposta a diversos patógenos invasores sem a exigência de uma exposição anterior a estes enquanto o sistema imune adaptativo gera uma resposta imune específica a um determinado agente infeccioso com produção de anticorpos a este patógeno. Qualquer molécula reconhecida pelo sistema imunológico adaptativo é denominada antígeno (Ag).

A geração de anticorpos (Abs) é feita pelos linfócitos B (ou células B). Estas células são capazes de desenvolver uma memória imunológica que permite a identificação de um estímulo antigênico caso este seja novamente exposto ao sistema imune, evitando assim uma possível nova infecção.

Os sistemas imunológicos artificiais inspiram-se nas definições acima citadas e, através das características básicas do sistema imune biológico, constroem ferramentas computacionais que auxiliam na resolução de complexos problemas de engenharia.

5.1 Algoritmo de Seleção Clonal

Na imunologia o princípio da seleção clonal define que células que reconhecem antígenos são selecionadas para proliferar, passando por um processo de clonagem através de sucessivas mitoses. Estes clones estão sujeitos a mutações somáticas a altas taxas e uma força seletiva formando o processo de maturação de afinidade, onde os níveis de afinidade das células são melhorados em relação aos antígenos reconhecidos. Outro mecanismo a ser considerado é a edição de receptores, onde células com baixo nível de afinidade são substituídas por células totalmente novas, visando manter a diversidade populacional [10].

Baseado nestes conceitos o algoritmo de seleção clonal utiliza conceitos básicos do funcionamento do sistema imunológico biológico para formulação de ferramentas para resolução de diversos problemas complexos de engenharia. O algoritmo de seleção clonal pode ser também considerado um algoritmo evolucionário, devido às características de diversidade, variações genéticas e seleção natural presentes nele.

O algoritmo CLONALG proposto em [11] demonstra uma aplicação computacional dos princípios de seleção clonal e maturação de afinidade aplicada inicialmente a tarefas de aprendizagem de máquina e reconhecimento de padrões, sendo posteriormente adaptada a problemas de otimização.

O algoritmo implementado neste trabalho tem por objetivo a resolução de problemas de otimização, utiliza uma representação binária para os Abs e consiste dos seguintes passos:

i. Geração de uma população inicial aleatória de Abs, denominada conjunto Ab.
ii. Avaliação da afinidade dos indivíduos presentes em Ab em relação à função objetivo.
iii. Seleção dos b% Abs com maior afinidade em Ab, compondo uma subpopulação Abn.
iv. Clonar os Abs presentes em Abn, formando um conjunto de clones C, sendo o número de clones de cada Ab proporcional à afinidade dos mesmos, onde Abs com maiores afinidades possuem um maior número de clones.
v. Submissão do conjunto de clones C ao processo de maturação de afinidade, onde sofrem mutações em altas taxas, proporcionais aos seus níveis de afinidade. Abs com maiores afinidades têm menores taxas de mutação. Ao fim deste processo é gerado um conjunto Cm de clones maturados.
vi. Avaliação dos Abs do conjunto Cm de clones maturados.
vii. Seleção dos Abs do conjunto Cm com maiores afinidades para compor a população Ab. Um determinado Ab presente em Cm que tenha afinidade maior que seu respectivo representante na população Ab substitui este.
viii. Substituir os w% Abs com menores afinidades em Ab por novos indivíduos gerados aleatoriamente.

Esta sequência de passos se repete a partir do passo 2 até que se alcance um critério de convergência para o algoritmo.

Após selecionar-se os b% Abs com maiores afinidades da população Ab (passo 3) o processo de clonagem (passo 4) é regido por:

$$N_c = \sum_{i=1}^{n} round\left(\frac{\beta . N}{i}\right) \quad (4)$$

onde Nc é o número de total de clones gerados na etapa de clonagem, β é um fator de multiplicação, N é o total de Abs da população Ab e *round*() é utilizado para arredondar o valor da função para o inteiro mais próximo. Cada parcela do somatório presente na função representa a quantidade de clones de um elemento Ab_i sendo que estes elementos estão ordenados de forma decrescente em relação à afinidade, onde i representa o índice destes elementos ordenados.

No processo de maturação de afinidade a taxa de mutação é proporcional à afinidade dos indivíduos. Em [11] o cálculo da taxa de mutação é dado por:

$$\alpha = \exp(-\rho . f) \quad (5)$$

onde α é o tamanho do passo, ρ é o fator que controla o decaimento da função e f é a afinidade do indivíduo normalizada no intervalo [0;1]. A utilização de limites mínimos e máximos para a taxa de mutação pode auxiliar numa busca mais eficiente. A

relação entre a afinidade dos indivíduos e a taxa de mutação pode ser visualizada na figura 1, onde nota-se claramente a forte influência do parâmetro ρ no desempenho do algoritmo.

A operação de mutação no algoritmo de seleção clonal consiste na operação chave para determinar o desempenho do algoritmo de otimização em relação à velocidade de convergência e eficácia na busca pela solução ótima [12]. A mutação proporcional à afinidade representa um processo de busca local na superfície da função.

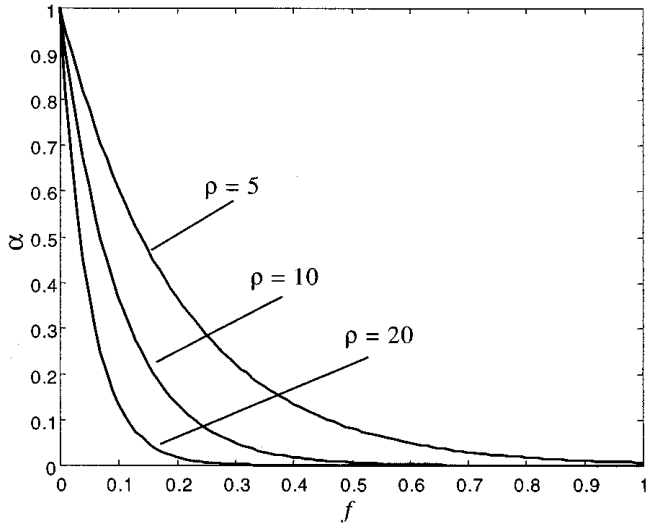

Figura 1 – Relação entre a taxa de mutação e a afinidade do indivíduo.

O processo de edição de receptores, representado pela substituição dos indivíduos com menores afinidades da população Ab por novos indivíduos gerados aleatoriamente, impõe a introdução e manutenção da diversidade populacional, efetuando uma busca global através da exploração de novas regiões na superfície total de busca.

Características como a aplicação a problemas de aprendizagem de máquina e otimização, busca por vários ótimos em funções multimodais e um pequeno número de parâmetros para se ajustar fazem com que o Clonal torne-se recomendável à aplicação em diversos problemas em diversas áreas de pesquisa.

6 O Método Proposto

A abordagem proposta é baseada na utilização de um método híbrido para seleção de características, onde os filtros uni-variados *F-Score* e Coeficiente de Correlação de *Pearson* são utilizados para rankear o conjunto total de características, sendo que as K-melhores características serão pré-selecionadas e submetidas a um método *wrapper*, que utiliza um algoritmo imunológico artificial, o Clonal, como estratégia para a realização da busca combinatória do menor subconjunto de características com melhor desempenho na classificação.

Cada anticorpo representa um subconjunto de características onde cada característica no subconjunto é representada por um bit de informação, sendo que o valor 0 (zero) indica ausência daquela característica e o valor 1 (um) indica sua presença.

Para avaliação da afinidade de um determinado anticorpo utiliza-se a função objetivo usada em [3] que busca a otimização de dois objetivos: a maximização da exatidão (taxa de classificações corretas) de um classificador e a minimização do tamanho do subconjunto de características, sendo definida por:

$$f(x) = w * c(x) + (1 - w) * \frac{1}{s(x)} \quad (6)$$

onde x é um vetor de características que representa um determinado anticorpo, c(x) é a exatidão de um classificador, s(x) é o tamanho do subconjunto de características e w ϵ [0,1] é um parâmetro utilizado para ponderar as duas partes da expressão, sendo que a definição de um valor adequado para este parâmetro conduzirá a um compromisso adequado entre exatidão e tamanho do subconjunto de características.

O classificador de *Bayes* foi o algoritmo de aprendizagem treinado utilizado para avaliação da afinidade de cada subconjunto de características.

7 Metodologia Utilizada

Inicialmente aplicaram-se os filtros *F-Score* e *Pearson* para rankear as 7129 características presentes na base de dados da leucemia. Após a obtenção deste rank as 50 primeiras características obtidas em cada método foram pré-selecionadas definindo assim o tamanho dos anticorpos, que é igual à quantidade de características utilizadas. No passo seguinte o algoritmo Clonal foi utilizado para se efetuar a busca do menor subconjunto presente nestas 50 características que conduzem à melhor

exatidão na classificação. Para o treinamento do classificador de *Bayes* utilizou-se a porção dos dados definida como conjunto de treinamento, com 38 amostras, enquanto que, o conjunto independente foi dividido proporcionalmente pela metade em dois subconjuntos, validação e teste, onde o subconjunto de validação é utilizado para definir a exatidão do classificador de *Bayes* no método *wrapper* e o subconjunto de testes é utilizado num momento posterior, após a seleção final de características, com o objetivo de testar o desempenho do método proposto. Após a obtenção do subconjunto de características mais adequado à classificação testou-se este subconjunto com o próprio classificador de *Bayes* e com o K-NN. Utilizou-se também o algoritmo *K-means* [7] para geração de clusters obtidos a partir dos dados de teste usando o subconjunto de características selecionado, com o objetivo de visualizar de forma geométrica a disposição dos padrões.

Para o algoritmo Clonal utilizou-se uma população com 80 anticorpos e 100 gerações. Os demais parâmetros utilizados foram: quantidade de indivíduos selecionados para clonagem (b) = 80%, percentual de piores indivíduos a serem substituídos (w) = 20%, ρ= 3,2 e β= 0,5. O valor 0,7 foi utilizado para o parâmetro w na ponderação dos objetivos na função afinidade, dando mais prioridade à taxa de classificações corretas do que para o tamanho do subconjunto de características.

8 Simulações e Resultados

Após o rankeamento das sondas através dos dois filtros uni-variados utilizados, as primeiras 50 sondas encontradas por cada método foram submetidas ao algoritmo Clonal, que retornou um subconjunto de 16 sondas referentes ao *F-Score* (S1), e um subconjunto de 15 sondas referentes ao método de *Pearson* (S2). Após a seleção realizada pelo clonal gerou-se um novo subconjunto S3 contendo as sondas que aparecem tanto em S1 quanto em S2. A Tabela 1 mostra a relação de sondas dos conjuntos S1, S2 e S3, indicando o índice de cada sonda dentre as 7129.

Tabela 1: Relação de sondas em cada subconjunto.

	Sondas Selecionadas
S1	2020 – 2288 – 3847 – 1882 – 4196 – 2402 – 6200 – 1674 – 6803 – 1807 – 3605 – 6405 – 5808 – 2001 – 4377 – 6919
S2	3320 – 2020 – 5039 – 1834 – 4196 – 2288 – 6201 – 1882 – 2121 – 6803 – 2402 – 3605 – 6677 – 6405 – 4377
S3	1882 – 2020 – 2288 – 2402 – 3605 – 4196 – 4377 – 6405 – 6803

As Tabelas 2 e 3 mostram a taxa classificações corretas para os dados de teste e conjunto independente (validação + teste) respectivamente, utilizando as sondas definidas em S1, S2 e S3 para o classificador de *Bayes* e o K-NN.

Tabela 2: Percentual de classificações corretas para o conjunto de dados de teste para cada classificador.

	Bayes	***K-NN***
S1	94,1176%	88,2353%
S2	94,1176%	94,1176%
S3	100%	94,1176%

Tabela 3: Percentual de classificações corretas para o conjunto de dados de independentes para cada classificador.

	Bayes	***K-NN***
S1	97,0588%	82,3529%
S2	97,0588%	85,2941%
S3	100%	91,1765%

Através das Figuras 2 e 3 podem ser visualizados os clusters formados pelo algoritmo *K-means* com o conjunto de dados de teste utilizando os subconjuntos S1 e S2 respectivamente, onde considerando estes clusters, obteve-se 94,1176% dos padrões agrupados em suas classes corretas.

Nas Figuras 4 e 5 são mostrados os clusters gerados pelo *K-means* para o subconjunto S3 utilizando o conjunto de dados de teste e conjunto total de dados (treinamento + independente) respectivamente. Neste caso obteve-se 94,1176% dos dados de teste agrupados em suas classes corretas e 91,6667% para o conjunto de dados total. Os clusters gerados pelo *K-means* são visualizados em relação às duas primeiras sondas de cada subconjunto.

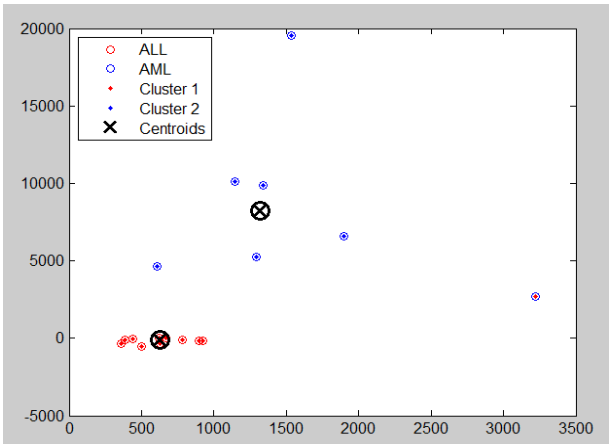

Figura 2 – Cluster gerado pelo *K-means* através do conjunto de dados de teste utilizando o conjunto S1.

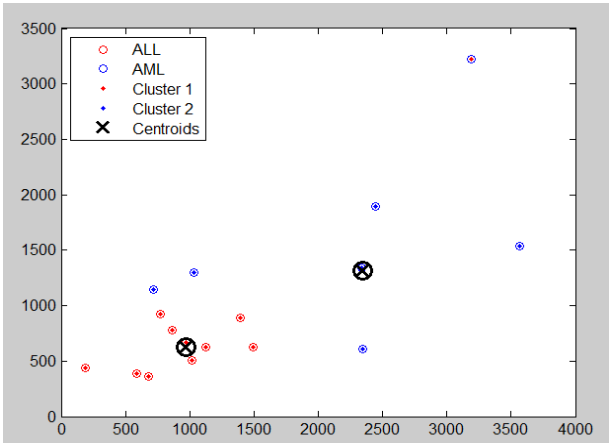

Figura 3 – Cluster gerado pelo *K-means* através do conjunto de dados de teste utilizando o conjunto S2.

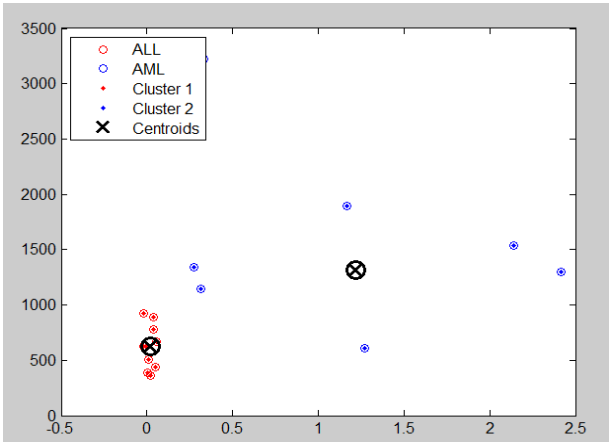

Figura 4 – Cluster gerado pelo *K-means* através do conjunto de dados de teste utilizando o conjunto S3.

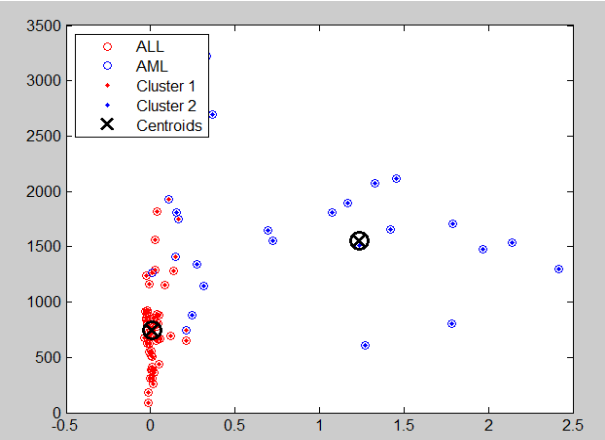

Figura 5 – Cluster gerado pelo *K-means* através do conjunto total de dados utilizando o conjunto S3.

Através dos resultados apresentados pode-se visualizar que os subconjuntos S1 e S2 têm efeitos muito semelhantes, sendo iguais quando se utilizou o classificador de *Bayes* tanto para o conjunto de testes como para o conjunto de dados independente e, verificou-se uma leve superioridade de S2 nos mesmos casos utilizando o *K-NN*. Quando se utilizou o conjunto S3 verificou-se resultados superiores em todos os casos em relação aos outros subconjuntos.

9 Conclusões

Este trabalho apresentou um método híbrido de seleção de características que realiza uma pré-seleção com filtros uni-variados e uma seleção multivariada através de um método *wrapper*. Na seleção multivariada utiliza-se o algoritmo Clonal como estratégia de busca e o classificador de *Bayes* para avaliação dos subconjuntos de características.

A classificação de tipos de leucemia em *ALL* e *AML* teve bons resultados com a utilização dos subconjuntos de sondas selecionadas pelo método utilizado. Percebeu-se que as melhores taxas de classificação foram alcançadas quando se combinou os subconjuntos resultantes retornados pelo método *wrapper*. A utilização de outros filtros uni-variados ou mesmo multivariados para pré-seleção das características poderia trazer maiores parâmetros para comparações. A combinação dos dados obtidos pelos filtros uni-variados num momento anterior à submissão ao método *wrapper* poderia gerar melhores resultados. A inclusão de novos tipos de gráficos para visualização da dispersão dos dados e agrupamento entre classes e a utilização de outros classificadores para avaliação das características selecionadas enriqueceriam trabalhos posteriores.

10 Agradecimentos

O presente trabalho foi realizado com o apoio financeiro da CAPES – Brasil.

11 Referencias

[1] T. R. Golub, at al., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, I**n Science**, 286(1999), 531–537.

[2] S. Cho, Exploring features and classifiers to classify gene expression profiles of acute leukemia, **International Journal of Pattern Recognition and Artificial Intelligence**, 16(2002), 831–844.

[3] F. Tan, X. Fu, Y. Zhang, A. G. Bourgeois, Improving feature subset selection using a genetic algorithm for microarray gene expression data, IEEE Congress on Evolutionary Computation, CEC 2006, (2006), 16–21.

[4] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techiniques in bioinformatics, **In Bioinformatics**, 23(2007), 2507–2517.

[5] I. Guyon, A. Elisseeff, An Introduction to Feature Extraction, **Springer**, (2006).

[6] Y. Chang, C. Lin, Feature ranking using linear SVM, **In WCCI2008 Workshop and Conference Proceedings 3**, (2008), 53–64.

[7] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2ª Edition , **John Wiley and Sons**, (2001).

[8] S. Haykin, Redes Neurais: principios e prática, 2ª Edition, **Bookman**, (2001).

[9] L. N. De Castro, Engenharia Imunológica: Desenvolvimento e Aplicação de Ferramentas Computacionais Inspiradas em Sistemas Imunológicos Artificiais, **Tese de doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação - Unicamp**, (2001).

[10] L. N. De Castro, J. Timmis, An Artificial Immune Network for Multimodal Function Optimization, **IEEE Congress on Evolutionary Computation (CEC'02)**, 1(2002), 699–674.

[11] L. N. De Castro, F. J. Von Zuben, Learning and Optimization Using the Clonal Selection Principle, **IEEE Transactions on Evolutionary Computation**, 6(2002), 239 - 251.

[12] L. Liang, G. Xu, D. Liu, S. Zhao, Immune Clonal Selection Optimization Method with Mixed Mutation Strategies, **Second International Conference on Bio-Inspired: Theories and Applications, BIC-TA**, (2007), 37 – 41.