Aplicação de Modelos Logísticos Univariados e Multivariados na Seleção de Genes para Problemas de Classificação de Leucemia

Alisson Marques Silva

alisson@cefetmg.br

Centro Federal de Educação Tecnológica de Minas Gerais - Campus Divinópolis

Rua Álvares de Azevedo, 400, Divinópolis, 35503-822, MG, Brasil

Resumo. *A classificação da Leucemia Aguda em seus subtipos Linfóide e Mielóide é geralmente realizada de acordo com a morfologia do tumor. No entanto, os subtipos podem ter aparência histopatológica similar, dificultando os procedimentos de diagnóstico. Este trabalho apresenta o uso de uma abordagem estatística baseada em análise univariada e multivariada, modelos logísticos e procedimentos de permutação para selecionar as variáveis estatisticamente mais relevantes. Esta abordagem foi empregada no problema de classificação de Leucemia. A base de dados utilizada nos experimentos é formada por perfis de expressão gênica, obtidas por meio de monitoramento de microarrays de DNA, de pacientes com Leucemia Aguda. Cada amostra é rotulada como ALL (Acute Lymphoid Leukemia - Leucemia Linfóide Aguda) ou AML (Acute Myeloid Leukemia - Leucemia Mielóide Aguda). Os dados possuem alta dimensionalidade e um pequeno número de amostras, o que torna necessário identificar e selecionar os genes mais importantes para esse diagnóstico. Para analisar o desempenho do processo foi utilizado, como agente classificador, uma rede neural MLP (MultiLayer Perceptron). Os resultados obtidos com a realização deste estudo foram comparados com os encontrados na literatura e se mostraram promissores.*

Palavras-chave: *Modelo Univariado, Modelo Multivariado, Expressão Gênica, Leucemia, Redes Neurais, AML e ALL*

1 INTRODUÇÃO

Os primeiros relatos de Leucemia datam do ano de 1845, desde então grandes avanços foram obtidos no seu entendimento, identificação e tratamento. A Leucemia é um grupo de doenças complexas e diferentes entre si que afetam a produção dos glóbulos brancos. De acordo com o Instituto Nacional do Câncer, a Leucemia é um tipo de Câncer que começa em tecidos formadores de sangue, como a medula óssea, e faz com que um grande número de células do sangue seja produzido. Mais especificamente, na Leucemia ocorre uma alteração genética adquirida nas células primitivas da medula óssea. O resultado desta alteração é um crescimento anormal dos glóbulos brancos, aumentando sua concentração no sangue. O principal desafio do tratamento da Leucemia e de outros tipos de Câncer é encontrar terapias específicas para o tratamento de tipos de tumores diferentes, a fim de maximizar a eficácia e minimizar a toxicidade. Portanto, a classificação do subtipo específico é essencial para o sucesso do tratamento.

Os diferentes tipos de Leucemia são agrupados pela velocidade em que a doença se desenvolve (crônica e aguda), e também são categorizados de acordo com quais células do sangue são afetadas (linfóide e mielóide) (Faria et al., 2015). As Leucemias crônicas permitem o crescimento de maior número de células já desenvolvidas e diferenciadas que, em geral, são capazes de exercer as suas funções normais e possuem uma progressão mais lenta. As Leucemias agudas são caracterizadas pelo crescimento rápido de células sanguíneas imaturas, que não são capazes de desempenhar as suas funções normais. Isso torna a medula óssea incapaz de produzir células sanguíneas saudáveis (Hamerschlak, 2005). A Leucemia Linfóide Aguda (ALL - *Acute Lymphoid Leukemia*) resulta na produção descontrolada de blastos de características linfóides e no bloqueio da produção normal de glóbulos vermelhos, brancos e plaquetas (Mughal et al., 2006). A Leucemia Mielóide Aguda (AML - *Acute Myeloid Leukemia*) se caracteriza pelo crescimento descontrolado e exagerado das células indiferenciadas chamadas blastos. Estas células não apresentam as funções normais dos glóbulos brancos. Além disso, existe um bloqueio na fabricação das células normais, ocorrendo uma deficiência de glóbulos vermelhos, plaquetas e glóbulos brancos (Mughal et al., 2006).

Ambos os tipos de células de Leucemia Aguda aparecem idênticos ao microscópio, o que, por muitos anos, levou-os a ser considerado uma única doença. No entando, eles são geneticamente muito distintos e podem seguir cursos clínicos significativamente diferentes. Mesmo com os avanços tecnológicos e o estabelecimento da distinção entre a AML e ALL, nenhum teste único é atualmente suficiente para estabelecer um diagnóstico preciso (Look, 1997; Rowley et al., 2000, Faria et al., 2015). Geralmente, esse diagnóstico não é efetuado de uma forma sistemática e necessita da intervenção de diferentes especialistas (Silva et al., 2003). Com isso, ainda há relatos de ocorrência de classificação incorreta da Leucemia (Golub et al., 1999). O diagnóstico correto do tipo de Leucemia é extremanente importante, pois possibilita ao médico a aplicação de um regime específico de tratamento, buscando maximizar a eficácia e minimizar a toxidade.

A tecnologia de micro*arrays* de DNA é um método de medição genômica de alto rendimento, que possibilita a identificação de assinaturas de expressão gênica associadas a subtipos clínicos distintos de Leucemia (Wouters et al., 2009). Um DNA micro*array*, ou DNA-chip, consiste num arranjo pré-definido de moléculas de DNA quimicamente ligadas a uma superfície sólida (lâminas de microscópio ou membranas de *nylon*) com carga positiva. Estes micro*arrays* são utilizados para a detecção e quantificação de ácidos nucléicos que passam pelo processo

de hibridação com o DNA fixado no *array*. Para que ocorra a detecção, as amostras são marcadas com fluorocromos cianina 3 (Cy3) ou cianina 5 (Cy5) quando utiliza-se micro*arrays* em lâminas de microscópio ou com o isótopo 33-P quando os *arrays* são preparados em membranas de *nylon*. A Figura 1 ilustra as principais etapas para obtenção dos perfis de expressão gênica por micro*arrays*.

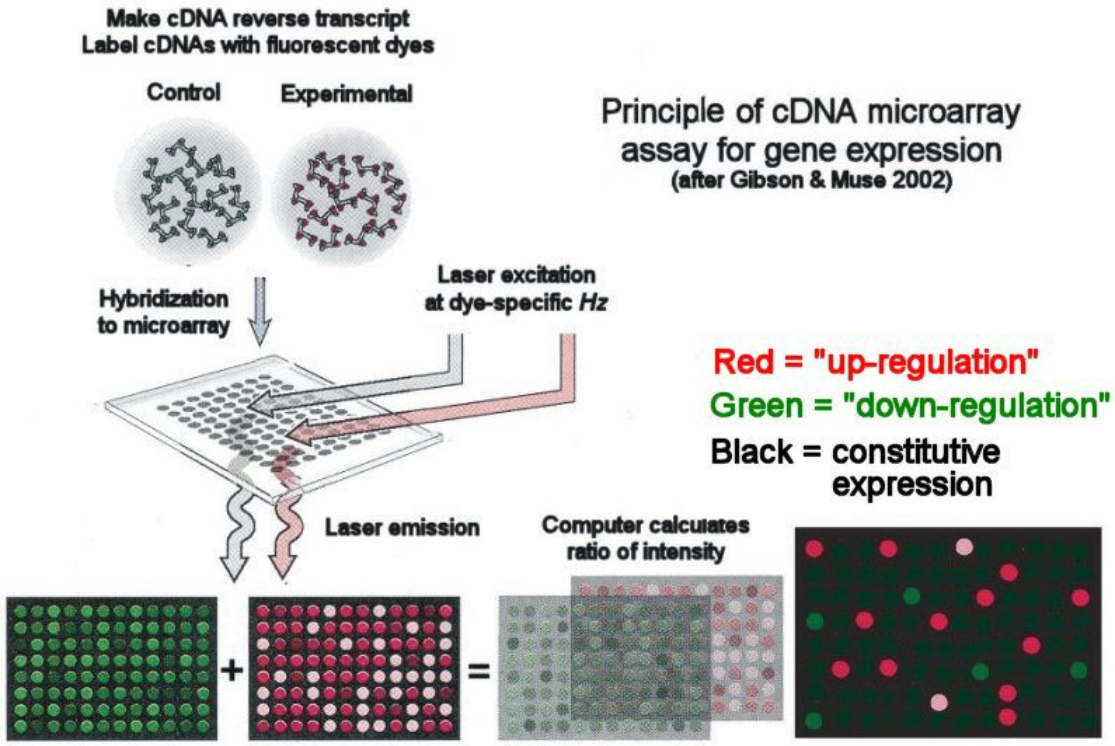

Figure 1: Etapas da técnica de micro*arrays*
Fonte: http://www.mun.ca/bioLogy/scarr/cDNA_microarray_Gene_Expression.html

Como os ensaios de micro*arrays* possibilitam analisar a expressão de múltiplos genes em paralelo, eles foram propostos como um teste robusto para o diagnóstico. Estes podem ser analisados para: medir os níveis de expressão de determinado conjunto de genes de uma célula sob efeito de uma determinada droga, descobrindo qual deles estão mais relacionados ao efeito da droga; comparar os níveis de expressão de uma célula saudável com uma doente; monitorar as mudanças nos níveis de expressão em resposta a um determinado tratamento (Silva et al., 2003). A classificação de tumores cancerígenos, no caso da Leucemia, é também uma das aplicações possíveis dos micro*arrays*, pois possibilita a criação de uma estratégia geral e sistemática que permite melhorar e facilitar o diagnóstico. O número e a variedade de aplicações da tecnologia dos microarrays de DNA são ilimitadas (Silva et al., 2003).

A tecnologia de micro*arrays* produz uma quantidade de informação onde o número de genes (variáveis) é muito maior que o número de amostras. Isso dificulta e até impossibilita o uso de ferramentas estatísticas habituais de classificação. Devido esse fato, e, principalmente, por saber que apesar do número elevado de genes, apenas alguns deles são relevantes na identificação de uma patologia, selecionar esses genes se torna um fator primordial no processo de classificação. É importante considerar que devido a atual tecnologia de micro*arrays* de DNA e também por se tratar de dados biológicos, esses dados possuem muito ruído (Costa et al., 2011). O objetivo principal dos estudos de micro*arrays* aplicados na classificação de

tumores é identificar os genes com maior potencial preditivo. A criação de mecanismos que permitem selecionar apenas estes genes é essencial, tanto a nível estatístico quanto a nível de interpretação biológica (Silva et al., 2003).

O grande desafio na modelagem de dados de micro*arrays* para seleção de genes relevantes na identificação de patologias está no elevado número de genes e na baixa quantidade de amostras. Associado ao problema de seleção de genes, um preditor robusto é essencial para a classificação correta dos pacientes. As redes neurais artificiais (RNAs) são algoritmos computacionais baseados na estrutura e no comportamento dos neurônios no cérebro humano e podem ser treinados para reconhecer e categorizar padrões eficientemente complexos (Bishop et al., 2006).

O objetivo deste trabalho é aplicar as RNAs em problemas de classificação de diagnóstico de Leucemia com base no perfil de expressão gênica. O problema abordado foi a classificação de pacientes com Leucemia Linfóide Aguda e pacientes com Leucemia Mielóide Aguda. Nos últimos anos, os métodos de seleção de características foram aplicados em dados de micro*arrays* (Dudoit et al., 2002; Lee et al., 2005; Li et al., 2004; Statnikov et al., 2005; Ben-Dor et al., 2000 Jafari e Azuaje, 2006, Shah et al., 2012, Bhalla e Agrawal, 2013), e não há acordo quanto ao método mais apropriado. Este artigo apresenta um estudo de metodologia estatística proposta por Costa et al. (2011) e sua aplicação ao problema de seleção de genes para distinção entre Leucemia Linfóide Aguda e Leucemia Mielóide Aguda. Nesta metodologia, primeiro são selecionadas as sondas por meio de modelos logísticos univariados. Para cada sonda selecionada, é utilizado um modelo multivariado que acrescenta, incrementalmente, variáveis até que se alcance um número máximo de variáveis definidas pelo usuário, ou quando o modelo fornece a classificação completa dos dados de treinamento. Para validar e medir o desempenho do método será empregada uma rede neural MLP (*MultiLayer Perceptron*).

De forma a alcançar os objetivos apresentados, este artigo está organizado da seguinte forma: na seção 2 são apresentados alguns dos trabalhos relacionados ao problema de classificação por meio de análise de expressão gênica; na seção 3 descreve-se os conceitos básicos de redução de dimensionalidade e o método de ranqueamento de característica empregado nos experimentos; a seção 4 discorre sobre o conjunto de dados e os experimentos realizados; na seção 5 são apresentados os resultados e discussões; e, por fim, na seção 6 são descritas as conclusões.

2 REVISÃO DA LITERATURA

O trabalho pioneiro na análise do perfil de expressão gênica, para identificação de Leucemias, foi publicado por Golub et al. (1999). Neste estudo, foi desenvolvida uma técnica denominada *neighborhood analysis*, pela qual identificou uma lista de 50 genes mais fortemente correlacionados com a diferenciação entre ALL e AML. Após este trabalho, muitas pesquisas foram aplicadas na identificação de doenças, por meio de análise da expressão gênica. Entre essas pesquisas pode-se citar Smith et al. (2000), Slonin et al. (2000), Nguyen e Rocke (2001), Gamberger et al. (2004), Lemos et al. (2006), que utilizam a mesma base de dados que Golub (Golub et al., 1999). Também pode-se citar Silva et al. (2003) e Nascimento (2007) que utilizaram outras bases de expressão gênica.

Em suas pesquisas Smith et al. (2000) utilizou *T-statistic* e *Stepwise Discrimination* (ANCOVA) para selecionar os genes mais relevantes. Diferentemente de Golub et al. (1999),

que utilizou um conjunto para treinamento com 38 amostras e um conjunto para testes com 34 amostras, Smith et al. (2000) uniu esses dois conjuntos e depois criou dois subconjuntos aleatórios com 36 amostras cada, sendo um para treinamento e o outro para testes. Por outro lado, no trabalho de Nguyen e Rocke (2001) foi introduzido o método PLS (*Partial Least Squares* - Mínimos Quadrados Parciais) na classificação de Leucemia. Neste trabalho, também se discutiu sobre o uso de PCA (*Principal Components Analysis* - Análise dos Componentes Principais), ADQ (*Quadratic Discriminant Analysis* - Análise Discriminante Quadrática) e LD (*Logistic Discrimination* - Discriminação Logística) aplicados na seleção dos genes mais relevantes no processo de classificação de Leucemia.

Várias metodologias para a redução de dimensionalidade aplicadas a micro*arrays* para a classificação de tumores cancerígenos foram apresentadas em Silva et al. (2003). Na seleção de genes para duas classes Silva et al. (2003) empregou a métrica de correlação e estatística T*. Na seleção para mais de duas classes, utilizou um contra todos, análise de variância, estatística F, Rácio BSS/WSS e método do centróide mais próximo. Também foram utilizados PCA, PLS e variáveis canônicas. Um algoritmo de classificação denominado TSP (*Top Scoring Pair*), foi proposto em Naiman et al. (2004) com o objetivo de gerar classificadores a partir de dados de expressão gênica, que além de precisos, fossem de fácil interpretação. Essa técnica consiste na geração de regras com poucos elementos em seu termo antecedente. Neste trabalho foram utilizadas três bases de dados: uma de células de mama; outra de células de próstata e; também uma de células de Leucemia (Golub et al., 1999). Já no trabalho de Lemos e Baranaukas (2006) foi empregada uma técnica que utiliza árvores de decisão com arrendondamento de valores para classificação nos dois tipos de Leucemia (ALL e AML). Na pesquisa de Araújo (2008) foi realizada uma análise de sete técnicas de agrupamento e quatro técnicas de medida de proximidade, analisadas em 35 conjuntos de expressão gênica.

Costa et al. (2011) utilizou um conjunto de expressão gênica de 22.283 sondas. Para cada paciente, os dados são o resultado do tratamento quimioterápico pré-operatório e indicam: resposta completa ao tratamento (PCR) ou resposta não completa (noPCR - resíduo da doença). Neste trabalho foi proposta uma metodologia hierárquica multivariada para seleção dos genes. Esta inclui permutação de Monte Carlo e teste de hipótese nula. A metodologia proposta por Costa et al. (2011) será detalhada na próxima seção.

3 REDUÇÃO DA DIMENSINALIDADE

Segundo Martins Junior (2004) a redução de dimensionalidade busca identificar um subespaço suficientemente reduzido de características, que seja capaz de representar qualquer padrão conhecido de acordo com um determinado critério. É recomendável que essa redução seja realizada de forma automática e, principalmente, sem sacrificar a precisão na classificação.

Basicamente, duas abordagens podem ser empregadas para tratar esse problema: a fusão de características e a seleção de características (Martins Junior, 2004). Os algoritmos de fusão de características criam novas características a partir de transformaçôes ou combinações do conjunto original, um exemplo deste tipo é o PCA. A seleção de características consiste na utilização de métodos estatísticos na extração das informações mais relevantes de um conjunto de dados, identificando os dados que melhor representam uma categoria. Estes algoritmos analisam cada característica individualmente e seleciona as n melhores. Porém, nem sempre o conjunto formado pelas n melhores variáveis individualmente formam o melhor conjunto.

A metodologia que será apresentada a seguir seleciona as n melhores variáveis individuais e depois acrescenta incrementalmente novas variáveis ao conjunto de forma que melhore seu desempenho.

Para seleção de características, isto é, conjunto de genes (níveis de expressão gênica) mais relevantes na classificação de Leucemia em ALL ou AML foi utilizada a metodologia proposta por Costa et al. (2011) aplicada ao conjunto de treinamento. Inicialmente o método de permutação foi empregado para identificar um conjunto de níveis de expressão gênica para distinção entre as duas respostas possíveis para o tratamento. Costa et al. (2011) utilizou uma combinação de análise univariada, multivariada e modelos logísticos como classificador. Este método pode ser dividido em 3 etapas, que são descritas a seguir.

Na primeira etapa, o modelo estatístico utilizado é uma regressão logística (distribuição de Bernoulli). O resultado do modelo se relaciona com um preditor linear e uma função logística, $\mu = exp(\beta_0 + \beta_1 x)/[1 + exp(\beta_0 + \beta_1 x)]$, onde μ é a probabilidade da Leucemia ser ALL ($y = 1$). Os parâmetros ($\beta' s$) foram obtidos aplicando o método $glmfit$ (MathWorks, 2015), com uma função binomial. Para teste estatístico do modelo, foi utilizado uma função desvio (*deviance*) obtida por meio da função $glmfit$. A *deviance* é calculada para cada sonda no modelo univariado e as sondas são classificadas por ordem crescente, sendo as com menores *deviances* possuem os melhores coeficientes.

Na segunda etapa, o procedimento de permutação busca encontrar um limite para os valores da *deviance* e selecionar as variáveis com valores abaixo desse limiar. Sendo que estas rejeitam a hipótese nula de não associação entre a variável preditora e a variável de resposta. Os passos para a permutação são definidos como se segue:

1. Ajuste do modelo logístico para cada k variável dependente.
2. Ordene as variáveis em ordem crescente de acordo com os valores das *deviances*.
3. Gere S simulações do teste estatístico T:
   (a) Gere aleatoriamente os elementos de cada vetor x_k^T;
   (b) Calcule uma nova *deviance*, D_{ik}, onde i é' a t^h simulação do total de S;
   (c) Faça $T_i = minD_{ik}$.
4. Defina o limiar $(100.\alpha)$ percentil da simulação $T_i' s$.
5. Selecione as sondas cujas *deviances* são inferiores ao limiar.

Por fim, na terceira etapa, na abordagem multivariada, cada uma da sondas (agora chamadas de sementes) selecionadas no modelo univariado são analisadas com as outras $n - 1$ sondas e avaliadas como um potencial segundo modelo. Apenas a sonda que oferece o menor *deviance* com a semente é selecionada para compor o novo conjunto. Após a formação deste conjunto (formado pela semente e pela sonda com menor *deviance*) o processo é repetido para as outras $n - 2$ sondas e assim sucessivamente, até que, o modelo alcance um número k de variáveis definido pelo usuário. Esse procedimento é executado para todas as variáveis selecionadas pelo método univariado. Pode ocorrer do número de variáveis ser menor do que k, isso ocorre quando o processo multivariado encontra um pequeno subconjunto que encontra uma *deviance* nula.

4 MATERIAS E MÉTODOS

O conjunto de dados utilizado nos experimentos foi o Leukemia[1], inicialmente empregado em Golub et al. (1999). Este conjunto é composto por 7.129 níveis de expressão gênica de 72 pacientes com Leucemia, 47 com ALL e 25 com AML. As amostras foram obtidas por monitoramento de micro*arrays* pela técnica *Affymetrix HuGeneFL Array* (Affymetrix, 2016). O conjunto de dados foi dividido em conjunto de treinamento com amostras de 38 pacientes (27 ALL e 11 AML) e conjunto de teste com 34 (20 ALL e 14 AML).

Os experimentos foram executados no conjunto de dados com e sem pré-processamento como descrito a seguir:

1. Conjunto de dados sem pré-processamento;
2. Conjunto de dados pré-processado pelo processo de normalização Quantile (Matworks, 2015).
3. Conjunto de dados pré-processado pela aplicação da função Logaritmo com base 10, seguido do processo de normalização Quantile.

Os experimentos foram realizados como descrito a seguir:

- O nível de significância foi variado de $\alpha = 0,01$ (1%) a $\alpha = 0,09$ (9%) para o procedimento de seleção univariado e multivariado. Com base nos resultados obtidos com a variação do valor de α optou-se por utilizar $\alpha = 0,02$.
- O número de simulação inicialmente foi testado[2] com os seguintes valores 10, 25, 50, 250, 500, 1.000. O tempo total de simulação é variável e depende da normalização ou não do conjunto de dados e também do tipo de normalização realizada. Optou-se pelo valor de total de simulações $S = 500$. A escolha se deu pelo elevado custo computacional para executar as simulações. Somente para o cálculo univariado com 500 simulações em computador com Windows 7 Professional, processador Core i5 2.4 GHz e 8 GB de RAM o tempo de simulação chegou a 7 horas. A Tabela 1 ilustra as sondas selecionadas pelo método univariado para serem as sementes do método multivariado.

Table 1: Sondas selecionadas pelo método univariado

Seq.	Sem Normalizar.	Quantile	Log10 + Quantile
1	X95735	X95735	X95735
2	U50136	U50136	U50136
3	M92287	M27891	M27891
4	M27891	M92287	M92287
5	Y12670	M31523	M31523
6	X74262	X74262	M27783
7	M27783	M27783	X74262
8		M31211	
9		D88422	

[1]Disponível em http://www.broadinstitute.org

[2]Os testes iniciais foram realizados com a base de dados *Heart*, que possui 13 características de 270 pacientes.

- O teste multivariado foi executado com k igual ao número de sondas selecionadas pelo método univariado. Devido a isso, para cada conjunto de dados será obtido k modelos com k variáveis cada um. As Tabelas 2, 3 e 4 apresentam respectivamente os modelos criados para os dados sem pré-processamento, normalizados pelo Quantile e pela aplicação do Log10 seguido do Quantile.

Table 2: Conjuntos formados por meio dos dados sem pré-processamento

Gene Mod.	1	2	3	4	5	6	7
1	X95735	BioB-5	BioC-3	BioB-M	BioB-3	BioC-5	BioDn-5
2	U50136	BioB-5	BioB-M	BioB-3	BioC-5	BioC-3	BioDn-5
3	M92287	HG1102	BioB-5	BioB-M	BioB-3	BioC-5	BioDn-5
4	M27891	M31158	AC00064-CDS1	BioB-5	BioB-M	BioB-3	BioC-5
5	Y12670	AB000462	BioB-5	BioB-3	BioB-M	BioC-3	BioDn-3
6	X74262	L36051	BioB-5	BioB-M	BioB-3	BioC-5	BioC-3
7	M27783	Y10871	M11507-M	BioB-5	BioDn-5	BioB-M	BioC-5

Table 3: Conjuntos formados por meio dos dados processados pelo Quantile

Gene Mod.	1	2	3	4	5	6	7	8	9
1	X95735	BioB-5	BioB-M	BioB-3	BioC-5	BioC-3	BioDn-5	BioDn-3	CreX-5
2	U50136	BioB-5	BioB-M	BioB-3	BioC-5	BioDn-3	BioC-3	BioDn-5	CreX-5
3	M27891	AF005043	BioB-5	BioB-M	BioB-3	BioC-5	BioC-3	BioDn-5	BioDn-3
4	M92287	D43767	BioB-M	CreX_3	CreX-5	BioB-5	BioB-3	BioC-5	BioC-3
5	M31523	M33197	BioB-5	BioB-M	BioB-3	BioC-5	BioC-3	BioDn-5	BioDn-3
6	X74262	M11507-3	BioB-M	BioB-3	BioC-5	BioB-5	BioC-3	BioDn-5	BioDn-3
7	M27783	X00351-5	BioB-5	BioB-M	BioB-3	BioC-3	BioDn-5	BioDn-3	CreX-5
8	M31211	HT4215	BioB-5	BioB-M	BioC-5	BioB-3	BioC-3	BioDn-5	BioDn-3
9	D88422	PheX_M	BioB-5	BioB-M	BioB-3	BioC-3	BioDn-5	BioDn-3	CreX-5

Table 4: Conjuntos formados por meio dos dados processados por Log10 + Quantile

Gene Mod.	1	2	BioB-3	4	5	6	7
1	X95735	BioB-5	BioB-M	BioB-3	BioC-5	BioC-3	BioDn-5
2	U50136	BioB-5	BioB-M	BioC-5	BioB-3	BioDn-3	BioC-3
3	M27891	AF00504	BioB-5	BioB-M	BioB-3	BioC-5	BioC-3
4	M92287	BioB-3	BioB-M	CreX_5	BioDn-3	BioC_3	CreX-5
5	MBioB	PheX_M	BioC-3	BioB-5	BioB-M	BioDn-5	BioC-5
6	M27783	BioB-3	BioB-5	BioB-M	BioB-3	BioC-3	BioDn-5
7	X74262	M115073	BioB-M	BioB-3	BioC-5	BioB-5	BioC-3

- O classificador baseado no modelo Logístico empregado por Costa et al.(2011) foi substituído por uma rede neural MLP. A rede neural foi treinada no conjunto de treinamento

e os resultados apresentados são os obtidos na base de teste. A rede possui uma camada oculta com cinco neurônios e função de ativação Logarítmica Sigmoidal e a camada de saída com um neurônio e função de ativação Tangente Sigmoidal. O algoritmo de treinamento utilizado foi o Levenberg-Marquart e cada experimento foi executado 10 vezes, sendo avaliado a média de acerto das 10 execuções.

5 RESULTADOS

Os resultados obtidos são apresentados na Tabela 5. Foram selecionados as médias das 10 execuções para cada um dos conjuntos de dados testados (sem pré-processamento, normalizados por Quantile e processados por Log10 + Quantile).

Table 5: Resultados

Conj. dados	N. Var.	Taxa de Erros - %
Quantile	9	7,17
Lemos e Baranaukas (2006)	*	8,46
Sem pré-process.	7	8,82
Log10 + Quantile	7	9,35
Gamberger et al. (2004)	4	13,25
Golub et al. (1999)	50	14,70

Os resultados apresentados na Tabela 5 ilustram que o melhor resultado foi obtido empregando a RNA e a normalização dos dados pelo Quantile. A segunda menor taxa de erros foi obtida no trabalho de Lemos e Baranaukas (2006)[3] e a terceira empregando a RNA e os dados sem pré-processamento. Os resultados desses três experimentos são similares e obtiveram desempenhos superiores aos apresentados nos trabalhos de Golub et al. (1999) e Gamberger et al. (2004). Deve destacar que nos experimentos realizados neste trabalho utilizou-se um número consideralmente menor de sondas e os resultados foram similares ou superiores aos encontrados na literatura.

6 CONCLUSÃO

O grande número de preditores potenciais e o pequeno tamanho da amostra de dados de micro*array* representam desafios adicionais para o desenvolvimento de modelos de aprendizagem. Se for encontrada evidência estatística de associação entre os níveis de expressão gênica e a discriminação de diferentes tipos de Câncer, os diagnósticos orientados para o paciente são viáveis usando modelos preditivos.

Este trabalho emprega métodos estatísticos univariados e multivariados na seleção de genes para classificação de Leucemia. A metodologia proposta por Costa et al. (2011) foi adaptada com a alteração nos índices e com a inclusão de diferentes métodos de tratamento de dados. Uma rede neural MLP foi utilizada como agente classificador. Os experimentos realizados mostraram que com a seleção de genes mais relevantes, essas metodologias se mostram adequadas para identificação de diferentes subtipos de Câncer.

[3]O número de variáveis não disponível.

REFERÊNCIAS

Affymetrix, 2006 *Affymetrix*. Disponvel em http://www.affymetrix.com/estore/. Acesso: junho, 2016

Araújo, D., 2008. *Algoritmos de agrupamento aplicados a dados de expressão gênica de câncer: Um estudo comparativo*. Dissertação de Mestrado, Universidade Federal do Rio Grande do Norte - Programa de Pós-Graduação em Sistemas e Computação.

Ben-Dor, A., Bruhn, L., Friedman, N., et al., 2000. Tissue classification with gene expression profiles. *J. Comput. Biol.* 7, 559-583.

Bhalla, A., Agrawal, R., 2013. Microarray gene-expression data classification using less gene expressions by combining feature selection methods and classifiers. *Int. J. Inf. Eng. Electron. Bus.* 5, 42-48.

Bishop, C.M., et al., 2006. *Pattern Recognition and Machine Learning*, Volume 1. Springer, New York.

Dudoit, S., Fridlyand, J., Speed, T.P., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* 97, 77-87.

Faria, A.; Silva, A. ; Rodrigues, T.; Costa, M.; Braga, A., 2015 . A Ranking Approach for Probe Selection and Classification of Microarray Data with Artificial Neural Networks. *Journal of Computational Biology*, v. 22, p. 953-961.

Gamberger, D.; Lavrac, N.; Zelezny, F.; Tolar, J., 2004. Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *Journal of Biomedical Informatics*, 37:269-284.

Golub, T.; Slonim, D.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.; Coller, H.; Loh, M.; Downing, J.; Caligiuri, M.; Bloomfield, C., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531-537.

Hamerschlak, N., 2005. Leucemia: Um doença potencialmente curável. *Prática Hospitalar*.

Jafari, P., and Azuaje, F., 2006. An assessment of recently published gene expression data analyses: Reporting experimental design and statistical factors. *BMC Med. Inform. Decis. Making* 6, 27.

Junior, D., 2004. *Redução de dimensionalidade utilizando entropia condicional média aplicada a problemas de bioinformática e de processamento de imagens*. Dissertação de Mestrado, Universidade de São Paulo.

Lee, J.W., Lee, J.B., Park, M., and Song, S.H., 2005. An extensive comparison of recent classification tools applied to microarray data. *Comput. Stat. Data Anal.* 48, 869-885.

Lemos, R.; Baranaukas, J., 2006. Uma avaliação de dados de expressão gênica em leucemias agudas utilizando árvores de decisão. *X Congresso Brasileiro de Informática em Sade - CBIS* - Florianópolis - SC.

Li, L., Umbach, D.M., Terry, P., and Taylor, J.A., 2004. Application of the ga/knn method to seldi proteomics data. *Bioinformatics* 20, 1638-1640.

Look, A.T., 1997. Oncogenic transcription factors in the human acute leukemias. Science 278,

1059-1064. Maimon, O.Z., and Rokach, L. 2005. *Data Mining and Knowledge Discovery Handbook*, Volume 1. Springer, New York.

MathWorks, 2015. Statistics Toolbox. *The MathWorks*, Natic, MA, USA.

Mughal, T., Goldman, J., and Mughal, S., 2006. *Understanding Leukemias, Lymphomas, and Myelomas.* Taylor & Francis, New York.

Naiman, D.; Geman, D., d'Avigon, Winslow, R., 2004. Classifying gene expression profiles from pairwise mRNA comparisons. *Statistical applications in genetics and molecular biology*, 2004.

Nascimento, E., 2007. *Análise dos microRNAs mir-223 e mir-155 em leucemia linfóctica crônica*. Dissertação de Mestrado, Departamento de Genética - Universidade de São Paulo.

Nguyen, D.; Rocke, D., 2001. Classification of acute leukemia based on dna microarray gene expressions using partial least squares. *Kluwer Academic*, 2001.

Rowley, J., et al., 2000. Molecular genetics in acute leukemia. *Leukemia* 14, 513-517.

Shah, M., Marchand, M., and Corbeil, J., 2012. Feature selection with conjunctions of decision stumps and learning from microarray data. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 174-186.

Silva, L., 2003. *Seleção de Variáveis em Microarrays de DNA*. Tese de Doutorado, Departamento de Matemática Aplicada - Faculdade de Ciências da Universidade do Porto - Portugal.

Slonim, D.; Tamayo, P.; Mesirov, J.; Golub, T.; Lander, E., 2000. Class prediction and discovery using gene expression data. In *RECOMB 00: Proceedings of the fourth annual international conference on Computational molecular biology*, pages 263-272, New York, NY, USA.

Smith, A.; Satagopan, J.; Gonen, M.; Begg, C., 2000. Exploring class prediction for leukemia gene expression data. *CAMDA*.

Statnikov, A., Aliferis, C.F., Tsamardinos, I., et al., 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* 21, 631-643.

Wouters, B.J., Lo wenberg, B., and Delwel, R., 2009. A decade of genomewide gene expression profiling in acute myeloid leukemia: Ashback and prospects. *Blood* 113, 291-298.