

Comitês de Agrupamento para Análise de Dados de Expressão Gênica

Shirly C. M. Silva, Daniel S. A. de Araújo, Raul B. Paradedda, Valmar S. Severiano Sobrinho,
Marcilio C. P. de Souto

*Departamento de Informática e Matemática Aplicada
Universidade Federal do Rio Grande do Norte - UFRN
Natal, RN, Brasil, 59072-970*

Email: {shirly, paradedda}@ppgsc.ufrn.br, {danielsa, xantho}@lcc.ufrn.br,
marcilio@dimap.ufrn.br

Abstract—Ensemble techniques have been successfully applied in the context of supervised learning to increase the accuracy and stability of classification. Recently, similar techniques have been proposed for clustering algorithms. In this context, we analyze the potential of applying cluster ensemble techniques to gene expression microarray data. Our experimental studies show that there is often a significant improvement in the results obtained with the use of ensemble when compared to those based on the clustering techniques used individually.

Resumo—O uso de técnicas de comitês (*ensembles*) vem sendo bem sucedido na abordagem de aprendizado de máquina supervisionado por aumentar a precisão e estabilidade da classificação. Recentemente, técnicas similares foram propostas para algoritmos de agrupamento (*clustering*). Nesse contexto, analisamos o potencial da aplicação das técnicas de comitês de agrupamento para dados de expressão gênica gerados por *microarrays*. Nossos experimentos mostraram uma melhoria significativa no resultado obtido com o uso do comitês comparado com técnicas de agrupamento usadas individualmente.

I. INTRODUÇÃO

Métodos de mineração de dados vêm sendo extensamente aplicados para descobrir padrões e relações complexas em uma grande variedade de dados [15]. Nesse artigo, destacamos o problema de analisar grupos gerados a partir de técnicas não-supervisionadas. Trabalharemos com a descoberta de diferenças entre os subgrupos que compõem uma população, onde os membros dessa população compartilham algumas características comuns que estão sendo consideradas no domínio do problema estudado [16],[20].

Como indicado em [20], a análise não-supervisionada é de grande interesse no campo de genomas funcionais e análise de dados de expressão gênica. Uma das razões para isso é a necessidade de um refinamento em nível molecular de classes biológicas, com implicações em diagnóstico, prognóstico e tratamento de doenças [20], [21],[22].

A análise de agrupamento, em geral, envolve repetidas execuções de diferentes algoritmos de agrupamento com inicializações aleatórias, seguida pela seleção de uma solução individual, de acordo com um critério de otimização definido pelo usuário [16]. Entretanto, em lu-

gar de apenas selecionar a “melhor” partição como uma solução, trabalhos recentes vêm mostrando que a combinação na forma de comitê (*ensemble*) das diferentes partições geradas pode, em geral, produzir melhores resultados.

Técnicas de comitês têm sido aplicadas com sucesso em aprendizado supervisionado para melhorar a exatidão e a estabilidade de algoritmos de classificação [1], [4]. Recentemente, tentativas desse tipo têm sido aplicadas para combinação de algoritmos não-supervisionados, tais como as técnicas de agrupamento [5], [7], [10], [12], [14], [23], [26].

Neste trabalho, analisaremos o potencial da aplicação de técnicas de comitês de agrupamento para dados de expressão gênica obtidos com *microarray*. Mais precisamente, desenvolveremos os experimentos com os métodos de comitês descritos em [5], [23]. Para isso serão selecionados como entrada para a construção dos comitês as partições geradas pelos algoritmos *k*-médias e EM (*Expectation Maximization*). Esses algoritmos têm sido amplamente usados na literatura para dados de expressão gênica [21], [22].

Nos experimentos, consideraremos três bases de dados bastante utilizadas, para as quais há uma distinção multi-classe (fenótipo) disponível: Leukemia, Novartis e St. Jude Leukemia [11], [20], [25], respectivamente. Estas bases de dados nos permitem, entre outras coisas, analisar a habilidade do método de agrupamento de encontrar refinamentos nas classes biológicas.

II. GERAÇÃO DE COMITÊS

De acordo com [24], o problema de geração de comitês (agrupamento consenso) pode ser definido formalmente como a seguir. Seja X um conjunto de N pontos de dados (objetos ou padrões) no espaço d -dimensional. Suponha que seja dado um conjunto de partições $\Pi = \{\pi_1, \pi_2, \dots, \pi_H\}$ de objetos. Cada partição de Π é um conjunto de grupos disjuntos, exaustivos e não-vazios $\pi_i = \{L_1^i, L_2^i, \dots, L_{K(i)}^i\}$, $X = L_1^i \cup \dots \cup L_{K(i)}^i$, $\forall \pi_i$ e $K(i)$ é o número de grupos da i -ésima partição. O problema de gerar o agrupamento consenso é encontrar uma nova partição $\sigma = \{C_1, \dots, C_2\}$ de X ,

dadas as partições em Π , tal que os objetos em um mesmo grupo de σ sejam mais similares entre si do que com os objetos em grupos diferentes σ .

A definição anterior nos leva as seguintes questões: como gerar diferentes partições? Como combinar múltiplas partições? Com relação a primeira pergunta, diversas abordagens similares àquelas usadas em aprendizado supervisionado têm sido propostas para introduzir instabilidades artificiais em algoritmos de agrupamento. Desta forma, diferentes agrupamentos do mesmo conjunto de dados podem ser produzidos, o que pode conduzir à melhoria da qualidade e robustez da saída do comitê [5], [7], [10], [12], [14], [20], [23], [26].

Por exemplo, em [7], [10], [20], algoritmos de fusão de múltiplas partições são apresentados, baseados na combinação de várias execuções de um mesmo algoritmo de agrupamento, tendo como resultado uma partição comum (partição consenso). Uma outra abordagem é a geração do comitê a partir da combinação de métodos de agrupamento diferentes, como em [5], [14], [23], [26].

A combinação de múltiplas partições é realizada em geral através da definição de uma função consenso. Uma função consenso mapeia um dado conjunto de partições $\Pi = \{\pi_1, \pi_2, \dots, \pi_H\}$ para uma partição alvo σ [24]. Alguns exemplos de funções consenso encontradas na literatura são as seguintes:

- *Matriz de co-associação* [9]: a similaridade entre dois objetos é estimada contando o número de grupos compartilhados. Então, o algoritmo hierárquico com ligação mínima é usado para encontrar a partição consenso.
- *Métodos baseados em hipergrafos* [23]: grupos em partições diferentes são representados por hiperarestas. A partição consenso é encontrada por meio de um *k-way min-cut* do hipergrafo.
- *Re-rotulagem e votação* [5]: se o problema de correspondência¹ da rotulagem entre os grupos para as partições dadas for resolvido, um simples procedimento de votação pode ser usado para atribuir objetos aos grupos.
- *Informação mútua* [24]: esses procedimentos maximizam a informação mútua entre as partições individuais e a partição consenso alvo.

Neste artigo, escolhemos analisar dois métodos de comitê de agrupamento apresentados em [5], [23]. O primeiro método [5] permite a combinação de diferentes partições em uma nova partição *fuzzy* que representa, de maneira ótima, as partições iniciais. As partições iniciais podem ser o resultado da aplicação de diferentes algoritmos de agrupamento ou o resultado da aplicação repetida de um algoritmo com diferentes inicializações aleatórias.

O algoritmo proposto combina as partições de forma seqüencial, escala linearmente com o número de padrões e o número de repetições, sendo factível para conjuntos de dados grandes. Ele melhora a habilidade de um algoritmo encontrar estruturas em um conjunto de dados, uma vez

¹ O problema de correspondência é caracterizado pela rotulagem dos grupos para diferentes partições de um comitê.

que pode encontrar grupos de qualquer formato. A idéia do algoritmo é buscar uma re-rotulação ótima dos grupos para identificar os grupos correspondentes em diferentes partições e buscar uma combinação ótima desses grupos re-rotulados. Para isso utiliza um esquema de votação.

O outro método de comitê utilizado nesse trabalho está descrito em [23]. Nesse contexto, todos os grupos das partições geradas pelo comitê podem ser representados como hiperarestas em um gráfico com N vértices. Cada hiperaresta descreve um conjunto de objetos pertencendo ao mesmo grupo. Uma função consenso é executada como uma solução do problema de particionamento de hipergrafo para *k-way mincut*.

O particionamento do hipergrafo é considerado um problema NP-difícil, mas várias heurísticas eficientes foram desenvolvidas para sua solução com complexidade proporcional ao número de hiperarestas $O(|E|)$. Em [23], três funções consensos são usadas baseadas em modelos hipergrafos: CSPA (*Cluster-based Similarity Partitioning Algorithm*), HGPA (*Hiper-Graph-Partitioning Algorithm*) e MCLA (*Meta-Clustering Algorithm*). A saída deste método é uma partição consenso, encontrada pelas funções consensos, com a informação mútua normalizada pela média máxima, como definido pelos autores.

Especificamente, neste trabalho, foram comparados os resultados obtidos pelo uso de dois algoritmos de agrupamento, representativos de paradigmas diferentes de agrupamento, *k*-médias e o algoritmo *Expectation-Maximization* (EM), vistos em [16], [15], respectivamente, com os resultados dos métodos de comitê, apresentados em [5], [23]. Estes algoritmos de agrupamento têm sido amplamente utilizados na literatura para dados de expressão gênica [21], [22].

III. METODOLOGIA DE AVALIAÇÃO

O objetivo da avaliação usada no trabalho é mostrar a eficiência dos métodos de agrupamento investigados através da recuperação de grupos conhecidos (classes) no conjunto de dados de expressão gênica. Consideramos três conjuntos de dados para os quais as classes (fenótipo) já são conhecidas.

Assim, como em outros trabalhos na literatura, cada um desses conjuntos de dados constitui o *padrão ouro* com o qual avaliamos os resultados dos agrupamentos [16], [20].

Baseado em [20], nos referimos à partição *padrão ouro* como classe, e reservamos a palavra *grupo* para partição retornada pelo algoritmo de agrupamento.

A. Índices Externos

A qualidade de agrupamento pode ser medida pelo grau de concordância entre duas partições (U e V), em que a partição U é o resultado do método de agrupamento e a partição V é formada por uma informação *a priori* independente da partição U , como um rótulo de classe [9], [6]. Há um número de índices externos definidos na literatura como Hubbert, Jacard, Rand e Rand corrigido (ou Rand ajustado) [16] que podem ser usados para essa medida.

Uma característica da maioria desses índices é que eles podem ser sensíveis ao número de classes na partição ou a distribuição de elementos nos grupos. Por exemplo, alguns índices têm a tendência de apresentar valores elevados para partições com muitas classes (Hubbert e Rand), outros para partições com um pequeno número de classes (Jaccard) [6].

O índice Rand corrigido não possui essas características indesejáveis [17]. Assim, o índice Rand corrigido (do inglês, *corrected Rand*), é o índice externo empregado na metodologia de avaliação usada neste trabalho. O índice externo Rand corrigido pode assumir valores entre -1 e 1, com 1 indicando uma concordância perfeita entre as partições, e valores próximos a 0 ou negativos correspondem a concordâncias encontradas ao acaso.

Formalmente, seja $U = \{u_1, \dots, u_r, \dots, u_R\}$ uma partição resultante do algoritmo de agrupamento, e $V = \{v_1, \dots, v_c, \dots, v_C\}$ a partição formada por uma informação *priori* obtida independente da partição U (*padrão ouro*). O Rand corrigido é definido por:

$$\frac{\sum_i^R \sum_j^C \binom{n_{ij}}{2} - \binom{n}{2}^{-1} \sum_i^R \binom{n_i}{2} \sum_j^C \binom{n_j}{2}}{\frac{1}{2} [\sum_i^R \binom{n_i}{2} + \sum_j^C \binom{n_j}{2}] \binom{n}{2}^{-1} \sum_i^R \binom{n_i}{2} \sum_j^C \binom{n_j}{2}},$$

em que (1) n_{ij} representa o número de objetos nos grupos u_i e v_j ; (2) n_i indica o número de objetos no grupo u_i ; (3) n_j indica o número de objetos no grupo v_j ; (4) n é o número total de objetos; e (5) $\binom{a}{b}$ é o coeficiente binomial $\frac{a!}{b!(a-b)!}$.

B. Validação Cruzada

A comparação de dois métodos de aprendizado supervisionado, em geral, é feita pela análise da significância estatística da diferença entre a taxa média do erro de classificação dos métodos avaliados em conjuntos de testes independentes. Entretanto, o número de base de dados avaliadas é freqüentemente limitado. Um meio de superar esse problema é dividir a base de dados em dados de treinamento e de testes pelo uso do procedimento de *k-fold cross-validation* (validação cruzada) [19].

No caso de aprendizado não-supervisionado (ou análise de agrupamentos), quando houver uma classificação *a priori* dos dados avaliados, a comparação entre dois métodos pode também ser feita detectando a significância estatística da diferença entre os valores médios de um determinado índice externo. Mas novamente, a quantidade de dados disponível para treinamento é limitada. Em [2], um método para contornar esse problema foi apresentado. Tal método, que será usado nesse trabalho, é uma adaptação do procedimento *k-fold cross-validation* para métodos não-supervisionados.

A base de dados no procedimento em [2] também é dividida em *k-folds*. Em cada interação do procedimento, um *fold* é usado como conjunto de teste, e os *folds* restantes como conjunto de treinamento. O conjunto de treinamento é apresentado ao método de agrupamento, produzindo uma partição como resultado (partição de treinamento). Em seguida, a técnica de centróide mais próximo é usada para

construir um classificador a partir da partição de treinamento. A técnica de centróide calcula a proximidade entre os elementos nos dados de teste e os centróides de cada grupo na partição de treinamento (a proximidade deve ser medida com o mesmo índice de proximidade usado pelo método de agrupamento avaliado).

Uma nova partição (partição teste) é então obtida atribuindo-se cada objeto no conjunto de teste para o grupo com o centróide mais próximo. Depois, a partição de teste é comparada com a partição *a priori* (padrão ouro) usando um índice externo (essa partição *a priori* contém apenas os objetos da partição teste). No final do procedimento, obtemos uma amostra de tamanho k para o valor do índice externo.

IV. BASE DE DADOS

Neste trabalho são utilizadas três bases de dados de expressão gênica que são: Leukemia [11], Novartis (amostras de vários tecidos cancerígenos) [20] e St. Jude Leukemia, que são apresentadas em [25].

- *Leukemia*[11]: os exemplos foram obtidos da medula óssea de pacientes de leucemia aguda. Os exemplos são formados por subgrupos de leucemia: 11 leucemia mielóide aguda (AML, do inglês *Acute Myeloid Leukemia*); 8 leucemia linfoblástica aguda tipo T (ALL-T, do inglês *Acute Lymphoblastic Leukemia T-lineage*); e 19 leucemia aguda linfoblástica aguda tipo B (ALL-B, do inglês *Acute Lymphoblastic Leukemia B-lineage*).
- *Novartis*[20]: exemplos de tecidos de quatro tipos de câncer foram coletados: 26 de próstata, 16 de mama, 28 de pulmão e 23 de cólo.
- *St. Jude Leukemia*[25]: amostras de diagnósticos de medula óssea de pacientes pediátricos com leucemia aguda, representando classes de prognósticos de seis sub-tipos de leucemia: 43 *T-lineage ALL*; 27 *E2A-PBX1*; 15 *BCR-ABL*; 79 *TEL-AML1*; 20 *MLL rearrangements*; e 64 "*hyperdiploid > 50 chromosomes*".

No trabalho apresentado em [20], uma razão simples sinal-ruído é usada para ranquear os genes de cada base de dados. O conjunto final de genes escolhidos para formarem os atributos de um padrão é obtido pela escolha dos genes mais *up-regulated* para cada classe, em que esse número depende da base de dados.

TABELA I
DESCRIÇÃO DA BASE DE DADOS

Base de Dados	Classes	Exemplos	Características
Leukemia	3	38	999
Novartis	4	103	1000
St. Jude Leukemia	6	248	985

Adicionalmente para as etapas de pré-processamento aplicadas por [20], escalamos os níveis de expressão para o intervalo entre 0 e 1 (em termos de técnicas de agrupamento, [18] mostrou que este tipo de transformação é mais eficiente do que transformações *z-score*).

Os experimentos foram executados com as três base de dados (Leukemia, Novartis, e St. Jude Leukemia), com os algoritmos individuais: o k -médias e o EM (ambos implementados com a distância Euclidiana). Inicialmente, cinco replicações do *2-fold cross-validation* foram aplicadas a cada base de dados, o que resultou em dez amostras para cada conjunto de dados.

Em termos de parâmetros de configuração, o número de grupos para o k -médias e para o EM foi definido como o número de classes da base de dados que está sendo considerada. Além disso, como esses algoritmos são dependentes da escolha das condições iniciais (por exemplo, centros e médias iniciais), repetimos cada execução dez vezes, cada uma com uma inicialização diferente e aleatória. Por exemplo, para cada uma das dez amostras da validação-cruzada de cada base de dados, os experimentos com k -médias (ou o algoritmo EM) geraram dez partições.

Baseado nisso, os comitês foram formados como a seguir. Foram criados dez comitês, onde cada comitê foi gerado a partir da combinação de dez partições do k -médias (ou EM) para um dado *fold* da validação-cruzada (dez amostras no total - *2x5 cross-validation*). Ou seja, formamos comitês homogêneos, onde cada partição usada para formar a partição consenso é resultado de várias execuções de um mesmo algoritmo de agrupamento.

Para a realização dos experimentos foram utilizadas duas técnicas de comitê vistas em [23], [5], a primeira é um módulo *ClusterEnsemble* do Pacote *Cluster* (uma coleção de funções do Matlab para análise de agrupamento), que pode ser encontrada no endereço: <http://www.lans.ece.utexas.edu/~strehl/soft.html>. O outro método de comitê é uma função *Voting*, implementada em R, apresentada em [5].

Para todos os experimentos, os valores médios do índice Rand corrigido (CR) para as partições de teste foram calculados. Em seguida, a média do CR obtida com os métodos individuais (bases) foram comparadas, duas a duas, com aquelas obtidas com os comitês.

VI. RESULTADOS

As Tabelas II, III e IV apresentam a média do Rand corrigido (e o desvio padrão) para os conjuntos de testes das bases de dados de Leukemia, Novartis e St. Jude Leukemia, respectivamente. As linhas das tabelas apresentam os resultados do algoritmo k -médias, do método *Voting* construído com o k -médias, do método de *Comitê* do k -médias, do algoritmo EM, do método de *Voting* gerado pelo algoritmo EM e o *Comitê* construído com o algoritmo EM.

Todas as comparações foram feitas por meio de um Teste t pareado, como descrito em [3], com o nível de significância $\alpha=0.05$ e o grau de liberdade da amostra igual a 9.

De acordo com a Tabela II, os comitês gerados para a base de dados de Leukemia das partições do k -médias (*Voting* e *Comitê* k -médias) obteve médias de acurácia maiores (0.6568 e 0.6182, respectivamente) do que a média alcançada com o método individual (0.3637). No caso

do *Voting* do k -médias foram encontradas diferenças estatísticas significantes, o que não ocorreu para o *Comitê* do mesmo. Da mesma forma, os comitês homogêneos gerados pelas partições do algoritmo EM com o método *Voting* apresentaram uma média de exatidão maior (0.5919) do que a média obtida com o método individual (0.3427), enquanto que não existem evidências estatísticas para afirmar com o *Comitê* do EM.

TABELA II
RESULTADOS DA BASE DE DADOS DE LEUKEMIA

Métodos	Média CR	Desvio Padrão
k -médias	0,3637	0,2783
<i>Voting</i> k -médias	0,6568	0,2216
<i>Comitê</i> k -médias	0,6182	0,3446
EM	0,3427	0,2779
<i>Voting</i> EM	0,5919	0,2216
<i>Comitê</i> EM	0,4512	0,25

TABELA III
RESULTADOS DA BASE DE DADOS DE NOVARTIS

Métodos	Média CR	Desvio Padrão
k -médias	0,7485	0,1970
<i>Voting</i> k -médias	0,8718	0,1582
<i>Comitê</i> k -médias	0,9614	0,0266
EM	0,7764	0,1833
<i>Voting</i> EM	0,9573	0,0413
<i>Comitê</i> EM	0,9614	0,0266

TABELA IV
RESULTADOS DA BASE DE DADOS DE ST. JUDE LEUKEMIA

Métodos	Média CR	Desvio Padrão
k -médias	0,8244	0,1049
<i>Voting</i> k -médias	0,7667	0,1289
<i>Comitê</i> k -médias	0,7759	0,0560
EM	0,7903	0,1239
<i>Voting</i> EM	0,7470	0,0998
<i>Comitê</i> EM	0,8676	0,0885

Para o conjunto de dados Novartis, como apresentado na Tabela III, os comitês gerados das partições k -médias também obtiveram médias de exatidão maiores (0.8718 e 0.9614, respectivamente) do que a média conseguida com o método individual (0.7485). Ao contrário dos resultados da base de dados vista anteriormente, a diferença entre a média de acurácia para o *Comitê* do k -médias e o método individual foi estatisticamente significativa. Isso não aconteceu para o caso do *Voting* do k -médias. Para os comitês homogêneos resultantes das partições do algoritmo EM, ambos os métodos (*Voting* e *Comitê*) apresentaram médias de exatidão maiores (0.9573 e 0.9614, respectivamente) quando comparadas com a média atingida pelo método individual.

Diferentemente das outras base de dados, no caso da base de St. Jude Leukemia, não foram encontradas evidências estatísticas para afirmar que as médias de acurácia dos comitês homogêneos, *Voting* e *Comitê* (tanto

para o k -médias quanto para o algoritmo EM), são significativamente diferentes daquelas obtidas com os métodos individuais.

VII. CONCLUSÃO

No caso da base de dados de Leukemia, o uso de métodos de comitê não implica em uma diminuição grande do desvio padrão da média de acurácia, se comparada com os métodos individuais. Uma das razões para isto é o fato de que o conjunto de dados é composto por poucos exemplos (38 exemplos de três classes diferentes), o que torna o aprendizado difícil na maioria dos problemas.

No entanto, para uma base de dados grande como a base Novartis, o uso dos métodos de comitês, na maioria dos casos, consegue uma diminuição no desvio padrão da média de exatidão com respeito aos métodos individuais. Deste ponto de vista, as técnicas de comitê têm melhorado com sucesso não somente a acurácia, mas também a estabilidade dos algoritmos de agrupamento individuais.

Uma possível explicação para os resultados obtidos pelos métodos de comitê para a base de St. Jude Leukemia seria a seguinte. Como indicado em [8], quando esta base de dados foi aplicada ao k -médias e ao SOTA com o número de grupos igual ao número de classes (seis), as partições geradas frequentemente não separam os padrões das classes BCR e *hyperdip*.

Por exemplo, as partições consistentemente apresentaram um pequeno grupo contendo poucos padrões de *hyperdip* e BCR agrupados juntos com padrões de outras três classes, e um grupo grande com a maioria dos padrões *hyperdip* e BCR. Neste caso, a maioria das partições apresentariam o mesmo tipo de erro. Consequentemente, sem informações extras, esse tipo de erro também seria apresentado nos métodos de comitês.

Uma solução para este problema seria a construção das partições de entrada, por exemplo, não somente pelo uso de diferentes inicializações aleatórias do mesmo algoritmo, mas também pela variação do número de grupos. Por exemplo, os autores em [8] mostraram que os algoritmos de agrupamento foram capazes de separar corretamente as classes *hyperdip* e BCR quando o número de grupos foi ajustado para 15. Outra possibilidade é construir as partições de entrada de diferentes algoritmos de agrupamento (comitês heterogêneos) [24][13].

REFERÊNCIAS

- [1] L. Breiman. Bagging predictors. *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.
- [2] I. G. Costa, F. A. T. de Carvalho, and M. C. P. de Souto. Comparative study on proximity indices for cluster analysis of gene expression time series. *Journal of Intelligent and Fuzzy Systems*, vol. 13, no. 3, pp. 133-142, 2003.
- [3] T. G. Dietterich. Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, vol. 10, no. 7, pp. 1895-1923, 1998.
- [4] T. G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume vol.1857 of *Lecture Notes in Computer Science*, pp. 1-15, 2000.
- [5] E. Dimitriadou, A. Weingessel, and K. Hornik. A cluster ensembles framework. In *Third International Conference on Hybrid Intelligent Systems (HIS)*, pp. 528-534, 2003.
- [6] R. Dubes. How many clusters are best? An experiment. *Pattern Recognition*, vol. 20, no. 6, pp. 645-663, 1987.
- [7] S. Dudoit and J. Fridlyand. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, vol. 19, no. 9, pp. 1090-1099, 2003.
- [8] K. Faceli, A. C. P. L. F. de Carvalho, and M. C. P. de Souto. Evaluation of the contents of partitions obtained with clustering gene expression data. In J. Setubal, editor, *Proc. of Brazilian Symposium on Bioinformatics - Lecture Notes on Bioinformatics*. Springer-Verlag, 2005. Aceito para publicação.
- [9] A. L. N. Fred and A. K. Jain. Data clustering using evidence accumulation. In *16th International Conference on Pattern Recognition*, pp. 276-280, 2002.
- [10] D. S. Frossyniotis, M. Pertselakis, and A. Stafylopatis. A multi-clustering fusion algorithm. In *Second Hellenic Conference on AI, SETN 2002*, *Lecture Notes in Computer Science*, pp. 225-236, 2002.
- [11] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, vol. 5439, no. 286, pp. 531-537, 1999.
- [12] D. Greene, A. Tsymbal, N. Bolshakova, and P. Cunningham. Ensemble clustering in medical diagnostics. In *17th IEEE Symposium on Computer-Based Medical Systems (CBMS2004)*, pp. 576-581, 2004.
- [13] S. T. Hadjitodorov, L. I. Kuncheva, and L. P. Todorova. Moderate diversity for better cluster ensembles. *Information Fusion*, 2005. Aceito para publicação.
- [14] X. Hu and I. Yoo. Cluster ensemble and its applications in gene expression analysis. In *Second Asia-Pacific Bioinformatics Conference (APBC 2004)*, pp. 297-302, 2004.
- [15] I. H. Witten and E. Frank. *Data mining: practical machine learning tools and techniques with Java implementation*. Morgan Kaufman Publishers, USA, 2000.
- [16] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice Hall, 1988.
- [17] G. W. Milligan and M. C. Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, vol. 21, pp. 441-458, 1986.
- [18] G. W. Milligan and M. C. Cooper. A study of standardization of variables in cluster analysis. *Journal of Classification*, vol. 5, pp. 181-204, 1988.
- [19] T. Mitchell. *Machine Learning*. McGraw Hill, New York, 1997.
- [20] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, vol. 52, pp. 91-118, 2003.
- [21] J. Quackenbush. Computational analysis of cDNA microarray data. *Nature Reviews*, vol. 6, no. 2, pp. 418-428, 2001.
- [22] D. Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics*, vol. 32, pp. 502-508, 2002.
- [23] A. Strehl and J. Ghosh. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research (JMLR)*, vol. 3, pp. 583-617, 2002.
- [24] A. P. Topchy, A. K. Jain, and W. F. Punch. Combining multiple weak clusterings. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003)*, pp. 331-338, 2003.
- [25] E. J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C. H. Pui, W. E. Evans, C. Naeye, L. Wong, and J. R. Downing. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, vol. 1, no. 2, pp. 133-143, 2002.
- [26] Y. Zeng, J. Tang, J. Garcia-Frias, and G. R. Gao. An adaptive meta-clustering approach: Combining the information from different clustering results. In *1st IEEE Computer Society Bioinformatics Conference (CSB 2002)*, pp. 276-287, 2002.