# Agrupamento Semi-supervisionado e Não-supervisionado para Análise de Dados de Expressão Gênica

Fabiana Mari Assao[1], Heloisa de Arruda Camargo[1]

[1] Programa de Pós-Graduação em Ciência da Computação
Universidade Federal de São Carlos
Rod. Washington Luiz, km. 235, São Carlos – SP - Brasil
fabiana.assao@gmail.com, heloisa@dc.ufscar.br

**Resumo.** Neste trabalho são investigados algoritmos de agrupamento semi-supervisionado e não supervisionado aplicados a dados de expressão gênica. O intuito é realizar uma investigação das vantagens e desvantagens da utilização destes métodos de agrupamento e, a partir disso, prover subsídios para obtenção de resultados significativos para a área de Biologia. Foram implementados e testados algoritmos de agrupamento com diferentes características, com o objetivo de verificar evidências de eventuais ganhos obtidos com a rotulação parcial dos genes com relação a técnicas não-supervisionadas. Os experimentos realizados consideraram conjuntos de dados do domínio de expressão gênica. Os resultados obtidos foram avaliados com medidas de validação usualmente aplicadas em contextos semelhantes. As análises desenvolvidas reforçam o importante papel das técnicas de agrupamento na analise de dados biológicos, que visam auxiliar a compreensão das estruturas e funções dos genes.

**Palavras Chave:** Agrupamento, Agrupamento Semi-supervisionado, Expressão Gênica, Aprendizado de Máquina, Bioinformática.

## 1 Introdução

O surgimento da tecnologia de microarray de DNA, que tem sido utilizada como uma importante metodologia na biologia molecular experimental, tornou possível a obtenção de grande volume de dados valiosos relativos ao perfil de expressão dos genes [1]. A identificação de estruturas presentes nesses dados é um importante mecanismo para melhorar a compreensão da genômica funcional. Um experimento de microarray obtém dados de expressão de genes sob condições que podem ser instantes de tempo durante um processo biológico ou diferentes amostras de tecidos. Entre as técnicas para análise de dados utilizadas neste contexto, destacam-se as técnicas de agrupamento de dados [2], que são o principal representante da categoria de métodos de aprendizado não supervisionado, isto é, métodos que não possuem informação prévia sobre a classe a que pertencem os dados.

Este trabalho focaliza o agrupamento baseado em genes, no qual a suposição fundamental é que genes com a mesma função tendem a ter expressões semelhantes, assim, agrupando genes pelos seus perfis de expressão, obtêm-se grupos de genes

com funções similares. Trabalhos relatados sobre esse tópico incluem tanto abordagens que utilizam métodos de agrupamento conhecidos [4, 5, 6] como o desenvolvimento de novos algoritmos especificamente projetados para tratar as questões de análise de dados de expressão gênica [7, 8].

O aprendizado semi-supervisionado [9, 10] utiliza dados rotulados e não rotulados durante o processo de treinamento para melhorar o resultado obtido.

O agrupamento semi-supervisionado aplicado à análise de dados de expressão gênica tira proveito da crescente informação biológica disponível sobre os genes, especialmente quanto às suas funções, o que pode ser obtido a partir das anotações de genes frequentemente divulgadas em bases de dados públicas disponíveis na web [14, 15]. O foco deste trabalho é no método semi-supervisionado [11, 12, 13] que visa melhorar o desempenho de algoritmos de agrupamento não-supervisionado com o uso de informação previamente disponível sobre o domínio do conhecimento considerado, que pode vir na forma de dados rotulados ou de restrições entre pares.

Muitos algoritmos de agrupamento não supervisionado e semi-supervisionado têm sido propostos para analisar dados de expressão gênica mas, a exemplo do que ocorre em todos os domínios quanto se aplicam esses algoritmos, pouca orientação está disponível para ajudar a escolher entre eles, bem como entre as possíveis medidas de similaridade utilizadas nesses algoritmos, as quais podem influenciar os resultados.

O objetivo deste trabalho é investigar o desempenho de diferentes algoritmos de agrupamento semi-supervisionado e não supervisionado aplicados a dados de expressão gênica, especificamente para a identificação e descoberta de novas funções de genes. Através desta análise, o propósito é contribuir para a criação de mecanismos que propiciem a geração de resultados significativos para a área de Biologia.

A implementação e testes dos algoritmos levaram em consideração os diferentes níveis de expressão gênica sob diferentes condições.

Na próxima seção são introduzidos os algoritmos de agrupamento semi-supervisionado. Na seção seguinte, os experimentos e resultados obtidos são apresentados e, em seguida, as conclusões e trabalhos futuros são discutidos.

## 2 Agrupamento Semi-supervisionado

Diversos algoritmos com o objetivo de melhorar o agrupamento de dados explorando algum tipo de supervisão foram propostos nos últimos anos. A informação disponível para rotulação dos dados tem sido utilizada em duas abordagens diferentes, chamadas de abordagem baseada em restrições e abordagem baseada em métrica.

Nas abordagens baseadas em restrições o próprio algoritmo de agrupamento é modificado tal que a informação disponível fornecida pelo usuário é usada para guiar o algoritmo a um particionamento dos dados mais apropriado [16, 13, 11].

O conhecimento disponível que permite a incorporação de supervisão parcial nessa classe de métodos pode surgir tanto na forma de restrições entre pares como na forma de rótulos para um subconjunto dos dados. As restrições entre pares podem ser da forma *must-link*, indicando que um par de dados deve pertencer ao mesmo *cluster* ou *cannot-link,* indicando que os exemplos do par devem pertencer a *clusters* distintos.

Nas abordagens baseadas em métricas um algoritmo conhecido que usa métricas de distância é utilizado, mas a métrica é treinada anteriormente para satisfazer os rótulos ou restrições dos dados rotulados [17, 18].

A seguir são apresentados alguns dos principais algoritmos baseados em restrições encontrados na literatura, de interesse para este trabalho.

- Algoritmo COP-K-means [13]- é uma variante do algoritmo não supervisionado *K-means*. O conhecimento prévio é descrito na forma de restrições do tipo *must-link* e *cannot-link*. Os *clusters* encontrados pelo *COP-K-means* devem respeitar todas as relações *must-link* e *cannot-link* impostas nos exemplos rotulados. Durante a construção dos *k clusters*, cada exemplo do conjunto de exemplos não rotulados é associado ao *cluster* mais próximo.

- Algoritmo *PCK-means* [12] - também é uma variante do algoritmo não supervisionado *K-means* e utiliza restrições entre pares dos tipos *must-link e cannot-link.* A medida de similaridade utilizada nesse algoritmo é composta pela medida de distância convencional, como por exemplo, a distância Euclidiana, entre dois exemplos, adicionada de dois fatores que avaliam o custo de violação das restrições conhecidas. O custo de violação de uma restrição do tipo *must-link* é dada por $w* \mathbf{1}(l_i \neq l_j)$ e o custo de violação de uma restrição do tipo *cannot-link* é dada por $w* \mathbf{1}(l_i = l_j)$ onde $w$ é o peso associado a violação de uma restrição e $\mathbf{1}$ é a função indicador, com $\mathbf{1}(true) = 1$ e $\mathbf{1}(false) = 0$ e $l_i$ denota o cluster a que o exemplo $i$ foi atribuído.

- Algortmo SEEDED-K-means - Proposto em [11], é um algoritmo variante do *K-means*, que utiliza exemplos inicialmente rotulados para calcular os centróides iniciais dos *clusters*, isto é, as sementes, ao invés de escolhê-los aleatoriamente. O algoritmo exige que para cada *cluster* seja atribuído, no mínimo, uma semente. A partição definida pelas sementes é usada apenas para inicialização e as sementes não são usadas nos passos seguintes do algoritmo.

- Algoritmo CONSTRAINED-K-means [11] - é uma melhoria do algoritmo *SEEDED-K-means.* A diferença está nos passos seguintes à inicialização dos centróides, nos quais os exemplos que fazem parte do conjunto das sementes, e que foram inicialmente associados a um dado *cluster* pelo usuário, não poderão ser associados a um outro *cluster*. Assim, apenas os exemplos não selecionados como sementes serão reagrupados, diferentemente do *SEEDED-K-means* em que as sementes podem vir a pertencer a *clusters* diferentes daqueles inicialmente associados.

- Algoritmo *Huang & Pan* [14] - considera funções conhecidas dos genes, explorando esse conhecimento pela incorporação das funções conhecidas em uma nova métrica de distância, que reduz a distância baseada na expressão entre dois genes até zero, apenas quando os dois genes compartilham da mesma função. Esse método é baseado no método *k-medoids* [23], que utiliza, como centro dos clusters, a mediana dos dados atribuídos a cada cluster ao invés da média, como no *k-means*.

Nessa proposta, é assumido que são conhecidas F categorias de genes e que cada gene pode ser atribuído a pelo menos um e possivelmente mais de um dos F+1 grupos $G_0, G_1, ..., G_F$, sendo $G_0$ formado por genes com funções desconhecidas e $G_1, ..., G_F$ formados por genes que tem a mesma função.

Há dois passos básicos neste método de agrupamento. No primeiro passo, é aplicado o *k-medoids* para os genes em $G_1, ..., G_F$ usando a nova matriz de distância $D^* = (d_{ij})$, obtendo *clusters* para os genes com funções conhecidas. O número de

cluster, $k_0$, é fornecido. No segundo passo, é aplicado o *k-medoids* modificado para a matriz de distância $D$ tal que os genes em $G_0$ podem ser associados a um dos $k_0$ *clusters* obtidos anteriormente ou para um dos $k_1$ novos *clusters*, enquanto os medoides e as atribuições dos genes em $G_1$, ..., $G_F$ aos clusters feitas anteriormente permanecem fixos. Os genes com funções desconhecidas podem assim ser agrupados em clusters novos, o que permite a descoberta de estruturas desconhecidas correspondentes a novas categorias de funções.

## 3 Métodos Estudados e Avaliação Experimental

Para os experimentos realizados, foram utilizados dados de expressão gênica que são variações dos dados de Saccharomyces cerevisiae apresentados em [20], como explicado a seguir, que registra dados de expressão durante a fermentação de *Saccharomyces cerevisiae*. Inicialmente os genes com dados faltantes foram eliminados e, em seguida, foi utilizada a ferramenta FunCat disponível em http://mips.gsf.de/proj/funcatDB/search_main_frame.html, para classificar os genes em suas respectivas funções biológicas. Foram obtidas 17 funções distintas para o conjunto de dados, sendo que apenas 10 delas foram utilizadas, por serem as que possuíam nove ou mais genes: *metabolism, energy, cell cycle and DNA processing, transcription, protein synthesis, protein fate, transport, defense, biogenesis* e *cell differentiation.* Este conjunto deu origem a dois conjuntos distintos, que serão chamados de *yeast1* e *yeast2*. O primeiro conjunto de dados (yeast1) contém os genes que possuem somente uma função associada, ou seja, pertencem a apenas uma classe. O segundo conjunto de dados (yeast2) é formado pelo conjunto de dados *yeast1* adicionado de 20 genes aleatórios que possuem mais de uma função e, portanto, pertencem a mais de uma classe. Esse conjunto foi usado em experimentos com algoritmos semi-supervisionados que aceitam dados com essa característica.

Os outros dois conjuntos de dados de levedura *Saccharomyces cerevisiae* foram obtidos de um conjunto de dados utilizado em [14]. As funções gênicas foram obtidas da base de dados MIPS [21]. Os conjuntos de dados usados neste trabalho foram construídos seguindo as mesmas considerações de Huang e Pan no seu trabalho [14], que usaram apenas três funções gênicas: *mitotic cell cycle and cell cycle control* (classe 1), *mitochondrion* (classe 2) e *c-compound and carbohydrate utilization* (classe3). O primeiro conjunto de dados usado neste trabalho, obtido do trabalho de Huang e Pan, foi chamado de yeast3 e contém somente genes da classe 1 e classe 2. O segundo conjunto de dados, chamado de *yeast4*, contém genes das três classes.

**Tabela 1:** Características dos conjuntos de dados

| Conjunto de Dados | Número de Objetos | Dimensão | Número de Classes |
|---|---|---|---|
| *Yeast1* | 545 | 80 | 10 |
| *Yeast2* | 585 | 80 | 10 |
| *Yeast3* | 630 | 300 | 2 |
| *Yeast4* | 845 | 300 | 3 |

### 3.1 Algoritmos Utilizados

Neste trabalho foram utilizados algoritmos de agrupamento não-supervisionado e semi-supervisionado. O algoritmo de agrupamento não-supervisionado utilizado foi o *K-means*. Os algoritmos semi-supervisionados considerados foram:

- baseados em sementes, como *Seeded K-means* e *Constrained K-means* [11] , que são extensões do K-means e foram propostos como algoritmos gerais de agrupamento, no contexto de aprendizado de máquina;
- baseados em restrições, como *Cop-Kmeans* [13] (extensão do K-means) e *PCKMeans* [12] (extensão do K-medoids) e também propostos como algoritmos gerais de agrupamento, no contexto de aprendizado de máquina;
- proposto especificamente para domínios de expressão gênica, como método de *Huang & Pan* [14].

Estes algoritmos foram utilizados para realizar uma análise do comportamento dos diferentes algoritmos aplicados à conjuntos de dados de domínio de expressão gênica.

### 3.2 Medidas de Avaliação

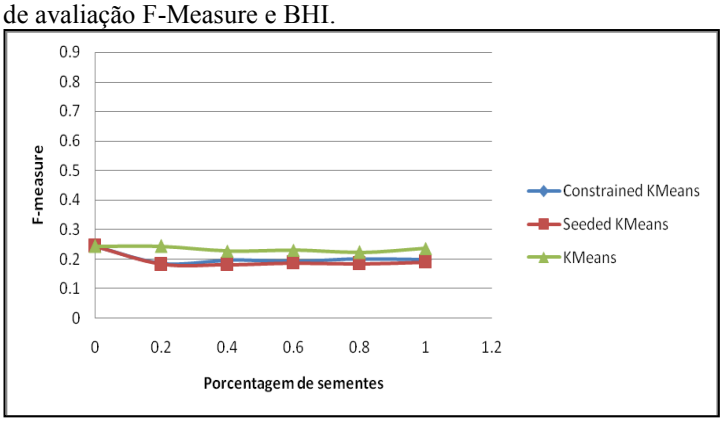
Uma das medidas de validação utilizadas foi o *F-measure* baseado em pares, que é definido como a média harmônica entre a precisão (relação entre o número de pares corretamente atribuídos ao mesmo cluster e o número total de pares atribuídos ao mesmo cluster) e o *recall* (relação entre o número de pares corretamente atribuídos ao mesmo cluster e o número total de pares realmente pertencentes ao mesmo cluster). Assim, é necessário conhecer previamente as classes dos dados agrupados. Um par de pontos é considerado como corretamente atribuído ao mesmo cluster se ambos os pontos do par possuem a mesma classe [12]. A medida F-measure foi selecionada para este trabalho por ser uma das medidas mais utilizadas pela maioria dos trabalhos que propõem algoritmos de agrupamento semi-supervisionado.

Outra medida de validação chamada *Biological Homogeneity Index*, BHI, proposta recentemente no domínio de expressão gênica com o objetivo de avaliar o resultado de algoritmos de agrupamento e sua habilidade de produzir clusters biologicamente significativos [22] também foi empregada para os métodos semi-supervisionados: *Seeded-K-Means, Constrained-K-Means, Huang & Pan e Boratyn*. Essa medida avalia a homogeneidade dos clusters e não foi aplicada aos algoritmos baseados em restrições de pares, por necessitar explicitamente do rótulo do dado para ser calculada.

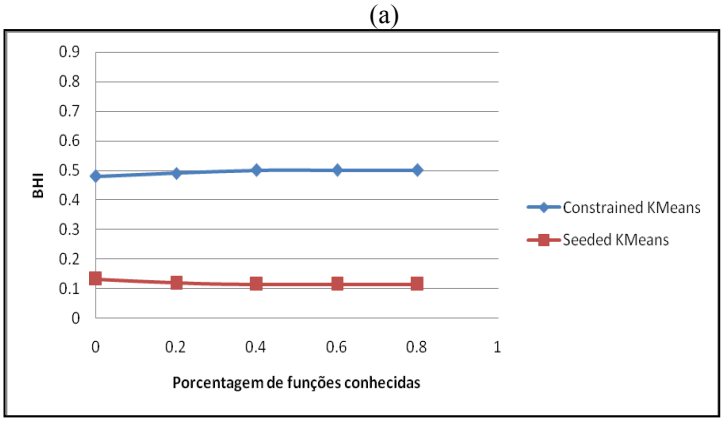
### 3.3 Resultados

A metodologia utilizada nos experimentos, na maioria dos casos, foi a validação cruzada com 5 partições – os conjuntos de dados foram particionados aleatoriamente em 5 partições, em cada uma das 5 execuções, uma das partições foi usada como conjunto de teste e as outras 4 como conjunto de treinamento. Os conjuntos de dados yeast1 e yeast2 foram executados com $k$ (número de cluster) igual a 10, o conjunto de dados yeast3 com $k$=2 e o conjunto de dados yeast4 com $k$=3.

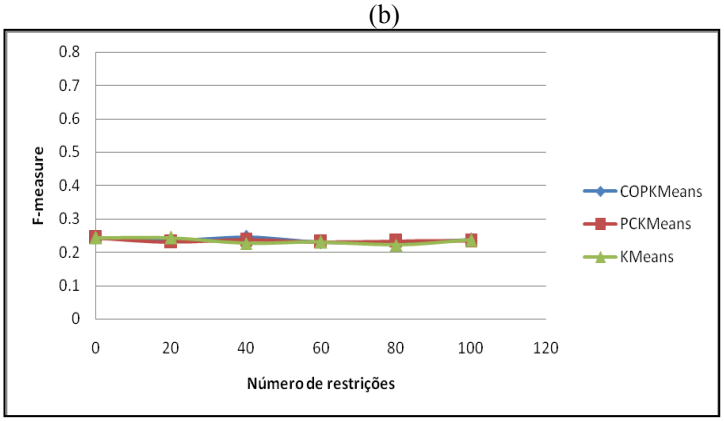
As Figuras de 1 a 4 mostram os resultados obtidos pela aplicação dos algoritmos nos conjuntos de dados yeast1, yeast2, yeast3 e yeast4, respectivamente, para as medidas de avaliação F-Measure e BHI.




(a)

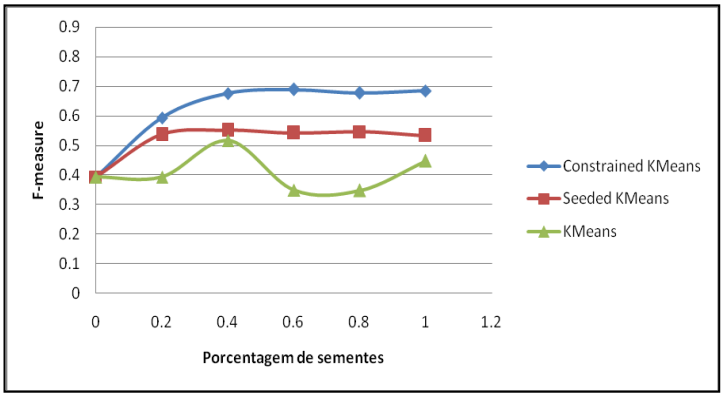



(b)

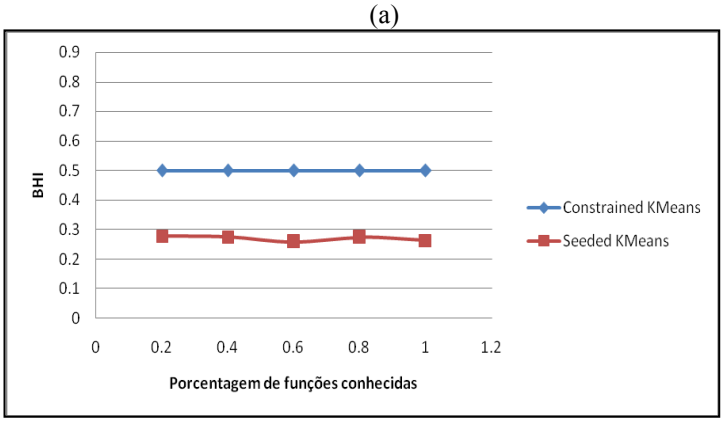



(c)

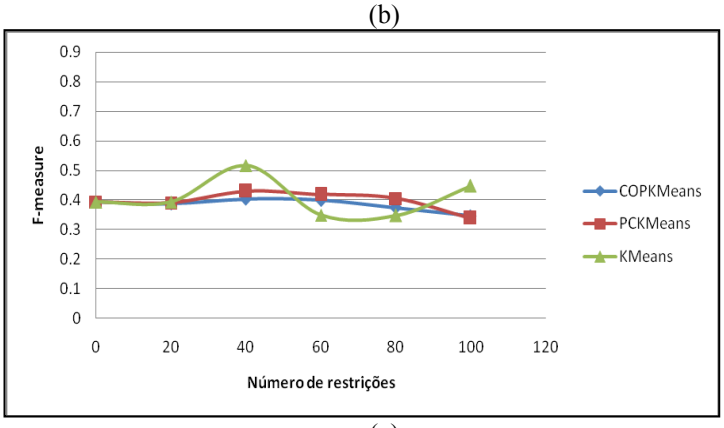
**Fig. 1.** Conjunto Yeast1 – (a) F-measure - Comparação K-Means, Seeded-Kmeans e Constrained-Kmeans; (b) BHI - Comparação Seeded-Kmeans e Constrained-Kmeans; (c) F-measure - Comparação K-Means, Cop-Kmeans e PCKMeans.




(a)

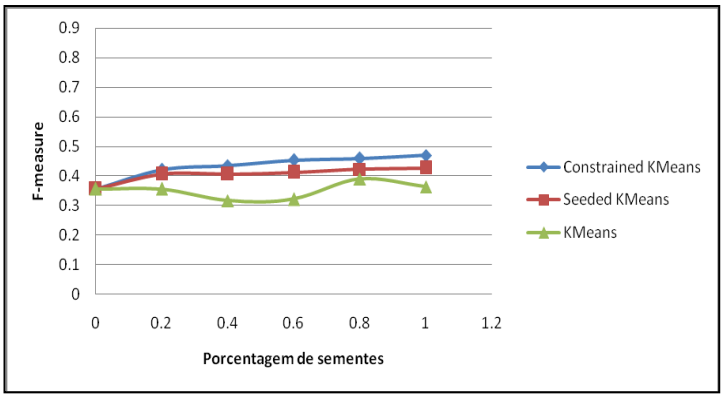



(b)

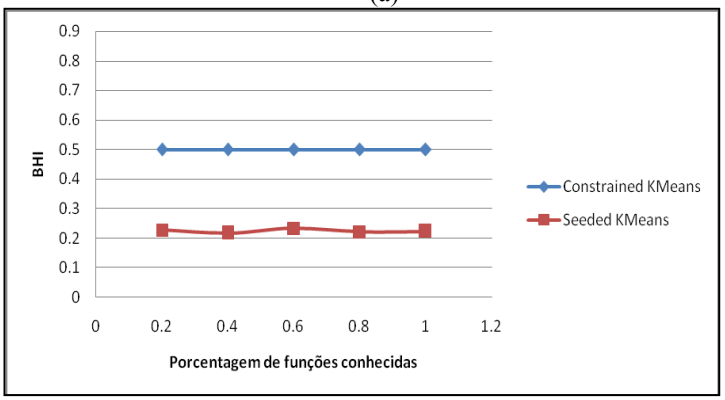



(c)

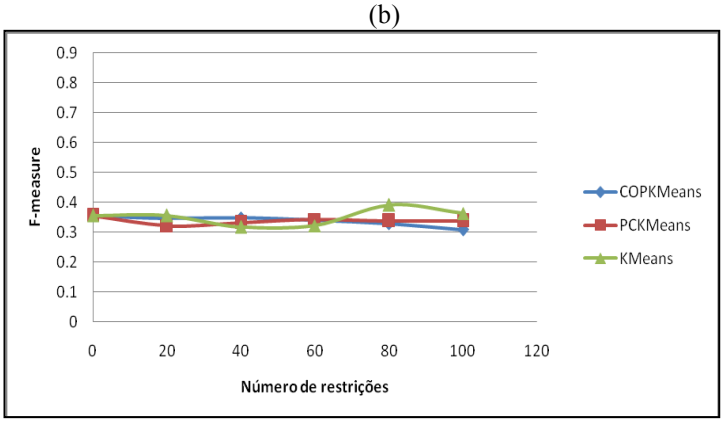
**Fig. 2.** Conjunto Yeast3 – (a) F-measure - Comparação K-Means, Seeded-Kmeans e Constrained-Kmeans; (b) BHI - Comparação Seeded-Kmeans e Constrained-Kmeans; (c) **F-measure** - Comparação K-Means, Cop-Kmeans e PCKMeans.




(a)




(b)




(c)

**Fig. 3.** Conjunto Yeast4 – (a) F-measure - Comparação K-Means, Seeded-Kmeans e Constrained-Kmeans; (b) BHI - Comparação Seeded-Kmeans e Constrained-Kmeans; (c) F-measure - Comparação K-Means, Cop-Kmeans e PCKMeans.

| Conjunto de dados *Yeast1* | | | | Conjunto de dados *Yeast2* | | | |
|---|---|---|---|---|---|---|---|
| | | $r$=1 | $r$=0.8 | | | $r$=1 | $r$=0.8 |
| F-measure | $k_1$=0 | 0.243 | 0.239 | F-measure | $k_1$=0 | 0.242 | 0.230 |
| | $k_1$=1 | 0.241 | 0.236 | | $k_1$=1 | 0.239 | 0.227 |
| BHI | $k_1$=0 | 0.090 | 0.239 | BHI | $k_1$=0 | 0.093 | 0.234 |
| | $k_1$=1 | 0.090 | 0.225 | | $k_1$=1 | 0.094 | 0.227 |
| Conjunto de dados *Yeast3* | | | | Conjunto de dados *Yeast4* | | | |
| | | $r$=1 | $r$=0.8 | | | $r$=1 | $r$=0.8 |
| F-measure | $k_1$=0 | 0.354 | 0.378 | F-measure | $k_1$=0 | 0.308 | 0.313 |
| | $k_1$=1 | 0.337 | 0.364 | | $k_1$=1 | 0.300 | 0.303 |
| BHI | $k_1$=0 | 0.278 | 0.416 | BHI | $k_1$=0 | 0.213 | 0.311 |
| | $k_1$=1 | 0.278 | 0.416 | | $k_1$=1 | 0.213 | 0.311 |

**Fig. 4.** Resultado do método *Huang & Pan* aplicado aos conjuntos de dados Yeast1, Yeast2, Yest3 e Yeast4. $\boldsymbol{k_1}$ é o número de novos clusters e **r** é o fator de redução.

## 4 Conclusões

Neste trabalho foram estudados diversos métodos para agrupamento de dados de expressão gênica.

Os resultados obtidos mostram que os algoritmos semi-supervisionados Seeded-Kmeans e Constrained-Kmeans tiveram, no geral, melhor desempenho quando comparado ao algoritmo não supervisionado K-means (figuras 1a, 2a e 3a). O mesmo acontece quando o K-means é comparado aos algoritmos COP-Kmeans e PCKmeans (figuras 1c, 2c e 3c). Os resultados do método proposto por Huang & Pan mostram (figura 4) que quando se tem informações prévias sobre as funções do conjunto, o processo de agrupamento apresenta melhor desempenho do que quando nenhuma informação prévia é conhecida.

Em resumo, os resultados mostraram que a utilização de conhecimento prévio disponível pode levar a uma melhora no desempenho dos algoritmos de agrupamento semi-supervisionados comparados aos algoritmos de agrupamento convencionais. Assim, o estudo realizado acrescenta novos indicadores ao cenário de análise de dados de expressão gênica quanto ao comportamento dos algoritmos, contribuindo com a relevante questão de prover informações que auxiliem a escolha de algoritmos com essa finalidade.

Como trabalhos futuros, os autores pretendem avaliar o comportamento dos algoritmos com diferentes medidas de similaridades quanto aplicados a outros conjuntos de dados. Um segundo estudo poderia ser feito ao redor da análise de métodos para classificação de amostras, como por exemplo, classificação de pessoas sadias versus pessoas doentes.

## Referências


1. Nguyen, D. V., Arpat, A. B. et al.: DNA Microarray Experiments: Biological and Technological Aspects. Biometrics, Blackwell Synergy 58, 701—717 (2002).
2. Jain, A. K. et al.: Data Clustering: a review. ACM Comp. Surveys 31, 264-323 (!999).
3. Jiang, D. et al.: Cluster Analysis for gene expresson data: a survey. IEEE Trans. On Knowledge and Data Engineering, 16, 1370—1386 (2004).
4. Eisen et al.: Cluster Analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Of Science, 95, 14863—14868 (1998).
5. Golub, T. R. et al.: Molecular Classification of cancer:class discovery and class prediction by gene expression. Science, 286, 531—537.
6. Yeung, K. Y. et al.: Model-based clustering and data transformation for gene expression data. Bioinformatics, 17, 977-987 (2001).
7. Madeira, S. C. & Oliveira, A. L.: Biclustering algorithms for biological data analysis: a survey. IEEE Trans. Comput. Bil. Informatics, 1, 24—45 (2004).
8. Hastie, T. et al.: Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. Genome Biology, 1, 1—21 (2000).
9. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. Proc. of the 11th Annual Conf. on Computational Learning Theory (1998).
10. Joachims, T.: Transductive inference for text classification using support vector machines. In: Proc. Of 16th Intl. Conf. on Machine Learning (1999).
11. Basu, S., Banerjee, A., Mooney, R. J.: Semi-supervised clustering by seeding. In: Proc. Of 19th Intl. Conf. on Machine Learning, 19—28, (2002).
12. Basu, S., Banerjee, A., Mooney, R. J.: Active semi-supervision for pairwise constrained clustering. In: Proc. of the 2004 SIAM Intl. Conf. on Data Mining, (2004).
13. Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S: Constrained K-Means clustering with background knowledge. In: Proc. of 18th Intl. Conf. on Machine Learning, 577—584, (2001).
14. Huang, D. & Pan, W.: Incorporationg biological knowledge into distance-based clustering analysis of microarray gene expression data. Bioinformatics, 22, 1259—1268 (2006).
15. Boratyn, G., Datta, S., Datta, S. Biolobically supervised hierarchical clustering alforithms for gene expression data. In: Proc. 28th IEEE EMBS Annual Intl. Conf. (2006).
16. Demiriz, A. et al.: Semi-supervised clustering using genetic algoritms, Artificial Neural Networks n Engineering, 809—814, (1999).
17. Klein et al.: From instance –level constraints to space-level constrains: Making the most of prior knowledge in data clustering. In: Proc. Of the 19th Intl. Conf. on Data Mining, 307—314, (2002).
18. Bilenko, M., Mooney, R. J. Adaptive duplicate detection using learnable string similarity measures. In: Proc. Of the 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 39—48, (2003).
19. Bilenko, M., Mooney, R. J. Adaptive duplicate detection using learnable string similarity measures. In: Proc. Of the 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 39—48, (2003).
20. Chu, S. et al.: The transcriptional program of sporulation in Dudding yeast, Science, 282, 699—705, (1998).
21. Mewes, H. W.: MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res., 32, D41—D44, (2004).
22. Datta, S., Datta, S.: Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. BMC Bioinformatics, 7, (2006).
23. van der Laan, M. J. et al. (2003) A new partitioning around medoids algorithm. *J. Stat. Comput. Sim.*, 73, 575-584.