Agrupamento Baseado em Análise de Filogenias

1

Abstract. *The discovery of patterns can reveal crucial information. The algorithms for clustering frequently require prior knowledge of the amount of groups, otherwise the parameter must be set and may not reflect the reality of the facts. This paper presents the DAMICORE-2, a correlation detection algorithm of mixed data types, that does not require number of groups. Using as metric the normalized compression distance, phylogenetic tree reconstruction techniques and community detection structure in complex networks. In tests with known bases of literature have been achieved high hit rates.*

Resumo. *A descoberta de padrões pode revelar informações de fundamental importância. Os algoritmos para agrupamento frequentemente exigem um conhecimento prévio da quantidade de grupos existente, caso contrário o parâmetro deve ser ajustado podendo, no final, não refletir a realidade dos fatos. Neste artigo apresentamos o DAMICORE-2, um algoritmo de detecção de correlação entre dados de tipos mistos, que dispensa a informação do número de grupos. Ele utiliza como métrica a distância por compressão normalizada, técnicas de reconstrução de árvores filogenéticas e detecção de estrutura de comunidade em redes complexas. Nos testes realizados com bases conhecidas da literatura foram alcançadas altas taxas de acertos.*

1. Introdução

Com o crescimento de bases de dados em diversos setores da economia e da sociedade, a armazenagem de uma diversidade de tipos de dados em uma mesma base (mesmo contexto) tornou-se relativamente comum. A organização desses dados para se inferir correlações em geral não tem sido um trabalho trivial. A dificuldade em se automatizar esse tipo de processamento ocorre, principalmente, porque a detecção de correlação entre dados de tipos diferentes (por exemplo: números e textos) é relativamente complexa.

O método DAMICORE [Delbem et al. 2010] (do inglês *DAta MIning of COde REpositories*) [Sanches et al. 2011] tem mostrado ser capaz de encontrar correlações em dados de diversos tipos que podem ser também mistos (registros com diferentes tipos de dados) e envolver instâncias relativamente grandes. O DAMICORE é um método que integra um conjunto de técnicas de várias áreas do conhecimento (Teoria da Computação, Bioinformática e Física) de forma a extrair informações relevantes através de uma matriz de distâncias calculada por meio de uma métrica universal, como a NCD (do inglês, *Normalized Compression Distance*).

A NCD é uma métrica que pode encontrar relações entre dados, determinando semelhança entre as variáveis com base em seus tamanhos compactados e descompactados. Essa abordagem, desenvolvida na Teoria da Informação [Cilibrasi e Vitányi 2005], não requer qualquer conhecimento específico do domínio de aplicação. De acordo com

[Cilibrasi e Vitányi 2005], a NCD é uma métrica universal e robusta que tem sido aplicada com sucesso em áreas como a genética, literatura, música e astronomia.

Os autores de [Sanches et al. 2011] mostraram que a NCD pode encontrar padrões em códigos-fonte de programas, quando combinada com técnicas como NJ (do inglês, *Neighbor Joining*) e FA (do inglês, *Fast Newman Algorithm*). A abordagem chamada DAMICORE [Delbem et al. 2010], que usa a NCD, mostrou que é possível trabalhar com diferentes tipos de representações de códigos. A NCD integrada ao DAMICORE gera Matrizes de Distância para um conjunto de códigos, possibilitando verificar níveis de semelhança entre eles e propriedades em comum desses códigos que podem ser exploradas, por exemplo, em projeto de *Hardware* de alto desempenho utilizando FPGAs (do inglês, Field Programmable Gate Array) [Sadrozinski e Wu 2011, Silva et al. 2014].

Neste trabalho apresenta-se uma nova versão do DAMICORE [Delbem et al. 2010], chamada DAMICORE-2 (descrito na Seção 2), investigando como este método possibilita a determinação de correlação entre dados, realizando o processo de *clustering* [Jain et al. 1999] através de redes obtidas por meio da reconstrução de filogenias.

Uma filogenia é uma representação, em forma de árvore, do relacionamento de espécies com a mesma origem. O termo árvore filogenética tem sido usado para ambas filogenias obtidas de dados morfológico e de sequências genéticas. Neste trabalho, filogenias de vários tipos de dados são reconstruídas com o objetivo de determinar grupos de objetos (filos) correlacionando esses dados.

Usualmente, árvores filogenéticas (um grafo acíclico conectado) são árvores binárias onde suas folhas representam espécies, assim, folhas são rotuladas com o nome da espécie correspondente, enquanto os nós internos indicam ancestrais hipotéticos ou espécies extintas.

Os testes foram realizados em contextos para os quais já se têm dados preliminares e que foram usados como parâmetro de comparação, são eles os conjuntos de dados: *iris*, *glass* e *seeds* [Bache e Lichman 2013].

2. Descrição do DAMICORE-2

O DAMICORE-2 necessita que os dados sejam dispostos na estrutura de uma rede, na qual os grupos serão identificados. Para tanto, uma Matriz de Distância entre os dados já pré-processados do conjunto de dados é calculada através da NCD. A matriz é por sua vez utilizada para a reconstrução de árvores filogenéticas, naturalmente hierárquicas, que proporcionarão uma relação mais precisa entre os elementos na rede formada ao final, fortalecendo as relações entre os elementos de um mesmo grupo. Cada árvore reconstruída dá origem a uma Rede Completa, as quais são combinadas pelo FA dando origem a Redes de Folhas. Estas por sua vez são convertidas em uma Rede Equivalente sobre a qual é aplicada uma técnica de detecção de comunidades (da área de Redes Complexas [Donetti e Muñoz 2004, Duch e Arenas 2005]), as quais representam os grupos formados pelo DAMICORE-2 ao final do processo.

O DAMICORE-2 tem oito passos principais e a Figura 1 resume o seu funcionamento como um todo.

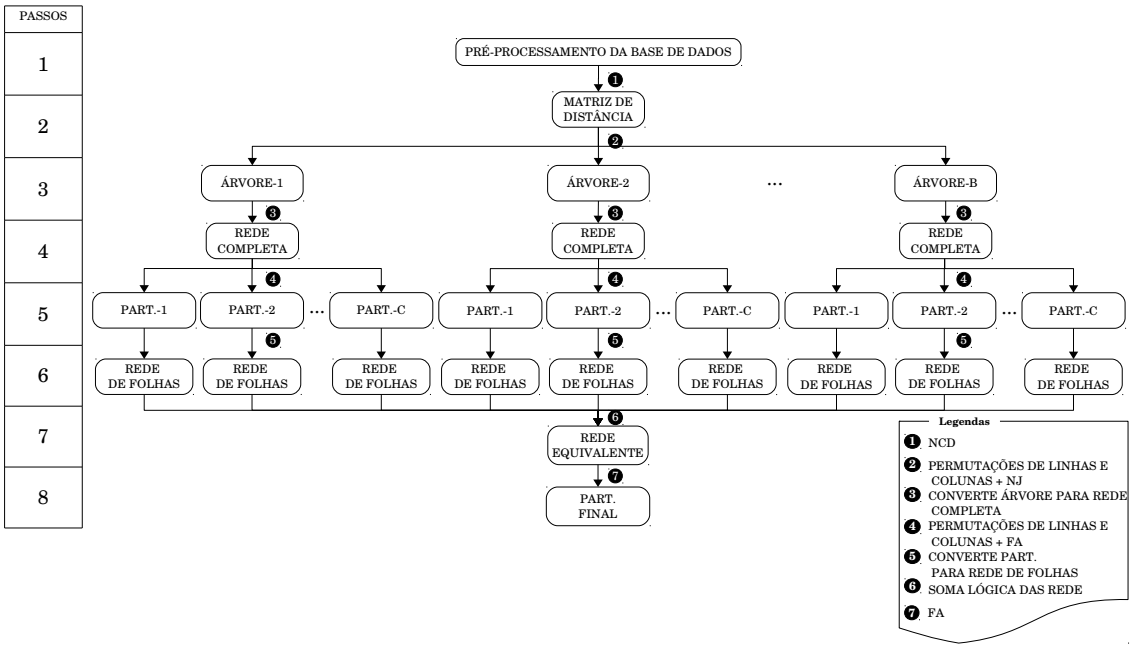

Figura 1. Diagrama resumindo o DAMICORE-2, destacando seu caráter de paralelização.

Passo 1: Pré-processamento da base de dados

O passo 1 consiste na discretização do conjunto de dados: para os atributos que podem ter diferentes valores dentre um domínio contínuo específico, novos valores discretos serão estabelecidos. Assim, o DAMICORE-2 estará apto a identificar mais facilmente um possível relacionamento entre os atributos mostrando melhores resultados em seu agrupamento, em troca de perda de informação. Os dois métodos mais comumente usados são Discretização por Largura Igual (EWD, do inglês *Equal Width Discretization*) e Discretização por Frequência Igual (EFD, do inglês *Equal Frequency Discretization*) [Lopes et al. 2014].

No modelo EWD, os intervalos de valores de cada atributo do conjunto de dados são divididos em intervalos de larguras iguais. Já no modelo EFD os intervalos de valores são divididos de modo que cada um contenha a mesma quantidade de valores distintos entre os elementos fornecidos.

Passo 2: Matriz de Distância

Uma matriz de distância é calculada usando distância por compressão normalizada (NCD [Cilibrasi e Vitányi 2005]) aplicada a cada par de variáveis de um problema (note que os valores de uma variável no conjunto de dados compõem um vetor). Esta matriz de distância será usada no Passo 3 para reconstrução de árvores filogenéticas.

Passo 3: Árvores

Para cada Matriz de Distância (Passo 2), o Passo 3 gera uma filogenia (que pode representar relacionamentos hierárquicos entre objetos, variáveis no DAMICORE-2) usando o algoritmo *Neighbor Joining* (NJ) [Saitou e Nei 1987, Studier e Keppler 1988]. NJ possui

eficiencia computacional cúbica ($O(\ell^3)$, onde ℓ é o número de amostras no conjunto de dados, ou o número de variáveis em um problema de otimização). Entretanto, ele não pode garantir a filogenia ótima já que é um algoritmo guloso. Quando colunas e linhas de uma Matriz de Distância são permutadas, NJ usualmente constrói diferentes árvores sub-ótimas para cada permutação, mas estas árvores diferentes preservam sub-estruturas em comum.

Passo 4: Rede Completa

O Passo 4 converte a saída do NJ do formato Newick [Felsenstein 2003] (usualmente aplicado em ferramentas de bioinformática como o PHYLIP [Felsenstein 2000]) para o formato de Matriz de Adjacência. O uso deste padrão de representação de redes faz possível a aplicação de algoritmos de detecção de comunidades do campo das Redes Complexas [Donetti e Muñoz 2004, Duch e Arenas 2005], como proposto no Passo 5.

O uso do formato Newick juntamente com esta conversão em Matriz de Adjacência também permite o uso de outros algoritmos de reconstrução de árvores (Passo 3). Além disso, outros algoritmos de detecção de comunidades podem ser aplicados diretamente a partir desta matriz.

Passo 5: Particionamento

Com base na rede obtida no Passo 4, o Passo 5 determina um conjunto de possíveis particionamentos de variáveis (um conjunto de sub-estruturas comuns, Passo 3) do problema. Aplica-se a técnica de detecção de comunidades chamada de *Fast Newman Algorithm* (FA), que tem um equilíbrio adequado entre qualidade das comunidades encontradas e o tempo de computação requerido [Crocomo 2012] ($O((a + \ell)\ell)$, onde a é o número de arestas na rede e ℓ o número de nós [Newman e Girvan 2004]). Similarmente ao NJ, FA também é um algoritmo guloso, para que permutações aleatórias de linhas e colunas da Matriz de Adjacência geralmente produzam particionamentos diferentes.

Destaca-se que o particionamento resultando diretamente do FA usa todos os nós da filogenia, nós internos e nós folha (as folhas representam variáveis do problema). O Passo 5 remove os nós internos, assim, a saída deste passo contém somente nós folha.

Passo 6: Redes de Folhas

O Passo 6 converte cada particionamento de nós folha (obtidos do Passo 5) para o formato de Matriz de Adjacência representando um grafo chamado de Rede de Folhas. É assumido que todos os nós na mesma partição (sub-estrutura comum) de um particionamento formam um grafo [Diestel 2006]. Este formato de matriz é requerido pelo passo 7, o qual também representa os índices de linhas e colunas em uma ordem padronizada. Neste caso, a ordem ascendente de índices é usada.

Passo 7: Rede Equivalente

Todas as Matrizes de Adjacência representam o particionamento encontrado em cada Rede de Folhas obtida da mesma base de dados são somadas (soma lógica ou operador booleano *OR* [Lipschutz e Lipson 2004]). O resultado é uma Matriz de Adjacência Equivalente ou Rede Equivalente.

Passo 8: Particionamento Final

Finalmente, o FA é aplicado à Rede Final, obtendo o Particionamento Final, somando as redes equivalentes resultantes do Passo 7. Nesta etapa o DAMICORE-2 revela, de certa forma, que de quantidade é possível extrair qualidade (informação relevante), uma vez que a inclusão de relações presentes em uma Rede de baixa qualidade (por exemplo com a ligação errônea de duas partições) tende a não interferir no resultado final das comunidades identificadas a partir da Rede Final, que representa o consenso entre as redes equivalentes.

Isso deve-se ao fato de a importância de uma possível relação entre nós que não é evidente ser relativamente pequena frente a importância das arestas corretamente incluídas que, em geral, ocorrem em maior número. A soma dessas Matrizes possibilita entender uma das propriedades do DAMICORE-2, a de que partições menores podem se tornar maiores (ou formar clados) ao se verificar consensos entre essas partições; ou simplesmente, podem se tornar mais representativas ou confiáveis por meio da inclusão de arestas entre seus nós.

A inovação do DAMICORE-2 em utilizar um conjunto de filogenias, para compor um particionamento dos grupos das variáveis, permite que o método tenda a fugir da aleatoriedade de resultados subótimos do NJ e FA, buscando um consenso desses resultados de onde se pode abstrair um resultado mais consistente e menos flutuante a cada execução. Com isso, descobrir novidades nas bases de dados.

2.1. Comparativo com o DAMICORE [Delbem et al. 2010]

O DAMICORE [Delbem et al. 2010] cria uma rede completa a partir da reconstrução de uma única árvore filogenética. O DAMICORE-2 torna flexível a configuração dos parâmetros de execução do método, permitindo a especificação do número de árvores que serão reconstruídas, bem como do número de redes equivalentes obtidas de cada árvore. Assim, a classificação final corresponde ao consenso entre os resultados de todas as redes (Rede de Folhas) obtidas, o qual fortalece as relações entre os indivíduos de um mesmo grupo. Um detalhe importante a ser destacado é que, assim como seu antecessor, o DAMICORE-2 pode - possivelmente - realizar agrupamento dentro de *clusters* (agrupamento hierárquico).

A seguir tem-se um comparativo entre os resultados obtidos pelo DAMICORE-2 e o DAMICORE [Delbem et al. 2010] (equivalente ao DAMICORE-2 configurado para reconstruir apenas uma árvore filogenética e uma rede completa). Foram utilizadas também as configurações com reconstrução de duas, cinco e dez árvores filogenéticas. Em todos os casos, o número de redes completas foi configurado para ser igual ao número de árvores e a base de dados selecionada foi a *iris* [Bache e Lichman 2013].

Na Tabela 1 é visualizado o resultado do agrupamento quando o DAMICORE-2 foi configurado para ser equivalente ao DAMICORE (uma árvore e uma rede). Foram formados dezesseis grupos, número bem superior a três (valor indicado pela literatura ([Bache e Lichman 2013])). Percebe-se que foram formados vários sub-grupos em cada grupo. Obteve-se boas taxas de acerto (*hit*) quando comparadas ao agrupamento original, com poucas exceções (*clusters* 7 e 11).

Ao se incrementar o número de árvores e redes em uma unidade cada, ou seja, duas árvores e duas redes, o número de *clusters* encontrados reduziu para oito, conforme

Tabela 1. Análise do agrupamento da base *Iris* com uma árvore e uma rede.

		Análise			
Cluster	# Elem.	Erros	*Hit* (%)	Total de erros	Total de *hit* (%)
0	7	0	100	0	100
1	7	0	100		
2	11	0	100		
3	11	0	100		
4	10	0	100		
5	4	0	100		
6	7	1	85.71	10	80.39
7	13	4	69.23		
8	8	1	87.50		
9	11	0	100		
10	12	0	100		
11	9	4	55.55	4	91.84
12	10	0	100		
13	12	0	100		
14	11	0	100		
15	7	0	100		

a Tabela 2. Embora as taxas de acerto tenha reduzido em alguns grupos individuais com relação à configuração anterior, a taxa de acerto para os grupos equivalentes da literatura alcançaram valores aceitáveis.

Tabela 2. Análise do agrupamento da base *Iris* com duas árvores e duas redes.

		Análise			
Cluster	# Elem.	Erros	*Hit* (%)	Total de erros	Total de *hit* (%)
0	21	0	100	0	100
1	29	0	100		
2	21	1	95.23	5	89.13
3	13	4	69.23		
4	12	0	100		
5	9	5	44.44	6	87.75
6	24	1	95.83		
7	16	0	100		

A seguir, na Tabela 3, tem-se o resultado para uma configuração com cinco árvores e cinco redes. Foram detectados sete grupos, que quando comparados aos grupos apontados pela literatura, possue, taxas de acerto que variam de 83,64% a 100%.

A Tabela 4 traz os resultados para os grupos formados com a utilização da configuração com dez árvores e dez redes, para a qual o método atingiu o ponto de saturação de *clusters*. À exceção do *cluster* 4, todos os demais alcançaram uma taxa de acerto de no mínimo 87,18%. Quando comparado aos *clusters* indicados pela literatura as taxas de acerto variaram de 83,33% a 100%.

Nota-se facilmente que o número de grupos formados converge até o momento

Tabela 3. Análise do agrupamento da base *Iris* com cinco árvores e cinco redes.

		Análise			
Cluster	# Elem.	Erros	*Hit* (%)	Total de erros	Total de *hit* (%)
0	17	0	100	0	100
1	29	0	100		
2	17	5	70.59	9	83.64
3	38	4	89.47		
4	9	4	55.55	4	89.80
5	17	0	100		
6	23	0	100		

Tabela 4. Análise do agrupamento da base *Iris* com dez árvores e dez redes.

		Análise			
Cluster	# Elem.	Erros	*Hit* (%)	Total de erros	Total de *hit* (%)
0	21	0	100	0	100
1	29	0	100		
2	39	5	87.18	10	83.33
3	21	5	87.18		
4	40	0	55.55	0	100

de ocorrência da saturação de *clusters*, onde o FA não consegue mais generalizar as características dos elementos, impossibilitando que grupos sejam combinados formandos outros maiores. Verifica-se ainda que a abordagem utilizada pelo DAMICORE-2 representa acréscimo importante com relação às capacidades do DAMICORE, uma vez que o consenso faz com que o número de grupos formados aponte para o indicado pelos especialistas para as bases de dados.

Com o aumento do número de árvores filogenéticas e redes geradas evidencia-se a performance superior do DAMICORE-2 e fica clara a convergência do número de *clusters* até um limite, o qual foi nomeado de saturação de *cluster*.

3. Resultados

Para os testes foram utilizados as bases de dados (*Iris*, *Glass* e *Seeds*)[1]. Sobre as bases são conhecidos *a priori* a quantidade de *clusters* que devem ser criados, segundo a literatura [Bache e Lichman 2013].

Testes preliminares apontaram que a utilização de uma configuração com sete árvores e sete redes é suficiente para que ocorra a saturação de *clusters* na maioria das bases de dados. Aqui foi utilizada a configuração com dez árvores e dez redes, a qual é suficiente para abranger um número ainda maior de bases de dados. A seguir são apresentados os melhores resultados obtidos para cada base de dados experimentada.

[1]Disponíveis em: http://archive.ics.uci.edu/ml/

Experimento 1

Classificação da base de dados *glass*

A Tabela 5 se refere ao melhor resultado obtido nos testes com a base *Glass* [Bache e Lichman 2013]. Também neste experimento a melhor metodologia de discretização foi a EFD. A base de dados *glass* possui sete classes, porém possui amostras de seis delas [Bache e Lichman 2013]. Entretanto nota-se a partir da Tabela 5 que o DAMICORE-2 atingiu o ponto de saturação de *clusters* com sete grupos detectados, para os quais as taxas de acerto variaram de 100%. Tal fato é devido à capacidade de agrupamento hierárquico do método proposto, que apontou a existência de dois subgrupos em um dos grupos existentes.

Tabela 5. Análise do agrupamento da base *Glass* com dez árvores e dez redes.

		Análise			
Cluster	# Elem.	Erros	*Hit* (%)	Total de erros	Total de *hit* (%)
0	41	14	65.85	14	65.85
1	26	5	69.23	5	69.23
2	27	2	92.59	7	86.79
3	26	5	80.77		
4	44	16	63.64	16	63.64
5	28	8	71.43	8	71.43
6	22	0	100	0	100

Experimento 2

Classificação da base de dados *Seeds*

No experimento 3, com a base de dados *seeds*, mais uma vez a discretização EFD se sobressaiu com relação ao EWD. Na tabela 6 são expostos o número de erros e a taxa de *hit* para cada *cluster* resultante. A Tabela 6 expõe uma taxa de *hit* um pouco mais elevada que as demais. Percebe-se ainda que cada um dos três grupos conhecidos da literatura [Bache e Lichman 2013] foi sub-dividido em dois grupos, o que levou a seis grupos resultantes do método DAMICORE-2, ao atingir o ponto de saturação de *cluster*.

Tabela 6. Análise do agrupamento da base *Seeds* com dez árvores e dez redes.

		Análise			
Cluster	# Elem.	Erros	*Hit* (%)	Total de erros	Total de *hit* (%)
0	56	13	76.79	20	74.68
1	23	7	69.56		
2	29	2	93.10	3	95.00
3	31	1	96.77		
4	17	6	64.71	9	87.32
5	54	3	94.44		

4. Conclusão

Algoritmos de classificação frequentemente necessitam da especificação de um parâmetro, normalmente chamado de k. A escolha do valor de k depende do conhecimento prévio que se tem sobre o problema investigado. Quando não se tem nenhuma informação acerca do problema, essa escolha se dá de forma intuitiva e muitas vezes precisa ser ajustada até se obter um valor adequado.

O DAMICORE-2 mostrou ser um método eficiente de classificação, capaz de agrupar com sucesso bases de dados de tipos mistos, graças à métrica de distância utilizada (NCD). As técnicas empregadas por ele permitem o agrupamento dos dados sem a necessidade de que seja informado o número de *clusters* nos quais devem ser classificados. O DAMICORE-2 se difere de outras alternativas, uma vez que algoritmos que possuem essa necessidade se tornam limitados, pois eventualmente podem não detectar padrões importantes.

O método proposto revelou ainda a capacidade de realizar agrupamento dentro de *clusters* (sub-grupos), o que pode ser utilizado para compreender mais a fundo os motivos pelos quais determinados elementos ficam juntos ou separados evidenciando alguma correlação entre eles. Além disso, essa característica também pode permitir a análise hierárquica dos *clusters*.

Nos testes realizados em base de dados já revisadas na literatura, a quantidade de grupos encontrada pelo DAMICORE-2 mostrou convergir para o valor real (ou próximo dele), já conhecido para as bases, à medida em que o número de árvores filogenéticas reconstruídas – e por consequência o número de redes – crescia, fazendo prevalecer o consenso entre a classificação realizada por cada uma delas permitindo, além disso, uma melhor compreensão dos grupos gerados.

Agradecimento

Os autores agradecem à CAPES e à FAPESP pelo suporte financeiro.

Referências

Bache, K. e Lichman, M. (2013). UCI machine learning repository. `http://archive.ics.uci.edu/ml/`.

Cilibrasi, R. e Vitányi, P. M. B. (2005). Clustering by compression. *IEEE Transactions on Information Theory*, 51:1523–1545.

Crocomo, M. K. (2012). *Algoritmo de otimização bayesiano com detecção de comunidades*. PhD thesis, Universidade de São Paulo (USP) - Instituto de Ciências Matemáticas e de Computação (ICMC).

Delbem, A. C. B., Melo, V. V., e Vargas, D. V. (2010). Algoritmo filo-genético. In *2ª Escola Luso-Brasileira de Computação Evolutiva (ELBCE)*. APDIO.

Diestel, R. (2006). *Graph Theory*. Electronic library of mathematics. Springer.

Donetti, L. e Muñoz, M. A. (2004). Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(10):P10012.

Duch, J. e Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2):027104+.

Felsenstein, J. (2000). Phylip (phylogeny inference package).

Felsenstein, J. (2003). *Inferring Phylogenies*. Sinauer Associates, 2 edição.

Jain, A. K., Murty, M. N., e Flynn, P. J. (1999). Data clustering: a review. In *ACM Computing Surveys (CSUR)*, volume 31, pags 264–323.

Lipschutz, S. e Lipson, M. (2004). *Matemática Discreta: Coleção Schaum*. BOOKMAN COMPANHIA ED.

Lopes, L. A., Machado, V. P., e de A. L. Rabêlo, R. (2014). Automatic cluster labeling through artificial neural networks. In *Internationa Joint Conference on Neural Networks*, pags 762–769.

Newman, M. E. J. e Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113. PT: J; PN: Part 2; PG: 15.

Sadrozinski, H. F.-W. H. F.-W. e Wu, J. (2011). *Applications of field-programmable gate arrays in scientific research*. Boca Raton, Fla. : CRC Press. Includes bibliographical references and index.

Saitou, N. e Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425.

Sanches, A., Cardoso, J., e Delbem, A. (2011). Identifying merge-beneficial software kernels for hardware implementation. In *Reconfigurable Computing and FPGAs (ReConFig), 2011 International Conference on*, pags 74–79.

Silva, B. d. A., Cuminato, L. A., e Delbem, A. C. B. (2014). An application-oriented cache memory configuration for energy efficiency in multi-cores. *Accept by IET Computers & Digital Techniques*, 0:0.

Studier, J. e Keppler, K. (1988). A note on the neighbor-joining algorithm of saitou and nei. *Molecular Biology and Evolution*, 5(6):729.