Mineração de Dados: Trabalho Prático 2

Artur Rodrigues

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)

artur@dcc.ufmg.br

1. INTRODUÇÃO

Problemas de agrupamento podem surgir de diversas aplicações, como mineração de dados e aprendizado de máquina, compressão de dados e classificação e reconhecimento de padrões. A noção do que constitui um bom agrupamento depende diretamente da aplicação e existem muitas maneiras de achar esses agrupamentos de acordo com diversos critérios, sejam eles *ad hoc* ou sistemáticos.

As técnicas de agrupamento surgem quando não se possui classes para serem preditas mas as observações devem ser separadas em grupos naturais. Esses grupos presumivelmente refletem um mecanismo que está em funcionamento que faz com que algumas instâncias possuam um grau de semelhança maior com certos elementos que com outros.

Os resultados de um agrupamento podem ser expressos de diversas maneiras. Os grupos que são identificados podem ser exclusivos: uma dada observação pertence a somente um grupo. Ou eles pode haver sobreposição: uma observação pode pertencer a diversos grupos. Ou eles podem ser probabilísticos: uma observação pertence a cada grupo com uma certa probabilidade. Ou eles podem ser hierárquicos: uma divisão grosseira das observações em grupos no nível superior e cada grupo sendo refinado posteriormente - talvez até o nível de observações individuais. Na realidade, a escolha entre essas possibilidades deve ser ditada pela natureza dos mecanismos que se acredita definirem o fenômeno do agrupamento. Todavia, como esses mecanismos são raramente conhecidos - a própria existência dos agrupamentos é, afinal de contas, algo que tentamos descobrir - e também por razões pragmáticas, a escolha é na maioria das vezes ditada pelas ferramentas disponíveis [Witten et al. 2011].

Nesse trabalho iremos examinar um algoritmo que trabalha em domínios numéricos, particionando as observações em grupos disjuntos. Faz parte do nosso trabalho um estudo cuidadoso sobre diferentes métodos de inicialização, avaliação exeperimental de valores para o número de grupos e análise de uma base de dados fornecida pela comissão avaliadora.

2. K-MEANS

Dentre as formulações de agrupamentos fundamentadas na minimização de uma função objetivo, talvez a mais amplamente utilizada e estudada seja o agrupamento *k*-means. Dado um conjunto de *n* pontos num espaço real *d*-dimensional, $\mathbb{R}^d$, e um inteiro k, o problema é determinar um conjunto de k pontos em $\mathbb{R}^d$, denominados centros, para minimizar a distância total quadrada de cada ponto para seu centro mais próximo. Esse tipo de agrupamento é enquadrado na categoria geral de agrupamentos baseados em variância [Inaba et al. 1994, Inaba et al. 1996].

Agrupamento baseada no k-means está relacionada com uma séria de outras técnicas de agrupamentos e problemas de localização, incluindo o *k-medians* Euclidiano, no qual o objetivo é minimizar a soma das distâncias para o centro mais próximo, e também o problema *k-centers*, onde o objetivo é minimizar a distância máxima de todos os pontos para o centro mais próximo.

Não existe solução eficiente para nenhum desses problemas e algumas formulações são NP-hard [Garey and Johnson 1979]. Uma das heurísticas mais populares para a solução do k-means é baseada em um simples esquema iterativo que nusca solução mínimas locais. Esse algoritmo é geralmente chamado de *algoritmo k-means* [Forgy 1965, MacQueen 1967], e especificamente nesse trabalho é utilizada a versão conhecida como *Algoritmo de Lloyd* [Lloyd 1982].

O algoritmo de Lloyd é baseado na simples observação de que o posicionamento ótimo de um centro é no centroide do agrupamento associado. Dado um conjunto de k centros Z, para cada centro $z \in Z$, seja $V(z)$ o conjunto de pontos onde z é o vizinho mais próximo. Em termos geométricos, $V(z)$ é o conjunto de pontos sobre a célula de Voronoi de z [Preparata and Shamos 1985]. Cada estágio do algoritmo de LLoyd move cada ponto central z para o centroide de $V(z)$ e depois atualiza $V(z)$ ao reavaliar a distância de cada ponto para o seu centro mais próximo. Esses passos são repetidos até que uma condição de convergência seja atingida. Em geral, especialmente se nenhum ponto é equidistante de dois centro, o algoritmo irá eventualmente convergir para um ponto que é um mínimo local para a distorção. Nesse trabalho a condição de parada é satisfeita quando $V(z)$ para cada ponto central z não se altera após um novo estágio de atualização.

2.1. Avaliação de Qualidade

Uma das maneiras mais diretas de se aferir a qualidade da solução encontrada é através da medida de Distância Quadrada Total, que mede a distância ao quadrado de cada observação até o centroide mais próximo. Uma outra medida, mais robusta, é o Índice de Jagota [Jagota 1991], que avalia a tensão ou homogeneidade dos objetos dentro dos agrupamentos. Ela é definida como:

$$Q = \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{x \in C_i} d(x, \mu_i)$$

Nesse trabalho essa foi a maneira utilizada para se aferir a qualidade da solução. Em [Dunn 1974] é apresentado o Índice de Dunn que define uma razão entra as distância mínima e máxima intra-agrupamentos. Já em [Davies and Bouldin 1979] é formalizado o Índice de Davies-Bouldin, onde valores baixos indicam que os agrupamentos são compactos e seus centros afastados uns dos outros - consequentemente, o número de agrupamentos que minimiza esse índice é tomado como ótimo. Apesar de representarem boas alternativas para a avaliação de qualidade, ponderou-se que sua utilização foge do escopo do trabalho.

2.2. Complexidade

Em termos de complexidade temporal, podemos dizer que a maior parte do tempo é gasta na computação das distâncias entre as observações e os centros. Essa operação tem custo $O(M)$, onde M é a dimensão dos vetores. O passo de atualização computa KN distâncias, dessa maneira, sua complexidade é $O(KNM)$. Para um número fixo de iterações I, a complexidade geral é $O(IKNM)$.

Dessa maneira, o k-means é linear em todos os fatores relevantes: iterações, número de agrupamentos, número de observações e dimensionalidade do espaço. Em [Inaba et al. 1994] é mostrado que se a dimensionalidade M e o número de agrupamentos K são fixados, o problema pode ser resolvido em $O(N^{MK+1} \log N)$.

3. ESCOLHA DOS CENTRÓIDES INICIAIS

Pode-se argumentar que o algoritmo k-means define um mapeamento determinístico a partir de um ponto inicial até a solução. Isso significa que o ótimo local encontrado como solução é sensível a escolha inicial dos agrupamentos. Arranjos completamente diferentes para a solução final podem surgir a partir de pequenas alterações na escolha inicial. De acordo com [Duda and Hart 1973] (p. 228):

> "One question that plagues all hill-climbing procedures is the choice of the starting point. Unfortunately, there is no simple, universally good solution to this problem."

"Repetição com diferentes escolhas aleatórias" [Duda and Hart 1973] é geralmente a estratégia mais utilizada. Nesse trabalho, foram avaliadas três diferentes maneiras de se escolher aleatoriamente os k pontos iniciais para serem centros dos agrupamentos. Além disso, foi implementada uma maneira de se utilizar k observações definidas pelo usuário como centroides iniciais.

3.1. *k* centroides Aleatórios

Como será apresentado na seção 4, cada uma das observações e consequentemente também cada um dos centroides é definido como um ponto em $[0, 1]^M$.

Assim, essa maneira de inicialização simplesmente define um valor real aleatório no intervalo $[0, 1]$ para cada uma das M dimensões, para cada um dos vetores. Um exemplo de vetor gerado para $M = 5$ é apresentado abaixo:

$$v = [0.43, 0.11, 0.78, 0.91, 0.32]$$

3.2. Inicialização de Forgy

Nesse método, apresentado em [Forgy 1965], são escolhidas k observações aleatórias para serem os os pontos centrais iniciais.

3.3. Partições Aleatórias

Essa alternativa assinala um agrupamento aleatório para cada uma das observações e em seguida procede para o passo de atualização do algoritmo k-means, obtendo pontos médios que serão utilizados como centroides iniciais.

4. BASE DE DADOS

A base de dados utilizada foi fornecida pela comissão avaliadora, onde cada observação é uma música identificada por um número, seguida de rótulos que foram associados a ela por usuários do sistema de onde foi extraída. Um exemplo de entrada é exibido abaixo:

```
14 jazz music chillout futuristica electronic hip-hop
```

Para a modelagem do problema foi efetuado o que é conhecido por desnormalização, onde cada um dos rótulos existentes na base de dados é representado por uma dimensão. No caso da base fornecida, existem 3869 rótulos distintos, implicando no mapeamento do problema para $M = 3869$ dimensões.

Como consequência dessa modelagem, os pontos que representam cada uma das observações são vetores com alto grau de esparsidade: as posições que representam rótulos da observação assumem valor 1 e as demais 0.

A figure 1 exibe os 10 rótulos mais frequentes.

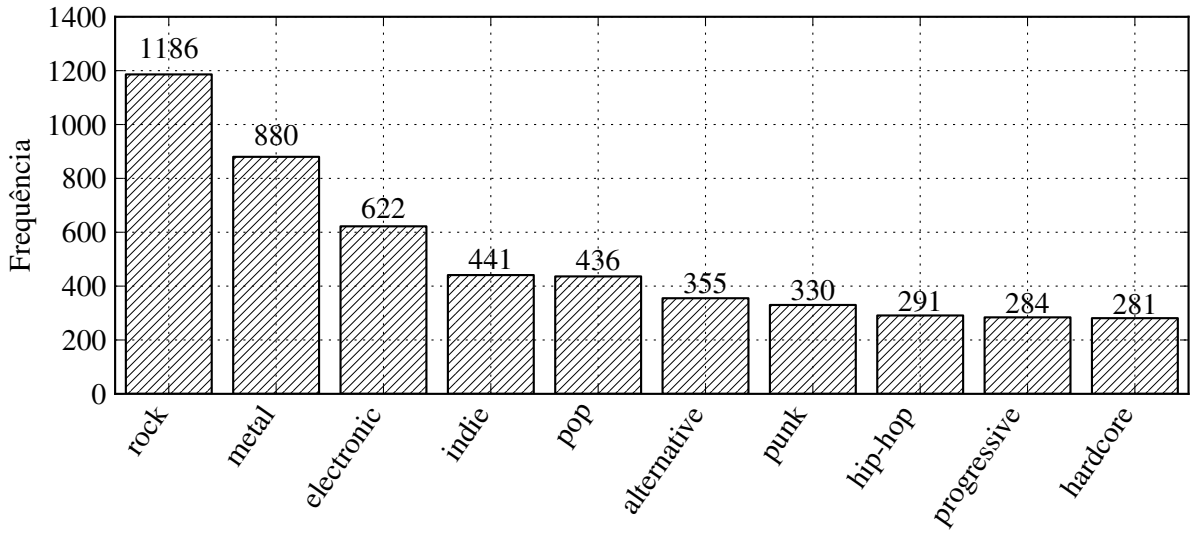

Figura 1: Os 10 rótulos mais frequentes

A figura 2 mostra como o número de rótulos (que irá definir a dimensionalidade do problema) evolui com o crescimento da base de dados. É interessante notar que esse crescimento não respeita a Lei de Zipf, ao crescer linearmente.

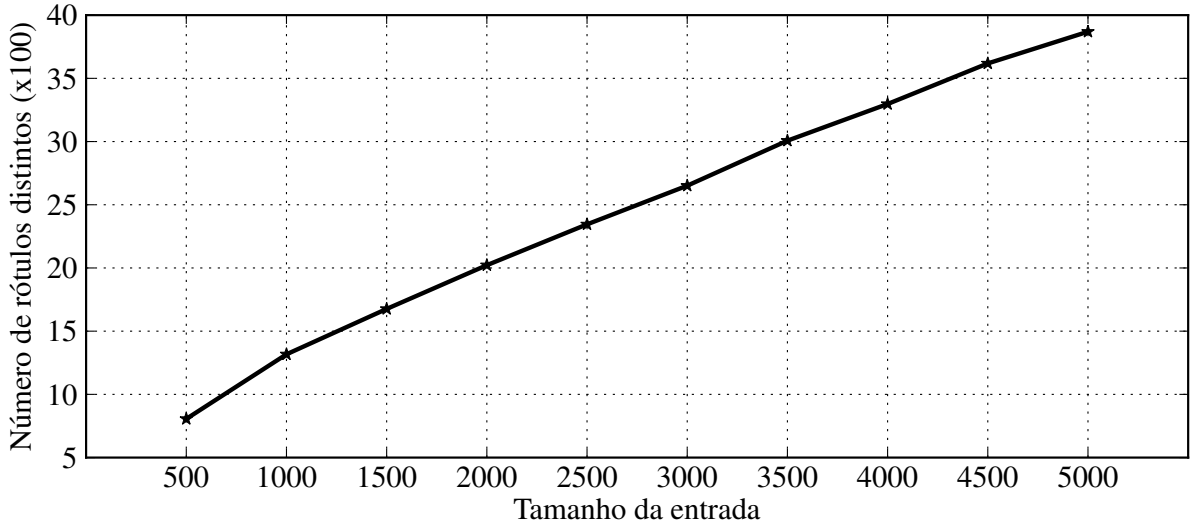

Figura 2: Número de rótulos com o crescimento da base de dados

5. AVALIAÇÃO EXPERIMENTAL

5.1. Procedimentos

Com o intuito de se obter testes mais consistentes, os experimentos foram executados em ambiente virtualizado, com capacidade de processamento e memória primária reduzidas, 50% da

capacidade da máquina hospedeira e 1024MiB, respectivamente. O sistema operacional do ambiente virtualizado era Ubuntu Server 12.04 64 bits e os softwares utilizados foram interpretador Python (2.7.2) e GCC versão 4.2.1. A máquina hospedeira possuía sistema operacional Mac OS X 10.8.2, processador *quad-core* de 2.3GHz e memória primária com capacidade de 16GiB.

Todos os testes foram realizados 3 vezes e o resultados médios para os valores aferidos foram considerados. Finalmente, certificou-se que a solução desenvolvida execute perfeitamente na estação `claro.grad.dcc.ufmg.br`.

5.2. Análise dos Métodos de Inicialização

Como apresentado na seção 3, foram implementadas três maneiras de obter os centroides iniciais. Cada uma dessas alternativas foi estudada, com valor de $k = 50$, valor esse obtido através da regra de ouro $k \approx \sqrt{N/2}$ [Mardia et al. 1979]. A tabela 1 apresenta os resultados para essas execuções.

Medidas	Valores	centroides Aleat.	Forgy	Partições Aleat.
Iterações	Exec. 1	57	13	19
	Exec. 2	56	14	15
	Exec. 3	61	17	22
	Média	**58.00**	**14.67**	**18.67**
Distância Quadrada Total	Exec. 1	17903.03	17789.62	16923.46
	Exec. 2	17989.05	17682.62	16926.98
	Exec. 3	17793.68	17745.91	16906.91
	Média	**17895.25**	**17739.38**	**16919.11**
Índice Jagota	Exec. 1	90.46	72.62	88.39
	Exec. 2	89.77	63.91	87.20
	Exec. 3	88.60	74.97	87.94
	Média	**89.61**	**70.50**	**87.84**
Tempo (s)	Exec. 1	721.09	164.63	249.11
	Exec. 2	721.42	177.61	192.54
	Exec. 3	796.81	219.69	291.86
	Média	**746.44**	**187.31**	**244.50**

Tabela 1: Comparação dos Métodos de Inicialização

Percebe-se que o método de Inicialização de Forgy produz os agrupamentos com os melhores índices de Jagota, através de menos iterações e consequentemente em menos tempo. Observa-se ainda que o método Partições Aleatórias, apesar de não produzir um valor para o índice de Jagota tão bom quanto o de Forgy, produz os agrupamentos com a menor distância quadrada total.

As avaliações seguintes utilizaram o método de inicialização de Forgy.

5.3. Análise do Valor de K

Para a análise do número de agrupamentos (K), utilizou-se um método conhecido como *Elbow Method*, creditado a [Thorndike 1953]:

> Esse método "utiliza o percentual de variância explicada como função do número de agrupamentos: deve ser escolhido um número de agrupamentos de

maneira que a adição de outro grupo não implique numa melhora significativa na modelagem dos dados."

Em termos práticos, a mesma análise pode ser feita através da avaliação da variação da soma média dos quadrados intra-agrupamentos, ou seja, a média da soma das distâncias de cada observação ao centroide mais próximo (ou *within sum of squares*), quando se aumento o número de agrupamentos.

A figura 3 mostra exatamente esse ponto, onde $k = 7$. Esse tipo de avaliação pode ser melhor realizada se levado em conta o Princípio da Descrição com Comprimento Mínimo (ou *Minimum Description Length Principle*), que determina que a melhor teoria para um corpo de dados é aquele que minimiza o tamanho da teoria somado à quantidade de informação necessária para especificar as exceções a teoria [Witten et al. 2011]. Esse tipo de estudo foge do escopo desse trabalho.

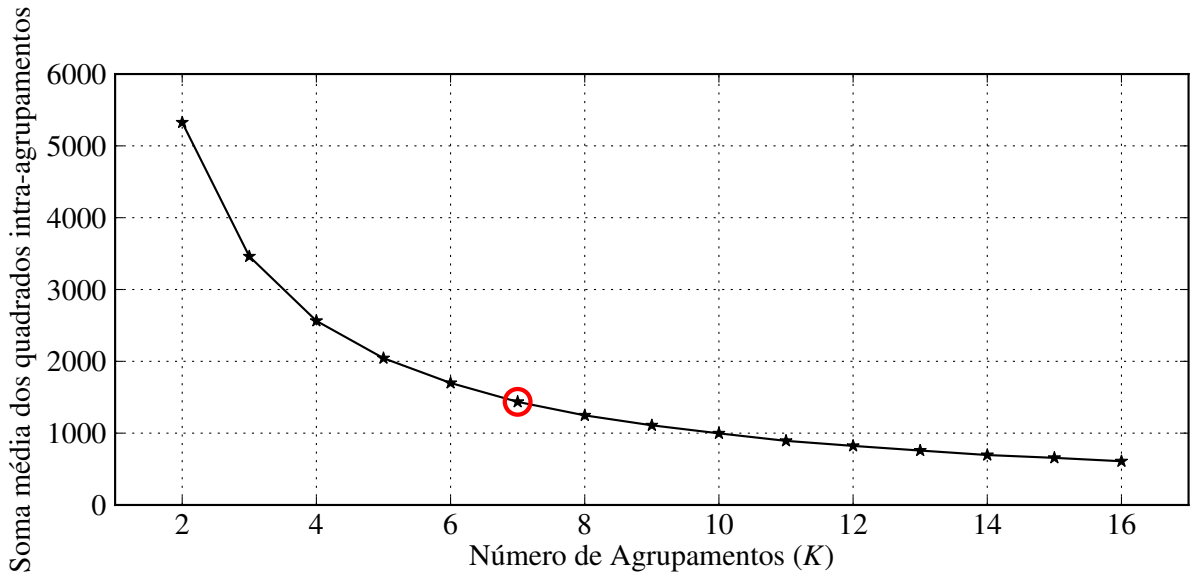

Figura 3: "Elbow" para o estudo dos valores de k

5.4. Análise do Tamanho da Entrada

Para essa avaliação experimental, variou-se o tamanho da entrada, utilizando-se somente as X primeiras observações da base de dados disponibilizada, e foi aferido o tempo de execução, em caráter excepcional, para 10 execuções. O valor médio foi considerado e o desvio padrão amostral podem ser apreciados na figura 4.

Nota-se que o experimento não foi de encontro com a análise de complexidade discutida na seção 2.2. De acordo com a tabela 2, a dimensão dos pontos que estão sendo representados cresce linearmente com o crescimento da base de dados. É esperado portanto, que o tempo de execução cresca de maneira quadrática com o aumento da base, o que não ocorre.

Avalia-se que esse comportamento se deve ao fato da inicialização dos centroides ser feita de maneira aleatória, o que influência consideravelmente, tanto na taxa de convergência, quanto na qualidade da solução encontrada. Alternativas que forneceriam resultados mais palpáveis, mas menos realistas, envolvem a utilização de uma mesma semente para a seleção dos centroides iniciais, ou o estabelecimento de um número de iterações comum para todas as execuções.

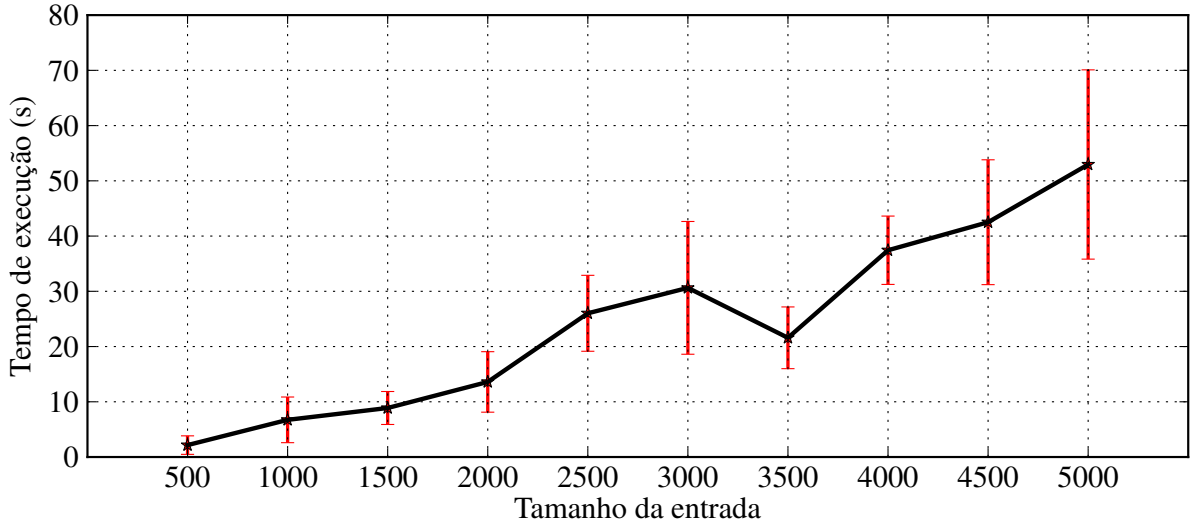

Figura 4: Tempo de execução em função do tamanho da entrada

5.5. Análise da Qualidade da Solução

A análise da qualidade da solução de certa forma já foi feita nas seções anteriores, como apresentado na tabela 1 e na figura 3. Adicionalmente, a figura 5 apresenta como as medidas Distância Total Quadrada e índice de Jagota se comportam com a evolução da execução do algoritmo.

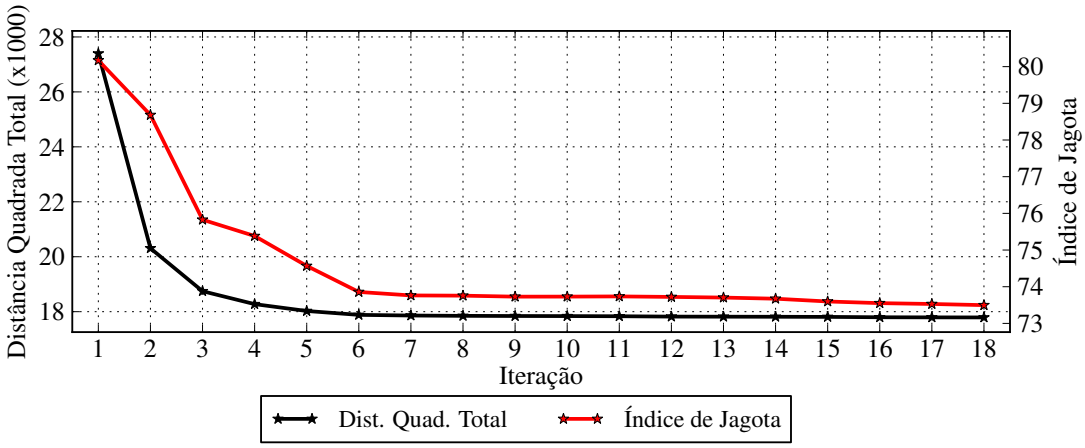

Figura 5: Evolução das medidas de qualidade com a execução do algoritmo

É interessante notar que após a iteração 7 os valores para esses índices pouco mudam, uma característica forte da convergência do algoritmo.

6. CONCLUSÃO

Nesse trabalho, foi feito um estudo sobre o algoritmo de agrupamento k-means. Foram avaliadas três alternativas de inicialização para os centroides, onde se identificou que o método de Forgy é a melhor alternativa. Além disso, foram expostas algumas métricas de avaliação de qualidade dos agrupamentos gerados, com um enfoque sobre o Índice de Jagota. Finalmente, foi feita uma análise para a variação do número de agrupamentos através do *Elbow Method*.

O trabalho atende os objetivos propostos ao promover um estudo mais cuidadoso sobre a técnica, além de exigir o auxílio à literatura, e como consequência, a familiarização com a tarefa de agrupar itens.

Referências

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1(2):224–227.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons Inc, 1 edition.

Dunn, J. C. (1974). Well separated clusters and optimal fuzzy-partitions. *Journal of Cybernetics*, 4:95–104.

Forgy, E. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–780.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA.

Inaba, M., Imai, H., and Katoh, N. (1996). Experimental results of randomized clustering algorithm. In *Proceedings of the twelfth annual symposium on Computational geometry*, SCG '96, pages 401–402, New York, NY, USA. ACM.

Inaba, M., Katoh, N., and Imai, H. (1994). Applications of weighted voronoi diagrams and randomization to variance-based k-clustering. In *Proceedings of the tenth annual symposium on Computational geometry*, SCG '94, pages 332–339, New York, NY, USA. ACM.

Jagota, A. (1991). Novelty detection on a very large number of memories stored in a hopfield-style network. In *Neural Networks, 1991., IJCNN-91-Seattle International Joint Conference on*, volume ii, page 905 vol.2.

Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J., editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

Preparata, F. P. and Shamos, M. I. (1985). *Computational geometry: an introduction*. Springer-Verlag New York, Inc., New York, NY, USA.

Thorndike, R. L. (1953). Who belong in the family? *Psychometrika*, 18.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3rd edition.

Zaki, M. and Junior, W. M. (2012). *Fundamentals of Data Mining Algorithms*. Cambridge, draft edition.