UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA

DEPARTAMENTO DE ESTATÍSTICA

USO DO ALGORITMO ROCK PARA AGRUPAR DADOS CATEGORIZADOS

Monografia apresentada ao Departamento de Estatística da Universidade Federal da Paraíba - UFPB para a obtenção do grau de Bacharel em Estatística

Por **ISIS MILANE BATISTA DE LIMA**

Orientador: **JOAB DE OLIVEIRA LIMA**

João Pessoa - PB, Brasil

Março/2013

Monografia de Projeto Final de Graduação sob o título “USO DO ALGORITMO ROCK PARA AGRUPAR DADOS CATEGORIZADOS”, defendida por Isis Milane Batista de Lima e aprovada em 26 de Março de 2013 em João Pessoa no Estado da Paraíba, sendo apreciada pela banca examinadora constituída pelos Professores:

Aprovada em ____/____/_____.

BANCA EXAMINADORA

Prof. Dr. Joab de Oliveira Lima

Universidade Federal da Paraíba

Prof[a]. Dra. Renata Patrícia Lima Jerônimo

Universidade Federal da Paraíba

Prof. Dr. Marcelo Rodrigo Portela Ferreira

Universidade Federal da Paraíba

Resumo

O objetivo da Análise de Agrupamento é atribuir observações a grupos (*cluster*), de modo que as observações dentro de cada grupo sejam o mais homogênea possível com relação às variáveis ou atributos de interesse. Em outras palavras, a Análise de Agrupamento procura o número e a composição dos grupos, sem exigir nenhuma estrutura paramétrica para os dados.

Na literatura há inúmeros métodos de agrupamentos, sendo a maioria deles dedicado ao estudo de variáveis contínuas. Para o caso de variáveis qualitativas, são poucos os procedimentos existentes. Entre eles está o algoritmo Robust Clustering using Links – ROCK. Diferente das metodologias tradicionais de agrupamento, o algoritmo ROCK utiliza o número de vizinhos comuns (*links*) entre os elementos para agrupá-los, sendo, por isso, chamado de método robusto de agrupamento.

Nesse trabalho utilizou-se o algoritmo ROCK para agrupar casos de duas bases de dados reais. No primeiro caso prático, que tratava do número de procedimentos ambulatoriais realizados pelos principais estabelecimentos de saúde de João Pessoa-PB, o método ROCK agrupou corretamente entre 78% e 91% das observações reais para cada um dos três graus de vizinhança utilizado. Já para a segunda base de dados que relacionava as condições dos homicídios com as características das vítimas dos crimes realizados entre 2008 e 2011 no município de Campina Grande-PB, o algoritmo ROCK apresentou um desempenho mediano, pois agrupou corretamente entre 54% e 59% dos homicídios para cada um dos três graus de vizinhança utilizado.

De um modo geral, o algoritmo ROCK se mostrou uma boa alternativa para o agrupamento de casos medidos através de atributos e poderá ser uma opção interessante para o estudo de dados segmentados.

Palavras-chave: Variáveis Categóricas, Análise de Agrupamento, ROCK.

Abstract

The goal of Cluster Analysis is to assign observations to groups (cluster), so that the observations within each group are as homogeneous as possible with respect to the variables or attributes of interest. In other words, Cluster Analysis seeks to discover the number and composition of the groups, without requiring any parametric structure for the data.

In literature there are numerous methods of groups, most of them dedicated to the study of continuous variables. For the case of qualitative variables, few existing procedures. Among them is the Robust Clustering using Links – ROCK algorithm. Unlike traditional clustering methods, the algorithm ROCK uses the number of common neighbors (*links*) between the elements to group, being therefore called robust method of grouping.

In this work we used the ROCK algorithm to group cases of two real datasets. In the first case study, which was the number of outpatient procedures performed by leading healthcare facilities of João Pessoa-PB, the ROCK method correctly grouped between 78% and 91% of actual observations for each of the three degrees of neighborhood used. For the second database that listed the conditions of homicides with the characteristics of the victims of crimes conducted between 2008 and 2011 in Campina Grande-PB, the ROCK algorithm showed an average performance, because grouped correctly between 54% and 59% of homicides for each of the three degrees of neighborhood used.

In general, the algorithm ROCK proved a good alternative to the grouping cases measured through attributes and could be an interesting option for the study of segmented data.

Keywords: Categorical Variables, Cluster Analysis, ROCK.

Dedico este trabalho a todos aqueles que, de forma impulsora, me auxiliaram na elaboração e desenvolvimento deste trabalho. Em especial ao Professor e Orientador Doutor Joab de Oliveira Lima, pela paciência, incentivo e dedicação a mim dispensados, na elaboração concreta das idéias, guiando os meus saberes e demonstrando a real importância deste trabalho.

Agradecimentos

A Deus, a Jesus Cristo e a Nossa Senhora Aparecida, por trilhar o meu caminho até aqui.

A todos que, direta ou indiretamente, contribuíram para a consecução deste empreendimento.

Aos colegas componentes da Comissão Especial de Avaliação, por aceitarem a incumbência desta tarefa e pela contribuição à melhoria desta monografia.

Aos meus companheiros de sala de aula Adeilda Fernandes de Melo Lima, Alisson de Oliveira Silva, Amaury Ceciliano Bandeira, Antônio Guedes Corrêa Filho, Anna Paola Bezerra e Silva, José Edson Rodrigues Guedes Gondim, Lídia Dayse Araújo de Souza, Luana Cecília Meireles da Silva, Natália Rodrigues Guedes Gondim, Rodrigo Cabral da Silva, Thiago Cavalcanti de Melo Lima, Wanessa Weridiana da Luz Freitas e a tantos outros, por todos os momentos maravilhosos de estudos.

Aos(s) grandes professores (as) e educadores (as) Abdoral de Souza Oliveira, Ana Flávia Uzeda dos Santos Macambira, Andréa Vanessa Rocha, Antônio Marcos Moreira, Eufrásio de Andrade Lima Neto, Gilmara Alves Cavalcanti, Hemílio Fernandes Campos Coêlho, Joab de Oliveira Lima, João Batista Parente, José Carlos de Lacerda Leite, Luciano da Costa Silva, Marcelo Rodrigo Portela Ferreira, Neir Antunes Paes, Renata Patrícia Lima Jerônymo, Sydney Gomes da Silva, Turíbio José Gomes dos Santos, Ulisses Umbelino dos Anjos.

À funcionária Maria de Fátima Lima Mesquita pelo apoio nas horas difíceis.

Em especial à minha mãe Irandí Batista de Lima, meu pai Aldecy Batista de Lima, meu irmão Aldeir Philype Batista de Lima e meu esposo Vyctor Rey da Costa Ramos Miranda que me incentivaram e acreditaram no meu potencial.

Lista de Símbolos

μ	Média da população
$\bar{x}$	Média da amostra
S	Desvio padrão da amostra
N	Tamanho da amostra
ROCK	Robust Clustering using Links
d	Distância Euclidiana
$Z_{ij} \sim (0, 1_j)$	Variável Normal padronizada
$\bar{d}$	Distância Euclidiana Média
D_a^2	Distância de Mahalanobis
X	Vetor de características de um item
$SQRes$	Soma de quadrados dos resíduos
θ	Parâmetro relacionado à medida de proximidade

Lista de Tabelas

Lista de Figuras

Sumário

1. INTRODUÇÃO

Análise de Agrupamentos engloba uma variedade de técnicas e algoritmos cujo objetivo é encontrar e separar objetos em grupos similares. Essa atividade pode ser observada, por exemplo, em um centro de distribuição de uma agência dos correios, em que os funcionários separam as cartas para a entrega. Nesses casos é comum os funcionários separá-las segundo uma ou mais características, tais como, por tipo, por bairro e por ruas. Nessa situação, os funcionários estão praticando uma Análise de Agrupamentos. Usar mais de uma característica para formar os grupos de cartas torna-se uma atividade mais trabalhosa, exigindo conceitos mais sofisticados de semelhança e procedimentos mais "científicos" para organizá-las (EVERITT, 1974).

Desse modo, a Análise de Agrupamento pretende resolver a seguinte questão: "dada uma amostra de *n* objetos (ou indivíduos), cada um deles medido segundo *p* variáveis, como é possível definir uma estratégia de classificação que agrupe os objetos em *g* grupos, de forma que os objetos agrupados sejam os mais similares possíveis?"

Essa técnica de análise multivariada é empregada em diversas áreas de conhecimento tais como medicina, administração, experimentos agronômicos, engenharia florestal, entre outros, e vem aumentando muito nos últimos anos. Jain (1999) traz ainda aplicações em análise de dados, reconhecimento de padrões, processamento de imagens e pesquisa de mercado.

O processo de agrupamento envolve a estimação de uma medida de dissimilaridade entre os indivíduos e a escolha de uma técnica de formação de grupos, também chamados de algoritmos de agrupamento. Esses algoritmos são de fácil aplicação, pois não exigem muitas manipulações de matrizes de incidência (DUDA; HART, 1973). As medidas de dissimilaridade são quantidades comparativas que medem a distância entre dois objetos, ou quantifica o quanto eles são diferentes. De forma geral, os algoritmos descrevem como o procedimento de agrupamento deve ser realizado e alguns deles possuem características especiais como a inicialização do algoritmo e a determinação de parâmetros (MATOS, 2007). Existe um grande número de métodos de agrupamentos disponíveis na literatura, dos quais o pesquisador deve escolher o mais adequado ao seu propósito, uma vez que as diferentes técnicas podem levar a diferentes soluções (SOUZA et al., 1997).

Adicionalmente, alguns autores oferecem soluções para tratar observações categóricas, como as apresentadas por Mingoti (2005) que, inicialmente, transforma as variáveis categóricas em contínuas, atribuindo números às categorias ou transformando-as em binárias, em que “zero” represente a ausência de certo atributo e “um” a presença deste atributo. Mas, essas transformações causam perda de informações, aumento indesejado do número de variáveis, impossibilidade de se consolidar resultados discordantes e dificuldade na determinação dos pesos da medida de proximidade conjunta. Para aperfeiçoar isso, algumas metodologias foram desenvolvidas, tais como: k-Modas e k-Protótipos (HUANG, 1997, 1998), STIRR: Sieving Through Iterated Relational Reinforcement (GIBSON et al.,2000), CACTUS: Clustering Categorical Data Using Summaries (GANTI et al., 1999), Fuzzy c-Modas (HUANG; NG, 1999), ROCK: Robust Clustering Using Links (GUHA et al., 2000), A Robustand Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment (CHIU et al., 2001), LIMBO (ANDRITSOS, 2004), QROCK: Quick ROCK (DUTTA et al., 2005), M-BILCOM (ANDREOPOULOS et al., 2005), BILCOM (ANDREOPOULOS et al., 2006), dentre outras.

Em especial, nessa monografia, os esforços serão concentrados no estudo e aplicações do algoritmo ROCK.

2. OBJETIVOS

2.1. OBJETIVO GERAL

Este trabalho tem como objetivo apresentar o uso do algoritmo RObust Clustering using Links– ROCK para o agrupamento de dados categorizados, enfatizando, principalmente, as suas aplicações.

2.2. OBJETIVOS ESPECÍFICOS

- Apresentar em detalhes a estrutura analítica do algoritmo ROCK;
- Implementar o algoritmo ROCK no software R;
- Aplicar o algoritmo ROCK para o agrupamento de dados reais.

3. REVISÃO DA LITERATURA

Segundo Hair et al.(2009), a análise de agrupamento é um grupo de técnicas multivariadas cuja finalidade principal é agregar objetos com base nas características que eles possuem. Essa técnica reúne indivíduos ou objetos em grupos tais que os objetos no mesmo grupo são mais parecidos uns com os outros do que com os objetos de outros grupos. A idéia é maximizar a homogeneidade de objetos dentro de grupos, ao mesmo tempo em que se maximiza a heterogeneidade entre os grupos.

Desse modo, o problema da análise de agrupamento pretende, dada uma amostra de n objetos (ou indivíduos), cada um deles medidos segundo p variáveis, procurar um esquema de classificação que agrupe os objetos em g grupos, exigindo-se daí conceitos científicos mais sofisticados de semelhança. Devem ser determinados também o número e as características desses grupos (BUSSAB; MIAZAKI; ANDRADE, 1990).

Sobre o estudo de técnicas de análise multivariada existem obras bastante significativas como as de Anderberg (1973), Hair et al. (2009), Mingoti (2005), dentre outras. Há, ainda, aquelas específicas sobre análise de agrupamento como Rosseeuw e Kaufman (2005) e Everitt (1980), por exemplo. Alguns autores fazem uso de medidas de distâncias para tratar variáveis contínuas, outros apresentam medidas de proximidade para variáveis categóricas, mas apenas Mingoti (2005) discute um método específico para agrupamento de variáveis categóricas. Dessa forma, essa monografia baseou-se, em grande parte, em artigos que citam algoritmos para agrupamento de variáveis categóricas.

Com o objetivo de avaliar se diferentes coeficientes de similaridade influenciam nas análises com marcadores moleculares dominantes, Meyer (2002) comparou vários coeficientes de similaridade e os resultados sugerem o uso do coeficiente de Jaccard como escolha dentre os que desconsideram ausência conjunta de atributos.

Albuquerque (2005), por sua vez, utilizou os algoritmos de agrupamento para dados de vegetação e os resultados indicaram que, em princípio, os algoritmos de agrupamento estudados são igualmente eficientes.

Já Matos (2007) propôs comparar cinco algoritmos de análise de agrupamento em variáveis categóricas e três metodologias aplicáveis a variáveis categóricas e contínuas. O objetivo foi analisar a extensão do método ROCK para o caso da mistura das

variáveis, chamado de Quick ROCK – QROCK. Os resultados mostraram que o algoritmo ROCK se destacou nos estudos de simulação realizados.

Por fim, para estudar sobre a criminalidade em Belo Horizonte, Souza (2012) verificou a existência de semelhança entre os indivíduos buscando evidenciar a ocorrência de agrupamentos formados por aqueles que possuíam alguma semelhança. O autor concluiu, por meio do ROCK, que é possível traçar o perfil dos indivíduos que fazem parte do mesmo agrupamento, encontrando indícios da ocorrência de crimes mais comuns cometidos por um mesmo grupo social em regiões especificas da cidade.

4. ANÁLISE DE AGRUPAMENTO

4.1. VISÃO GERAL

A Análise de Agrupamento é uma técnica útil para agrupar objetos (itens ou indivíduos), tais que os elementos dentro de um determinado grupo sejam similares entre si, ou seja, tenham características semelhantes, enquanto que elementos em grupos distintos sejam diferentes (GUHA et al., 2000).

Os algoritmos de agrupamentos discutidos na literatura podem ser classificados como Técnicas de Agrupamentos Não-hierárquicos ou Hierárquicos (DUDA; HART, 1973). Os algoritmos não-hierárquicos, também chamados de algoritmos particionais, como o próprio nome sugere, dividem os pontos amostrais dentro de *k* grupos, de modo que otimize uma certa função objetivo. A função objetivo mais tradicional para espaços métricos é como segue:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} d(x; m_i) \qquad [1]$$

em que m_i é o centroide do grupo C_i e $d(x; m_i)$ é a distância Euclidiana o ponto amostral e o centróide do grupo C_i.

Então, intuitivamente, a meta é minimizar a função objetivo E, isto é, tentar minimizar a distância de todos os pontos à média (centróide) do grupo, aos quais esses pontos pertencem (JAIN; DUBES, 1988). Assim, de um modo geral, os métodos de agrupamento visam organizar objetos em grupos homogêneos. Na prática, a aplicação de técnicas como estas tendem a reduzir o grande número de informações normalmente contidas em extensas massas de dados transacionais.

Já os algoritmos de agrupamento hierárquicos possuem dois tipos básicos de procedimentos, os aglomerativos e divisivos. Nos métodos aglomerativos cada objeto começa como seu próprio agrupamento, já nos métodos divisivos todas as observações

começam com um único agrupamento e são divididas até que cada uma seja um único agrupamento.

Os métodos de agrupamento utilizam, em suma, alguma medida de similaridade ou de dissimilaridade (CHU; TSAI, 1973). Esses algoritmos, quando comparados a outros algoritmos de agrupamento, são de fácil aplicação, pois não exigem muitas manipulações de matrizes de incidência (DUDA; HART, 1973). As medidas de similaridade mais conhecidas são: Coeficiente de Russel e Rao (RUSSEL; RAO, 1940), Dice (DICE, 1945), Sokal e Sneath I(SOKAL; SNEATH, 1963), Sokal e Sneath II (SOKAL; SNEATH1973), Rogers e Taminoto (ROGERS; TAMINOTO, 1960), Simple Matching (SOKAL; MICHENER, 1958) e Jaccard (DUDA; HART, 1973; EVERITT, 1993). Os algoritmos aglomerativos mais utilizados são: Método do vizinho mais próximo, método do vizinho mais distante, método do centróide, método das distâncias médias e ROCK(GUHA et al., 2000). Esse último será foco do assunto estudado nesse trabalho e será visto em mais detalhe nas seções que seguem.

4.2. MEDIDAS DE SIMILARIDADE E DE DISSIMILARIDADE

O primeiro algoritmo que utilizou medidas de similaridade como critério de agrupamento foi publicado em 1973 (CHU; TSAI, 1990). Existem inúmeras medidas de similaridade e com isso um grande número de algoritmos que se baseiam nelas (FILHO; MAESTRELLI; BATOCCHIO, 2002). Esses algoritmos, quando comparados a outros algoritmos de agrupamento, são de fácil aplicação, pois não exigem muitas manipulações de matrizes de incidência.

As medidas de distância podem ser entendidas com uma medida de dissimilaridade, ou seja, que representam as disparidades entre as observações. A maioria dos algoritmos de análise de agrupamento é baseada em medidas de similaridade ou de dissimilaridade. Algumas dessas medidas são descritas a seguir.

4.2.1. DISTÂNCIA EUCLIDIANA

É a medida de dissimilaridade mais utilizada na prática e representa a distância, em linha reta, de duas entidades vetoriais. Desse modo, define-se a Distância Euclidiana entre os indivíduos *a* e *b*, por:

$$d_{ab} = \left[\sum_{j=1}^{p}\left(X_{aj} - X_{bj}\right)^2\right]^{1/2} \qquad [2]$$

em que,

p é o número de componentes do vetor X;
X_{aj} é o valor da *j-ésima* variável para o indivíduo *a;*
X_{bj} é o valor da *j-ésima* variável para o indivíduo *b.*

A expressão (2) na forma matricial torna-se:

$$d_{ab} = [(X_a - X_b)^t(X_a - X_b)] \qquad [3]$$

Sendo,

$X_a = \left[X_{a1}X_{a2} \ldots X_{ap}\right]^t$ vetor de características do indivíduo *a;*
$X_b = \left[X_{b1}X_{b2} \ldots X_{bp}\right]^t$ vetor de características do indivíduo *b.*

Em geral, na maioria das situações práticas, as variáveis são observadas a partir de unidades de medida diferentes. Isso dificulta de certa forma, a identificação de unidades vetoriais similares. Para contornar esse problema, normalmente, lança-se mão de alguma técnica de padronização de dados antes de se calcular as medidas de distância.

Abaixo é mostrada uma possibilidade de padronização de variáveis para o intervalo (0; 1).

$$Z_{ij} = \frac{X_{ij} - \overline{X_j}}{S_j}, \quad Z_{ij} \sim (0, 1_j)$$

ou [4]

$$Z_{ij} = \frac{X_{ij}}{S\,(\,X_j)}, \quad Z_{ij} \sim (\overline{Z_j}, 1)$$

4.2.2. DISTÂNCIA EUCLIDIANA MÉDIA

A distância Euclidiana aumenta à medida que o número de variáveis cresce. Uma maneira de eliminar esse efeito é dividir o valor da distância Euclidiana pela raiz quadrada do número de variáveis. Desse modo,

$$\bar{d}_{ab} = \frac{1}{\sqrt{p}} d_{ab} \qquad [5]$$

em que,

$\bar{d}_{ab}$ é a distância Euclidiana média entre *a* e *b;*

p é o número de variáveis;

d_{ab} é a distância Euclidiana entre *a* e *b.*

4.2.3. DISTÂNCIA DE MAHALANOBIS

A principal desvantagem da distância Euclidiana é que essa pondera as variáveis de maneira uniforme, ou seja, igualitária. Assim, uma medida de distância mais geral seria aquela que ponderasse as variáveis envolvidas no cálculo de forma diferente, isto é, que, pelo menos, fosse proporcional às variabilidades de cada variável.

A medida de distância que desempenha esse papel é denominada de distância estatística ou distância de Mahalanobis (ANDERBERG, 1973; JOBSON, 1991-92; REIS, 1997) que é calculada como segue:

$$D_{ab}^2 = [X_a - X_b]^{\mathrm{t}} S^{-1} [X_a - X_b] \qquad [6]$$

em que,

D_{ab}^2 é a distância de Mahalanobis entre os indivíduos *a* e *b;*

X_a é o vetor de características do individuo *a;*

X_b é o vetor de características do individuo *b;*

S é a matriz de variância-covariância amostral da população.

Além das medidas de distância para variáveis quantitativas, existem algumas propostas de medidas de similaridade e/ou de dissimilaridade apropriadas para variáveis qualitativas (DUDA; HART, 1973; EVERITT, 1993). As mais comuns são: Coeficiente de Concordância Simples ou de Sokal e Coeficiente de Jaccard.

Essas medidas são calculadas com base em uma tabela cruzada de freqüência. Assim, suponha que duas observações sejam caracterizadas por duas variáveis x e y dicotômicas (1 se a variável possuir a característica desejada e 0 caso contrário). A TABELA 1 apresenta tais freqüências.

TABELA 1: Representação das freqüências das variáveis categóricas binárias.

Variável X	Variável Y		**Total**
	1	**0**	
1	a	b	**a + b**
0	c	d	**c + d**
Total	**a + c**	**b + d**	***n* = a + b + c +d**

Para essa representação tabular, o coeficiente de concordância simples ou de Sokal (SOKAL; MICHENER, 1958) é definido por:

$$S = \frac{a + d}{a + b + c + d} \qquad [7]$$

O objetivo deste coeficiente é representar as freqüências das características concordantes de ambas as variáveis. Neste caso, são atribuídos pesos iguais aos pares concordantes “1–1” (presença da característica) e “0–0” (ausência da característica).

Outra alternativa é o Coeficiente de Jaccard que relaciona o par concordante “1–1” com os pares discordantes “1–0” e “0–1”, conforme mostra a expressão abaixo (MEYER, 2002).

$$J = \frac{a}{a + b + c} \qquad [8]$$

Segundo Guha et al. (2000), o Coeficiente de Jaccard mede o grau de concordância entre as observações. Para isso, desconsidera a quantidade d que se refere à soma de todos os pares negativos “0–0”, ou seja, exclui as freqüências em que as características das variáveis de interesse possuem ausência conjunta de atributos. Isto se deve ao fato de que o coeficiente leva em consideração a quantidade de pares coincidentes em relação aos indivíduos que possuem, pelo menos, uma das características desejadas.

Além dos coeficientes citados, a literatura especializada apresenta outras medidas de similaridade para dados categorizados, como mostra a TABELA 2.

TABELA 2: Outros coeficientes de similaridade para dados qualitativos binários.

Coeficiente	Fórmula
Russel–Rao:	$a/(a+b+c+d)$
Czekanowski–Sorensen–Dice:	$2a/(2a+b+c)$
Taminoto–Rogers:	$(a+d)/[(a+d)+2(b+c)]$
Sokal–Sneath:	$2(a+d)/[2(a+d)+b+c]$

4.3. MÉTODOS DE AGRUPAMENTO

As técnicas de agrupamento podem ser classificadas em dois grandes grupos de procedimentos: hierárquicos e não-hierárquicos. Esses procedimentos, naturalmente, produzem resultados ou agrupamentos diferentes e são aplicados em situações específicas.

4.3.1. TÉCNICAS DE AGRUPAMENTO NÃO-HIERÁRQUICO

O que essas técnicas procuram na verdade é a coesão interna dos objetos e isolamento externo entre os grupos (BUSSAB; MIAZAKI; ANDRADE, 1990). As técnicas não hierárquicas foram desenvolvidas para agrupar elementos em k grupos (ALBUQUERQUE, 2005). Uma partição de n objetos exige critérios que produzam medidas de qualidade dos aglomerados formados e pressupõe-se ainda o conhecimento prévio do número k de partições desejadas.

Desse modo, o problema recai, basicamente, em procurar por uma partição dos n objetos em k grupos, de modo que o critério de partição seja adequado. Se forem construídas todas as possíveis partições, seria possível determinar o valor de cada medida e selecionar a melhor partição, mas isso seria inviável devido ao grande número

de partições. O que as técnicas de partição pretendem, então, é investigar algumas dessas partições, procurando uma partição ótima ou quase ótima.

Um dos principais representantes dos métodos não-hierárquicos ou particionais é o Método das K-Médias. Esse método é baseado em análise de variância, cujo critério mais usado é a soma de quadrados residual. Suponha, então, um conjunto de *n* objetos medidos a partir de *p* variáveis x_{ij} com *i*=1, 2, ..., *n* e *j*=1, 2, ..., *p*. Assim, o algoritmo das k-médias aloca cada objeto a um dos k grupos (protótipos), de forma a minimizar a soma dos quadrados dentro dos grupos, ou seja:

$$SQE(k) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij} - \bar{x}_{kj}\right)^2 \qquad [9]$$

em que $\bar{x}_{kj}$ é o centróide do grupo *k*;

O fato da escolha de k ser definido pelo usuário costuma ser um problema, tendo em vista que não se sabe quantos grupos existem a priori (LINDEN, 2009).

Além das K-médias existem outros métodos de agrupamento não-hierárquicos, entre eles, os métodos dos K-medóides, Fuzzy e Redes Neurais Auto-organizáveis (JAIN, 1999).

4.3.2. TÉCNICAS DE AGRUPAMENTO HIERÁRQUICO

As técnicas hierárquicas de agrupamento podem ser subdivididas em dois grupos de procedimentos: os divisivos e os aglomerativos. Os algoritmos divisivos, como o próprio nome sugere, começam com todos os objetos aglomerados em um grupo único que, seqüencialmente, são separados até que cada objeto represente, individualmente, o seu próprio grupo (SOUZA, 2012).

Segundo Kaufman e Rousseeuw (1990), o processo dos algoritmos divisivos inicia-se dividindo todas as observações em dois grupos e, para que isso ocorra, devem-se

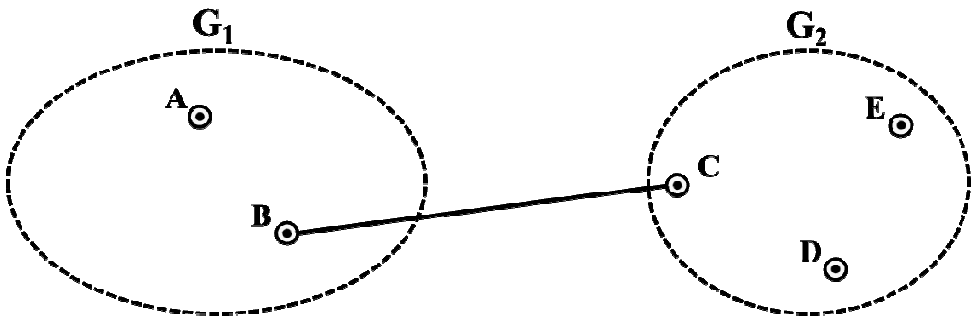
considerar todas as possíveis divisões, o que cria um custo computacional elevado (REIS, 1997). Em grandes bancos de dados essas divisões tornam-se impossíveis de serem realizadas devido ao grande número de combinações.

Everitt (1993) descreve dois tipos de métodos divisivos por meio de exemplos, o monotético e o politético. No primeiro a divisão é feita com base em um único objeto, já no politético a divisão é feita com todos os atributos usados conjuntamente.

Por outro lado, os algoritmos aglomerativos iniciam-se com *n* grupos individuais, de modo que, seqüencialmente, os objetos vão sendo agrupados até que reste um único grupo contendo todos os *n* objetos (TOTTI; VENKOVSKY; BATISTA, 2011). Por serem mais famosos, os algoritmos hierárquicos aglomerativos merecem um detalhamento maior sobre as técnicas utilizadas para a realização das amalgamações, como mostram as explanações que seguem.

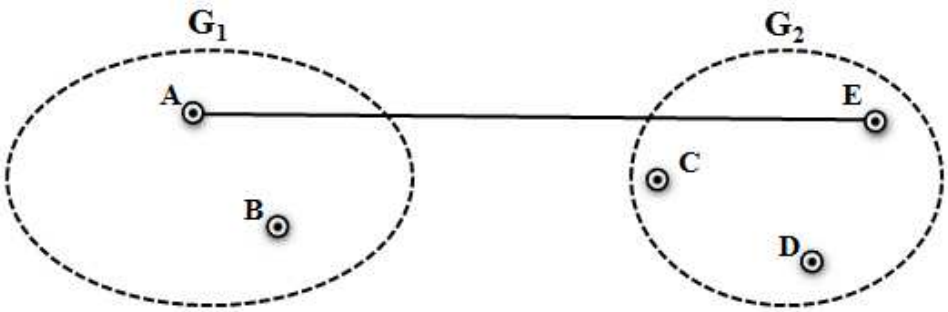
a) Método da Ligação Simples: É também conhecido como método de encadeamento simples "*Single Linkage Method*". Nesse método calcula-se a matriz de distâncias entre os n objetos do grupo, em seguida os objetos mais próximos são agrupados. A FIGURA 1 revela que a distância entre os grupos G_1 e G_2 é igual à distância entre as observações B e C, que são as mais próximas (os vizinhos mais próximos).

FIGURA 1: Método da Ligação Simples ou vizinho mais próximo



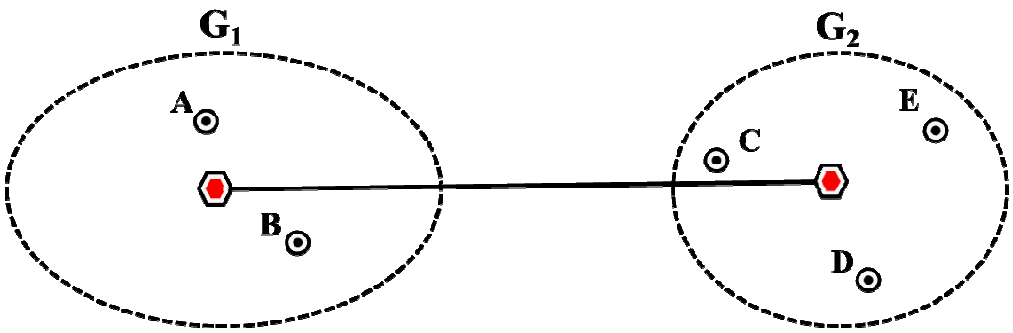
b) Método da Ligação Completa: É também conhecido como método de encadeamento completo "*Complete Linkage Method*". Nesse método calcula-se a matriz de distâncias entre os n objetos do grupo, em seguida os objetos mais dissimilares são agrupados. A FIGURA 2 revela que a distância entre os grupos G_1 e G_2 é igual à distância entre as observações A e E, que são as mais distantes (os vizinhos mais dissimilares).

FIGURA 2: Método da Ligação Completa ou vizinho mais distante



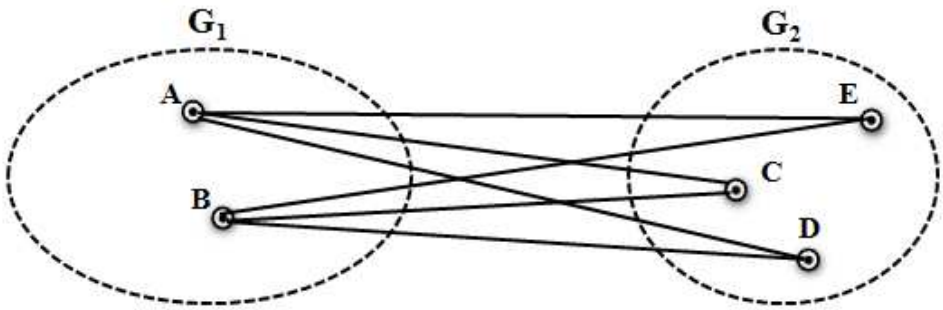
c) Método do Centróide: Esse algoritmo é o mais direto dos métodos, pois ele atribui a cada grupo em um único ponto representado pelas coordenadas de seu centro. A distância entre os grupos é justamente a distância entre os centros. Por fim, procura-se unir grupos que tenham a menor distância entre os seus respectivos centros. A FIGURA 3 mostra que a distância entre os grupos G_1 e G_2 é representada pela distância entre os seus centros.

FIGURA 3: Método Centróide



d) Método das Distâncias Médias: Nesse método a proximidade entre dois grupos é representada pela média das medidas de proximidade entre todas as combinações desses grupos. A FIGURA 4 mostra que a distância entre os grupos G_1 e G_2 é representada pela média das distâncias entre os elementos de ambos os grupos.

FIGURA 4: Método das Distâncias Médias



Como discutido acima, existe uma vasta quantidade de obras e procedimentos para agrupar dados contínuos, mas pouco se estudou sobre técnicas para agrupar dados qualitativos de forma eficiente. Desse reduzido conjunto de procedimentos, o algoritmo ROCK é um dos mais promissores. Por isso, convém discuti-lo mais detalhadamente, o que será feito na próxima seção.

5. ROCK – ROBUST CLUSTERING USING LINKS

5.1. INTRODUÇÃO

O algoritmo RObust Clustering using linKs – ROCK é um procedimento de agrupamento hierárquico aglomerativo que emprega a idéia de *links* para agrupar os elementos em cada fase do processo(JAIN; DUBES, 1988). Esse algoritmo, criado por Guha et al.(2000), é um método de análise multivariada de dados utilizado para agrupar dados descritos por variáveis categorizadas e se baseia no uso de uma medida particular de similaridade para agrupar as observações. Esse procedimento utiliza alguns conceitos importantes, entre eles: as idéias de vizinhança e *links*, função critério e medida de qualidade. Para propiciar um melhor entendimento desses conceitos, uma explicação mais cuidadosa será apresentada nas seções que seguem.

5.2. MEDIDA DE SIMILARIDADE

Para medir o quão próximos são os objetos avaliados, utiliza-se o Coeficiente de Jaccard. Como já comentado, é possível reescrever a equação (8), que se refere ao coeficiente de Jaccard, de uma forma alternativa usando a idéia de união e interseção de conjuntos (equação 10). Assim, sejam T_i e T_j dois vetores de componentes booleanas indicando a escolha de dois indivíduos sobre os atributos de um conjunto de k itens. Assim, o coeficiente de Jaccard, que mede o padrão de similaridade de escolha de dois indivíduos quaisquer (T_i, T_j), poderá ser redefinido da seguinte forma:

$$Sim(\mathrm{T}_i; \mathrm{T}_j) = \frac{|\mathrm{T}_i \cap \mathrm{T}_j|}{|\mathrm{T}_i \cup \mathrm{T}_j|} = \frac{|\mathrm{T}_i \cap \mathrm{T}_j|}{|\mathrm{T}_i| + |\mathrm{T}_j| - |\mathrm{T}_i \cap \mathrm{T}_j|} \qquad [10]$$

em que $0 \leq Sim(\mathrm{T}_i; \mathrm{T}_j) \leq 1$.

Como exemplo, suponha que num conjunto de seis itens (1, 2, 3, 4, 5 e 6), quatro indivíduos escolheram suas preferências, como por exemplo:

$\boldsymbol{I_1}$ – Indivíduo 1**:** {1, 2, 3, 5}; $\boldsymbol{I_2}$ – Indivíduo 2**:** {2, 3, 4, 5};

$\boldsymbol{I_3}$ – Indivíduo 3**:** {1, 4}; $\boldsymbol{I_4}$ – Indivíduo 4**:** {6};

Neste caso, a representação dessas escolhas na escala *booleana* ficaria:

$\boldsymbol{I_1}$: {1, 1, 1, 0, 1, 0}; $\boldsymbol{I_2}$: {0, 1, 1, 1, 1, 0};

$\boldsymbol{I_3}$: {1, 0, 0, 1, 0, 0}; $\boldsymbol{I_4}$: {0, 0, 0, 0, 0, 1};

Para essa situação, o coeficiente de Jaccard seria calculado como:

$$Sim(I_1; I_2) = \frac{|I_1 \cap I_2|}{|I_1 \cup I_2|} = \frac{|I_1 \cap I_2|}{|I_1| + |I_2| - |I_1 \cap I_2|} \quad [11]$$

Desse modo, a medida de similaridade entre os vetores transacionais I_1 e I_2 seria:

$$Sim(I_1; I_2) = \frac{3}{4 + 4 - 3} = \frac{3}{5} = 0{,}6 \quad [12]$$

Para esse exemplo fictício conclui-se que as escolhas dos indivíduos I_1 e I_2 são similares a um grau de 60%.

5.3. VIZINHANÇA E LINKS

Duas observações i e j são consideradas vizinhas se a sua medida de similaridade $Sim(T_i; T_j) \geq \theta \in [0; 1]$, em que θ é um parâmetro ou ponto de corte que representa o padrão de similaridade desejado. Desse modo, a escolha de um valor muito pequeno para θ geraria um número elevado de vizinhos. Por outro lado, a escolha de um valor de θ próximo de 1 exigiria uma semelhança muito forte entre dois objetos para que esses fossem considerados vizinhos. Portanto, valores elevados para θ reduzem o número de objetos vizinhos.

Existem algumas discussões a respeito do valor mais adequado para o parâmetro θ. Alguns autores concordam que esse parâmetro deve pertencer ao intervalo [0, 1], mas não sugerem qual valor adequado para θ. Eles ressaltam ainda que o desempenho do algoritmo ROCK dependa desse parâmetro.

É possível definir *links* como sendo o número de vizinhos comuns entre dois objetos genéricos X_ie X_j, de modo que quanto maior for o número de vizinhos comuns, mais provável é que X_ie X_j pertençam ao mesmo grupo, ou seja, também sejam considerados vizinhos. Esse algoritmo é considerado robusto porque não utiliza a idéia de distância para maximizar as dissimilaridades entre os grupos, mas sim uma função objetivo que se baseia na quantidade de vizinhos.

5.4. FUNÇÃO CRITÉRIO

O interesse é maximizar a soma de *links* ou o número de vizinhos pertencentes a um mesmo grupo e, ao mesmo tempo, minimizar o número de vizinhos em grupos distintos, ou seja, garantir simultaneamente a homogeneidade interna aos grupos e a heterogeneidade entre grupos.

Para isso, a função critério utilizada para maximizar os k grupos é definida por:

$$\mathrm{E}_l = \sum_{i=1}^{\mathrm{k}} n_i \times \frac{\sum_{(\mathrm{X}_i, \mathrm{X}_j) \in \mathrm{G_i}} link(\mathrm{X}_i, \mathrm{X}_j)}{n_i^{1+2\,\mathrm{f}(\theta)}} \qquad [13]$$

em que G_i denota o grupo *i* de tamanho n_i.

Essa função critério garante que as observações similares são atribuídas em um mesmo grupo, embora não obrigue que observações pouco similares sejam divididas em grupos distintos.

5.5. MEDIDA DE QUALIDADE

A medida de qualidade $g(\mathrm{G}_i; \mathrm{G}_j)$ avalia se um dado par de objetos deve ou não ser unido. Assim, aquele par de observações que apresentar o maior valor de $g(\mathrm{G}_i, \mathrm{G}_j)$ será amalgamado. Do ponto de vista matemático, a função gé calculada como segue:

$$g(\mathrm{G}_i; \mathrm{G}_j) = \frac{link[\mathrm{G}_i; \mathrm{G}_j]}{(n_i + n_j)^{1+2f(\theta)} - {n_i}^{1+2f(\theta)} - {n_j}^{1+2f(\theta)}} \qquad [14]$$

em que,

$$link[\mathrm{G}_i; \mathrm{G}_j] = \sum_{\mathrm{X}_i \in \mathrm{G}_i,\ \mathrm{X}_j \in \mathrm{G}_j} link(\mathrm{X}_i; \mathrm{X}_j) \qquad [15]$$

Assim, o denominador representa o número esperado de *links* cruzados entre os grupos comparados. Isto se faz necessário para que os grupos com poucos *links* sejam mantidos separados.

5.6. PASSOS PARA A EXECUÇÃO DO ALGORITMO ROCK

O algoritmo é executado a partir seguintes etapas:

1º PASSO: Definir os parâmetros de entrada: Fixar o parâmetro θ e definir o número de grupos (k) a ser obtido;

2º PASSO: Calcular a matriz de proximidade D baseada no coeficiente de similaridade de Jaccard;

3º PASSO: Definir a matriz de vizinhos (A), comparando cada observação da matriz de proximidade com θ;

4º PASSO: Definir a matriz de *links* (L), que pode ser obtida facilmente fazendo $L = A'A$;

5º PASSO: Determinar medida de qualidade G para cada par de pontos;

6º PASSO: As observações ou grupos com maior valor de G serão unidas em um único grupo;

7º PASSO: Recalcular a matriz de *links*, com $n - 1$ grupos em relação ao passo anterior;

8º PASSO: Repetir a partir do 5º passo até que se consiga agrupar adequadamente todas as observações ou grupos, atendendo ao número desejado de grupos ou até que a função critério seja nula.

5.7. VANTAGENS E DESVANTAGENS DO ROCK

Todo e qualquer procedimento de agrupamento tem suas vantagens e desvantagens. Entre as vantagens, o algoritmo ROCK trabalha bem com dados categorizados, utiliza o conceito de links ao invés de uma medida de distância e gera agrupamentos de melhor

qualidade que outros métodos. Por outro lado, dentre as suas limitações estão: o uso de matrizes esparsas para armazenar as ligações dos grupos, o que provoca uma queda da eficiência do algoritmo; a medida de similaridade é calculada a partir do Coeficiente de Jaccard e, por fim, o cálculo da função de adequação (g) é fortemente dependente do tamanho da base de dados.

5.8. EXEMPLO PRÁTICO DE APLICAÇÃO DO ALGORITMO

O exemplo a seguir é baseado no apresentado em Pantuzzo (2002). Suponha uma matriz X contendo 6 observações e 5 variáveis binárias. O objetivo é utilizar o algoritmo ROCK para agrupar as observações em três grupos considerando $\theta = 0{,}5$.

$$X = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

1º Passo– Fixando $\theta = 0{,}5$ e k = 6.

2º Passo– Matriz de proximidade utilizando o coeficiente de Jaccard

$$D = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 1{,}0000000 & 0{,}6666667 & 0{,}3333333 & 0{,}6666667 & 0{,}0000000 & 0{,}3333333 \\ 0{,}6666667 & 1{,}0000000 & 0{,}2500000 & 0{,}5000000 & 0{,}2000000 & 0{,}6666667 \\ 0{,}3333333 & 0{,}2500000 & 1{,}0000000 & 0{,}2500000 & 0{,}2500000 & 0{,}3333333 \\ 0{,}6666667 & 0{,}5000000 & 0{,}2500000 & 1{,}0000000 & 0{,}2000000 & 0{,}2500000 \\ 0{,}0000000 & 0{,}2000000 & 0{,}2500000 & 0{,}2000000 & 1{,}0000000 & 0{,}2500000 \\ 0{,}3333333 & 0{,}6666667 & 0{,}3333333 & 0{,}2500000 & 0{,}2500000 & 1{,}0000000 \end{pmatrix}$$

3º Passo– Matriz de Vizinhos

$$A = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

4º Passo– Matriz de *links*

$$L = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 0 & 3 & 0 & 3 & 0 & 1 \\ 3 & 0 & 0 & 3 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 3 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 \end{pmatrix}$$

5º Passo– Medida de qualidade

$$G = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 0{,}0000000 & 2{,}4702910 & 0{,}0000000 & 2{,}4702910 & 0{,}0000000 & 0{,}8234303 \\ 2{,}4702910 & 0{,}0000000 & 0{,}0000000 & 2{,}4702910 & 0{,}0000000 & 1{,}6468606 \\ 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 \\ 2{,}4702910 & 2{,}4702910 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}8234303 \\ 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 \\ 0{,}8234303 & 1{,}6468606 & 0{,}0000000 & 0{,}8234303 & 0{,}0000000 & 0{,}0000000 \end{pmatrix}$$

6º Passo- União de grupos

Grupos unidos: 1 (*observação* 1) e 2 (*observação* 2)

7º Passo– Recalculando a matriz de *links*, agora com 5 grupos

$$L = \begin{matrix} \{1;2\} \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 0 & 0 & 6 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 \end{pmatrix}$$

8º Passo– Repetindo o 5º passo e redefinindo a nova matriz G

$$G = \begin{matrix} \{1;2\} \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 0{,}0000000 & 0{,}0000000 & 2{,}7910490 & 0{,}0000000 & 1{,}3955245 \\ 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 \\ 2{,}7910490 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}8234303 \\ 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 \\ 1{,}3955245 & 0{,}0000000 & 0{,}8234303 & 0\ {,}0000000 & 0{,}0000000 \end{pmatrix}$$

- Unindo novamente os grupos: Grupo 1 (*observações* 1 e 2) e 3 (*observação* 4)

- Recalculando a matriz de *links*, agora com 4 grupos

$$L = \begin{matrix} \{1;2;4\} \\ 3 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{pmatrix}$$

Como ainda existem grupos para agrupar voltamos ao 5º passo:

- Matriz de qualidade G

$$G = \begin{matrix} \{1;2;4\} \\ 3 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 1{,}3475224 \\ 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 \\ 0{,}0000000 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 \\ 1{,}3475224 & 0{,}0000000 & 0{,}0000000 & 0{,}0000000 \end{pmatrix}$$

- Grupos unidos: 1 (*observações* 1, 2 e 4) e 4 (*observação* 6)

- Recalculando a matriz de *links*

$$L = \begin{matrix} \{1;2;4;6\} \\ 3 \\ 5 \end{matrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

- Por fim,

$$G = \begin{matrix} \{1;2;4;6\} \\ 3 \\ 5 \end{matrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Repete-se o algoritmo até que a função G seja nula ou até que o número desejado de grupos seja atingido. O resultado final do algoritmo ROCK contemplou três grupos com as seguintes observações:

G_1: {1; 2; 4 e 6}; **G_2**: {3}; **G_3**: {5}

5.9. IMPLEMENTAÇÃO DO ALGORITMO ROCK

Uma versão do algoritmo ROCK foi implementada no R (R Development Core Team, 2009) e se baseou na seguinte estrutura computacional:

Dados de Entrada:
S: Matriz de dados amostrais (booleanos);
θ: Parâmetro de similaridade;
k: Número desejado de grupos;

Procedimento ROCK Clustering (S; θ; k)

INÍCIO

1. Computar os vizinhos: ***link***<- compute *links*(S);
2. Para cada s ∈ S faça:
    3. Construir a lista local de vizinhos: **q[s]**<- (link [s,j]>0;s);
4. Construir a lista global de vizinhos: **G**<- *g*(s ;max(q[s]));
5. Enquanto Tamanho(G)> k faça
   {
    6. Extrair o máximo de G: **u**<- max(G);
    7. Encontrar o melhor objeto para amalgamar: **v**<- max(q[u]);
    8. Excluir o elemento v de G: deletar (Q ; v);
    9. Agrupar u com v: **w**<- agrupar(u ; v);
    10. Para cada $x \in$ q[u]∪q[v] faça # atualiza q para cada *x*
    {
    11. *link*[x,w] <- *link*[x,u] + *link*[x,v];

12. Excluir *link*[q[x],u] e *link*[q[x],v];
13. Inserir (q[x],w,*g*(x,w)) e (q[w],x,*g*(x,w)); # criar q para w
14. Atualizar a matriz Q: (Q, *x*, q[x]); # Atualiza Q
}
15. Inserir (Q, w, q[w]); # Atualiza Q
16. Retirar q[u] e q[v] da lista local e atualizá-la;
}

FIM

6. **BASES DE DADOS**

O emprego do algoritmo ROCK será efetuado em duas bases de dados bem distintas. A primeira base de dados se refere a uma lista de estabelecimentos de saúde de João Pessoa. Esse banco de dados, que contemplou 7 procedimentos (TABELA 3), foi gerado a partir do pacote computacional Tabwin que integra o sistema DATASUS – Ministério da Saúde. Como arquivo produzido considerou o número de procedimentos realizados em 2012, para habilitar o uso do algoritmo ROCK, o seguinte processo de dicotomização foi empregado para cada procedimento (codificado como P_j):

$$X_j = \begin{cases} 1 \ ; se P_j > 0 \\ \\ 0 \ ; \ c.c. \end{cases} \quad ; \ j = 1,2,\cdots 7 \qquad [16]$$

As informações filtradas da base de dados, e de interesse desse estudo, foram somente àquelas relativas à cidade de João Pessoa no ano de 2012, por procedimentos. Foram selecionados 271 estabelecimentos de saúde que realizam procedimentos ambulatoriais. Os procedimentos foram subdivididos como descrito na TABELA 3 e a arquitetura geral do banco de dados está representada na FIGURA 5.

Além disso, para viabilizar a avaliação da qualidade do algoritmo ROCK, foi sugerido um agrupamento natural dos estabelecimentos de saúde em três categorias: PSF, Clínicas e Hospitais. Assim, uma vez gerados os agrupamentos pelo algoritmo ROCK, será possível compará-los com os grupos naturais da base de dados.

TABELA 3: Procedimentos e codificação

Procedimentos	**Código**
Ações de promoção e prevenção em saúde	X_1
Procedimentos com finalidade diagnóstica	X_2
Procedimentos clínicos	X_3
Procedimentos cirúrgicos	X_4
Transplantes de órgãos, tecidos e células	X_5
Órteses, próteses e materiais especiais	X_6
Ações complementares da atenção à saúde	X_7

FIGURA 5: Estrutura geral da primeira base de dados

ID	ESTABELECIMENTO	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	2341417 NEFRUZA	0	1	1	1	0	1	0
2	2343142 UNIDADE DO PSF COMENDADOR SANTOS COELHO	1	1	1	1	0	0	1
3	2343479 FUNAD	0	1	1	0	0	0	0
4	2356651 UNIDADE DO PSF PEDRA BRANCA I	1	1	1	1	0	0	1
5	2356678 UNIDADE DO PSF MUSSUMAGO I	1	1	1	1	0	0	1
6	2356775 USF ALDEIA SOS	1	1	1	1	0	0	1
7	2357135 UNIDADE DO PSF PADRE HILDON BANDEIRA	1	1	1	1	0	0	1
8	2357186 UNIDADE DO PSF MANGABEIRA POR DENTRO	1	1	1	1	0	0	1
9	2357194 UNIDADE DO PSF TIMBO I	1	1	1	1	0	0	1
10	2357224 UNIDADE MOVEL DE OFTALMOLOGIA	0	1	1	0	0	1	0
11	2357240 UNIDADE DO PSF RIACHO DOCE MARCONARIA	1	1	1	1	0	0	1
12	2357348 CONE CENTRO MEDICO DO NORDESTE	0	0	0	1	0	0	0
13	2357364 CLINICA DE OLHOS PARAIBANA	0	0	1	1	0	0	0
14	2357402 CENTRO MEDICO AUDIOVISUAL LTDA	0	1	1	1	0	0	0
15	2357518 CEDIPA	0	1	0	1	0	0	0
16	2357569 LABORATORIO DR WALDEREDO NUNES	0	1	0	0	0	0	0
17	2357623 ECOCLINICA	0	1	0	0	0	0	0
18	2398915 CAIS CRISTO REDENTOR	1	1	1	1	0	0	1
19	2398923 UNIDADE DO PSF GROTAO II	1	1	1	1	0	0	1
20	2398931 UNIDADE DO PSF BELA VISTA I	1	1	1	1	0	0	1
21	2398974 LACEN MUNICIPAL LABORATORIO CENTRAL	0	1	0	0	0	0	0
22	2399008 CLINICA DE REUMATOLOGIA REAB LTDA	0	0	1	0	0	0	0
23	2399016 UNIDADE DO PSF GROTAO I	1	1	1	1	0	0	1
24	2399024 LABORATORIO PASTEUR	0	1	0	0	0	0	0
25	2399032 LABORATORIO VALDEVINO	0	1	0	0	0	0	0
26	2399075 UNIDADE DO PSF SAO JOSE II	1	1	1	1	0	0	1
27	2399180 LABORATORIO IVAN RODRIGUES	0	1	0	0	0	0	0
28	2399202 LABORATORIO GILSON GUEDES	0	1	0	0	0	0	0
29	2399210 UNIDADE DO PSF MANGABEIRA IV AMBULANTES I	1	1	1	1	0	0	1
30	2399229 CENTRO DE SAUDE DE MANDACARU	1	1	1	1	0	0	1
31	2399237 HOSPITAL SAO LUIZ	0	0	1	0	0	0	0
32	2399253 UNIDADE DE SAUDE DAS PRAIAS MARIA ALICE M B CAVALCANTI	1	1	1	1	0	0	1
33	2399261 CENTRO DE SAUDE LOURIVAL GOUVEIA MOURA	1	0	1	1	0	0	1
34	2399318 HOSPITAL INFANTIL ARLINDA MARQUES	0	1	1	1	0	0	0
35	2399326 UNIDADE DO PSF JARDIM MIRAMAR I	1	1	1	1	0	0	1
36	2399334 HOSPITAL JOAO SOARES	0	0	1	0	0	0	0
37	2399342 CENTRO DE SAUDE MARIA LUIZA TARGINO	1	1	1	1	0	0	1
38	2399350 LACEN ESTADUAL LABORATORIO CENTRAL DE SAUDE PUBLICA	0	1	0	0	0	0	0
	[illegible] HOSPITAL RODRIG[illegible]		0	1				
	[illegible]			1				
	[illegible]							
238	[illegible] II	1			1	0		1
239	3446182 UNIDADE DO PSF JOSE AMERICO III	1	1	1	1	0	0	1
240	3446190 UNIDADE DO PSF MANDACARU IX	1	1	1	1	0	0	1
241	3446204 UNIDADE DO PSF MANGABEIRA VI 1 ETAPA	1	1	1	1	0	0	1
242	3446212 UNIDADE DO PSF MANGABEIRA VII A	1	1	1	1	0	0	1
243	3446220 UNIDADE DO PSF MANGABEIRA VII B	1	1	1	1	0	0	1
244	3446239 UNIDADE DO PSF MANGABEIRA VII C	1	1	1	1	0	0	1
245	3446247 UNIDADE DO PSF PARATIBE II	1	1	1	1	0	0	1
246	3446255 UNIDADE DO PSF PEDRO LINS	1	1	1	1	0	0	1
247	3446263 UNIDADE DO PSF PROJETO MARIZ	1	1	1	1	0	0	1
248	3446271 UNIDADE DO PSF PROSIND II	1	1	1	1	0	0	1
249	3446298 UNIDADE DO PSF RANGEL IV	1	1	1	1	0	0	1
250	3446301 UNIDADE DO PSF VALE VERDE	1	1	1	1	0	0	1
251	3446328 UNIDADE DO PSF VALENTINA I	1	1	1	1	0	0	1
252	3446336 UNIDADE DO PSF VALENTINA II	1	1	1	1	0	0	1
253	3446344 UNIDADE DO PSF VALENTINA III	1	1	1	1	0	0	1
254	3446352 UNIDADE DO PSF VALENTINA IV	1	1	1	1	0	0	1
255	3524566 CENTRO DE ESPECIALIDADES ODONTOLOGICAS CEO	1	1	1	1	0	1	0
256	3617742 UNIDADE DO PSF CRUZ DAS ARMAS IX	1	1	1	1	0	0	1
257	3651118 SERVICO DE ATENDIMENTO MOVEL DE URGENCIA SAMU 192	0	0	1	0	0	0	0
258	3854612 UNIDADE DO PSF FUNCIONARIOS II 1 ETAPA	1	1	1	1	0	0	1
259	3871002 CENTRO DE REF MUNICIPAL DE INCLUSAO P PESSOAS C DEFICIENCIA	1	1	1	1	0	0	0
260	5283477 USF BAIRRO DAS INDUSTRIAS III CIDADE VERDE II	1	1	1	1	0	0	1
261	5326672 CEREST	0	0	1	0	0	0	0
262	5631742 CENTRO DE ATENCAO INTEGRAL A SAUDE DO IDOSO	1	1	1	1	0	0	0
263	5654319 INSTITUTO DO CORACAO DO ESTADO DA PARAIBA	0	1	1	0	0	0	0
264	5842026 OFTALMOCLINICA SAULO FREIRE LTDA	0	0	1	1	0	0	0
265	5971136 CAPSI INFANTO JUVENIL CIRANDAR	0	0	1	0	0	0	0
266	6352677 AMIP PRAIA	0	1	1	1	0	1	0
267	6442390 ASSOCIACAO PESTALOZZI DA PARAIBA	0	1	1	0	0	0	0
268	6442862 CENTRO DE OLHOS DA PARAIBA	0	0	1	0	0	0	0
269	6509258 CAPS AD DAVID CAPISTRANO	1	1	1	1	0	0	0
270	6525938 SECRETARIA MUNICIPAL DE SAUDE DE JOAO PESSOA	1	0	0	0	0	0	0
271	6940315 UPA OCEANIA	0	1	1	1	0	0	0

Já a segunda base de dados se refere a um recorte dos registros de homicídios, de autoria inicialmente desconhecida, ocorridos em Campina Grande - PB no período de 2008 a 2011. Nesse banco de dados foram levantadas as condições de ocorrência dos crimes, bem como as características das vítimas. O banco de dados contém 130 observações, sendo a motivação do homicídio a nossa variável de grupo.

Essa base de dados é um pouco diferente da primeira, pois é composta apenas de variáveis qualitativas (FIGURA 6) e, nesse caso, o processo de dicotomização das mesmas será realizado através da construção de variáveis tipo *dummy*.

FIGURA 6: Estrutura geral da segunda base de dados

ID	MOTIVAÇÃO (MT)	TIPO DE ARMA (TA)	NÚMERO DE LESÕES (NL)	VÍTIMA ALCOOLIZADA (VA)	VÍTIMA DROGADA (VD)	ANTECEDENTES CRIMINAIS (AC)	SITUAÇÃO CARCERÁRIA (SC)
1	Fúteis	Arma branca	Múltipla	Não	Não	Não	Nunca fora preso
2	Fúteis	Arma branca	Múltipla	Não	Não	Não	Nunca fora preso
3	Fúteis	Arma branca	Única	Não	Não	Não	Nunca fora preso
4	Fúteis	Arma branca	Única	Não	Não	Não	Nunca fora preso
5	Fúteis	Arma de fogo	Única	Não	Não	Não	Nunca fora preso
6	Fúteis	Arma de fogo	Múltipla	Não	Não	Não	Nunca fora preso
7	Fúteis	Arma de fogo	Múltipla	Sim	Não	Não	Nunca fora preso
8	Fúteis	Arma de fogo	Múltipla	Não	Não	Não	Nunca fora preso
9	Fúteis	Arma de fogo	Múltipla	Sim	Não	Não	Nunca fora preso
10	Fúteis	Arma branca	Múltipla	Sim	Não	Não	Nunca fora preso
11	Fúteis	Arma de fogo	Múltipla	Não	Sim	Não	Nunca fora preso
12	Fúteis	Arma de fogo	Múltipla	Não	Sim	Não	Nunca fora preso
13	Fúteis	Arma de fogo	Múltipla	Sim	Não	Não	Nunca fora preso
14	Fúteis	Arma de fogo	Múltipla	Sim	Não	Não	Nunca fora preso
15	Fúteis	Arma de fogo	Múltipla	Não	Não	Não	Nunca fora preso
16	Fúteis	Arma de fogo	Múltipla	Sim	Sim	Sim	Nunca fora preso
17	Fúteis	Arma de fogo	Múltipla	Não	Não	Sim	Presidiário/Ex-presidiário
18	Fúteis	Arma de fogo	Única	Sim	Não	Não	Nunca fora preso
	Fúteis		Múltipla		Não	Não	fora preso
107	Vingança	Arma de fogo	Múltipla	Não	Não	Não	Nunca fora preso
108	Vingança	Arma de fogo	Múltipla	Não	Não	Não	Nunca fora preso
109	Vingança	Arma de fogo	Múltipla	Não	Sim	Sim	Presidiário/Ex-presidiário
110	Vingança	Arma de fogo	Múltipla	Não	Não	Não	Nunca fora preso
111	Vingança	Arma de fogo	Múltipla	Não	Não	Sim	Presidiário/Ex-presidiário
112	Vingança	Arma de fogo	Múltipla	Não	Não	Não	Nunca fora preso
113	Vingança	Arma de fogo	Múltipla	Não	Sim	Sim	Presidiário/Ex-presidiário
114	Vingança	Arma de fogo	Múltipla	Não	Sim	Sim	Presidiário/Ex-presidiário
115	Vingança	Arma de fogo	Múltipla	Não	Não	Sim	Presidiário/Ex-presidiário
116	Vingança	Arma de fogo	Múltipla	Não	Não	Não	Nunca fora preso
117	Vingança	Arma de fogo	Múltipla	Não	Sim	Sim	Presidiário/Ex-presidiário
118	Vingança	Arma de fogo	Múltipla	Não	Sim	Sim	Presidiário/Ex-presidiário
119	Vingança	Arma de fogo	Múltipla	Não	Não	Sim	Nunca fora preso
120	Vingança	Arma de fogo	Múltipla	Não	Sim	Sim	Presidiário/Ex-presidiário
121	Vingança	Arma de fogo	Múltipla	Não	Não	Sim	Presidiário/Ex-presidiário
122	Vingança	Arma de fogo	Múltipla	Não	Sim	Sim	Presidiário/Ex-presidiário
123	Vingança	Arma de fogo	Múltipla	Não	Não	Não	Nunca fora preso
124	Vingança	Arma de fogo	Múltipla	Não	Não	Não	Nunca fora preso
125	Vingança	Arma de fogo	Única	Não	Não	Não	Nunca fora preso
126	Vingança	Arma de fogo	Única	Não	Sim	Sim	Presidiário/Ex-presidiário
127	Vingança	Arma de fogo	Múltipla	Não	Não	Não	Nunca fora preso
128	Vingança	Arma de fogo	Múltipla	Não	Sim	Sim	Presidiário/Ex-presidiário
129	Vingança	Arma de fogo	Múltipla	Não	Não	Sim	Presidiário/Ex-presidiário
130	Vingança	Arma de fogo	Múltipla	Não	Sim	Sim	Menor

7. RESULTADOS E DISCUSSÃO

7.1. APLICAÇÃO 1

No sentido de compreender a disposição dos grupos existentes faz-se necessária a realização de uma breve análise descritiva das informações. Assim, da base de dados original, destaca-se a unidade do PSF Feirinha e Cidade Verde II que realizaram a mesma quantidade de ações de promoção e prevenção em saúde, 10.470 cada. O Centro Médico do Nordeste – CONE foi o único estabelecimento que realizou apenas um procedimento cirúrgico enquanto que os outros estabelecimentos realizaram até 13.398 cirurgias, de um total de 183.367. Observou-se também que o Hospital Unimed João Pessoa, São Vicente de Paula e o NEO foram os únicos que realizaram transplante de órgãos, tecidos e células, com 763, 470 e 161 transplantes, respectivamente.

Desse modo, conforme TABELA 4, foram realizadas, em média, aproximadamente 10.700 ações de promoção e prevenção em saúde nos PSF's, totalizando a maioria dos estabelecimentos que realizam esse procedimento. Fica fácil entender esse resultado, sabendo que a finalidade do PSF é realizar esse tipo de ação.

A especialidade dos procedimentos com finalidade diagnóstica é dos Hospitais e das Clínicas. Os Hospitais realizam, em média, três vezes mais procedimentos em relação às Clinicas e, quando comparados a média dos PSF's, realizam mais de 40% dos procedimentos.

Para os procedimentos clínicos, é possível observar que os PSF's realizaram, cerca de 2.000 procedimentos a mais que as Clínicas. Neste caso, destacam-se a média de procedimentos clínicos realizados pelos Hospitais, 95.356,05 procedimentos. Nos procedimentos cirúrgicos, os Hospitais realizaram uma média de 4.702,17 cirurgias, enquanto que a média dos procedimentos realizados em Clínicas foi de aproximadamente 1.062 cirurgias e, a metade disso, cerca de 585 cirurgias foram realizadas em PSF.

Como já comentado, apenas uma Clínica realizou 161 transplantes de órgãos, tecido e células e dois Hospitais realizaram, em média, 616,50 transplantes.

Para órteses, próteses e materiais especiais, quem mais realizou esse tipo de procedimento foram os Hospitais, com média de 5.408,75 procedimentos, seguido das

Clínicas, com uma média de 1.046,33 órteses e, finalmente, os PSF's que realizaram em média 557,67 procedimentos.

Por fim, foram realizadas, em média, 23,59 ações complementares da atenção a saúde nos PSF's e, também em média, 125 procedimentos realizados em Hospitais.

TABELA 4: Resumo estatístico do número de procedimentos em 2012

Procedimentos	Tipo de Estabelecimento						Total	
	PSF		Clínicas		Hospitais			
	Média	DP	Média	DP	Média	DP	Média	DP
Ações de promoção e prevenção em saúde	10696,34	23278,84	1118,00	--	2485,00	3242,79	10565,02	23126,57
Procedimentos com finalidade diagnóstica	2846,59	9976,06	32976,28	80912,73	114815,94	105568,96	13637,16	48078,90
Procedimentos clínicos	16782,94	29314,48	14598,08	14179,89	95356,05	116496,28	22809,63	47049,11
Procedimentos cirúrgicos	585,60	1138,68	1062,42	1913,44	4702,17	4836,23	837,29	1844,04
Transplantes de órgãos, tecidos e células	--	--	161,00	--	616,50	207,18	464,67	301,04
Órteses, próteses e materiais especiais	557,67	727,15	1046,33	987,12	5408,75	9621,02	2644,70	6073,53
Ações complementares da atenção à saúde	23,59	12,57	--	--	125,00	--	24,13	14,55

Para as analises dos grupos formados pelo algoritmo ROCK foram considerados três quantidades distintas de graus de vizinhança: $\theta = 0{,}4$; $\theta = 0{,}6$ e $\theta = 0{,}8$. Lembrando que quanto maior o grau de vizinhança (θ) mais difícil será a formação de grupos compostos por muitos vizinhos, uma vez que o coeficiente de Jaccard mede o grau de similaridade entre as observações, em um intervalo [0, 1], e θ é o ponto de corte para se determinar quais observações serão consideradas vizinhas.

Assim, a TABELA 5 apresenta os resultados obtidos segundo o algoritmo ROCK para a primeira base de dados. Percebe-se que para $\theta = 0{,}4$ o ROCK, no geral, agrupou corretamente cerca de 86% das observações, chegando a acertar 100% dos estabelecimentos classificados originalmente como PSF. Além disso, os estabelecimentos classificados como Clínicas, o algoritmo proposto também demonstrou um bom desempenho, acertando cerca de 45% das observações. Entretanto, para os Hospitais foi que o método não se mostrou eficiente, pois agrupou corretamente apenas 5% dos casos.

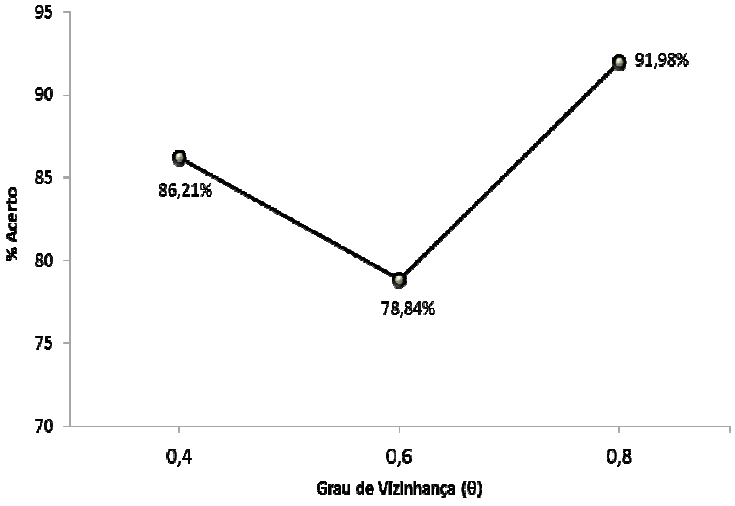
No segundo grau de vizinhança escolhido ($\theta = 0{,}6$) o percentual de acerto diminuiu para aproximadamente 78% no agrupamento geral dos casos, embora tenha agrupado corretamente cerca de 90% dos PSF's.

Por fim, para $\theta = 0{,}8$, observa-se um aumento substancial no desempenho do algoritmo ROCK, já que agrupou corretamente mais de 96% dos psf's e quase 81% das clinicas. Isso demonstra claramente que o grau de vizinhança $\theta = 0{,}8$ gerou uma melhor qualidade de agrupamento para esse banco de dados especificamente. A FIGURA 8 mostra o percentual de acertos classificados pelo ROCK nos três graus de vizinhança estudados.

TABELA 5: Análise do desempenho do Algoritmo ROCK para base de dados 1

Vizinhança	Classificação Original	Classificação ROCK				% Acerto
		PSF	Clínica	Hospital	**Total**	
θ =0,4	PSF	210	0	0	**210**	**86,21%**
	Clínica	14	15	4	**33**	
	Hospital	17	1	0	**18**	
	Total	**241**	**16**	**4**	**261**	
θ =0,6	PSF	183	22	0	**205**	**78,84%**
	Clínica	0	6	21	**27**	
	Hospital	0	8	1	**9**	
	Total	**183**	**36**	**22**	**241**	
θ =0,8	PSF	193	0	7	**200**	**91,98%**
	Clínica	2	21	3	**26**	
	Hospital	6	1	4	**11**	
	Total	**201**	**22**	**14**	**237**	

FIGURA 7: Comparação do % de acerto – Base de Dados 1



7.2. APLICAÇÃO 2

Inicialmente, foi realizada uma análise descritiva da caracterização das variáveis estudadas.

A TABELA 6 apresenta as freqüências de cada variável caracterizada. Observa-se que 31,30% dos homicídios são motivados pelo tráfico de drogas, seguidos de vingança, motivos fúteis e passionais, respectivamente 28,24%, 22,90% e 17,56%.

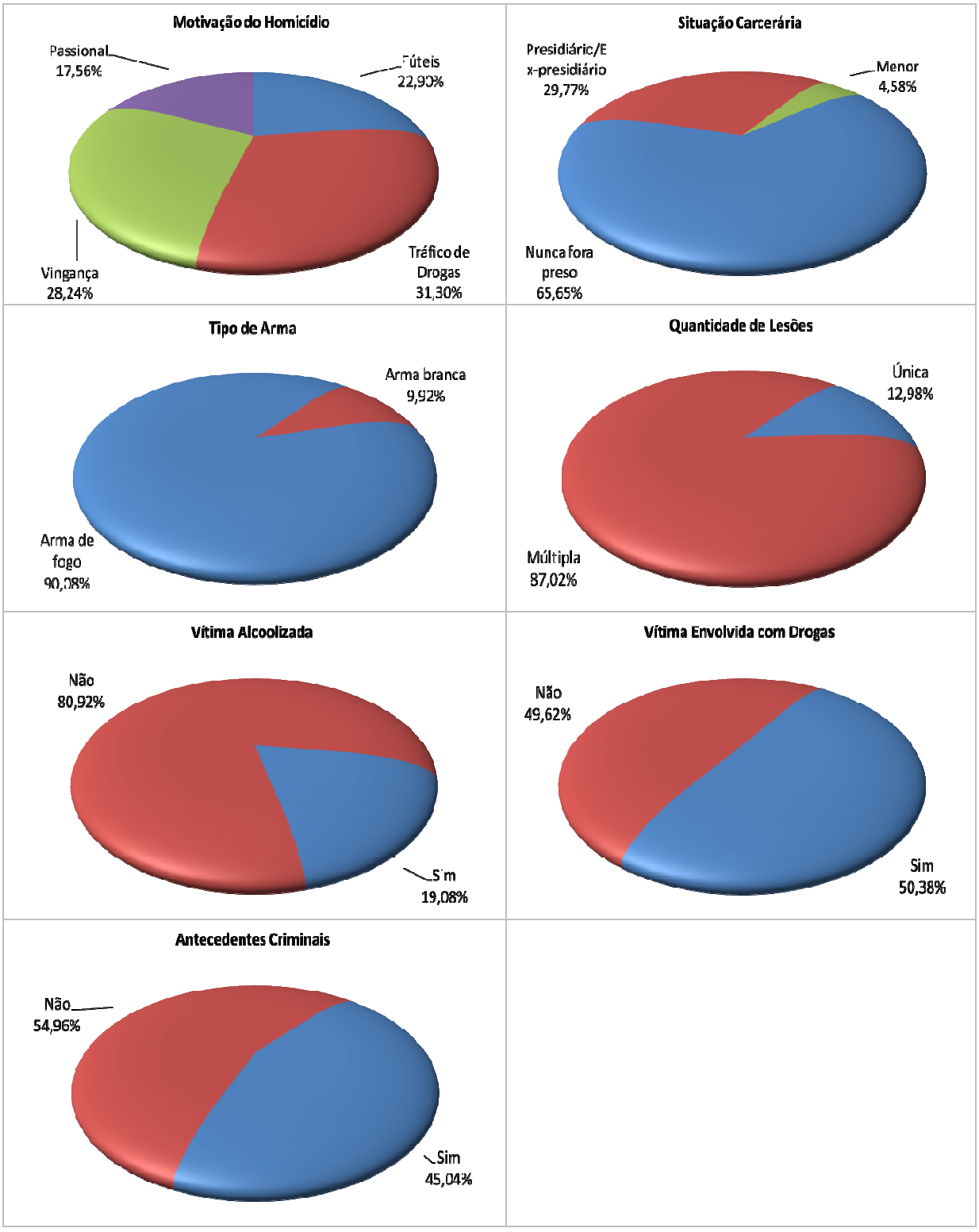
Da situação carcerária, a maioria dos crimes foi praticado por pessoas que nunca foram presas, 65,65%. Presidiários/Ex-presidiário cometeram quase 30% dos homicídios e o restante foram cometidos por menores de idade.

A grande maioria dos homicídios é cometida com o uso de armas de fogo, deixando múltiplas lesões e as vítimas não estavam alcoolizadas, mas 50% delas estão envolvidas com drogas. É possível afirmar ainda que 55% das pessoas que cometeram homicídio não têm antecedentes criminais. Fica fácil a interpretação utilizando a FIGURA 9 como visualização gráfica para tal estudo.

TABELA 6: Caracterização das variáveis para a segunda base de dados

Caracterização	Freq.	%
Motivação		
Fúteis	30	22,90
Tráfico de Drogas	41	31,30
Vingança	37	28,24
Passional	23	17,56
Situação Carcerária		
Nunca fora preso	86	65,65
Presidiário/Ex-presidiário	39	29,77
Menor	6	4,58
Tipo da Arma		
Arma de fogo	118	90,08
Arma branca	13	9,92
Quantidade de Lesões		
Única	17	12,98
Múltipla	114	87,02
Vítima Alcoolizada		
Sim	25	19,08
Não	106	80,92
Vítima envolvida com Drogas		
Sim	66	50,38
Não	65	49,62
Antecedentes Criminais		
Sim	59	45,04
Não	72	54,96

FIGURA 8: Visualização gráfica da caracterização das variáveis da base de dados2



Após a apresentação do resumo descritivo dos dados relacionados aos homicídios, é o momento de investigar a qualidade dos agrupamentos formados. Assim, análogo ao

que foi realizado para a primeira base de dados, foi aplicado o algoritmo ROCK para agrupar os casos de homicídios baseado nos sete atributos booleanos mostrado na FIGURA 9.

FIGURA 9: Estrutura geral da segunda base de dados – versão dicotomizada

ID	MT	TA	NL	VA	VD	AC	SC1	SC2
1	Fúteis	0	1	0	0	0	1	0
2	Fúteis	0	1	0	0	0	1	0
3	Fúteis	0	0	0	0	0	1	0
4	Fúteis	0	0	0	0	0	1	0
5	Fúteis	1	0	0	0	0	1	0
6	Fúteis	1	1	0	0	0	1	0
7	Fúteis	1	1	1	0	0	1	0
8	Fúteis	1	1	0	0	0	1	0
9	Fúteis	1	1	1	0	0	1	0
10	Fúteis	0	1	1	0	0	1	0
11	Fúteis	1	1	0	1	0	1	0
12	Fúteis	1	1	0	1	0	1	0
13	Fúteis	1	1	1	0	0	1	0
14	Fúteis	1	1	1	0	0	1	0
15	Fúteis	1	1	0	0	0	1	0
	Fúteis	1			1		1	
118	Vingança		1	0		1	0	1
119	Vingança	1	1	0	0	1	1	0
120	Vingança	1	1	0	1	1	0	1
121	Vingança	1	1	0	0	1	0	1
122	Vingança	1	1	0	1	1	0	1
123	Vingança	1	1	0	0	0	1	0
124	Vingança	1	1	0	0	0	1	0
125	Vingança	1	0	0	0	0	1	0
126	Vingança	1	0	0	1	1	0	1
127	Vingança	1	1	0	0	0	1	0
128	Vingança	1	1	0	1	1	0	1
129	Vingança	1	1	0	0	1	0	1
130	Vingança	1	1	0	1	1	0	0

Para facilitar a interpretação dos resultados, a variável “Motivação do Homicídio” que é representada pelas categorias “Fúteis”, “Tráfico de Drogas”, “Vingança” e

“Passional”, foi reagrupada em três classes: “Fúteis”, “Vingança/Passional” e “Tráfico de Drogas”. Essa recategorização foi necessária para tornar compatível a comparação do desempenho do algoritmo ROCK, pois, como já falado anteriormente, o método ROCK apresenta dois critérios de parada que são o número de grupos desejado ou a nulidade da matriz G. Por isso, no caso da segunda base de dados, o algoritmo ROCK não conseguiu agrupar todas as 130 observações para os três grupos sugeridos, já que a matriz G tornou-se nula prematuramente.

Desse modo, conforme mostra a TABELA 7, observa-se que o desempenho do algoritmo ROCK foi considerado mediano, uma vez que conseguiu agrupar corretamente entre 55% e 59% dos casos de homicídios.

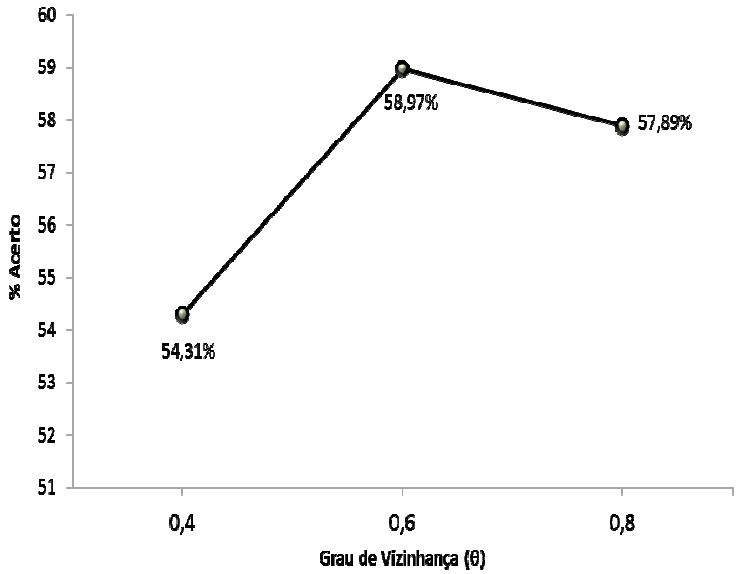
Além disso, como pode ser notado, o algoritmo ROCK agrupou com mais facilidade os casos de homicídios motivados por Vingança ou por relacionamentos passionais; e teve mais dificuldade para agrupar os crimes, cuja motivação fora Tráfico de Drogas.

Por fim, percebeu-se que as inter-relações ou similaridades existentes entre as variáveis do banco de dados sobre criminalidade não estavam muito evidentes, o que pode ter comprometido o desempenho do algoritmo ROCK especialmente para essa aplicação. Porém, essa condição não tira, de modo algum, os méritos e a qualidade dessa ferramenta. A FIGURA 10 ilustra os percentuais de acerto para os três graus de vizinhança (θ) avaliados.

TABELA 7: Análise do desempenho do algoritmo ROCK para os dados de homicídios

Vizinhança	Classificação Original	Classificação ROCK				% Acerto
		Fúteis	Vingança/Passional	Tráfico de Drogas	**Total**	
θ =0,4	Fúteis	6	8	8	**22**	**54,31%**
	Vingança/Passional	1	57	0	**58**	
	Tráfico de Drogas	0	36	0	**36**	
	Total	**7**	**101**	**8**	**116**	
θ =0,6	Fúteis	15	0	6	**21**	**58,97%**
	Vingança/Passional	0	53	2	**55**	
	Tráfico de Drogas	28	12	1	**41**	
	Total	**43**	**65**	**9**	**117**	
θ =0,8	Fúteis	0	0	6	**6**	**57,89%**
	Vingança/Passional	7	18	2	**27**	
	Tráfico de Drogas	8	1	15	**24**	
	Total	**15**	**19**	**23**	**57**	

FIGURA 10: Comparação do % de acerto – Base de Dados 2



8. CONSIDERAÇÕES FINAIS E SUGESTÕES DE TRABALHOS FUTUROS

Os resultados discutidos nessa monografia apresentaram evidências consistentes da qualidade e da versatilidade do algoritmo ROCK, principalmente no que se refere ao agrupamento de dados com variáveis do tipo atributos.

Por outro lado, ficou claro também algumas limitações desse procedimento, entre elas, a dificuldade computacional em se trabalhar com matrizes esparsas e a escolha do melhor ponto de corte (θ) para a definição dos objetos vizinhos. Ainda assim, o emprego do algoritmo ROCK para estudar as duas aplicações práticas propostas indicaram um desempenho razoável no agrupamento dos casos.

O estudo de uma versão mais rápida, do ponto de vista computacional, bem como uma investigação mais minuciosa sobre a influência do parâmetro θ na qualidade dos agrupamentos formados são expectativas de trabalhos futuros para algoritmo ROCK.

9. REFERÊNCIAS BIBLIOGRÁFICAS

[1] ALBUQUERQUE, M. A. Estabilidade em análise de agrupamento (cluster analysis). Universidade Federal Rural de Pernambuco, 2005.

[2] ANDERBERG, M. R. Clustering analisys for application. London: Academic Press, 1973. 359p.

[3] ANDREOPOULOS, B.; AN, A.; WANG, X. Bi-levelclusteringofmixedcategoricalandnumericalbiomedical data. InternationalJournalof Data Mining andBioinformatics, v. 1, n. 1, p. 19–56, 2006.

[4] ANDREOPOULOS, B.; AN, A.; WANG, X. Clusteringmixednumericalandlowqualitycategorical data: significancemetricson a yeastexample. In: IQIS ’05: Proceedingsofthe 2nd international workshop onInformationquality in information systems. New York: ACM Press, p. 87–98, 2005.

[5] ANDRITSOS, P. ScalableClusteringofCategorical Data andApplications. 2004. Tese (Doutorado) — Universidade de Toronto. Disponível em: <http:// www.dit.unitn.it/˜periklis/ publications.html>.

[6] BUSSAB, W. O.; MIAZAKI, E. S.; ANDRADE, D. F. Introdução a análise de agrupamentos. São Paulo: Associação Brasileira de Estatística, 1990. 105p.

[7] CHIU, T. et al. A robustandscalableclusteringalgorithm for mixedtypeattributes in largedatabaseenvironment. In: KDD ’01: Proceedingsoftheseventh ACM SIGKDD internationalconferenceonKnowledgediscoveryand data mining. New York: ACM Press, 2001. p. 263–268.

[8] CHU,H.C.; TSAI,M. A Comparison of Three Array Based Clustering Techniques for Manufacturing Cell Formation. InternationalJournalofProduction Research, 28(8), 1417-1433, 1990.

[9] DICE, L. R. Measures of the amount of ecologic association between species. Ecology, n.26, p.297-302, 1945.

[10] DUDA, R. O.; HART, P. E. Pattern of classification and Scene analysis. A Wiley-InterscicencePublication, New York, 1973.

[11] DUTTA, M.; MAHANTA, A. K.; PUJARI, A. K. QROCK: A quickversionofthe rock algorithm for clusteringofcategorical data. PatternRecognition, v. 26, p. 2364–2373, 2005.

[12] EVERITT, B. Clusters Analysis. 2ª ed. London: HeinemannEducational Books,1980.

[13] FILHO, A. N. C.; MAESTRELLI, N. C.; BATOCCHIO, A. Uso do Produto Escalar como Medida de Similaridade para Identificação de Agrupamentos em Manufatura Celular. Artigo Técnico, ENEGEP, ABEPRO, 2002.

[14] GANTI, V.; GEHRKE, J.; RAMAKRISHNAN, R. CACTUS: Clusteringcategorical data usingsummaries. In: KDD '99: Proceedingsofthefifth ACM SIGKDD internationalconferenceonKnowledgediscoveryand data mining. New York: ACM Press, 1999. p.73–83.

[15] GIBSON, D.; KLEINBERG, J. M.; RAGHAVA, P. Clusteringcategorical data: Anapproachbased in dynamical systems. The VLDB Journal, Springer-Verlag New York,Inc., v. 8, n. 3-4, p. 222–236, 2000.

[16] GUHA, S.; RASTOGI, R.; SHIM, K. ROCK: A Robust Clustering Algorithm for Categorical Attributes.Information System, v. 25, p. 345-366, 2000.

[17] HAIR JR,J.F; BLACK, W.C.; BABIN, B.J.;ANDERSON,R.E.;TATHAM,R.L. Análise Multivariada de Dados. Porto Alegre: Bookman, 2009.

[18] HUANG, Z. Clusteringlarge data sets withmixednumericandcategoricalvalues. In: ProceedingsoftheFirst Pacific AsiaKnowledge Discovery and Data Mining Conference. Singapore: World Scientific, 1997. p. 21–34.

[19] HUANG, Z. Extensionstothe k-meansalgorithm for clusteringlarge data sets withcategoricalvalues. Data Mining andKnowledge Discovery, v. 2, p. 283–304, 1998.

[20] HUANG, Z.; NG, M. K. A fuzzy k-modesalgorithm for clusteringcategorical data. IEEE TransactionsonFuzzy Systems, v. 7, p. 446–452, 1999.

[21] JAIN, A. K.; DUBES, R. C. Algorithms for clustering data. Prentice Hall. Englewood Cliffs, New Jersey, 1988.

[22] KAUFMANN, L.; ROSSEEEUW, P. J. Find groups in data: na introduction to cluster analysis. New York: John Wiley, 1990. 342p.

[23] LINDEN, R. Técnicas de Agrupamento. Revista de Sistemas de Informação da FMSA n. 4, pp. 18-36, 2009.

[24] MATOS, R.A. Comparação de Metodologias de Análise de Agrupamentos na Presença de Variáveis Categóricas e Contínuas. 2007. Dissertação (Mestrado) — ICEX - Departamento de Estatística, Universidade Federal de Minas Gerais.

[25] MEYER, A. S. Comparação de coeficientes de similaridade usados em análises de agrupamento com dados de marcadores moleculares dominantes. Escola Superior de Agricultura “Luiz Queiroz”, Universidade de São Paulo, 2002.

[26] MINGOTI, S. A. Análise de Dados Através de Métodos de Estatística Multivariada:Uma Abordagem Aplicada. Belo Horizonte: Editora UFMG, 2005.

[27] PANTUZZO, A. E. Comparação de Métodos de Análise de Cluster na Presença de Dados Categóricos e a Aplicação no Contexto de Data Mining: Estudo de Casos. 2002. Dissertação (Mestrado) — Departamento de Engenharia de Produção, Universidade Federal de Minas Gerais.

[28] R DEVELOPMENT CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3900051-07-0, URL http://www.R-project.org, 2009.

[29] ROGERS, F. J.; TAMINOTO, T. T. A computer program for classying plants. Science, n132, p. 1115-1118, 1960.

[30] RUSSEL, P. F.; RAO, T. R. On habitat and association of species of anopheline larvae in south-eastern Madras. JournalofMalariaIndiaInstitute, n3, p.153-178, 1940.

[31] SNEATH, P. H. A.; SOKAL, R. R. Numerictaxonomy: theprinciplesandpracticeofnumericalclassification. San Francisco: W. H. Freeman, 1973. 573p.

[32] SOKAL, R. R.; SNEATH, P. H. A. PrincipleofNumericalTaxonomy. San Francisco: W. H. Freeman, 1963. 359p.

[33] SOKAL, R.R.; MICHENER, C.D. A statisticalmethod for evaluatingsystematicrelationships. BulletinoftheSocietyUniversityof Kansas, n.38, p.1409-1438, 1958.

[34] SOUZA, A. L.; FERREIRA, R. L. C.; XAVIER, A. Análise de agrupamento aplicada à ciência florestal. Viçosa: SIF, 1997. 109 f. (Documento SIF, 16).

[35] SOUZA, C. R. S. Análise de Clusters via algoritmo ROCK aplicada ao problema da mineração de dados sobre criminalidade em Belo Horizonte. Artigo Técnico, Centro Federal de Educação Tecnológica de Minas Gerais, 2012.

[36] TABWIN/DATASUS. Disponível em:<http://datasus.gov.br/tabwin>. Acesso em 09 de janeiro de 2013.

[37] TOTTI, R.; VENKOVSKY, R.; BATISTA, L. A. R. Utilização de métodos de agrupamentos hierárquicos em acessos de Paspalum (Graminea(Poaceae)). Semina: Ci. Exatas Tecnol., Londrina, v. 22, p. 25-35, dez. 2011.