Universidade Federal de Pernambuco
Centro de Informática

Graduação em Ciência da Computação

Análise de séries temporais através de algoritmos de agrupamento

André Luis dos Santos Alves

Trabalho de Graduação

Recife, 10 de Junho de 2009

Universidade Federal de Pernambuco
Centro de Informática

Análise de séries temporais através de algoritmos de agrupamento

André Luis dos Santos Alves

Monografia apresentada ao Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do Grau de Bacharel em Ciência da Computação.

Orientador: Tsang Ing Ren

Recife, 10 de junho de 2009

“Nossas dúvidas são traidoras e nos fazem perder o bem que poderíamos conquistar, senão fosse o medo de tentar.”

William Shakespeare

Ao meu pai, José Carlos. (*in memoriam)*

À minha irmã, Carol (*in memoriam)*

AGRADECIMENTOS

Em primeiro lugar, gostaria de agradecer ao meu pai, José Carlos, que mesmo ausente na reta final da minha graduação sempre me deu a força necessária para que eu conquistasse meus objetivos. Ele sempre foi fonte de inspiração e um exemplo a ser seguido por mim. Tenho certeza que ele estaria orgulhoso neste momento, por ver o seu filho superar mais essa etapa da vida.

À minha mãe, Maria de Fátima pela força e por sempre se fazer presente, à minha irmã Karla, por me suportar nos piores dias e sempre me aconselhar, à minha irmã Carol, que mesmo ausente, continua sendo inspiração para mim.

Ao meu orientador, Professor Tsang Ing Ren, pela confiança depositada em mim, seja na confecção deste trabalho, seja na monitoria.

Aos meus amigos que ficaram em João Pessoa, mas que mesmo de longe sempre torcem por mim e se fazem presentes. Também agradeço aos diversos amigos que fiz em Recife e sem perceber se tornaram essencial para mim.

Aos diversos professores do Centro de Informática que seja em sala de aula ou nas conversas de corredor me ajudaram e contribuíram muito para minha formação. Gostaria de destacar alguns nomes como Alexandre Vasconcelos, Fábio Queda, Alex Sandro Gomes, Geber Ramalho e George Darminton.

Aos funcionários do Centro de Informática pelo trabalho diário com a finalidade de nos proporcionar o melhor ambiente de estudo.

Por fim, gostaria de agradecer a todos que acreditaram em mim e que, mesmo sem saber, são exemplos para mim.

RESUMO

Os investimentos em geral possuem três aspectos básicos: retorno, prazo e proteção. O investidor deve analisar qual é a rentabilidade, liquidez e grau de risco de cada aplicação antes de colocar o seu dinheiro. O lucro está proporcionalmente relacionado com o grau de risco, ou seja, quanto maior for o rendimento, mais arriscada será a aplicação. O desafio principal para os investidores é otimizar essa equação. Uma carteira de investimento é um conjunto de ativos que pertence a um investidor. Logo o investidor deve montar sua carteira de maneira que ela seja eficiente.

Clustering é uma classificação não-supervisionada de padrões em grupos. Esses grupos são construídos tomando como base a similaridade. Assim, intuitivamente, conclui-se que padrões que estão dentro de um mesmo grupo apresentam maior similaridade do que padrões que estão em grupos diferentes [2].

O objetivo desse trabalho é apresentar um sistema que dará suporte a formação de uma carteira de investimentos otimizada. Para alcançar esse objetivo, serão usados diferentes algoritmos de agrupamento, aplicando-os nos dados históricos de alguns ativos da Bovespa com a finalidade de encontrar as similaridades entre eles e assim agrupá-los da maneira mais adequada. Em seguida, apresentaremos uma análise comparativa entre os resultados obtidos por cada um dos algoritmos.

É importante deixar claro que o objetivo, nesse primeiro momento, não é desenvolver um sistema de predição e sim um sistema que crie carteiras de investimentos e deixe claro para o investidor o grau de riscos inerente a cada uma delas. A partir disso, cabe ao investidor, de acordo com o seu perfil, optar pela carteira mais arriscada ou menos arriscada.

Palavras-chave: Agrupamento Hierárquico, Séries Temporais, Reconhecimento de Padrões, Séries Financeiras, Mercado de Ações e Bovespa.

ABSTRACT

The investments, in general, have three basics characteristics: return, term and protection. The investor should analyse the profitable, the liquidity and the risk of each application before put their money. The profit is proportional to the risk, which means that how bigger is the rental, more risked will be the application. The main challenge to the investors is to optimize this equation. A portfolio of investment is a group of assets that belongs to a investor. So he must make his portfolio in a way that it be efficient.

Clustering is a non-supervisioned classification of patterns in group. These groups are built based on similarity. Therefore, intuitively, we deduce that patterns which are inside the same group present more similarity than patterns that are in different groups.

Our purpose in the research is to present a system that will give support to the creation of an optimized portfolio. To get this purpose we will use differents algorithms of clustering, applying them in historical datas of some assets of Bovespa with the purpose of find the similarities between them and just like that group them in the most appropriate way. Then, we will present a comparative analysis between the results achieved by each one of the algorithms.

It´s important to emphasize that our purpose, at this first moment, it´s not develop a prediction system but a system that creates portfolios and make clear to the investor the risks of each one of them. From this point, it up to the investor, according to his profile, choose the most risked portfolio or the less risked one.

Keywords: Hierarchical Clustering, Time Series, Pattern Recognition, Financial Series, Stock Market and Bovespa.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

Antes de começarmos a expor nossos experimentos computacionais, julgamos relevante abordar alguns tópicos relacionados à economia tais como: ações, investidores, conceitos relacionados à bolsa de valores, dentre outros.

1.1 A BOLSA DE VALORES

Bolsa de valores é um local onde se negociam ações de empresas de capital aberto (públicas ou privadas) e outros instrumentos financeiros, como opções e debêntures.

Empresa de capital aberto é uma sociedade anônima cujo capital social é formado por ações livremente negociadas no mercado sem necessidade de escrituração pública de propriedade (por parte da pessoa física compradora). A diferença principal entre empresas de capital fechado e aberto de tamanho semelhantes é apenas contábil. Na grande maioria das empresas um ou mais sócios controlam a maioria do capital da empresa e a gerenciam como uma empresa de capital efetivamente fechado. Em grandes empresas onde o controle do capital é diluído, os executivos diretores são subordinados ao conselho dos acionistas para que a empresa tome a direção que lhes parece melhor. Uma empresa lança ações, por exemplo, com a finalidade de captar novos recursos para investir em crescimento e modernização a médio e longo prazos, sem ficar refém dos altos juros cobrados quando se faz empréstimos. [3]

Tradicionalmente, os negócios aconteciam fisicamente na própria localização física da bolsa de valores: pregão viva-voz, porém, atualmente, as transações são cada vez mais realizadas por meios eletrônicos em tempo real, onde são colocadas as ordens pelos compradores e vendedores, isso se chama pregão eletrônico. Os movimentos dos preços no mercado ou em uma seção do mercado são capturados por meio de índices chamados índices da bolsa de valores.

No Brasil, as principais bolsas são a Bovespa (Bolsa de Valores de São Paulo), cujo principal índice é o Ibovespa ou apenas Ibov, o qual negocia principalmente ações, opções, debêntures e termos, e a BM&F (Bolsa de Mercadorias e Futuro), que lida com a negociação de contratos de mercadorias (commodities) e derivativos, à vista ou para pagamento futuro. No primeiro semestre de 2008, ocorreu a fusão da Bovespa com a BM&F o que deu origem à terceira maior Bolsa do mundo, com valor de mercado de 20 bilhões de dólares, ficando atrás somente das Bolsas Mercantil de Chicago (CME) e alemã, e a sexta entre as BM&F.

Como foi citado anteriormente o movimento dos preços no mercado é obtido através de índices, esses índices representam a variação média dos preços de um conjunto de bens em relação a um determinado intervalo de tempo. O Ibov compreende uma carteira teórica de ações e existe desde 1968. Essa carteira teórica é formada das ações que, em conjunto, representaram 80% do volume à vista nos 12 meses anteriores à formação da carteira. A atualização da carteira do Ibov se dá a cada quatro meses, segundo os critérios da Bovespa. No entanto, é importante ressaltar que existem outros índices além do Ibov, o IBX, por exemplo, mede o retorno de uma carteira teórica composta pelas cem ações mais negociadas na Bovespa.

Na Bolsa podem ser realizadas diversas ações, tais como:

- Mercado a vista: São operações de compra e venda de ações emitidas pelas empresas de capital aberto registradas Bolsa. Todos os negócios são liquidados à vista, ou seja, a transferência da titularidade dos ativos e o acerto financeiro ocorrem em um curto período de tempo;
- Mercado a termo: Negócios com um ativo com vencimento em determinada data futura. Compra-se ou vende-se o ativo hoje para ser liquidado no futuro. Nessa situação, o preço é formado pelo preço à vista mais uma taxa de juros;
- Mercado de opções: Negócios com direitos de compra e venda de um lote de ações, ativos financeiros ou commodities com preços e prazos predeterminados.;

- Mercado futuro: Negociações de compra e venda de contratos autorizados pela BM&F, para liquidação em data futura prefixada.

1.2 AÇÕES

A definição de ação dada pela Bovespa é a seguinte:

"Valor imobiliário, emitido pelas companhias, representativo de parcela do capital. Representa a menor parcela em que se divide o capital da companhia. Título negociável em mercados organizados".

Assim, podemos inferir que o investidor que compra uma ação de uma determinada empresa, torna-se sócio desta. No entanto, os poderes conferidos a ele são limitados pelo tipo de ação que comprou e também pela quantidade de ações que possui. Portanto, o comportamento das suas ações vai refletir o comportamento da empresa que as emitiu, ou seja, se a empresa vai mal, ações provavelmente também irão mal. Se a empresa estiver bem, com boas perspectivas e fundamentos, assim também estarão suas ações. No curto prazo, os preços das ações refletem a situação econômica do país naquele momento (notícias, especulações, dentre outros), mas no longo prazo, ações provavelmente irão refletir os fundamentos da empresa, tais como: o seu modelo de gestão, a cultura organizacional, o histórico da empresa dentre outras coisas.

Existem dois tipos de ações, as ações Ordinárias Nominativas (ON) dão direito à participação nos resultados econômicos da empresa e a voto em assembléia. A cada ação ordinária corresponde um voto nas deliberações da Assembléia Geral. O outro tipo de ação são as ações Preferenciais Nominativas (PN) que dão prioridade ao seu detentor no recebimento de dividendos. No caso de dissolução da empresa, há também prioridades no reembolso de capital [4].

Todas as ações têm um código que no Brasil, normalmente é composto por quatro letras e um número. Por exemplo, a ação preferencial da Petrobrás recebe o código PETR4, já a ação ordinária da Vale do Rio Doce recebe o código VALE3. É muito comum usar o 3 para ordinárias e o 4 para

preferenciais, mas isso não é obrigatório. Muitas preferenciais, como a Vale do Rio Doce e a Usiminas, por exemplo, usam o número 5. Nesse trabalho nós usaremos, majoritariamente, ações ordinárias. No entanto também haverá ações preferenciais como é o caso da ARCZ6. A lista completa das ações usadas nesse trabalho está no anexo 1.

Normalmente, as ações são negociadas em lotes, o que corresponde a uma quantidade de ativos ou títulos de características idênticas. Até esse momento, no Brasil, as ações normalmente têm um lote padrão ou lote mínimo. É a quantidade mínima de ações que pode ser comprada como um lote no mercado à vista. O preço do lote corresponde ao preço unitário da ação (que é divulgado pela Bolsa) multiplicado pela quantidade de ações no lote. Por exemplo, se a PETR4 está a R$50,00 e seu lote mínimo ou padrão é de 100 ações, então o lote custa R$ 5.000,00.

Apesar de as ações, normalmente, serem negociadas em lotes, existe uma outra possibilidade que é o investidor operar no mercado fracionário, essa é a opção que os pequenos investidores, geralmente, fazem, pois nesse tipo de mercado as ações são negociadas em qualquer quantidade, podendo até ser apenas uma. A maneira que ocorre a compra e a venda de ações no mercado fracionário é a mesma que acontece no mercado que negocia lotes.

O preço das ações é determinado pelo mercado de forma muito simples. O preço corresponde sempre ao último negócio realizado. Se há um comprador e um vendedor no mesmo preço, o negócio é fechado e aquele passa a ser o preço da ação até que um novo negócio seja realizado em outro preço. O mercado vai determinando o preço das ações a cada momento. Entretanto há circunstâncias especiais em que leilões são utilizados para definir o preço da ação. Nesse trabalho, nós utilizamos o preço de fechamento de cada uma das ações ou seja, utilizamos o preço do último papel negociado em cada dia.

1.3 CARTEIRA DE INVESTIMENTOS

Carteira de investimentos é um grupo de ativos que pertence a um investidor, pessoa física ou jurídica. Esses ativos podem ser ações, fundos, títulos públicos, debêntures, aplicações imobiliárias, entre outros. No nosso caso, usaremos apenas ações, ou seja, nossa carteira de investimentos será constituída apenas de ações por isso chamaremos de carteira de ações. Uma carteira de ações deve ser montada baseada no perfil de cada investidor, nesse trabalho, nós iremos propor uma maneira de montar carteiras de ações [5].

Existem dois tipos mais comuns de análises de mercado com a finalidade de montar carteiras de ações:

- **Análise Técnica:** esse tipo de análise é baseado na interpretação de gráficos de preços, volumes e outros indicadores. É uma metodologia que visa estudar o movimento de preços das ações, relacionados aos volumes negociados, para determinar tendências de alta, de estabilidade ou de baixa, em busca da oportunidade de comprar e vender ações a preços compensadores. Ela mostra como os preços se comportaram no passado e projeta uma série de expectativas de movimentos de preços no futuro.

- **Análise fundamentalista:** esse tipo de análise tenta definir se o preço a que a ação está sendo negociada naquele momento está caro ou barato, ou seja, ela tenta precificar a empresa e assim, ter um preço justo para a ação. Segunda a Bovespa, essa metodologia *"é uma análise de mercados baseada nos fatores econômicos, dependendo de estatísticas, projeções, condições de oferta e demanda de bens e serviços, e nos fundamentos da economia e das empresas. Metodologia para determinar o preço justo de uma ação, que se fundamenta na expectativa de lucros futuros*". Existem uma série de indicadores objetivos para a análise fundamentalista, no entanto, além de observar esses indicadores, os investidores devem aprender a analisar o setor ao

qual pertence a empresa: se é um setor promissor ou uma seção decadente da economia [6].

O objetivo desse trabalho será formar carteiras de ações otimizadas, utilizando técnicas computacionais. Serão construídas diversas carteiras de ações e cada uma terá um risco associado. Assim, baseado nesse risco, os investidores decidirão, baseado no seu perfil, em qual carteira deverá investir.

No capítulo 2 será abordada a técnica de *clustering* (agrupamento) que será muito importante para a formação das nossas carteiras de ações, bem como os algoritmos de agrupamento utilizados.

No capítulo 3 será mostrado como foram realizados os nossos experimentos, como tratamos nossa base de dados e também serão apresentados os grafos e dendrogramas obtidos.

No capítulo 4 serão analisados os resultados obtidos com os experimentos realizados.

No capítulo 5 será apresentada uma conclusão a respeito do trabalho realizado e possibilidades de trabalhos futuros.

2. AGRUPAMENTO

Agrupar é uma das mais primitivas atividades mentais realizadas pelos seres-humanos, usada para organizar a grande quantidade de informação que é recebida diariamente. O trabalho de processar toda essa informação individualmente, tornaria essa atividade inviável. Assim, nós categorizamos as entidades em *clusters* (grupos). Cada c*luster* é então caracterizado pelas características comuns que os elementos do *cluster* possuem [7]. Em suma, *clustering* é o termo que descreve métodos para agrupar dados não-rotulados.

2.1 PROCEDIMENTOS BÁSICOS E TERMINOLOGIA

Segundo Theodoridis, a tarefa de agrupamento (nesse trabalho nós usaremos a palavra “agrupamento” como sinônimo de *clustering)* possui alguns passos básicos que devem ser seguidos, são eles:

- *Seleção de característica:* Características devem ser bem selecionadas, essa etapa visa identificar o mais eficiente subconjunto de características originais para usar no processo de agrupamento.

- *Medida de semelhança:* É a medida que vai quantificar o quão semelhante ou diferente dois vetores de características são. É importante assegurar que todas as características contribuem igualmente para o cálculo da media de semelhança e que não há características que dominem outras.

- *Critério de agrupamento:* Dependendo do tipo de *cluster* que o analista deseja obter, ele alterará o critério de agrupamento para poder satisfazer suas expectativas.

- *Algoritmos de agrupamento:* Uma vez que já estão escolhidos a medida de semelhança e os critérios de agrupamento, esse passo será responsável por escolher um algoritmo específico que irá montar a estrutura de agrupamento do conjunto de dados.

- *Validação dos resultados:* Com os resultados gerados pelo algoritmo de agrupamento em mãos, é necessário verificar se aqueles dados estão corretos, isso é feito por meio de testes.

- *Interpretação dos resultados:* Em muitos casos, o analista deve integrar os resultados do agrupamento com outras evidências experimentais e analisar para chegar às suas conclusões.

Antes de detalharmos as duas grandes categorias de algoritmos de agrupamento, é interessante que nos familiarizemos com alguns termos usados na literatura sobre agrupamento, então, agora nós iremos contrapor algumas desses termos, tais como:

Aglomerativo VS. Divisível – Este aspecto diz respeito à estrutura e operação algorítmica. Uma abordagem aglomerativa começa com cada padrão num *cluster* distinto, e sucessivamente, une *clusters* até que o critério de parada seja satisfeito. Método divisível começa com todos os padrões num *cluster* simples e este vai sendo dividido até que o critério de parada seja satisfeito.

Monotético VS. Politético – Este aspecto diz respeito ao uso seqüencial ou simultâneo das características no processo de agrupamento. A maioria dos algoritmos são politéticos, ou seja, todas as características entram no cálculo das distâncias entre os padrões e as decisões são baseadas nessas distâncias. Já os algoritmos monotéticos tratam as características sequencialmente para dividir a coleção de padrões dadas. O grande problema da abordagem monotética é que ela vai gerar 2^d *clusters* onde d é a dimensionalidade dos padrões, o que pode acarretar complicações para valores muito grandes de d.

Incremental VS Não-Incremental – Esse aspecto surge quando o conjunto de padrões a ser agrupado é grande, e restrições de tempo de execução ou espaço de memória afetam a arquitetura do algoritmo. A recente história da metodologia de agrupamento não contem muitos exemplos de algoritmos de agrupamento construídos para trabalhar com grande conjuntos de dados, mas o advento do data mining tem fomentado o desenvolvimento de algoritmos de agrupamento que minimizem o numero de varreduras através do conjunto de padrões, reduzindo o numero de padrões examinados durante a execução, ou reduzam o tamanho das estruturas de dados usadas nas operações do algoritmo.

2.2. ALGORITMOS DE AGRUPAMENTO

Há duas categorias principais de algoritmos de agrupamento, são elas os algoritmos de agrupamento seqüenciais e os algoritmos de agrupamento hierárquico, nesse momento, iremos detalhar essas duas classes de algoritmos.

- *Algoritmos de agrupamento seqüenciais:* Esses algoritmos produzem um agrupamento simples. Eles são métodos rápidos. Na maioria deles, todos os vetores de características são apresentados para o algoritmo uma vez ou poucas vezes. O resultado final é, usualmente, dependente da ordem que o vetor é apresentado ao algoritmo. Nesse trabalho, nós usamos, majoritariamente, algoritmos de agrupamento hierárquico que é a categoria que veremos em seguida.

- *Algoritmos de agrupamento hierárquico:*. Algoritmos hierárquicos geram dendrogramas que representam um agrupamento aninhado de padrões e os níveis de similaridade que mudam o agrupamento. A maioria dos algoritmos de agrupamento hierárquico são variantes do *single linkage*, *complete linkage* e do algoritmo de mínima variância, também conhecido como algoritmo de *ward*. Nesse trabalho nós usaremos os três algoritmos anteriormente citados e ainda usaremos outros, tais como o *average linkage*, o *weighted linkage*, o *median linkage* e o *centroid linkage*. É através desses algoritmos (funções de *linkage*) que nós vamos obter as medidas de distância entre os objetos para que seja formado os *clusters* baseado no critério adotado. A diferença básica entre os algoritmos de agrupamento é a maneira como eles calculam as distâncias entre os *clusters.* Agora, vamos discorrer mais detalhadamente sobre cada um desses algoritmos (funções de *linkage*).

2.2.1 - SINGLE LINKAGE:

É um método de calcular distâncias entre *clusters* (grupos) no agrupamento hierárquico. No *single linkage*, a distância entre dois *clusters* é computada como a distância entre os dois elementos mais próximos nos dois *clusters*. [7]

Matematicamente, nós podemos demonstrá-la da seguinte maneira:

A função de linkage, ou seja, a distância D(X,Y) entre dois clusters X e Y, é descrita pela seguinte expressão:

$$D(X,Y) = min(d(x,y))$$

Onde d(x,y) é a distância entre os elementos x ϵ X e y ϵ Y.

X e Y são dois conjuntos de elementos *(clusters)*.

Abaixo temos um algoritmo para o *single linkage*:

O seguinte algoritmo é um esquema aglomerativo que apaga linhas e colunas na matriz de proximidade quando velhos clusters são fundidos em novos clusters. A matriz de proximidade D de tamanho NxN contem todas as distâncias d(i,j). Aos *clusters* são atribuídos seqüências de números 0,1,2,...,(n-1) e L(k) é o nível do k-ésimo *cluster*. Um *cluster* com um número de seqüência m é denotado m e a proximidade entre os *cluster* r e s é denotado d[(r),(s)]. Dito isso, segue o algoritmo:

1 – Comece com o agrupamento disjunto tendo nível L(0) = 0 e o número de seqüência m = 0;

2 – Achar o par de clusters mais similar no agrupamento atual, seguindo a fórmula d[(r),(s)] = min d[(i),(j)].

3 – Incremente o número de seqüência: m=m+1. Junte os clusters r e s em um só cluster, formando o próximo cluster m. Sete o nível do cluster para L(m) = d[(r),(s)].

4 – Atualize a matriz de proximidade, D, deletando as linhas e colunas correspondente aos clusters r e s e adicione uma linha e uma coluna correspondente

ao novo cluster formado. A proximidade entre o novo cluster, denotado por (r,s) e o cluster antigo k é definido como d[(k),(r,s)] = min d[(k,r)],d[(k,r)].

5 – Se todos os objetos estão em um cluster, pare. Caso contrário, vá para o passo dois.

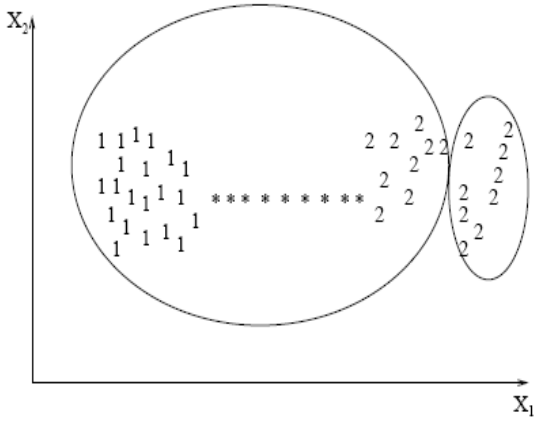

FIGURA 1 - AGRUPAMENTO COM *SINGLE LINKAGE*

2.2.2 - COMPLETE LINKAGE:

O *complete linkage* também chamado de método do vizinho mais distante, é um método para calcular distâncias entre *clusters*. A função de *linkage* diz que a distância entre dois *clusters* é computada como a máxima distância D(x,y) entre dois objetos onde o objeto x pertence ao primeiro *cluster* e o objeto y pertence ao segundo *cluster*. Em outras palavras, a distância entre dois *clusters* é computada como a distância entre os dois objetos mais distantes em dois *clusters.* Como foi citado anteriormente a diferença entre os algoritmos de agrupamento reside no cálculo das distâncias. Então, poderíamos aproveitar o algoritmo utilizado no *single linkage* no *complete linkage,* alterando o modo como a distância é calculada.

Matematicamente, nós podemos representar a função de *linkage* da seguinte maneira:

$$D(X,Y) = Max\ d(x,y),\ onde\ x \epsilon X\ e\ y \epsilon Y.$$

d(x,y) é a distância entre os objetos x e y.

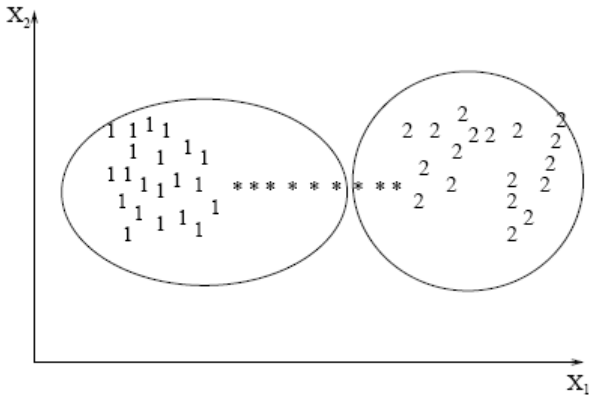

FIGURA 2 - AGRUPAMENTO COM *COMPLETE LINKAGE*

2.2.3 - AVERAGE LINKAGE:

O *Average linkage* é um método para calcular a distância entre *clusters* no agrupamento hierárquico. A função de *linkage* diz que a distância entre dois *clusters* é computada como a distância média entre os objetos do primeiro *cluster* e os objetos do segundo *cluster*. A média é executada em todos os pares (x,y) de objetos, onde x é um objeto do primeiro *cluster* e y é um objeto do segundo *cluster.* Mais uma vez, nós podemos utilizar o algoritmo mostrado no *single linkage* para o *average linkage*, alterando o modo como as distâncias entre os *clusters* são calculadas.

Matematicamente, nós podemos mostrar essa função de *linkage* como:

$$d(X,Y) = \frac{1}{Nx*Ny} \sum_{i=1}^{Nx} \quad \sum_{j=1}^{Ny} d(x,y);$$

Onde,

- x ϵ X e y ϵ Y

- d(x,y) é a distância entre os objetos x ϵ X e y ϵ Y;

- X e Y são dois conjunto de objetos (*clusters);*

- Nx e Ny são os números de objetos nos *clusters* X e Y, respectivamente.

2.2.4 - WARD´S LINKAGE:

Esse tipo de *linkage* usa a soma incremental dos quadrados, ou seja, aumenta a soma total dos quadrados dentro de um cluster como o resultado da junção de dois *clusters*. A soma dos quadrados dentro de um *cluster* é definida como a soma dos quadrados das distâncias entre todos os objetos no *cluster* e o *centroid* do *cluster.* Novamente, nós podemos utilizar o algoritmo mostrado no *single linkage* para o *Ward´s linkage*, alterando o modo como as distâncias entre os *clusters* são calculadas.

A distância equivale a:

$$d^2(r,s) = n_r n_s \frac{\|\overline{x_r} - \overline{x_s}\|^2}{(n_r + n_s)}$$

Onde,

- || || é a distância euclidiana

- $\overline{x_r}$ e $\overline{x_s}$ são os *centroids* dos *clusters* r e s, assim definidos:

$$\overline{x_r} = \frac{1}{n_r}\sum_{i=1}^{n} x_{ri}$$

2.2.5 - CENTROID LINKAGE:

O *Centroid linkage* é um método para calcular a distância entre *clusters* no agrupamento hierárquico. A função de *linkage* diz que a distância entre dois *clusters* é dada pela distância euclidiana entre os *centroids* de dois *clusters.* Mais uma vez, nós podemos utilizar o algoritmo mostrado no *single linkage* para o *centroid linkage*, alterando o modo como as distâncias entre os *clusters* são calculadas.

Matematicamente, pode representá-lo da seguinte maneira:

$$\mathrm{d(r,s)} = \|\overline{x_r} - \overline{x_s}\|$$

Onde,

- $\overline{x_r} = \frac{1}{n_r}\sum_{i=1}^{n} x_{ri}$

Ao utilizar o *Centroid Linkage*, encontramos algumas peculiaridades no seu dendrograma que serão abordadas mais na frente.

2.2.6 - MEDIAN LINKAGE:

O *Median linkage* é um método para calcular a distância entre clusters no agrupamento hierárquico. A função de *linkage* diz que a distância entre dois *clusters* é dada pela distância euclidiana entre os *centroids* com ponderados desses dois *clusters*. Mais uma vez, nós podemos utilizar o algoritmo mostrado no *single linkage* para o *median linkage*, alterando o modo como as distâncias entre os *clusters* são calculadas.

Matematicamente, nós podemos defini-lo como:

$$\mathrm{d(r,s)} = \|\widetilde{x_r} - \widetilde{x_s}\|$$

Onde,

$\widetilde{x_r}$ e $\widetilde{x_s}$ são os centroids ponderados para os clusters r e s. Se o cluster r foi criado pela combinação dos clusters p e q, $\widetilde{x_r}$ é definido recursivamente como:

$$\widetilde{x_r} = \frac{1}{2}(\widetilde{x_p} + \widetilde{x_q})$$

2.2.7 - WEIGHTED LINKAGE:

O *Weighted linkage*, também chamado de WPGMA, é um método para calcular a distância entre clusters no agrupamento hierárquico. Mais uma vez, nós podemos utilizar o algoritmo mostrado no *single linkage* para o *weighted linkage*, alterando o modo como as distâncias entre os *clusters* são calculadas.

A função de linkage diz que a distância entre dois clusters é dada pela seguinte fórmula:

$$\mathrm{d}(C_{novo}, C_{antigo}) = \frac{1}{2}\left(\mathrm{d}(C_i, C_{antigo}) + \mathrm{d}(C_j, C_{antigo})\right)$$

3. EXPERIMENTOS

Para efetuar nossos experimentos, nós utilizamos os dados históricos contidos no site da Bovespa, que é a Bolsa de Valores de São Paulo. Nesse site encontra-se muitas informações a respeito das empresas que negociam papéis na bolsa, há dicas de como investir, informações sobre corretoras, entre outras. No entanto, o que nos interessava mais eram os arquivos que possuíam o comportamento das ações no decorrer do tempo. A Bovespa disponibiliza as séries históricas das ações desde o ano de 1986. Essas séries históricas estão disponíveis no formato .txt, no entanto podem ser importados para o Microsoft Excel. Esse arquivos de texto possuem diversas informações sobre as ações tais como: o nome resumido da empresa emissora do papal, o código de negociação do papel, a data do pregão, o preço de abertura no pregão, o preço médio do papel no pregão, o preço do último papel negociado, o preço da melhor oferta de compra, o preço da melhor oferta de venda, o volume total negociado, dentre outros. Para possibilitar a interpretação do arquivo .txt, a Bovespa disponibiliza em [1] um arquivo com o layout que diz a posição de cada uma dessas informações no .txt.

Entretanto, nem todas as informações nos interessam nesse primeiro momento, então tivemos que elaborar um algoritmo que extraísse apenas as informações desejadas nesse momento. Essas informações são a data do pregão, o código do papel usado na negociação e o preço de fechamento da ação. Lembrando que o preço de fechamento é para um lote de ações, normalmente, um lote possui 100 ações.

Outra coisa que tivemos que fazer foi inspecionar toda a base de dados pois em algumas ações ocorreu o fenômeno chamado *split* que é a divisão de uma ação em várias outras. Por exemplo, em um split de um para três, que possui uma ação passa a ter 3, sendo que cada ação passa a valer 1/3 do que valia antes. Não há lucro nem prejuízo para o investidor quando uma ação sofre um *split* [9]. Então no caso de ocorrência desse fenômeno nós temos que ajustar os preços para o número de ações possuídas. O processo inverso

recebe o nome de *join*, ou seja, ao invés de transformar uma ação em duas, você agruparia dez ações em uma, por exemplo.

Nossa base de dados é composta pelos anos de 2005, 2006 e 2007. Nós escolhemos para fazer parte da nossa base de dados 20 ações que são negociadas na Bovespa, a lista com o nome e o respectivo código consta no Anexo 1. Os critérios utilizados para escolher as ações foram a presença destas no Ibov e a existência de informações sobre essas ações em todos os dias desses três anos. Algumas ações não apresentam cotação para alguns dias do ano e essa ausência poderia ter algum efeito negativo no nosso trabalho.

3.1 O PROCESSO DE AGRUPAMENTO

Nós aplicamos diversos algoritmos de *linkage* na nossa base de dados de séries históricas de alguns ativos da Bovespa. Como já foi dito, nossa base contem o preço de fechamento de 20 ações, que são negociadas na Bovespa e fazem parte do Ibov, nos anos de 2005, 2006 e 2007. Nós escolhemos o Ibov, principalmente, por dois motivos. Primeiro, pela facilidade de obtenção desses dados. Segundo, pela confiabilidade dos preços.

Primeiramente, nós olharemos para a série temporal da diferença logarítmica dos preços de fechamento. Essa diferença é dada por: [9]

$$Y_i\,(t) = \ln P_i(t) - \ln P_i(t\text{-}1)$$

Onde,

$P_i(t)$ é preço de fechamento do i-ésimo ativo no dia t. No entanto, tanto P_i como Y_i são funções muito irregulares no tempo. Então, para quantificar o grau de similaridade entre duas séries temporais e usar nosso algoritmo de linkage, nós adotamos a seguinte função de métrica, que quantifica a sincronia na evolução do tempo. [8]

$$d_{ij} = \sqrt{2(1 - c_{ij})}\ ,$$

Onde,

$$c_{ij} = \frac{<Y_i Y_j> - <Y_i><Y_j>}{\sqrt{(<Y_i^2> - <Y_i>^2)(<Y_j^2> - <Y_j>^2)}}$$

Os sinais de “maior que” (>) e “menor que” (<) denotam a média no intervalo de tempo desejado. O anexo 2, mostra a matriz de correlação obtida e o anexo 3 mostra a matriz de distâncias.

De posse da matriz de distância, nós conseguimos montar os dendrogramas que são um tipo específico de diagrama que organiza determinado fatores e variáveis. Ele resulta de uma análise estatística de determinados dados, em que se emprega um método quantitativo que leva a agrupamentos e à sua ordenação hierárquica ascendente. Em suma, o dendrograma ilustra o arranjo derivado da aplicação de um algoritmo de agrupamento. Então, nós temos um dendrograma para cada um dos algoritmos de agrupamento (funções de *linkage*) que nós utilizamos. A seguir mostraremos cada um dos dendrogramas:

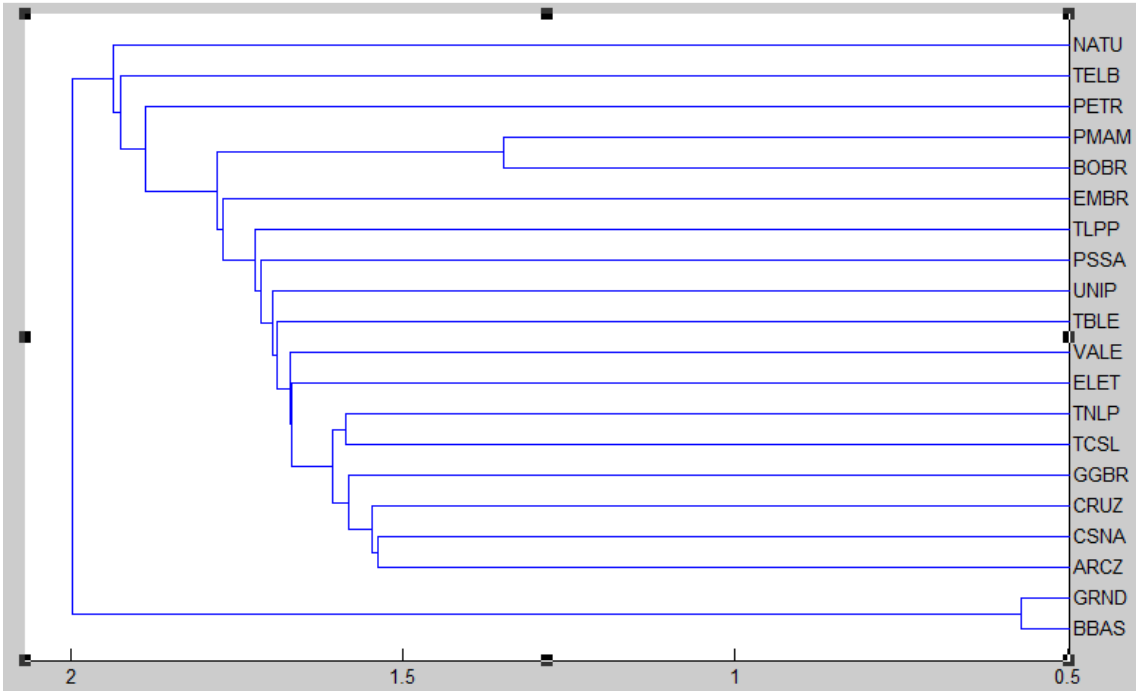

FIGURA 3 - DENDROGRAMA GERADO PELO *SINGLE LINKAGE*

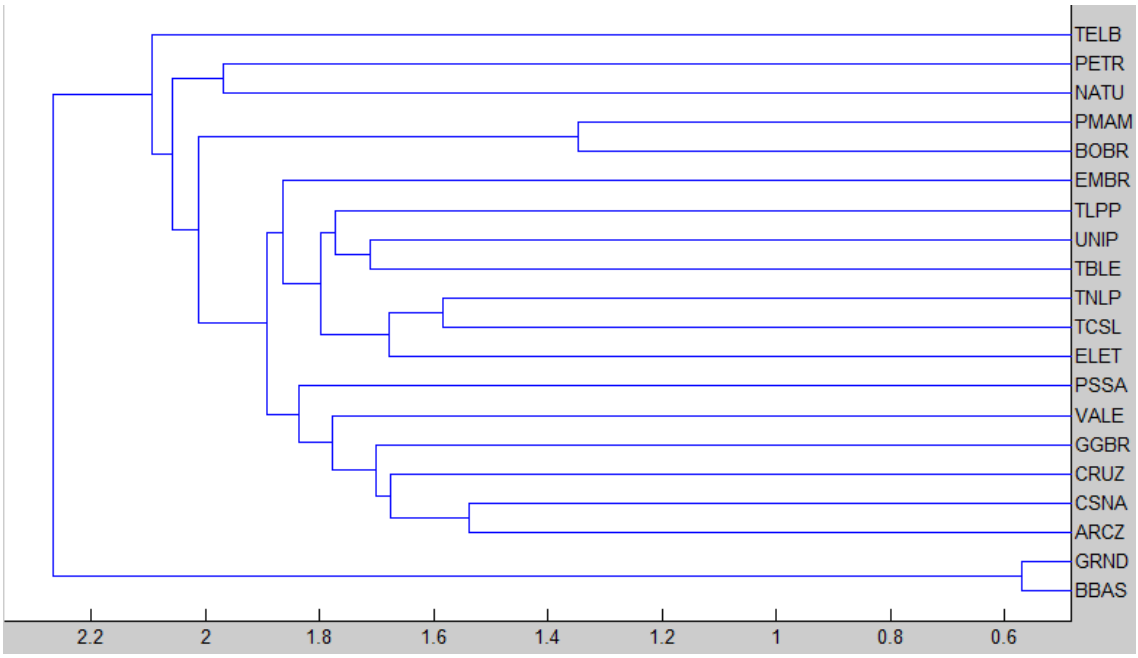

FIGURA 4 - DENDROGRAMA GERADO PELO *COMPLETE LINKAGE*

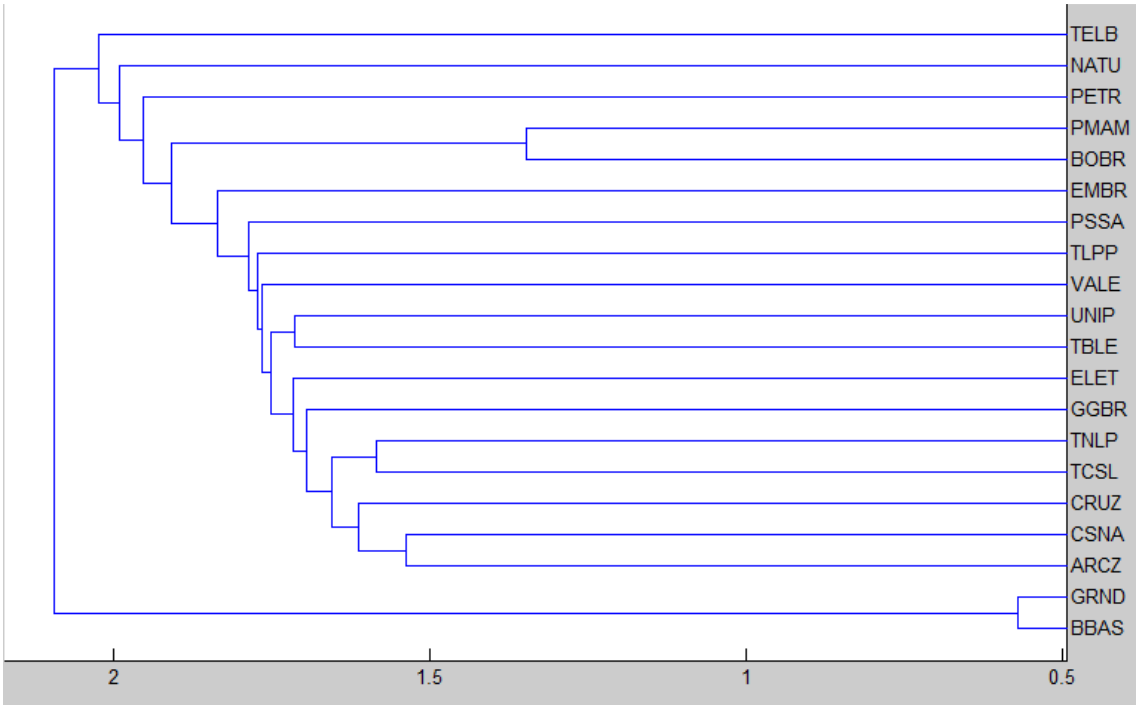

FIGURA 5 - DENDROGRAMA GERADO PELO *AVERAGE LINKAGE*

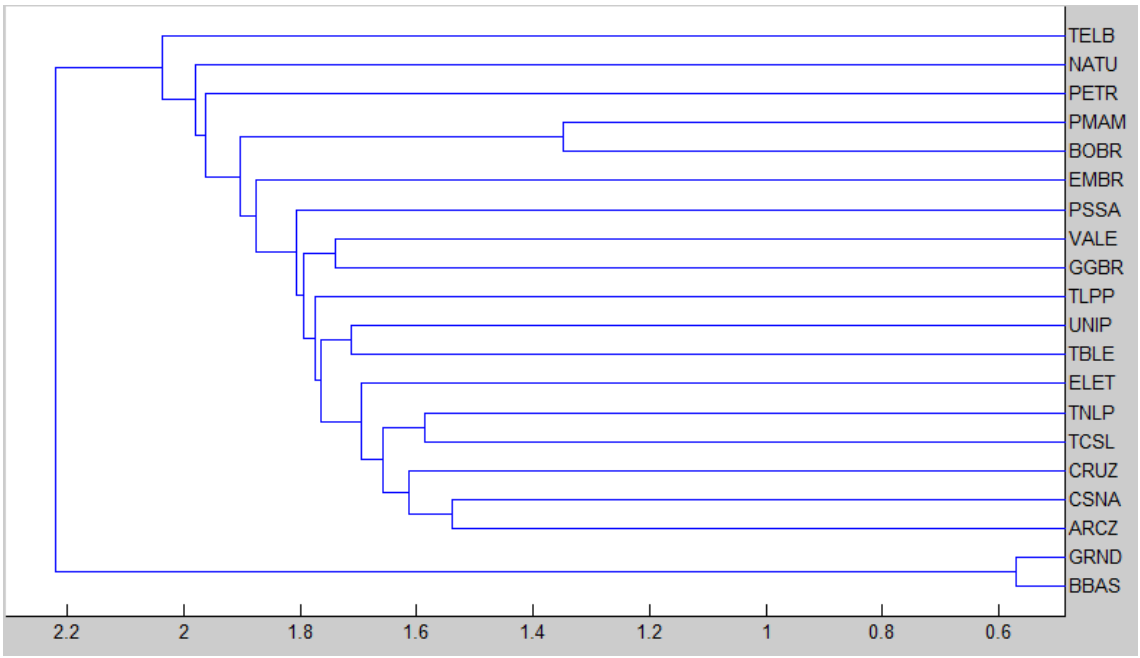

FIGURA 6 - DENDROGRAMA GERADO PELO *WEIGHTED LINKAGE*

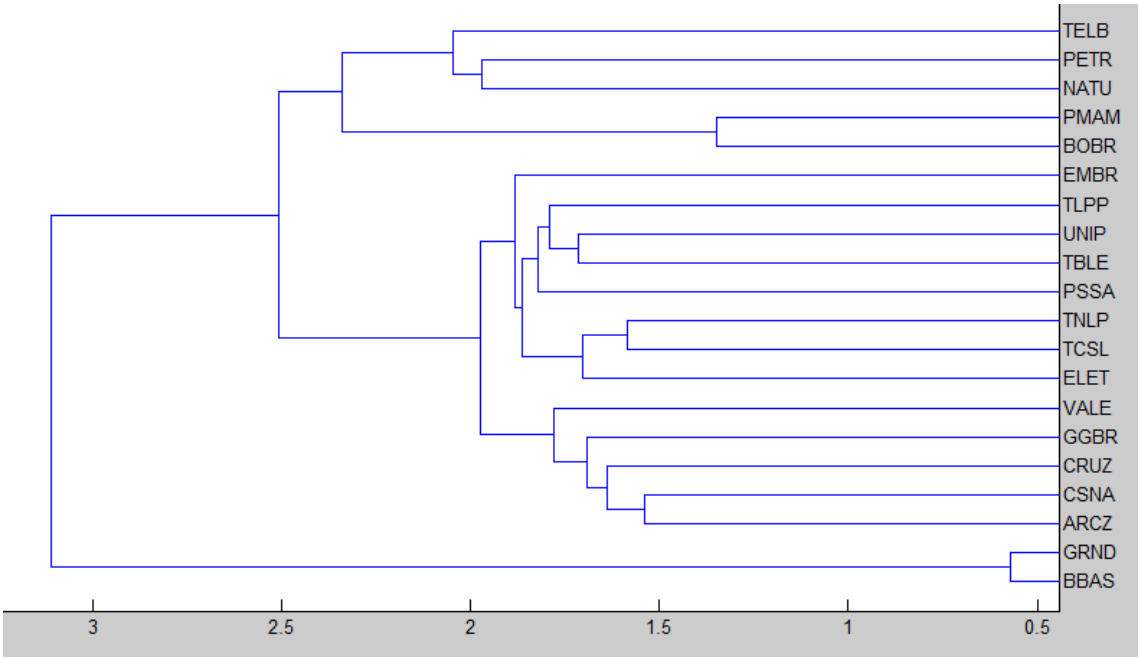

FIGURA 7 - DENDROGRAMA GERADO PELO *WARD LINKAGE*

Nos dendrogramas gerados pelo centroid linkage e pelo median linkage, nos deparamos com um fenômeno chamado de *crossover.* Esse fenômeno acontece quando um *cluster* é formado com um grau de semelhança maior do

que o *cluster* formado anteriormente, ou seja, com a distância menor [7]. Com isso, os dendrogramas formados ficam com a aparência a seguir:

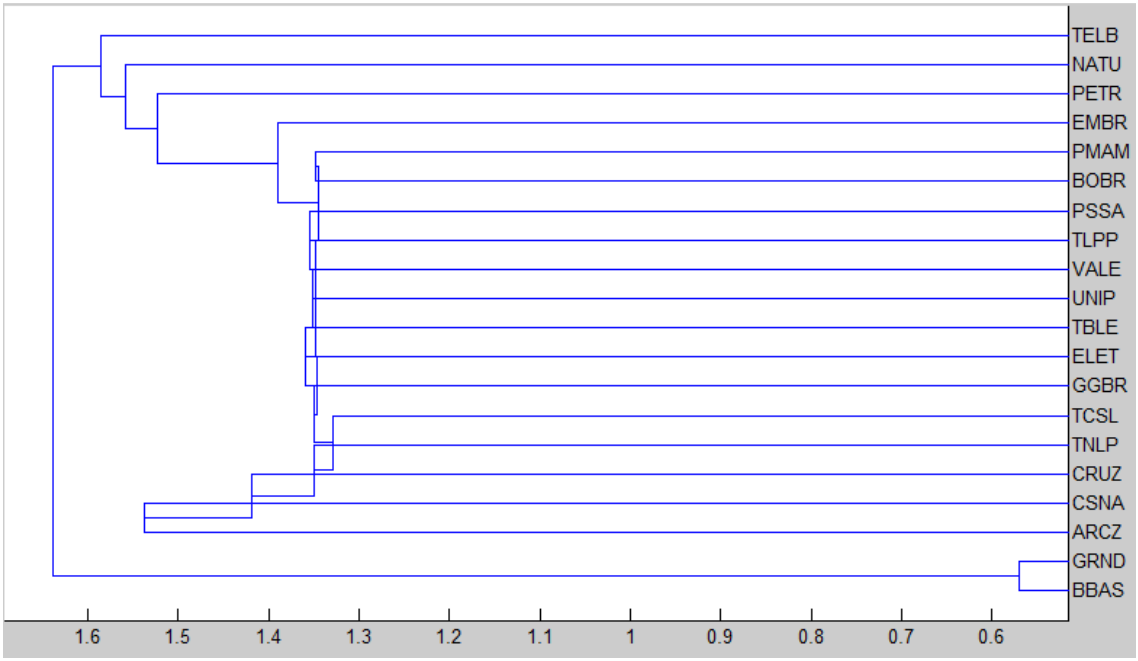

FIGURA 8 - DENDROGRAMA FORMADO COM O *CENTROID LINKAGE*

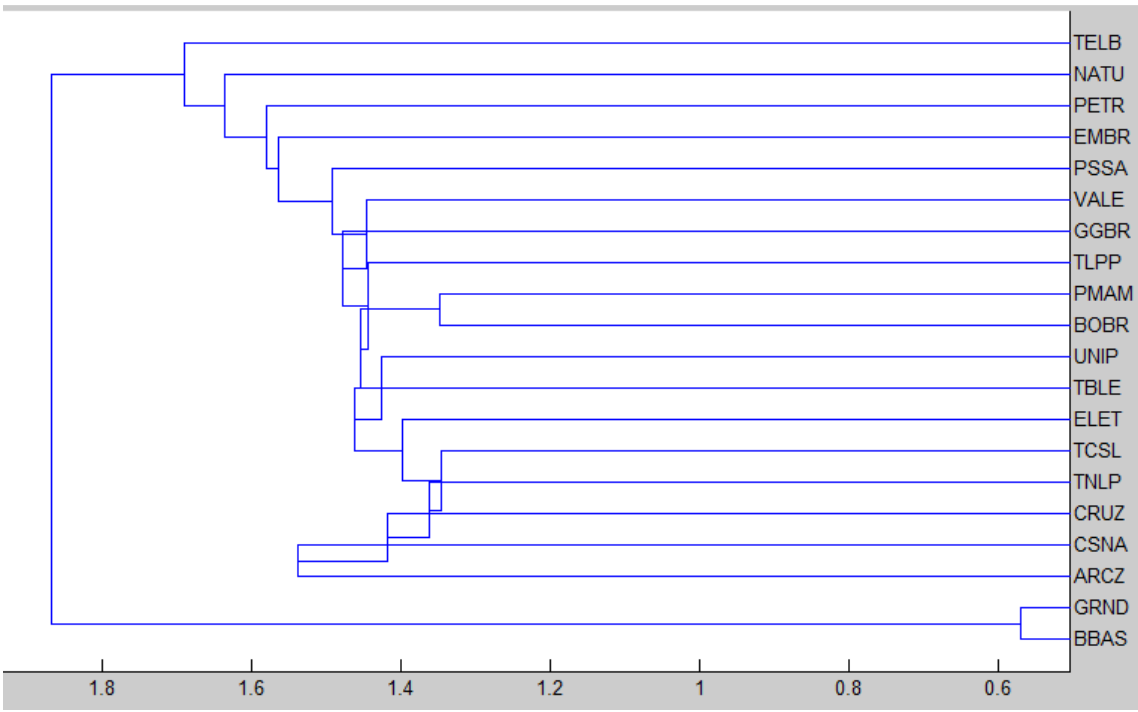

FIGURA 9 - DENDROGRAMA FORMADO COM O *MEDIAN LINKAGE*

Depois de gerar cada um dos dendrogramas, nós geramos grafos para cada um deles com a finalidade de observar melhor os clusters formados. Nós só não formamos grafos para o centroid linkage e o median linkage que são justamente os algoritmos que apresentaram o fenômeno chamado de *crossover.* A seguir veremos cada um dos grafos gerados pelos dendrogramas.

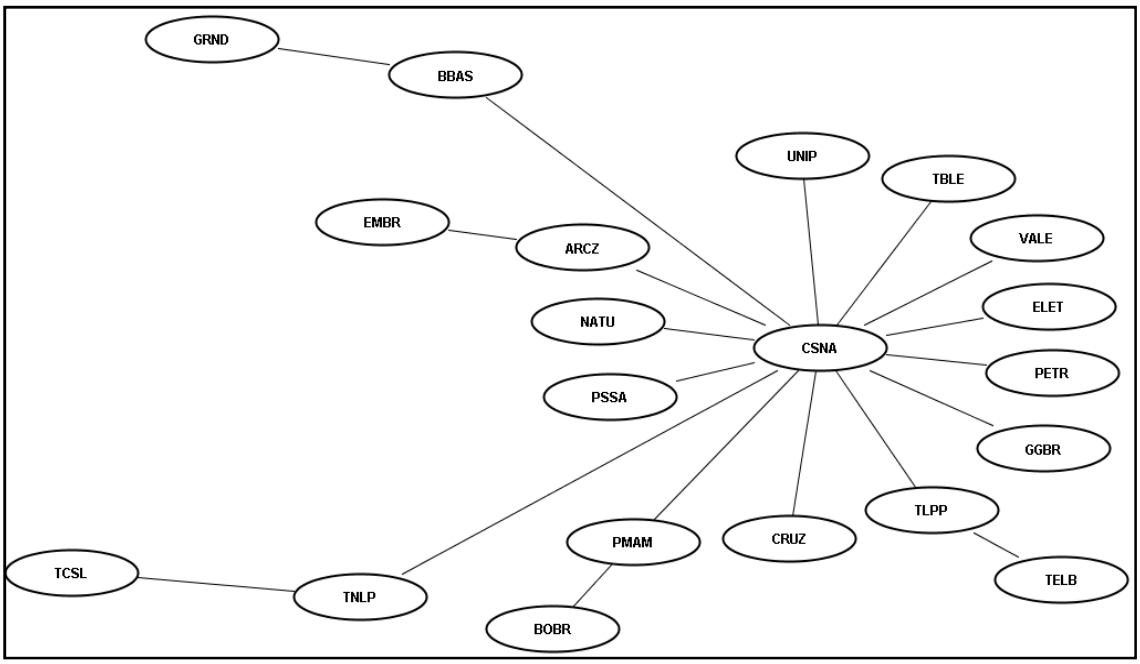

FIGURA 10 - GRAFO GERADO PELO DENDROGRAMA DO *SINGLE LINKAGE*

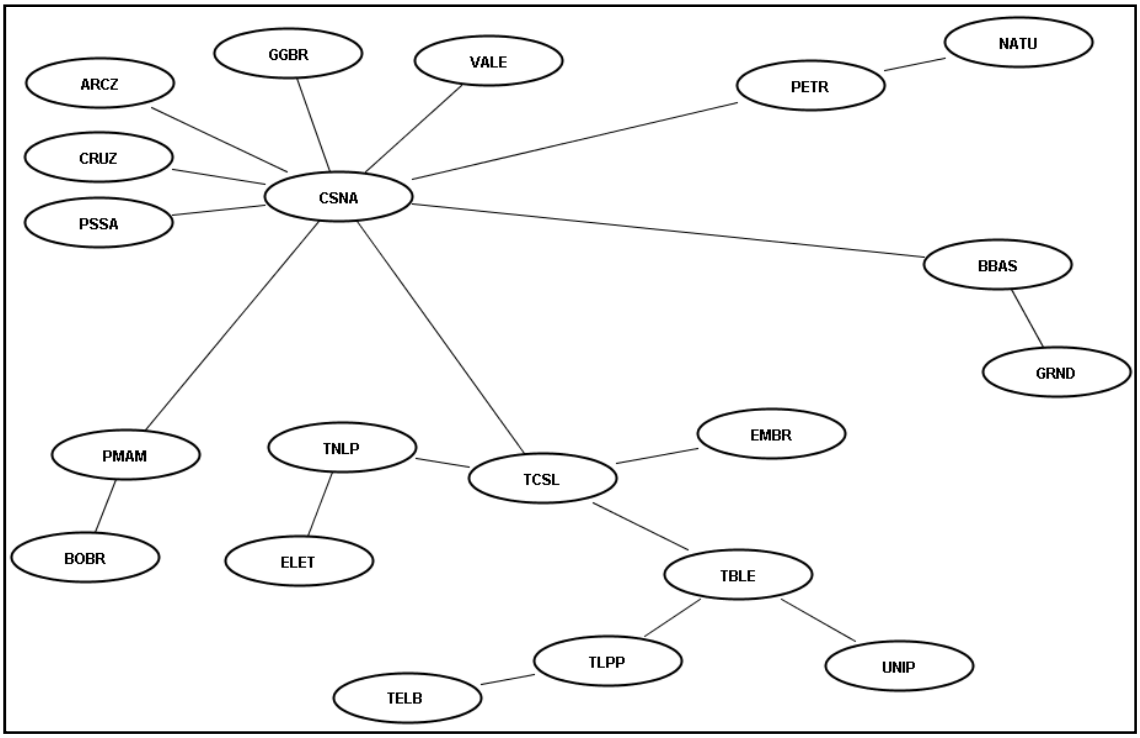

FIGURA 11 - GRAFO GERADO PELO DENDROGRAMA DO *COMPLETE LINKAGE*

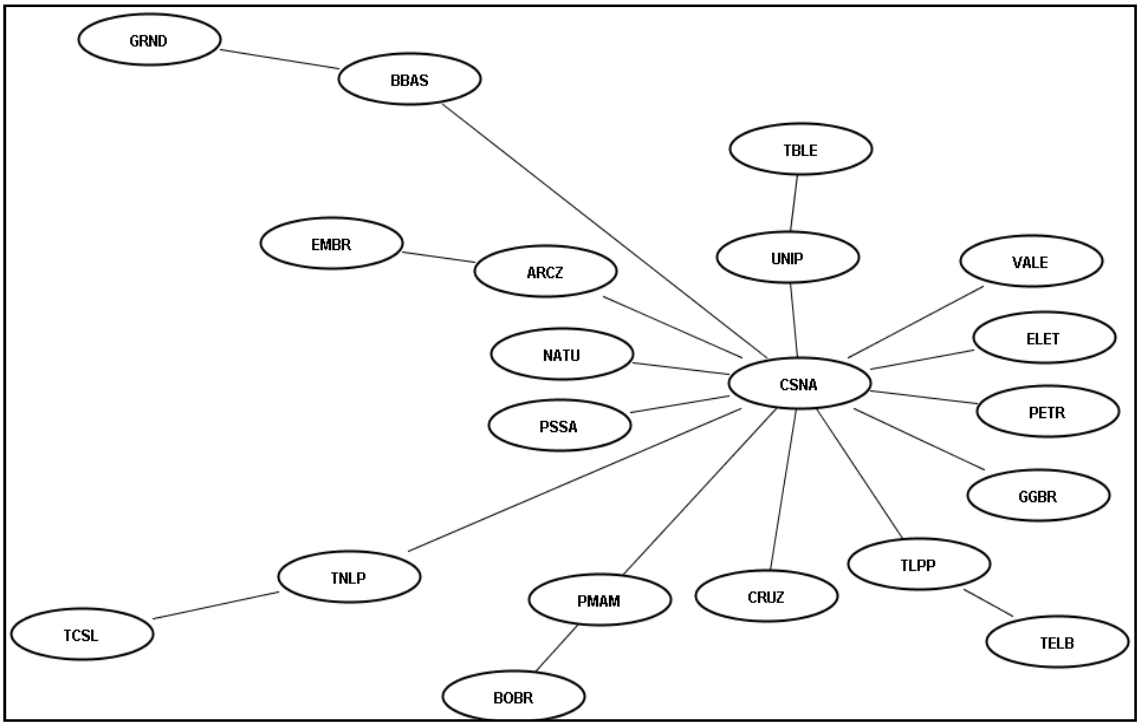

FIGURA 12 - GRAFO GERADO PELO DENDROGRAMA DO *AVERAGE LINKAGE*

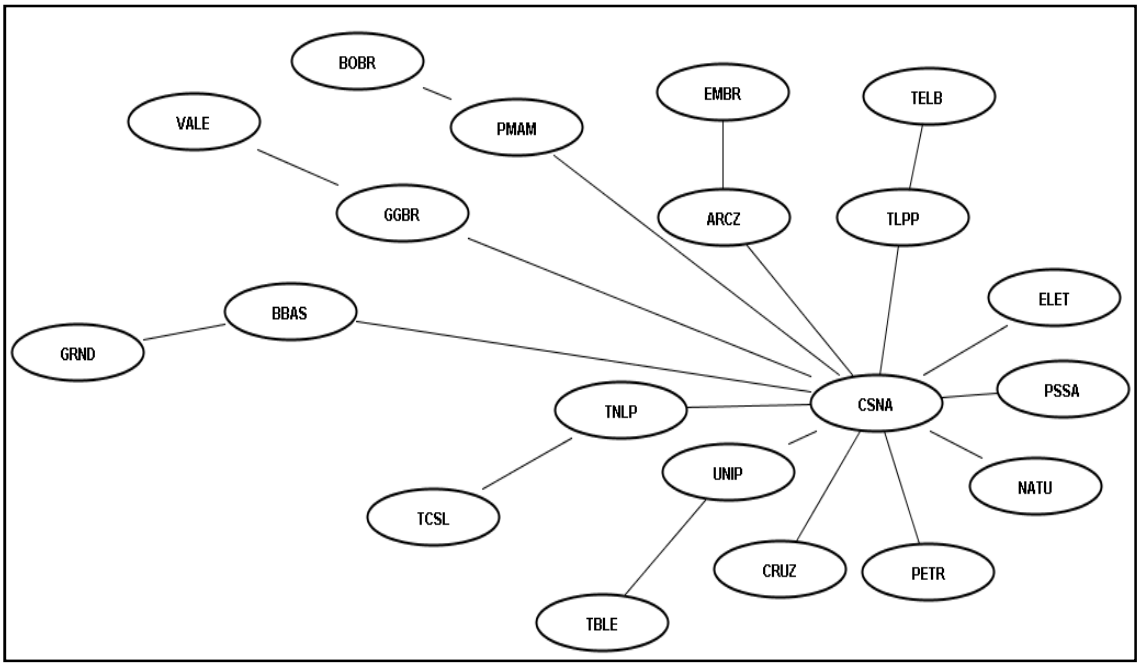

FIGURA 13 - GRAFO GERADO PELO DENDROGRAMA DO *WEIGHTED LINKAGE*

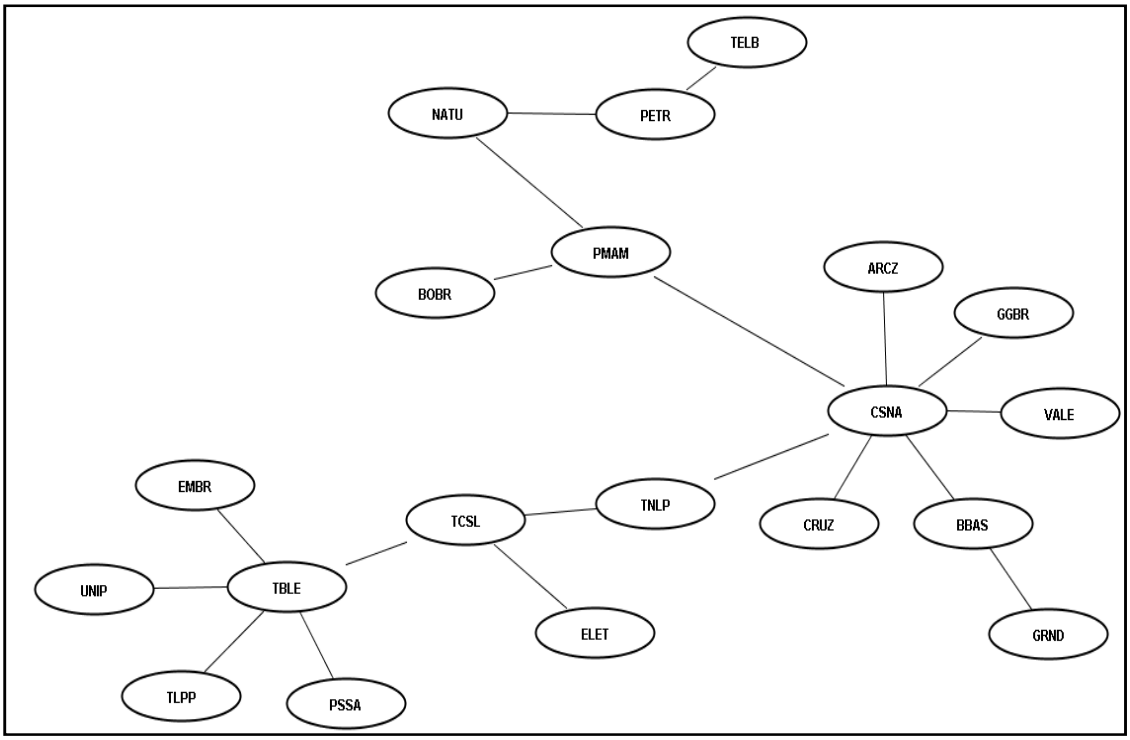

FIGURA 14 - GRAFO GERADO PELO DENDROGRAMA DO *WARD LINKAGE*

Lembrando que os grafos foram montados com base nos dendrogramas e as conexões entre os *clusters* foram feitas baseada na matriz de distância.

4. ANÁLISE E RESULTADOS

Para analisarmos o comportamento de cada um dos algoritmos de agrupamento, nós montamos carteiras de ações baseadas nos *clusters* formados por cada um desses algoritmos. Formamos uma carteira que chamamos de fortemente relacionada, uma carteira que chamamos de fracamente relacionada e uma carteira que chamamos de moderada, foi montada uma carteira de cada tipo para cada um dos algoritmos de agrupamento (*Single Linkage, Complete Linkage, Average Linkage, Weighted Linkage e Ward Linkage*). A carteira fortemente relacionada significa que todas as ações escolhidas para compor a carteira estão localizadas dentro do mesmo *cluster* baseado num determinado critério de distância. Já a carteira fracamente relacionada significa que as ações estão em *clusters* diferentes e a carteira moderada possui ações de um mesmo *cluster* e também de *clusters* diferentes. Também montamos uma carteira de ações aleatória, ou seja, uma carteira na qual as ações que a compõem foram escolhidas ao acaso. Feito isso, nós montamos uma tabela com os melhores resultados de cada um dos algoritmos. Como pode ser vista na tabela 1, o algoritmo de agrupamento que obteve o melhor resultado foi o algoritmo de *Ward*. Usaremos ele para demonstrar a metodologia que utilizamos para encontrar a distância "d" que otimiza a formação dos *clusters,* ou seja que nos fornece o melhor rendimento*,* pois nos outros algoritmos a metodologia aplicada foi a mesma.

Nossa base de dados contém o preço de fechamento de 20 ações no período de 2005 a 2007. Então, usaremos os dados do ano de 2008 para ter uma idéia de qual rendimento teríamos caso investíssemos nos resultados de cada um dos algoritmos de agrupamento. O cálculo do rendimento foi feito da seguinte maneira, nós compramos as ações no dia 28/12/2007 (último dia de negociação do ano de 2007) e montamos duas estratégias de venda, uma vendendo todas as ações antes da crise, ou seja, no dia 05/06/2008 cujo resultado consta na tabela 1 e outra vendendo as ações durante a crise, no dia 30/12/2008 cujo resultado consta na tabela 2. O capital inicial disponível para montar cada uma das carteiras é de R$ 50.000,00, então dentro desse valor nós vemos quantos lotes de cada ação podemos comprar e no dia da venda

vemos quanto eles estão valendo, a diferença entre o preço no dia da compra e o preço no dia da venda é o nosso lucro ou prejuízo.

A seguir mostraremos as ações contidas em cada uma das carteiras bem como a quantidade de lotes de cada uma para cada um dos algoritmos de agrupamento. A carteira 1 é a fortemente relacionada, a carteira 2 é a fracamente relacionada e a carteira 3 é a moderada.

Single Linkage	
Carteira 1	Dois lotes de CSNA, ARCZ e CRUZ
	Onze lotes de UNIP
	Um lote de GGBR
Carteira 2	Dois lotes de CSNA, NATU e TELB
	Três lotes de BOBR
	Um lote de PETR
Carteira 3	Dois lotes de CSNA, ELET e PMAM
	Um lote de PSSA e TBLE

TABELA 1 - CARTEIRAS FORMADAS COM *SINGLE LINKAGE*

Complete Linkage	
Carteira 1	Dois lotes de CSNA e ARCZ
	Um lote de CRUZ, PSSA e GGBR
Carteira 2	Onze lotes de TELB
	Cinco lotes de BOBR e NATU
	Quatro lotes de VALE e ELET
Carteira 3	Dois lotes de CSNA, ARCZ e TNLP
	Um lote de GRND e TBLE

TABELA 2 - CARTEIRAS FORMADAS COM O *COMPLETE LINKAGE*

Average Linkage	
Carteira 1	Dois lotes de PETR
	Um lote de CSNA, TLPP, ELET e CRUZ
Carteira 2	Cinco lotes de TELB
	Três lotes de TCSL, TBLE e BBAS
	Dois lotes de CSNA
Carteira 3	Dois lotes de CSNA, BOBR e GRND
	Um lote de PETR e TCSL

TABELA 3 - CARTEIRAS FORMADAS COM O *AVERAGE LINKAGE*

Weighted Linkage	
Carteira 1	Dois lotes de PETR
	Um lote de CSNA, TLPP, ELET e CRUZ
Carteira 2	Quatro lotes de GRND, VALE, BOBR, TBLE e TCSL
Carteira 3	Dois lotes de CSNA
	Um lote de CRUZ, VALE, PMAM e TNLP

TABELA 4 - CARTEIRAS FORMADAS COM O *WEIGHTED LINKAGE*

Ward Linkage	
Carteira 1	Três lotes de ARCZ
	Dois lotes de CSNA
	Um lote de GGBR, VALE e CRUZ
Carteira 2	Dois lotes de CSNA e PSSA
	Um lote de ELET, GRND e TCSL
Carteira 3	Dois lotes CSNA, NATU e BOBR
	Um lote de VALE e PSSA

TABELA 5 - CARTEIRAS FORMADAS COM O *WARD LINKAGE*

Então, no dia 28/12/2007 nós compramos cada uma dessas carteiras de ações e no dia 05/06/2008 vendemos as carteiras e montamos a tabela 1, a tabela 2 foi montada com base nos rendimentos caso as carteiras tivessem sido vendidas no dia 30/12/2008. É interessante ressaltar que os resultados apresentadas nas tabelas 1 e 2 foram os melhores rendimentos obtidos para cada um dos algoritmos de linkage, ou seja, colocamos nessas tabelas a configuração de *clusters* que possuem o valor ótimo da distância "d". Esse valor ótimo de "d" foi adquirido analisando o intervalo de tempo de 2005 a 2007 e depois aplicado a formação da nossa carteira de investimento para o ano de 2008.

	Carteira 1 (Altamente)	**Carteira 2 (Fracamente)**	**Carteira 3 (Moderada)**	**Carteira 4**
Single linkage	39%	11%	33%	x
Complete linkage	38,5%	25,8%	31,3%	x
Average linkage	20%	34,5%	11%	x
Weighted linkage	21%	-1,6%	31,6%	x
Ward linkage	40,1%	31,2%%	35%	x
Aleatório	x	x	x	-10%

TABELA 6 - RENDIMENTO EM 05/05/2008 (ANTES DA CRISE DE 2008)

	Carteira 1 (Altamente)	Carteira 2 (Fracamente)	Carteira 3 (Moderada)	Carteira 4
Single linkage	-43%	-49,5%	-41,5%	x
Complete linkage	-45%	-33,1%	-42,6%	x
Average linkage	-37,5%	-42%	-46%	x
Weighted linkage	-37,5%	-45,7%	-43%	x
Ward linkage	-47%	-41%	-43%	x
Aleatório	X	x	x	-55%

TABELA 7 - RENDIMENTO EM 30/12/2008 (DURANTE A CRISE DE 2008)

Como foi dito anteriormente, nós explicaremos a metodologia que aplicamos para escolher a distância "d" que otimiza a formação dos *clusters* utilizando o algoritmo de *ward* que foi o que obteve o melhor rendimento, segundo a tabela 1. Gostaríamos de ressaltar que para os outros algoritmos foi aplicada a mesma metodologia para encontrar os valores que constam nas tabelas 1 e 2 que contém os maiores rendimentos de cada um dos algoritmos de agrupamento.

Primeiramente, nós atribuímos alguns valores para "d" e daí observamos os *clusters* que eram formados.

- Para d < 1.8, segundo a figura 9, temos os seguintes *clusters:*

Clusters	
Cluster 1	CRUZ, CSNA, ARCZ, VALE E GGBR
Cluster 2	GRBD E BBAS
Cluster 3	TNLP, TCSL E ELET
Cluster 4	TLPP, UNIP, TBLE E PSSA
Cluster 5	PMAM E BOBR
Cluster 6	TELB
Cluster 7	PETR
Cluster 8	NATU
Cluster 9	EMBR

TABELA 8 - *CLUSTERS FORMADOS COM D < 1.8*

Como nós citamos anteriormente, nós montamos carteiras de investimento que chamamos de fracamente relacionada, fortemente relacionada e moderada. Então, uma carteira fracamente relacionada seria uma que possui ações que estivessem dentro de um mesmo *cluster*. Assim, formamos as seguintes carteiras para d < 1,8:

Carteira 1 (Fortemente) – 3 lotes de CSNA e CRUZ

2 lotes de ARCZ, VALE e GGBR

Com essa carteira nós obtivemos um rendimento de **59,5%,** caso comprássemos as ações no dia 03/01/2005 e vendêssemos em 28/12/2007.

Carteira 2 (Fracamente) – 2 lotes de PETR, CRUZ, TLPP, GRND e TNLP.

Com essa carteira nós obtivemos um rendimento de **129%**, caso comprássemos as ações no dia 03/01/2005 e vendêssemos em 28/12/2007.

Carteira 3 (Moderada) – 4 lotes de TCSL

3 lotes de ELET

2 lotes de CSNA, PETR e CRUZ.

Com essa carteira, nós obtivemos um rendimento de **186%**, caso comprássemos as ações no dia 03/01/2005 e vendêssemos em 28/12/2007.

- Para d < 1,9, segundo a figura 9, teremos os seguintes *clusters.*

Clusters	
Cluster 1	VALE, GGBR, CRUZ, CSNA E ARCZ
Cluster 2	EMBR, TLPP, UNIP, TBLE, PSSA, TNLP, TCSL E ELET
Cluster 3	PMAM E BOBR
Cluster 4	PETR
Cluster 5	NATU
Cluster 6	TELB
Cluster 7	GRND E BBAS

TABELA 9 - *CLUSTERS FORMADOS COM D < 1.9*

Assim, formamos as seguintes carteiras para d < 1.9

Carteira 1 (Fortemente) – 3 lotes de VALE

2 lotes de GGBR, CRUZ, CSNA e ARCZ

Com essa carteira, nós obtivemos um rendimento de **77%**, caso comprássemos as ações no dia 03/01/2005 e vendêssemos em 28/12/2007.

Carteira 2 (Fracamente) – 4 lotes de VALE, EMBR, TELB, GRND E PMAM

Com essa carteira, nós obtivemos um rendimento de **42%**, caso comprássemos as ações no dia 03/01/2005 e vendêssemos em 28/12/2007.

Carteira 3 (Moderada) – 2 lotes de VALE, EMBR E PETR

1 lote de CRUZ E NATU

Com essa carteira, nós obtivemos um rendimento de **149%**, caso comprássemos as ações no dia 03/01/2005 e vendêssemos em 28/12/2007.

- Para d < 2.04, segundo a figura 9, teremos os seguintes *clusters:*

Clusters	
Cluster 1	TELB, PETR E NATU
Cluster 2	PMAM E BOBR
Cluster 3	EMBR, TLPP, UNIP, TBLE, PSSA, TNLP, TCSL, ELET, VALE, GGBR, CRUZ, CSNA E ARCZ.
Cluster 4	GRND E BBAS

TABELA 10 - *CLUSTERS* FORMADOS COM D < 2.04

Assim, formamos as seguintes carteiras para d < 2.04:

Carteira 1 (Fortemente) – 1 lote de EMBR, VALE, CRUZ, PSSA, ARCZ, TLPP, UNIP, TBLE, TCSL E ELET.

2 lotes de CSNA, GGBR e TNLP.

Com essa carteira, nós obtivemos um rendimento de **66%**, caso comprássemos as ações no dia 03/01/2005 e vendêssemos em 28/12/2007.

Carteira 2 (Fracamente) – 7 lotes VALE, BBAS, NATU E CSNA

2 lotes PMAM

Com essa carteira, nós obtivemos um rendimento de **63%**, caso comprássemos as ações no dia 03/01/2005 e vendêssemos em 28/12/2007.

Carteira 3 (moderada) – 3 lotes de PSSA

2 lotes de CSNA e NATU

1 lote de VALE E PETR

Com essa carteira, nós obtivemos um rendimento de **146%**, caso comprássemos as ações no dia 03/01/2005 e vendêssemos em 28/12/2007.

Com isso, temos os valores dos rendimentos para cada uma dessas três distâncias “d” utilizadas, agora, nos resta observar qual a distância nos proporcionou o maior rendimento.

Analisando os resultados, percebemos que quando a $d < 1.8$, nós obtivemos, na média os melhores resultados. Com isso, nós utilizamos essa distância “d” como critério para criar os nossos *clusters* para criar as carteiras de investimentos para o ano de 2008 cujos rendimentos constam nas tabelas 1 e 2.

5. CONCLUSÃO E TRABALHOS FUTUROS

Como foi dito no inicio do trabalho, o nosso objetivo é otimizar o processo de formação de carteiras de investimentos, ou seja, formar carteira para o investidor informando a ele os riscos inerentes a cada um daqueles investimentos. Por exemplo, a carteira 1, do algoritmo de ward, é fortemente relacionada, isso significa dizer que caso a ação da companhia siderúrgica nacional (CSNA) suba (e foi isso que aconteceu), a tendência é que as outras ações que estão não mesmo *cluster* que ela também subam. No entanto, isso é arriscado, pois, nesse caso, você está investindo em ações que se comportam de uma mesma maneira sem saber ao certo se ela tem uma tendência maior de queda ou subida. Então, é menos arriscado (mas acarretará rendimentos menores) você investir em uma carteira moderada, pois ela mescla ações de diferentes *clusters*, é um intermediário entre a carteira fortemente relacionada (que possui ações de um mesmo *cluster*) e a fracamente relacionada (que possui ações de *clusters*, obrigatoriamente, diferentes). Isso é comprovado pela tabela 1, olhando para os resultados do algoritmo de ward, vemos que a carteira 1 obteve cerca de 40% de rendimento, a carteira 2 obteve 31% e a carteira 3 obteve 35%, como nós dissemos a carteira 1, que é fortemente relacionada, obteve o maior rendimento pois as ações do *cluster* estavam numa tendência de subida, no entanto, se elas estivessem numa tendência de descida o rendimento teria sido muito baixo ou inexistente, fato que não aconteceria numa carteira moderada pois a queda de um *cluster* seria "compensada" pela subida de outro, então concluímos que na carteira moderada o rendimento atingido pode ser menor mas o risco também é menor.

Um fato interessante que podemos notar é que na época da crise, o ward linkage não foi o que obteve os melhores rendimentos, no entanto, nós não levamos isso muito em consideração na nossa análise pois crise é um fato extraordinário, não é o comum. Então, julgamos, que é muito mais relevante, nós analisarmos o desempenho de um algoritmo nas condições normais.

Obtivemos esses resultados a partir dessa primeira análise do comportamento dos ativos da Bovespa, no entanto percebemos que ainda existem algumas coisas que podemos melhorar para conseguir rendimentos

cada vez maiores, algumas delas são utilizar não só o preço de fechamento mas também o volume negociado de cada uma das ações, utilizar alguma técnica para otimizar o valor da distância “d” que utilizamos para formar os *clusters,* utilizar a distância de hausdorff para agrupar as ações e elaborar uma estratégia de investimento que seja mais condizente com a realidade e que não seja tão simples como a que nós utilizamos nesse trabalho. Nós achamos que ao utilizar essas técnicas e talvez juntarmos a análise matemática que fizemos aqui com uma análise fundamentalista podemos obter resultados melhores. Outro trabalho que iremos executar é usar essa nossa metodologia para detectar séries exógenas [10].

6. ANEXOS

ANEXO 1 – AÇÕES UTILIZADAS

Ações utilizadas com o respectivo código e nome:

ARCZ6 – Aracruz (Preferencial)

BBAS3 – Banco do Brasil (Ordinária)

BOBR4 – Bombril (Preferencial)

CRUZ3 – Souza Cruz (Ordinária)

CSNA3 – Companhia Siderúrgica Nacional (Ordinária)

ELET3 – Eletrobrás (Ordinária)

EMBR3 – Embraer (Ordinária)

GGBR 3 – Gerdau (Ordinária)

GRND3 – Grendene (Ordinária)

NATU3 – Natura (Ordinária)

PETR3 – Petrobrás (Ordinária)

PMAM4 – Paranapanema (Ordinária)

PSSA3 – Porto Seguro (Ordinária)

TBLE3 – Tractebel (Ordinária)

TCSL3 – TIM (Ordinária)

TELB3 – Telebrás (Ordinária)

TLPP3 – Telesp (Ordinária)

TNLP3 – Telemar (Ordinária)

UNIP6 – Unipar (Preferencial)

VALE3 - Vale do Rio Doce (Ordinária)

ANEXO 2 – MATRIZ DE CORRELAÇÃO

	ARCZ	BBAS	BOBR	CRUZ	CSNA	ELET	EMBR	GGBR	GRND	NATU	PETR	PMAM	PSSA	TBLE	TCSL	TELB	TLPP	TNLP	UNIP	VALE
ARCZ	1	0,2032	0,1149	0,3025	0,4328	0,2343	0,2462	0,2924	0,2004	0,0833	0,1044	0,1676	0,2078	0,2314	0,2888	0,0664	0,1973	0,3091	0,2001	0,315
BBAS	0,2032	1	0,0568	0,1552	0,2334	0,1662	0,0879	0,1525	0,923	-0,0042	0,0585	0,1255	0,1438	0,1254	0,1722	0,0344	0,1767	0,1834	0,1378	0,1443
BOBR	0,1149	0,0568	1	0,095	0,2213	0,0513	0,1251	0,0908	0,1193	0,0411	0,031	0,5586	0,1501	0,1546	0,0978	0,0091	0,1437	0,1473	0,1179	0,0439
CRUZ	0,3025	0,1552	0,095	1	0,4247	0,2728	0,1624	0,2826	0,2276	0,1142	0,1069	0,1623	0,2264	0,249	0,2733	0,1089	0,2688	0,3484	0,2444	0,2206
CSNA	0,4328	0,2334	0,2213	0,4247	1	0,3448	0,2312	0,4111	0,2239	0,1209	0,1854	0,3057	0,3076	0,3226	0,3713	0,1038	0,3008	0,3719	0,327	0,3301
ELET	0,2343	0,1662	0,0513	0,2728	0,3448	1	0,2035	0,1939	0,1513	0,0863	0,0977	0,2009	0,1316	0,196	0,3038	0,0444	0,1987	0,316	0,1959	0,1958
EMBR	0,2462	0,0879	0,1251	0,1624	0,2312	0,2035	1	0,122	0,0626	0,0911	0,126	0,1534	0,1167	0,2101	0,2095	0,0645	0,1415	0,2084	0,196	0,1307
GGBR	0,2924	0,1525	0,0908	0,2826	0,4111	0,1939	0,122	1	0,178	0,0732	0,1227	0,176	0,2208	0,1888	0,2107	0,0937	0,1958	0,2718	0,2266	0,2486
GRND	0,2004	0,923	0,1193	0,2276	0,2239	0,1513	0,0626	0,178	1	0,0749	0,038	0,1497	0,1824	0,1211	0,1716	-0,0042	0,1918	0,1891	0,2269	0,1508
NATU	0,0833	0,0042	0,0411	0,1142	0,1209	0,0863	0,0911	0,0732	0,0749	1	0,0376	0,0758	-0,0032	0,0957	0,0719	-0,0728	0,0716	0,0853	0,079	0,0856
PETR	0,1044	0,0585	0,031	0,1069	0,1854	0,0977	0,126	0,1227	0,038	0,0376	1	0,068	0,0912	0,0448	0,1715	0,0322	0,0708	0,1207	0,0734	0,0986
PMAM	0,1676	0,1255	0,5586	0,1623	0,3057	0,2009	0,1534	0,176	0,1497	0,0758	0,068	1	0,1542	0,1809	0,1937	0,0585	0,1772	0,1909	0,2082	0,1559
PSSA	0,2078	0,1438	0,1501	0,2264	0,3076	0,1316	0,1167	0,2208	0,1824	-0,0032	0,0912	0,1542	1	0,2125	0,2814	0,0133	0,1863	0,2599	0,1891	0,1638
TBLE	0,2314	0,1254	0,1546	0,249	0,3226	0,196	0,2101	0,1888	0,1211	0,0957	0,0448	0,1809	0,2125	1	0,2949	0,0586	0,2196	0,293	0,2709	0,1544
TCSL	0,2888	0,1722	0,0978	0,2733	0,3713	0,3038	0,2095	0,2107	0,1716	0,0719	0,1715	0,1937	0,2814	0,2949	1	0,0685	0,2342	0,3749	0,2494	0,1865
TELB	0,0664	0,0344	0,0091	0,1089	0,1038	0,0444	0,0645	0,0937	-0,0042	-0,0728	0,0322	0,0585	0,0133	0,0586	0,0685	1	0,1357	0,0774	0,0145	0,0034
TLPP	0,1973	0,1767	0,1437	0,2688	0,3008	0,1987	0,1415	0,1958	0,1918	0,0716	0,0708	0,1772	0,1863	0,2196	0,2342	0,1357	1	0,2524	0,2192	0,1894
TNLP	0,3091	0,1834	0,1473	0,3484	0,3719	0,316	0,2084	0,2718	0,1891	0,0853	0,1207	0,1909	0,2599	0,293	0,3749	0,0774	0,2524	1	0,254	0,2541
UNIP	0,2001	0,1378	0,1179	0,2444	0,327	0,1959	0,196	0,2266	0,2269	0,079	0,0734	0,2082	0,1891	0,2709	0,2494	0,0145	0,2192	0,254	1	0,2246
VALE	0,315	0,1443	0,0439	0,2206	0,3301	0,1959	0,1307	0,2486	0,1508	0,0856	0,0986	0,1559	0,1638	0,1544	0,1865	0,0034	0,1894	0,2541	0,2246	1

1

2 ANEXO 3 – MATRIZ DE DISTÂNCIA

	ARCZ	BBAS	BOBR	CRUZ	CSNA	ELET	EMBR	GGBR	GRND	NATU	PETR	PMAM	PSSA	TBLE	TCSL	TELB	TLPP	TNLP	UNIP	VALE
ARCZ	0	1,2623	1,3304	1,1811	1,065	1,2374	1,2278	1,1896	1,2645	1,354	1,3383	1,2902	1,2587	1,2398	1,1926	1,3664	1,267	1,1754	1,2648	1,1704
BBAS	1,2623	0	1,3734	1,2998	1,2382	1,2913	1,3506	1,3019	0,3924	1,4117	1,3722	1,3224	1,3085	1,3225	1,2867	1,3896	1,2831	1,2779	1,3131	1,3082
BOBR	1,3304	1,3734	0	1,3453	1,2479	1,3774	1,3228	1,3484	1,3271	1,3848	1,3921	0,9395	1,3037	1,3003	1,3432	1,4077	1,3086	1,3059	1,3282	1,3828
CRUZ	1,1811	1,2998	1,3453	0	1,0726	1,2059	1,2942	1,1978	1,2428	1,331	1,3364	1,2943	1,2438	1,2255	1,2055	1,3349	1,2092	1,1415	1,2293	1,2485
CSNA	1,065	1,2382	1,2479	1,0726	0	1,1447	1,24	1,0852	1,2458	1,3259	1,2764	1,1783	1,1767	1,1639	1,1213	1,3388	1,1825	1,1208	1,1601	1,1574
ELET	1,2374	1,2913	1,3774	1,2059	1,1447	0	1,2621	1,2697	1,3028	1,3518	1,3433	1,2641	1,3178	1,268	1,18	1,3824	1,2659	1,1696	1,2681	1,2682
EMBR	1,2278	1,3506	1,3228	1,2942	1,24	1,2621	0	1,3251	1,3692	1,3482	1,3221	1,3012	1,3291	1,2569	1,2573	1,3678	1,3103	1,2582	1,268	1,3185
GGBR	1,1896	1,3019	1,3484	1,1978	1,0852	1,2697	1,3251	0	1,2821	1,3614	1,3246	1,2837	1,2483	1,2737	1,2564	1,3463	1,2682	1,2068	1,2437	1,2258
GRND	1,2645	0,3924	1,3271	1,2428	1,2458	1,3028	1,3692	1,2821	0	1,3602	1,387	1,304	1,2787	1,3258	1,2871	1,4171	1,2713	1,2734	1,2434	1,3037
NATU	1,354	1,4112	1,3848	1,331	1,3259	1,3518	1,3482	1,3614	1,3602	0	1,3873	1,3595	1,4164	1,3448	1,3624	1,4647	1,3626	1,3525	1,3572	1,3523
PETR	1,3383	1,3722	1,3921	1,3364	1,2764	1,3433	1,3221	1,3246	1,387	1,3873	0	1,3653	1,3481	1,3821	1,2872	1,3912	1,3632	1,3261	1,3613	1,3426
PMAM	1,2902	1,3224	0,9395	1,2943	1,1783	1,2641	1,3012	1,2837	1,304	1,3595	1,3652	0	1,3006	1,2799	1,2698	1,3722	1,2828	1,272	1,2584	1,2993
PSSA	1,2587	1,3085	1,3037	1,2438	1,1767	1,3178	1,3291	1,2483	1,2787	1,4164	1,3481	1,3006	0	1,2549	1,1988	1,4047	1,2756	1,2166	1,2734	1,2932
TBLE	1,2398	1,3225	1,3003	1,2255	1,1639	1,268	1,2569	1,2737	1,3258	1,3448	1,3821	1,2799	1,2549	0	1,1875	1,3721	1,2493	1,1891	1,2075	1,3004
TCSL	1,1926	1,2867	1,3432	1,2055	1,1213	1,18	1,2573	1,2564	1,2871	1,3624	1,2872	1,2698	1,1988	1,1875	0	1,3649	1,2375	1,1181	1,2252	1,2755
TELB	1,3664	1,3896	1,4077	1,3349	1,3388	1,3824	1,3678	1,3463	1,4171	1,4647	1,3912	1,3722	1,4047	1,3721	1,3649	0	1,3114	1,3583	1,4039	1,4118
TLPP	1,267	1,2831	1,3086	1,2092	1,1825	1,2659	1,3103	1,2682	1,2713	1,3626	1,3632	1,2828	1,2759	1,2493	1,2375	1,3147	0	1,2227	1,2496	1,2732
TNLP	1,1754	1,2779	1,3059	1,1415	1,1208	1,1696	1,2582	1,2068	1,2734	1,3525	1,3261	1,272	1,2166	1,1891	1,1181	1,3583	1,2227	0	1,2214	1,2213
UNIP	1,2648	1,3131	1,3282	1,2293	1,1601	1,2681	1,268	1,2437	1,2434	1,3572	1,3613	1,2584	1,2734	1,2075	1,2252	1,4039	1,2496	1,2214	0	1,2453
VALE	1,1704	1,3082	1,3828	1,2485	1,1574	1,2681	1,3185	1,2258	1,3032	1,3523	1,3426	1,2993	1,2932	1,3004	1,2755	1,4118	1,2732	1,2213	1,2453	0

7. REFERÊNCIA BIBLIOGRÁFICA

[1] **BOVESPA BM&F**. BM&F BOVESPA a Nova Bolsa. *BM&F Bovespa a Nova Bolsa.* [Online] [Citado em: 15 de 05 de 2009.] www.bovespa.com.br.

[2] Jain, A. K., Murty, M. N., and Flynn, P. **Data Clustering: a review**. ACM Computing Surveys, 1999.

[3] ELDER, Alexander. **Aprenda a operar no mercado de ações**. 1. Ed. São Paulo: Campus, 2005. 340 p.

[4] HISSA, Maurício, **Sobreviva na bolsa de valores**. 1. Ed. São Paulo: Campus, 2008. 234 p.

[5] CERBASI, Gustavo. **Investimentos inteligentes.** 1. Ed. São Paulo: Thomas Nelson Brasil, 2008. 272 p.

[6] PIAZZA, Marcelo. **Bem-vindo à Bolsa de Valores**. 7. Ed. São Paulo: Novo Conceito, 2008. 200 p.

[7] THEODORIDIS, Sergios; KOUTROUMBAS, Konstantinos. **Pattern Recognition**. 2. Ed. San Diego: Academic Press, 2003. 685 p.

[8] MANTEGNA, R. N. **Hierarchical structure in financial markets,** The European Physical Journal, 1999.

[9] BASALTO, N.; BELLOTI, R.; DE CARLO, F. **Hausdorff Clustering of Financial Time Series,** 2007.

[10] NETO, Manoel Christovam de Amorim. **Previsão de séries temporais usando séries exógenas e combinação de redes neurais aplicada ao mercado financeiro.** Dissertação de mestrado. Recife, PE, Brasil: Centro de Informática – UFPE, Dezembro de 2008.

ASSINATURAS

Tsang Ing Ren

Orientador

André Luis dos Santos Alves

Aluno