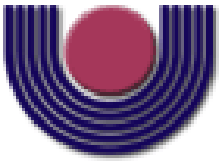


UNIOESTE – Universidade Estadual do Oeste do Paraná

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

Colegiado de Ciência da Computação

Curso de Bacharelado em Ciência da Computação

Agrupamento e Visualização de Dados: Estudo e Implementações para a Ferramenta YADMT

Mateus Felipe Teixeira

CASCAVEL

2013

MATEUS FELIPE TEIXEIRA

AGRUPAMENTO E VISUALIZAÇÃO DE DADOS: ESTUDO E IMPLEMENTAÇÕES PARA A FERRAMENTA YADMT

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação, do Centro de Ciências Exatas e Tecnológicas da Universidade Estadual do Oeste do Paraná - Campus de Cascavel.

Orientador: Prof. Dr. Clodis Boscarioli

Co-Orientadora: Prof[a]. Dr[a]. Rosangela Villwock

CASCAVEL

2013

MATEUS FELIPE TEIXEIRA

AGRUPAMENTO E VISUALIZAÇÃO DE DADOS: ESTUDO E IMPLEMENTAÇÕES PARA A FERRAMENTA YADMT

Monografia apresentada como requisito parcial para obtenção do Título de *Bacharel em Ciência da Computação*, pela Universidade Estadual do Oeste do Paraná, Campus de Cascavel, aprovada pela Comissão formada pelos professores:

Prof. Dr. Clodis Boscarioli (Orientador)
Colegiado de Ciência da Computação,
UNIOESTE

Prof.ª Dr.ª Rosangela Villwock (Co-Orientadora)
Colegiado de Ciência da Computação,
UNIOESTE

Prof. MEng. Carlos José Maria Olguín
Colegiado de Ciência da Computação,
UNIOESTE

Prof. Dr. Jerry Adriani Johann
Centro de Ciências Exatas e Tecnológicas,
UNIOESTE

Cascavel, 17 de outubro de 2013.

DEDICATÓRIA

Aos meus pais, João e Claudia, que me apoiaram e sempre me deram todas as condições para que eu pudesse aproveitar todas as oportunidades que eu tive. À Caroline, que também sempre me apoiou e aguentou os momentos difíceis ao meu lado me fazendo continuar.

AGRADECIMENTOS

Aos meus pais, João Luis e Claudia, que sempre me apoiaram nas minhas decisões e sempre que precisei estiveram ao meu lado me oferecendo tudo o que podiam para eu alcançar meus objetivos.

À minha namorada, Caroline, que em todos estes anos de graduação sempre esteve ao meu lado me ajudando e dando apoio nas horas mais difíceis, me ajudando também nas minhas escolhas e decisões, e mesmo longe, grande parte das vezes, nunca me deixou sozinho, e que seu amor e carinho a mim sempre foram um grande motivador para continuar.

Aos professores Clodis e Rosangela que caminharam ao meu lado durante toda a graduação, sendo professores de disciplinas e orientadores dos meus projetos de iniciação científica e também orientadores deste trabalho. Obrigado pelos conhecimentos que me transmitiram além de toda a seriedade, compromisso, responsabilidade e modo de trabalho.

Aos professores do Curso de Ciência da Computação pelas disciplinas ofertadas e também todo o conhecimento passado.

Aos meus colegas de turma, que viraram amigos, Thiago, Gustavo, Leandro, Astério, Wilson e Eduardo que sempre nos ajudamos nas horas de provas e trabalhos para vencermos juntos esta etapa, e também nossos momentos de risadas e descontrações que nos davam forças para continuar. Também a todos meus colegas que fizeram disciplinas e trabalhos junto comigo, sempre nos ajudando.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

KDD	*Knowledge Discovery in Databases*
YADMT	*Yet Another Data Mining Tool*
ACO	*Ant Colony Optimization*
SQE	Soma do Quadrado do Erro
2D	Bi-dimensional
3D	Tri-dimensional
MS	*MicroSoft*
SGBD	Sistema Gerenciador de Banco de Dados
ARFF	*Attribute-Relation File Format*
RNA	Rede Neural Artificial
MLP	*Multilayer Perceptron*
RBF	*Radial Basis Function*
LVQ	*Learning Vector Quantization*
SOTA	*Self-Organizing Tree Algorithm*
ROC	*Receiver Operating Characteristics*
SOM	*Self-Organizing Maps*
GIA	Grupo de Pesquisa em Inteligência Aplicada

Lista de Símbolos

P_{pick}	Probabilidade de carregamento de Padrão
P_{drop}	Probabilidade de descarregamento de Padrão
f(i)	Estimativa da fração de padrões localizados na vizinhança
$d(i, j)$	Função de dissimilaridade entre padrões
N_{cell}	Número de células analisadas para uma posição
α	Porcentagem de padrões na grade classificados como semelhantes
σ	Raio de percepção da vizinhança
F	Medida F
R	Índice Aleatório
V	Variância Intra-Grupos

Sumário

Resumo

Atualmente, há uma grande quantidade de dados sendo produzida a cada instante, sendo que estes volumes de dados podem conter informações que aplicando uma decisão sobre estes dados pode se agregar grande valor para estes dados. Porém, é necessária a análise e extração de conhecimentos contidos nesses dados. Para um analista humano esta função é praticamente inviável, inicialmente pelo tempo necessário e também pela qualidade dos conhecimentos gerados. Com isso, surge a mineração de dados como uma ferramenta para extração de conhecimento mais rápida e precisa. A mineração de dados faz parte de um processo chamado de Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Databases - KDD*). Os métodos de mineração de dados contidos no processo *KDD* extraem conhecimento sobre uma base de dados de uma maneira muito mais eficiente que um analista humano. Sendo assim, métodos de mineração de dados, como os métodos de agrupamento de dados se tornam importantes na extração de conhecimento e tomada de decisão sobre um volume de dados. Muitas vezes, no entanto, há dificuldade de interpretação dos resultados dos métodos. Emerge ai as técnicas de representação visual de dados, que aliadas à percepção humana, melhoram o entendimento e interpretação dos resultados de um determinado método de mineração de dados, como os de agrupamento de dados. Neste trabalho, fez-se a implementação de métodos de agrupamento de dados e de métodos de visualização de dados para agrupamento de dados, todos acoplados no Módulo de Agrupamento de Dados da Ferramenta YADMT desenvolvida no Curso de Ciência da Computação da Unioeste. A descrição deste módulo e os resultados obtidos seguem também apresentados.

Palavras-chave: Visualização de dados, Agrupamento de dados, YADMT.

Capítulo 1
Introdução

A análise de dados exige a combinação de conhecimentos de diferentes áreas, como a Matemática e a Estatística, bem como a experiência prévia do responsável por essa análise. Esta tarefa pode consumir muito tempo, o que pode inviabilizar sua realização. Por este motivo, o uso de técnicas e ferramentas computacionais é de suma importância para o auxílio na análise de dados.

Atualmente, há uma grande produção de dados, que contêm informações úteis que muitas vezes não estão facilmente disponíveis ou identificadas, de forma que a aplicação do processo de *KDD* (*Knowledge Discovery in Databases* ou Descoberta de Conhecimento em Bases de Dados) se faz cada vez mais necessária, facilitando a tomada de decisão.

Segundo (FAYYAD *et al.*, 1996), o processo *KDD* é um processo não trivial de descobertas de padrões válidos, novos, úteis e acessíveis. A principal vantagem do processo de descoberta é que não são necessárias hipóteses, sendo que o conhecimento é extraído dos dados sem conhecimento prévio sobre a base de dados. É um conjunto de atividades contínuas que são compostas, basicamente, por cinco etapas: seleção dos dados, pré-processamento, formatação, mineração de dados e interpretação dos resultados, como ilustra a Figura 1.1.

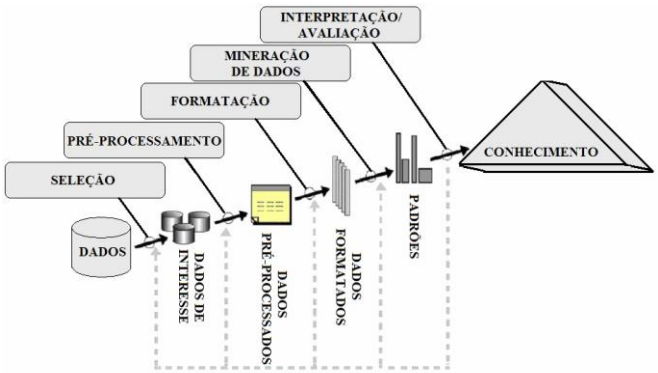


Figura 1.1: Etapas de um processo típico de *KDD*

Fonte: Adaptada de (FAYYAD *et al.*, 1996)

A primeira etapa do processo é a seleção dos dados, necessária à execução do processo como um todo, considerando que estes dados serão minerados no processo. Para a segunda etapa, pré-processamento, é feita a verificação dos dados, analisando se há dados ausentes, duplicados ou inconsistentes e, de acordo com alguma heurística, estes dados deverão sofrer algum processamento para que a base de dados seja composta corretamente.

Na etapa de formatação há uma preparação dos dados para os algoritmos de mineração de dados, que é a próxima etapa. Por exemplo, utilização de métodos de transformação de tipo de dados, de numérico para nominal.

A mineração de dados é o núcleo do processo *KDD*, que tem como principal objetivo a extração do conhecimento a partir de informações contidas nos dados que sejam úteis nas tomadas de decisões. Esta extração ocorre por meio de métodos de diferentes áreas científicas, que, de acordo com (TAN; STEINBACH; KUMAR, 2005), incluem Estatística, Inteligência Artificial, Aprendizagem de Máquina e Reconhecimento de Padrões.

E, finalmente, na etapa de Interpretação dos resultados é feita a validação do conhecimento extraído. As etapas do processo *KDD* são validadas tornando possível serem tomadas as decisões sobre a base de dados em que o processo foi aplicado.

A mineração de dados pode ser preditiva ou descritiva. As preditivas usam algumas variáveis para prever os valores desconhecidos ou futuros de outras variáveis, enquanto que as descritivas encontram padrões para descrever os dados. As principais tarefas de mineração de dados, de acordo com (FAYYAD *et al.*, 1996), são: Classificação, Regras de Associação e Agrupamento de padrões, sendo esta última foco deste trabalho juntamente com a visualização de seus resultados.

A utilização de métodos de Agrupamento de Dados auxilia na extração rápida de conhecimento de uma base de dados. Porém, a interpretação de seus resultados ainda pode ser demorada ou difícil, pelo fato de que estes resultados podem ainda não ser claros o suficiente. Neste cenário, métodos de visualização de dados se tornam importantes, pois visam melhorar a interpretação dos resultados obtidos.

1.1 Motivação

Considerando o cenário de grande produção de dados e a necessidade de analisa-los é que se faz válida a utilização de métodos de mineração de dados, com especial interesse neste trabalho, pelos métodos de agrupamento de dados, cuja interpretação dos resultados pode ser difícil e custosa muitas vezes, pela saída do próprio algoritmo, ou ferramenta, que foram aplicados sobre a base de dados. O retorno do conhecimento gerado pelos algoritmos muitas vezes é composto somente por dados que, dependendo de quem o está analisando, não os compreende.

Existe, portanto, a necessidade de implementação de métodos computacionais que proporcionem uma visualização mais significativa dos resultados gerados sobre uma base de dados, a exemplo da recuperação de agrupamento de dados e dos métodos de representação gráfica.

1.2 Objetivos

O objetivo principal deste trabalho foi estudar e implementar métodos que permitam a visualização de resultados obtidos por algoritmos de agrupamentos de dados presentes na ferramenta YADMT – *Yet Another Data Mining Tool*, de forma a melhorar a compreensão do usuário final em relação ao resultado gerado por esses métodos. Como objetivos específicos houve:

- Aplicar a metodologia desenvolvida a bases de dados públicas, como as disponíveis em http://archive.ics.uci.edu/ml/;
- Acoplar os métodos de visualização gráfica implementados ao módulo de Análise de Agrupamentos da YADMT, e comparar qualitativamente sua aplicabilidade frente a outras ferramentas existentes.

1.3 Resultados Obtidos

- Implementação de métodos de agrupamento de dados para o módulo de agrupamento de dados da YADMT;
- Implementação de métodos de visualização de dados aplicáveis a bases de dados de entrada e a resultados de agrupamento de dados;

- Uma avaliação comparativa da ferramenta desenvolvida juntamente a outras disponíveis na literatura.

1.4 Organização do Trabalho

Além do Capítulo 1 que contextualizou o presente trabalho, este documento esta assim dividido:

- O Capítulo 2 provê uma visão geral sobre a tarefa de Agrupamento de Dados, com ênfase nas categorias e métodos estudados e implementadas na YADMT.
- O Capítulo 3 traz uma introdução a Visualização de Dados e Grupos, assim como a apresentação de cada classe e métodos de visualização incluídos nestas classes. Também apresenta os métodos de visualização implementados na YADMT, contendo a justificativa da implementação e a avaliação dos mesmos, esta avaliação sendo feita de acordo com a visualização e a interpretação que o método proporciona.
- O Capítulo 4 faz a apresentação da YADMT, mais especificamente do módulo de Agrupamento de dados desenvolvido neste trabalho, apresentando os métodos nele contidos, tanto de agrupamento de dados quanto de visualização de dados.
- O Capítulo 5 sintetiza as execuções de testes com os métodos de agrupamento de dados explicitando os resultados por estes obtidos, também provê a avaliação comparativa da YADMT com outras ferramentas disponíveis na literatura.
- No Capítulo 6 constam as principais considerações deste trabalho, os resultados obtidos e os possíveis trabalhos futuros.

Capítulo 2

Agrupamento de Dados

O Agrupamento de Dados (*Clustering*) procura por padrões tal que padrões pertencentes a um mesmo grupo são mais similares uns aos outros e dissimilares a padrões em outros grupos. Para a determinação desta similaridade, ou dissimilaridade, são usadas medidas[1], que a partir dos próprios dados, encontram uma relação entre si, determinando, de acordo com alguma condição, o grau de relação, formando assim, grupos de dados.

Segundo (CORMACK, 1971), a ideia básica de agrupamento de dados pode ser definida como a coesão interna dos objetos e isolamento externo entre grupos. Ou seja, a similaridade para padrões de um mesmo grupo será grande e a dissimilaridade destes padrões para padrões de um segundo grupo também será grande.

Para (EVERITT, LANDAU, MORVEN, 2001), o agrupamento de dados é uma denominação geral para métodos computacionais que analisam dados visando a descoberta de conjuntos de observações homogêneas. Considerando uma base de dados com n padrões, cada um destes padrões medido segundo p variáveis, o objetivo é encontrar uma relação que os agrupe em g grupos. Com isto espera-se a visualização da relação entre os dados que anteriormente à aplicação do agrupamento de dados não era explícita.

O agrupamento de dados é um processo de aprendizado não supervisionado, pois neste caso não há uma classe que evidencie algum tipo de relação entre os dados, ou até mesmo não há a presença de agentes que supervisionem o aprendizado para o agrupamento (BOSCARIOLI, 2008). O resultado obtido pela aplicação de determinado método de agrupamento de dados será o número ótimo de grupos, definido pelo usuário ou definido pelo próprio método de agrupamento, assim como a semelhança entre padrões de um mesmo grupo, características de grupos, diferença entre grupos e outros dados relevantes à análise de agrupamento.

[1]As medidas de distâncias estudadas neste trabalho constam no Anexo A.

2.1 Técnicas de Agrupamento

Segundo (BOSCARIOLI, 2008), existem vários algoritmos para agrupamento de dados, que utilizam de maneiras diferentes para a identificação e a representação dos resultados do algoritmo. A escolha de cada algoritmo depende do tipo de dado que se pretende explorar, da aplicação e objetivos de análise. Tem-se também a possibilidade da aplicação de vários métodos sobre a mesma base de dados para posterior avaliação.

Ainda segundo (BOSCARIOLI, 2008), as técnicas de agrupamento podem ser divididas nas seguintes categorias:

- Métodos baseados em Particionamento;
- Métodos Hierárquicos;
- Métodos baseados em Densidade;
- Métodos baseados em Grades;
- Métodos baseados em Modelos;
- Métodos baseados em Redes Neurais Artificiais.

Dentre estas categorias, somente as duas primeiras categorias são mutuamente excludentes, e serão detalhadas a seguir. Alguns métodos possuem características que os classificam em mais de uma categoria, o que ocorre com um dos métodos utilizados neste trabalho.

2.1.1 Métodos Baseados em Particionamento

Os algoritmos de Agrupamento de Dados baseados em Particionamento, também conhecidos por Métodos Não Hierárquicos, procuram pela formação de um grupo sem a necessidade da associação hierárquica. A partir de n objetos procura-se formar k grupos otimizando algum critério de particionamento (BOSCARIOLI, 2008).

Outra importante característica que classifica algoritmos de agrupamento é a utilização de grades. Os métodos de agrupamento baseados em grade têm como principal característica a subdivisão do espaço em células. Métodos de particionamento são vantajosos em aplicações que envolvem grandes séries de dados.

O método baseado em particionamento mais conhecido, e estudado neste trabalho, é o *k-means* ou *k*-médias (HERNÁNDEZ *et al.*, 2012). Outro método estudado neste trabalho, o Algoritmo de Agrupamento baseado em Colônia de Formigas, pode ser classificado como método de particionamento, mas também é um método baseado em grade.

Nas seções que seguem, são apresentados os métodos de agrupamento de dados implementados no módulo de Análise de Agrupamento de Dados da YADMT – *Yet Another Data Mining Tool*.

2.1.1.1 Algoritmo *k-means*

Apresentado inicialmente por (MACQUEEN, 1967), o algoritmo de agrupamento de dados *k-means* é um dos mais simples e mais utilizados para a tarefa de agrupamento de acordo com (HERNÁNDEZ *et al.*, 2012). O Algoritmo 2.1 ilustra o processo.

Algoritmo 2.1 – Algoritmo *k-means*

```
Passo 1: Selecionar K centroides iniciais
Passo 2: repeat
Passo 3:     Atribuir cada padrão para o centroide mais próximo
Passo 4:     Recalcular centroides
Passo 5: until ( até que os grupos permaneçam estáveis)
```

Inicialmente, definem-se os k grupos a serem agrupados, após, são definidos os k centroides iniciais, que podem ser definidos por diversas heurísticas, baseando-se na base de dados. Por exemplo, números randomizados, a média de uma porcentagem x de padrões pertencentes à base de dados e, ainda, segundo (JOHNSON, WICHERN, 1998), pode-se definir de forma direta, ou seja, definir manualmente o valor dos centroides.

Para a etapa seguinte cada padrão da base de dados associa-se ao centroide mais próximo, definido por uma medida de distância, e ao final recalculam-se os k centroides calculando a média para cada atributo entre os padrões associados ao centroide analisado no momento. Este processo repete-se até que não haja mais mudança nos grupos formados.

Segundo (BOSCARIOLI, 2008), o algoritmo de *k-means* tem um funcionamento bom para grupos esféricos e a escolha dos k centroides iniciais é de suma importância para um bom resultado, além de ter a necessidade do conhecimento dos k grupos existentes.

2.1.1.2 Agrupamento baseado em Colônia de Formigas

O algoritmo baseado em Colônia de Formigas foi proposto inicialmente por (DENEUBOURG *et al.*, 1991), sendo implementado na YADMT o algoritmo consideradas algumas das modificações propostas por (VILLWOCK, 2009).

A capacidade mais reconhecida das formigas é a do trabalho em grupo para realizar uma tarefa que não poderia ser realizada somente por um indivíduo, tal que os resultados de uma tarefa realizada por um grupo são melhores que os resultados de uma tarefa realizada individualmente. O grande número de indivíduos em colônia de formigas e a abordagem descentralizada para tarefas realizadas simultaneamente significam que colônias de formigas mostram graus altos de paralelismo, auto-organização e tolerância a falhas. Segundo (BORYCZKA, 2009), essas são características desejáveis em técnicas modernas de otimização.

O agrupamento usando a técnica baseada em colônia de formigas utiliza grades fazendo a subdivisão do espaço em células. Segundo (HANDL; MEYER, 2007), existem dois tipos principais de agrupamento baseado em formigas. O primeiro grupo imita diretamente o comportamento observado no agrupamento de colônias de formigas reais, enquanto que o segundo é indiretamente inspirado pela natureza, pois a tarefa de agrupamento é reformulada como uma tarefa de otimização e, geralmente, heurísticas de otimização baseada em formigas são utilizadas para encontrar agrupamentos bons ou próximos do ótimo.

Segundo (TAN; TING; TENG, 2011) o algoritmo de agrupamento baseado em colônia de formigas é inspirado na atividade de formigas reais. A união das formigas para carregar outros indivíduos mortos até uma espécie de cemitério dentro de seu ninho demonstra uma forma de comunicação destas formigas. Cada formiga trabalha individualmente tendo apenas informações locais do seu trabalho, porém mesmo com esta situação a colônia como um todo trabalha de forma coletiva para o alcance do seu objetivo, formar um cemitério com as formigas mortas.

O Algoritmo 2.2, proposto por (DENEUBOURG *et al.*, 1991), ilustra o processo de Agrupamento baseado em Colônia de Formigas. Inicialmente são espalhados os padrões da base de dados na grade, após cada formiga carrega um destes padrões de forma aleatória. Para o laço de repetição tem-se que cada formiga irá andar na grade fazendo as decisões de carregar ou descarregar um padrão em um determinado local, tudo isto de maneira isolada a outra formiga. O algoritmo terminará quando um dado número de iterações for atingido.

Algoritmo 2.2 – Algoritmo para Agrupamento baseado em Colônia de Formigas

Passo 1: Espalhar aleatoriamente os padrões na grade
Passo 2: Cada formiga escolhe aleatoriamente um padrão
Passo 3: repeat
Passo 4: Selecionar formiga aleatoriamente
Passo 5: Formiga selecionada executa passo de comprimento para uma direção determinada aleatoriamente
Passo 6: Formiga decide probabilisticamente se descarrega padrão na posição
Passo 7: Decisão negativa para descarregar
Passo 8: Escolhe-se aleatoriamente outra formiga e volta para passo 5
Passo 9: Decisão positiva para descarregar
Passo 10: Verifica se posição atual esta livre
Passo 11: Caso esteja descarrega nesta posição
Passo 12: Se não estiver livre
Passo 13: Descarrega padrão em uma posição imediatamente vizinha, buscando esta posição de forma aleatória
Passo 14: repeat
Passo 15: Formiga procura, aleatoriamente, por novo padrão para carregar
Passo 16: Avalia probabilisticamente se carrega este padrão
Passo 17: until(padrão carregado pela formiga)
Passo 18: Retorna para o passo 4
Passo 19: until(número de interações seja satisfeita)

No agrupamento proposto por (DENEUBOURG *et al.*, 1991), as formigas são representadas como agentes que se moviam aleatoriamente em uma grade quadrada. Os padrões são dispersos nesta grade e podem ser carregados, transportados e descarregados pelas formigas, sendo que as decisões de carregar e descarregar são tomadas pelas probabilidades P_{pick} e P_{drop} dadas pelas Equações 2.1 e 2.2, respectivamente. Estas operações

são baseadas na similaridade e na densidade e padrões na grade. Padrões isolados ou cercados por dissimilares tem maior probabilidade de serem carregados e então descarregados numa vizinhança de similares.

$$P_{pick} = \left(\frac{k_p}{k_p + f(i)} \right)^2 \tag{2.1}$$

$$P_{drop} = \left(\frac{f(i)}{k_d + f(i)} \right)^2 \tag{2.2}$$

Nestas equações, *f(i)* é uma estimativa da fração de padrões localizados na vizinhança que são semelhantes ao padrão atual da formiga e k_p e k_d são constantes reais, que em (DENEUBOURG *et al.*, 1991) valem, respectivamente, 0,1 e 0,3. Em Handl, Knowles e Dorigo (2006) a função *f(i)* é dada pela Equação 2.3.

$$f(i) = \begin{cases} \frac{1}{\sigma^2} \sum_{j \in L} \left[1 - \frac{d(i,j)}{\alpha} \right] & , \quad se \quad \forall j \left(1 - \frac{d(i,j)}{\alpha} \right) > 0 \\ 0 \quad , \text{ caso contrário} \end{cases} \tag{2.3}$$

Em que *d(i, j)* é a função de dissimilaridade entre padrões *i* e *j* pertencentes ao intervalo [0, 1]; α é um parâmetro escalar dependente dos dados pertencente ao intervalo [0, 1]; L é a vizinhança local de tamanho igual a σ², onde σ é o raio da vizinhança. Caso $\left(1 - \frac{d(i,j)}{\alpha}\right)$ seja menor ou igual a zero a função *f(i)* recebe zero para, segundo os autores, penalizar dissimilaridades elevadas.

Segundo Handl, Knowles e Dorigo (2006), alfa (α) determina a porcentagem de padrões na grade classificados como semelhantes. A escolha de um valor pequeno para alfa impede a formação de grupos na grade, enquanto a escolha de um número alto para alfa ocasiona a fusão de dois grupos na grade.

O parâmetro sigma (σ) é o raio de percepção da vizinhança, ou seja, quantos padrões vizinhos serão analisados na posição em questão. Segundo Handl, Knowles, Dorigo (2006), é desejado um valor de sigma alto para que se possa empregar maior qualidade no agrupamento e na distribuição da grade. Porém, este procedimento é mais caro computacionalmente (porque o número das células a ser considerado para cada ação cresce quadraticamente com o

raio), e ainda inibe a formação rápida dos grupos durante a fase de distribuição inicial. Um raio de percepção que aumenta gradualmente com o tempo acelera a dissolução de grupos pequenos preliminares.

Tendo como base as modificações propostas por (VILLWOCK, 2009) foram realizados modificações na função *f(i)*, que é uma estimativa da fração de padrões localizados na vizinhança que são semelhantes ao padrão atual da formiga. Anteriormente, esta função *f(i)* era dada pela Equação 2.3, já apresentada, sendo que no algoritmo final implementado foi dada pela Equação 2.4. Em que N_{cell} é o número de células analisadas para aquela determinada posição da grade em que o padrão será inserido.

$$f(i) = \begin{cases} \frac{1}{N_{Cell}} \sum_{j \in L} \left[1 - \frac{d(i,j)}{\alpha}\right] & , \quad se \ \ \forall j \left(1 - \frac{d(i,j)}{\alpha}\right) > 0 \\ 0 \quad , \text{ caso contrário} \end{cases} \qquad (2.4)$$

2.1.2 Métodos Hierárquicos

Os métodos de agrupamento de dados hierárquicos procuram formar grupos de forma hierárquica, admitindo assim vários níveis de agrupamento. Estes níveis podem ser representados por árvores, que são formadas durante o processo de agrupamento. De acordo com (JAIN; DUBES, 1988) um dendrograma pode representar os níveis de agrupamento, assim como os níveis de similaridade.

A Figura 2.1 representa inicialmente a disposição de grupos e a relação que cada um tem entre si, apresentando a proximidade entre eles e também a formação de novos grupos, como A e B unindo-se com C. O dendrograma é capaz de fazer a mesma representação, porém também mostra a formação hierárquica dos grupos de uma forma mais clara.

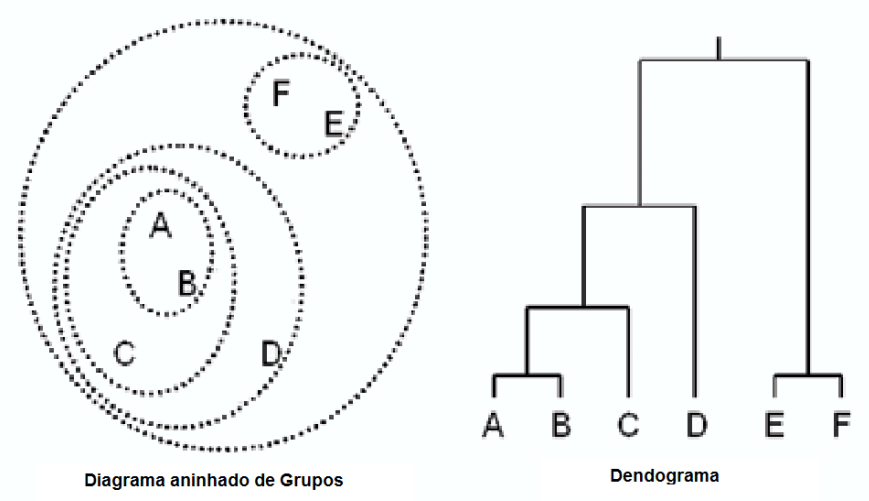


Figura 2.1: Duas possíveis representações de agrupamentos

Fonte: Adaptada de (BOSCARIOLI, 2008).

Os métodos de agrupamento de dados hierárquicos ainda podem ser divididos em duas classes: Aglomerativos e Divisivos. Os algoritmos aglomerativos iniciam-se com objetos individuais, tal que cada um desses objetos é um grupo. Seguindo a convergência do algoritmo, estes grupos unem-se formando um único grupo ao final. Os algoritmos divisivos trabalham de forma oposta. Inicialmente todos os objetos pertencem ao mesmo grupo, e são divididos até que o número de grupos seja igual ao número de objetos da base de dados (JOHNSON; WICHERN, 1998).

De acordo com (JAIN; MURTY; FLYNN, 1999), a maioria dos métodos de agrupamento de dados hierárquicos são variantes dos algoritmos *single-linkage*, *complete-linkage* e *minimum-variance*, estes métodos são descritos em (JOHNSON; WICHERN, 1998), sendo os métodos mais populares.

2.1.2.1 Métodos Hierárquicos Implementados na Ferramenta YADMT

Nesta seção são apresentados brevemente alguns algoritmos hierárquicos de agrupamento de dados.

- **Ligação Simples (*single-linkage*):** é um método de agrupamento de dados hierárquico aglomerativo que considera a menor distância entre pares de objetos i e j, sendo que estes pertencem a grupos distintos. Um grupo será formado pela fusão destes dois objetos. Após isto será refeita a distância entre todos os objetos da base de dados

agrupando novos objetos entre si, ou agrupando um único objeto a grupos já formados. O processo repete-se até que todos os objetos sejam um único grupo.

- **Ligação Completa (*complete-linkage)*:** é um método de agrupamento de dados hierárquico aglomerativo que considera a maior distância entre pares de objetos. É um método similar ao de ligação simples, diferenciando-se somente na análise da distância entre os objetos.

- **Ligação Média (*average-linkage)*:** é um método de agrupamento de dados hierárquico aglomerativo que considera a média da distância entre pares de objetos. É um método similar ao de ligação simples e de ligação completa, porém considerando a distância entre dois conjuntos como a distância média entre todos os pares de itens de outro grupo.

- **Método *Ward*:** é um método de agrupamento de dados hierárquico aglomerativo que faz a junção dos grupos baseando-se na perda de informação na junção de dois objetos. Geralmente, esta perda de informação é a soma do quadrado do erro (*SQE*, do inglês *sum of squared of errors*). O *SQE* é a soma do quadrado do erro de cada padrão do grupo em relação à média do grupo, o centroide. A junção de objetos a um grupo será feito quando se obtiver o menor aumento da soma do *SQE*.

2.2 Considerações Finais

A análise de agrupamento é uma atividade importante para o entendimento de inúmeras informações presentes no dia-a-dia. Esta atividade esta sendo usada em diversas aplicações como reconhecimento de padrões e análise de dados. A aplicação destes métodos auxilia na extração de conhecimento sobre uma determinada base de dados e garantem uma tomada de decisão de forma mais rápida e precisa quando comparada com a análise manual.

Neste capítulo foram apresentados, de forma introdutória, a tarefa de agrupamento de dados e os métodos de agrupamento de dados acoplados na YADMT, importantes para a compreensão do restante deste trabalho.

Capítulo 3
Visualização de Dados

Toda a quantidade de dados produzidos pode ser analisada por métodos de mineração de dados. Estes métodos auxiliam nas análises de uma forma rápida e com maior qualidade do que se fossem analisados por um analista humano. Porém, mesmo utilizando-se destes métodos para a análise de uma determinada base de dados, o resultado ainda pode não estar claro para o usuário final, ou seja, há necessidade de interpretar o resultado.

É neste cenário que o desenvolvimento de técnicas de visualização de resultados após a mineração de dados se encaixa, de forma a permitir ao usuário final interpretar os resultados de um método de agrupamento, por exemplo, identificando características de um determinado grupo, relação entre padrões e distinções entre grupos, distribuição espacial de padrões, entre outros.

As capacidades de flexibilidade, criatividade, conhecimento geral e percepção, do ser humano, também são muito importantes neste processo (KEIM; WARD, 2002). Estes são requisitos essenciais para que a visualização seja eficiente.

A ideia principal da visualização de dados é integrar o usuário final ao processo, dando ao usuário uma representação gráfica da base de dados. O usuário também poderá interagir com a representação gráfica e assim interpretar de uma melhor maneira o resultado gerado pelo método de agrupamento de dados (WONG, 1999).

Entende-se por visualização o processo de mapeamento de dados e informações em um formato gráfico, baseando-se em representações visuais e em mecanismos de interações. O propósito da visualização é a percepção do que esta sendo representado e não somente figuras. No processo de mineração de dados tem-se a mineração visual de dados e a visualização de agrupamento (*Clustering Visualization*), sendo esta última o foco deste trabalho.

A mineração visual de dados é a integração de métodos de mineração de dados com métodos de visualização. Para (RABELO, 2007) somente a visualização da informação não substitui os métodos convencionais da mineração de dados, porém, as técnicas unidas podem potencializar a exploração da informação. Ainda, segundo o mesmo autor, considerando um grande volume de dados, o usuário pode selecionar porções da base de dados do seu interesse

utilizando métodos de visualização, diminuindo assim a execução do método de mineração de dados, bem como o entendimento de seus resultados.

Na visualização de agrupamento, por sua vez, métodos de visualização são aplicados aos resultados de um agrupamento de dados para análise da qualidade do agrupamento, por exemplo, a visualização da distribuição espacial dos padrões por meio dos atributos de um determinado grupo.

3.1 Técnicas de Visualização de Dados

Há diferentes técnicas de visualização de dados, porém, a grande maioria é limitada pela dimensionalidade da base de dados a ser explorada. Com o passar do tempo, várias técnicas foram desenvolvidas para diferentes tipos de dados e também para diferentes dimensionalidades. Segundo (SHNEIDERMAN, 1996) os tipos de dados que podem ser visualizados são:

- Unidimensional;
- Bidimensional;
- Multidimensional;
- Texto e Hipertexto;
- Grafos e Hierárquicos e;
- Algoritmos e Softwares.

Para os dados unidimensionais tem-se um atributo por padrão, um exemplo são os dados temporais em que para cada ponto de tempo pode-se ter vários valores associados. Os dados bidimensionais são aqueles em que o dado tem duas variáveis distintas, por exemplo, dados geográficos (longitude e latitude). Dados multidimensionais são aqueles que se têm três ou mais variáveis e que para sua representação podem ser utilizadas somente até três dimensões, por exemplo, a longitude, latitude e altura.

Os dados de texto, hipertexto e grafos são aqueles dificilmente descritos por números, de forma que a aplicação de técnicas de visualização necessita de uma transformação prévia, geralmente feita para estruturas de dados, como vetores. Os dados de algoritmos e softwares são aqueles que ajudam ao entendimento de um determinando software, como, diagrama de fluxo de dados.

Para alguns tipos de dados pode haver relações entre eles, e podem ser descritos como Hierárquico e Grafos. Um grafo pode ser utilizado para mostrar interdependência entre os dados, como uma comunicação por e-mail entre várias pessoas. Dados hierárquicos são aqueles que apresentam hierarquia entre si, como *hiperlinks* contidos em outros *hiperlinks.* A Tabela 3.1 exemplifica alguns tipos de dados das categorias apresentadas.

Tabela 3.1: Caracterização de dados com base em critérios e classes

Critério	Classe	Exemplo
Classe de Informação	Categoria	Gênero
	Escalar	Temperatura
	Vetorial	Grandezas físicas associadas a dinâmicas dos fluídos
	Tensorial	
	Relacionamento	*Link* num hiperdocumento
Tipos de Valores	Alfa-numérico	Gênero
	Numérico	Temperatura
	Simbólico	*Link* num hiperdocumento
Natureza do Domínio	Discreto	Marcas de automóveis
	Contínuo	Superfícies de um terreno
	Contínuo-discretizado	Anos (tempo discretizado)
Dimensão do Domínio	1D	Fenômeno ocorrendo no tempo
	2D	Superfície de um terreno
	3D	Volume de dados médicos
	nD	Dados de uma população

Fonte: (FREITAS; WAGNER, 1995).

Segundo (KEIM; WARD, 2002), as técnicas de visualização de dados podem ser divididas em:

- Técnicas de visualizações em duas dimensões (2D) e três dimensões (3D);
- Técnicas de visualizações de Projeções Geométricas;
- Técnicas de visualizações baseada em Ícones - Iconográficas;
- Técnicas de visualizações orientadas a Pixels e;
- Técnicas de visualizações Hierárquicas.

Ainda de acordo com (KEIM; WARD, 2002), a visualização de dados é um processo composto por três dimensões, em que as duas primeiras dimensões são compostas pelas duas classes já apresentadas, tipos de dados e técnicas de visualização. A terceira dimensão é composta por Técnicas de Interação. Estas técnicas de Interação são operações que podem ser feitas durante a visualização dos dados, como *zoom*, distorções e outros. A relação entre estas três dimensões pode ser visualizada na Figura 3.1.

A abordagem para interagir com os dados visualmente pode ser projeções, filtragem, *zoom*, distorção e Seleção/Ligação. Ainda, segundo (KEIM; WARD, 2002), estas três dimensões são ortogonais, ou seja, deve ser usadas juntas, com qualquer tipo de combinação entre tipos de dados, técnica de visualização e técnica de interação.

Segundo (SILVA NETO, 2008), existem mais classes de técnicas de visualização, que são as Técnicas de Empilhamento de Dimensões e Híbridas. As duas técnicas podem ser juntadas devido à sua grande semelhança.

Figura 3.1: Relação entre as Três Dimensões do Processo de Visualização de Dados
Fonte: Adaptada de (KEIM; WARD, 2002).

3.1.1 Técnicas de Visualização em 2D e 3D

Esta classe de técnicas de visualização é a mais genérica, composta por Gráfico de Pizza, Dispersão, Linhas e outros, muito utilizados para a exibição de dados em duas e três dimensões. A planilha eletrônica *Excel*, que faz parte do pacote de ferramentas *Office* da *Microsoft* (*MS*), oferece estes tipos de gráficos para a visualização de dados. A Figura 3.2 mostra os tipos mais comuns presentes nesta ferramenta.

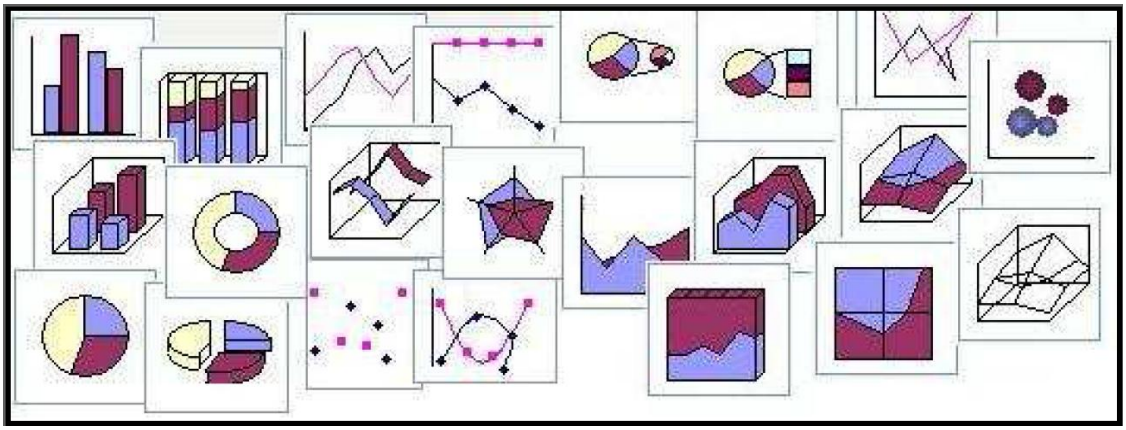


Figura 3.2: Gráficos presentes na ferramenta *MS* Office

Fonte: (SILVA NETO, 2008).

Os gráficos de pizza são indicados para a representação de dados e porcentagens, pela sua representação em fatia equivalente à sua porcentagem. Porém, quando se tem porcentagens muito pequenas cria-se fatias pequenas, o que dificulta sua visualização. Uma alternativa dada pela ferramenta é a criação de um sub gráfico de pizza que permite melhor visualização daquela fatia. A Figura 3.3 ilustra este processo, onde inicialmente têm uma fatia que representa três valores distintos, essa fatia é dividida em um gráfico de pizza com três fatias. Esse novo gráfico representa as fatias que tinham porcentagem muito pequena e que não teria uma visualização muito clara. O que a ferramenta faz é transformar estas fatias em uma única e desta gerar um segundo gráfico, não mudando os valores e nem a legenda destas fatias.

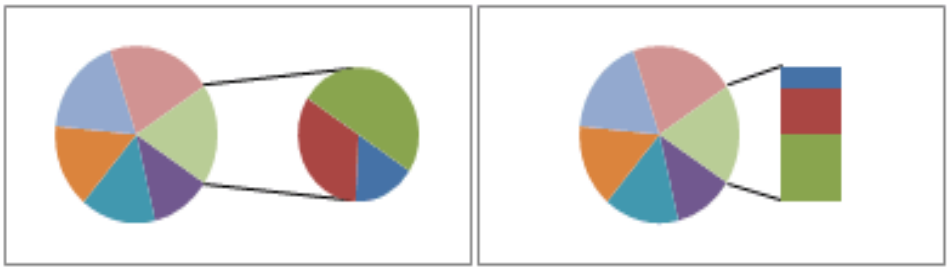


Figura 3.3: Sub Gráficos de Pizza

É importante ressaltar que não somente o pacote de ferramentas *Office* da *MS* é capaz de oferecer este e outros recursos, mas também outras ferramentas *Office* disponíveis.

Há também Gráficos de Barras, em que os dados são representados em barras retangulares, que podem ser horizontais e verticais. Para estes gráficos, dados qualitativos são bem ilustrados em que a altura da barra é igual à frequência. A representação deste tipo de gráfico

em 3D é denominada *Cityscapes* (CHUAH *et al.*, 1995). A Figura 3.4 traz um exemplo de gráfico *Cityscapes*.

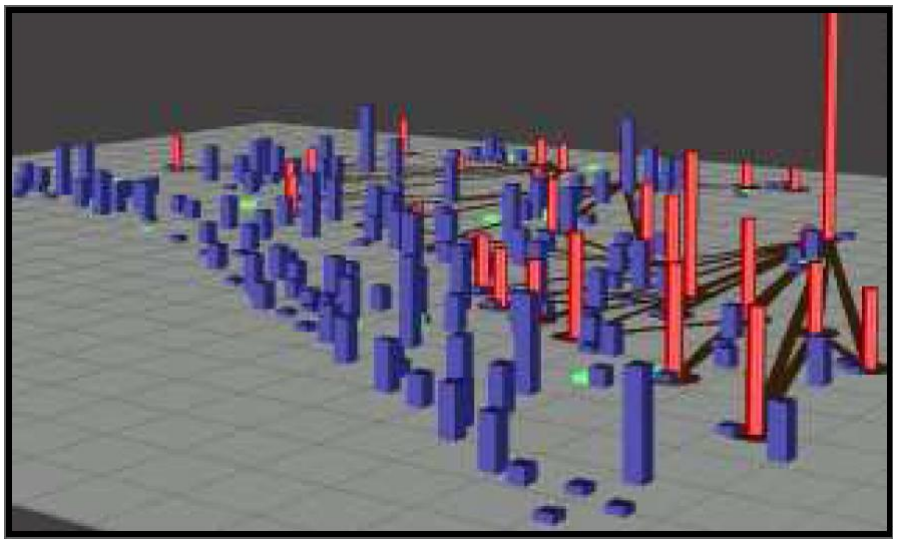


Figura 3.4: Exemplo de gráfico *Cityscape*
Fonte: (CHUAH *et al.*, 1995).

O gráfico *Cityscape* faz a representação dos dados de maneira igual à representação feita pelos gráficos de barras, porém, esta representação é feita em 3D, o que permite uma melhor visualização de um conjunto de dados e também uma melhor dispersão destes dados, já que haverá uma dimensão a mais para a representação.

A técnica de gráficos de dispersão oferece apoio eficaz para a análise visual permitindo detectar distribuição, correlação entre os atributos e outras informações, pois plota o comportamento das variáveis no espaço bidimensional, permitindo analisar a dispersão dos dados. Quando estes dados estão dispersos aproximando-se de uma reta, diz-se que os dados são altamente correlacionáveis. Caso esta reta seja crescente, tem-se uma correlação positiva, caso contrário, sua correlação será negativa. Caso os dados estejam dispersos, não tendo a formação próxima de uma reta, esta correlação será perto de zero e os dados não apresentam correlação.

3.1.2 Técnicas de Visualização de Projeções Geométricas

Técnicas de visualizações de projeções geométricas visam encontrar transformações que retornem um bom resultado ao conjunto de dados multidimensionais. Para esta classe os métodos de exibição geométrica incluem técnicas de estatísticas exploratórias, como matrizes S*catterplot* e técnicas de busca de projeção (ANDREWS, 1972) e (CLEVELAND, 1993). O

que os métodos de projeção geométrica fazem é projetar dados multidimensionais em um espaço bidimensional.

Outros exemplos de técnicas bastante utilizadas desta classe são Coordenadas Paralelas (*Parallel Coordinates*) e Gráfico Estrela (*Star Graph*), conforme (CARVALHO, 2001).

Um ponto negativo para esta classe de técnicas é a alta dimensionalidade dos dados, reduzindo a área de representação. Uma alternativa de resolver este problema é a utilização de interação com *zoom* (SILVA NETO, 2008).

Coordenadas Paralelas é uma técnica de visualização onde as dimensões são apresentadas como uma série de eixos paralelos uns aos outros e com igual espaçamento entre eles nos quais os valores estão representados. Nesta técnica, a relação entre as variáveis pode ser extraída analisando pares consecutivos de atributos. Um grupo de linhas projetadas bastante próximas uma das outras e sem muitos cruzamentos indicam um grau de relacionamento positivo entre a tupla, enquanto o contrário apresenta um grau de relacionamento negativo entre a tupla (SILVA NETO, 2008).

Também é possível ter a representação de todos os vetores em um mesmo gráfico, podendo fazer a comparação visual entre os vetores. Para os objetos analisados na mineração de dados isto representa uma padrão, objeto, formado pelos seus n atributos (SILVA NETO, 2008).

A Figura 3.5 ilustra um gráfico de Coordenadas Paralelas, que apresenta entre os padrões, um grau de relacionamento baixo, considerando que as linhas representando cada padrão são apresentadas distantes uma das outras e também, há vários cruzamentos entre as linhas dos padrões.

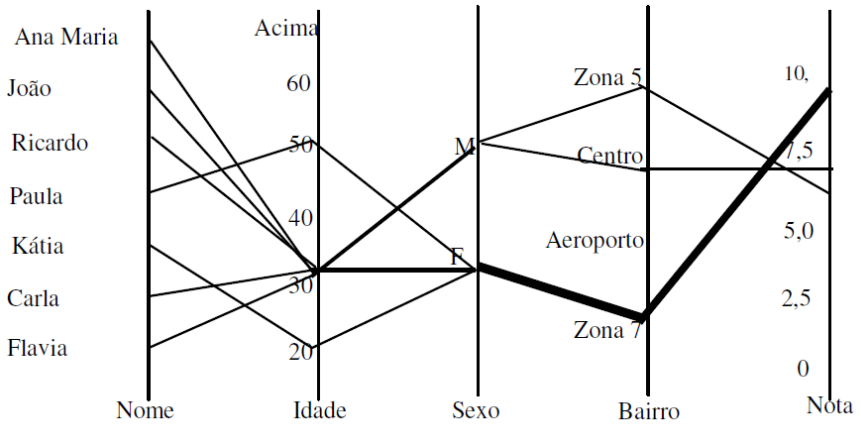


Figura 3.5: Exemplo de Coordenadas Paralelas com dados fictícios

Fonte: (RABELO, 2007).

Ainda nesta técnica pode-se identificar a formação de grupos (*clusters*). A formação destes grupos é identificada quando um conjunto de dados sai de um mesmo ponto e segue para as demais variáveis (Figura 3.6).

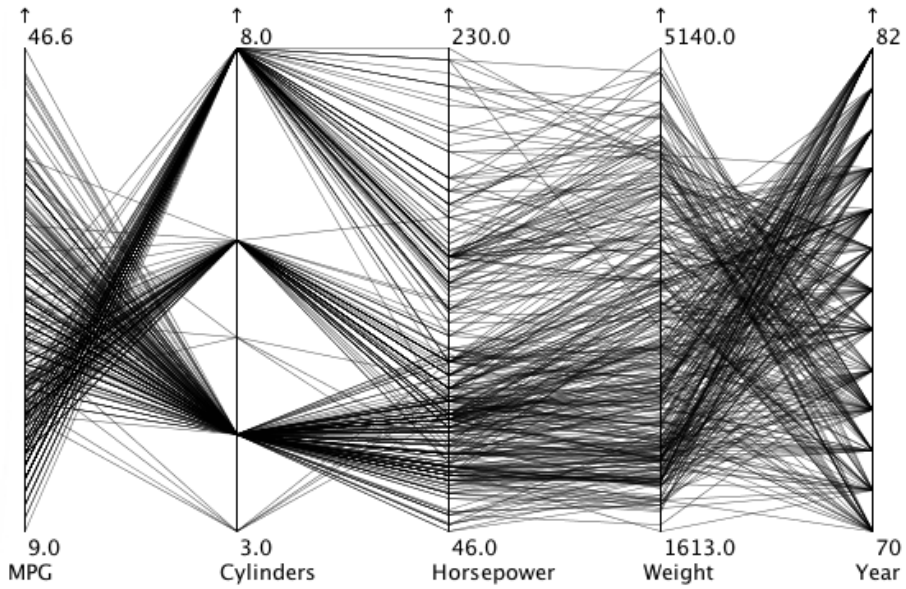


Figura 3.6: Coordenadas Paralelas para Agrupamento

Fonte: (KOSARA, 2010).

Como pode ser observado pela Figura 3.6, as coordenadas paralelas proporcionam a visualização da formação de grupos dentro de um determinado grupo ou de uma base de dados. Essa formação de grupos se dá pela formação de pontos onde as linhas que fazem a ligação de um atributo (linha na vertical) a outro se conectam, ou seja, quando ocorre junção dos atributos em um único ponto, ocorre a formação de grupos (SILVA NETO, 2008).

As desvantagens desta técnica apresentam-se quando há a presença de muitas variáveis. As linhas irão se sobrepor umas às outras. Outro problema é a dimensão das telas e o número de dimensões dos dados, pois quanto maior for o número de dimensões dos dados mais próximos irão ficar os eixos de dimensões do gráfico.

Outra técnica da classe de Projeções Geométricas é a *Radviz* (HOFFMAN, 1997), que faz a visualização de coordenadas radias e segue um princípio bastante semelhante à técnica de Coordenada Paralelas. Na técnica as *n* linhas que correspondem às dimensões de uma base de dados saem do centro de um círculo e terminam no perímetro deste centro, sendo espaçadas igualmente. Este espaçamento se dá por meio de molas imaginárias em que a posição dos pontos é determinada onde há equilíbrio das forças associadas em cada dimensão. Esta técnica possui como características de visualização:

- Itens de dados com valores de atributos muito próximos são mapeados próximos ao centro do círculo, devido à ação das molas imaginárias;
- Itens com valores similares, porém em dimensões diferentes também apresentam o mesmo posicionamento de valores próximos (centro do círculo);
- Os maiores valores de uma determinada dimensão atraem o ponto para próximo do eixo que equivale à dimensão.
- Como ponto negativo, também presente em outras técnicas, se houver repetição de valores de atributo para padrões diferentes haverá a sobreposição de pontos pintados em tela.

A Figura 3.7 apresenta a técnica de visualização *Radviz*.

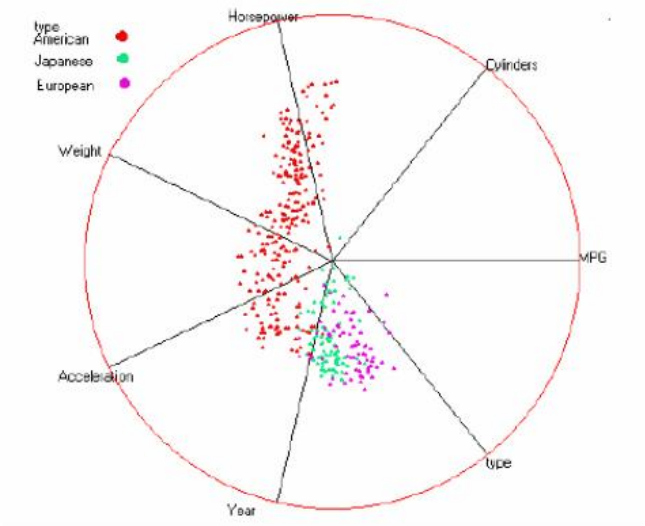


Figura 3.7: Representação visual da técnica *Radviz*

Fonte: (VALIATI, 2004).

3.1.3 Técnicas de Visualização Iconográficas

A ideia básica desta técnica é mapear os atributos dos dados para as características de um ícone. Cada característica do ícone representa um atributo dos dados multidimensional. Estes ícones podem ser definidos de forma arbitrária, podem ser pequenas faces (CHERNOFF, 1973) ou estrelas (WARD, 1994). Em caso de dados multidimensionais as duas primeiras dimensões são mapeadas em tela, dimensão espacial, e as demais são mapeadas para propriedades visuais de um ícone, como formato de boca, nariz e olhos.

Segundo (CHERNOFF, 1973), esta técnica é eficiente uma vez que as pessoas estão habituadas a distinguir expressões faciais, entre outros. Porém, devido à dificuldade de distinguir diferenças muito pequenas nas imagens resultantes, não é adequada para a identificação de agrupamento.

A Figura 3.8 ilustra um exemplo dessa técnica, as Faces de Chernoff, que mapeia os dados para características faciais com intenção de utilizá-las como identificador comum do dado. No exemplo é feita a representação longitudinal de oito atributos, e para cada um deles é feita a relação destes com objetos equivalentes ao seu significado. Por exemplo, o atributo qualidade é representado pelo nariz do boneco presente na imagem, para algo com grande qualidade utiliza-se o nariz de maior tamanho, para uma baixa qualidade utiliza-se um nariz com menor tamanho.

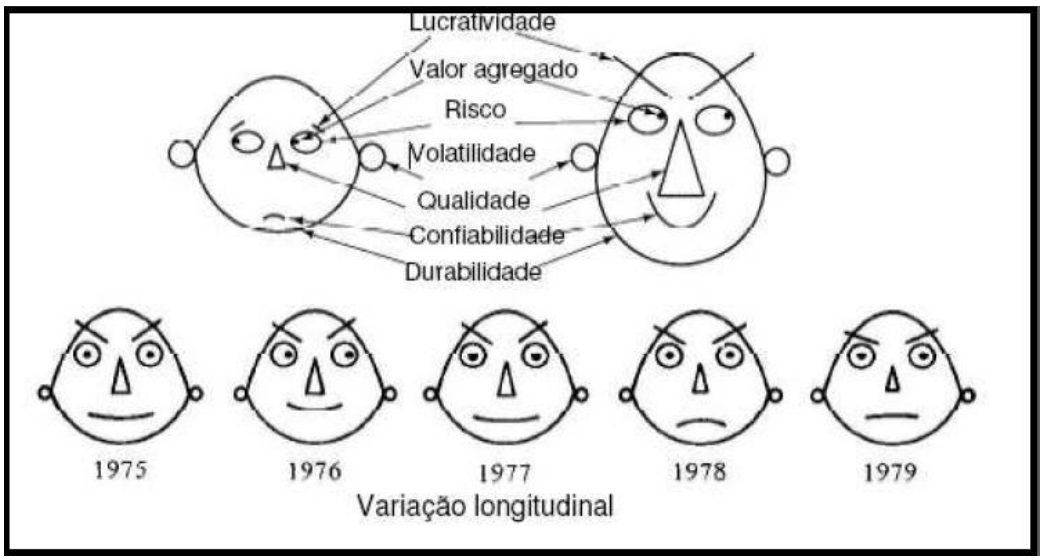


Figura 3.8: Ilustração da Técnica de Visualização Iconográfica

Fonte: (SILVA NETO, 2008).

3.1.4 Técnicas de Visualização orientadas a Pixels

A ideia é mapear cada dimensão de um dado para uma cor de pixel e agrupar estes pixels em uma área em comum. Segundo Keim e Kriegel (1996), definem que cada atributo é apresentado em uma janela individual de forma que para exibir m atributos a janela deverá ser dividida em m janelas. Cada pixel desta janela será a representação visual de um dos atributos dos dados. Estes pixels são coloridos por um mapa de cores previamente definidos.

A utilização desta técnica permite a determinação de grupos (*clusters*) dos dados, correlações e dependência funcional entre os atributos. A Figura 3.9 exemplifica o caso de correlação existente entre atributos.

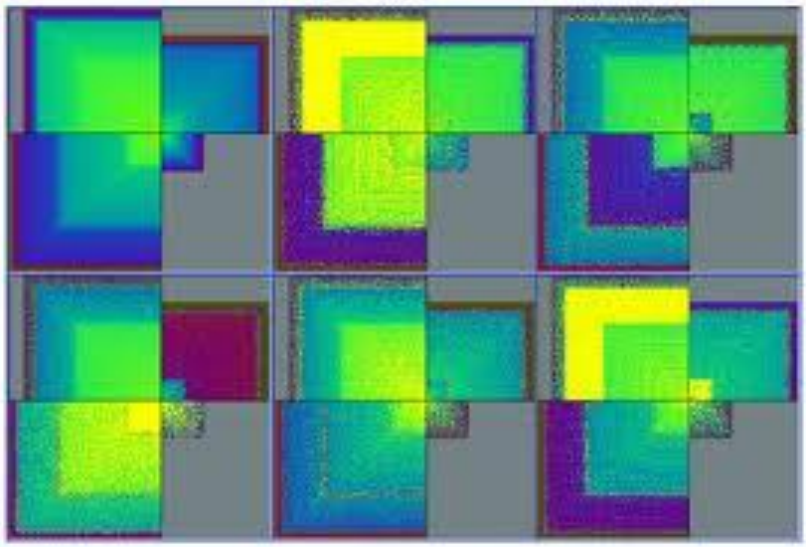


Figura 3.9: Formação de Agrupamento pela Técnica de Visualização Orientada a Pixels na ferramenta VisDB
Fonte: (KEIM, KRIEGEL, 1994).

Também fazendo parte desta classe de técnicas tem-se a técnica de segmentos circulares (*Circle Segments*) (ANKERST *et al.*, 1996). Esta técnica é constituída da ideia básica das técnicas orientadas a Pixels mudando somente a janela de visualização dos pixels. Como o nome já indica, os pixels são apresentados em segmentos circulares. Na Figura 3.10 tem-se a representação de ambos os segmentos de janelas, retangular e circular.

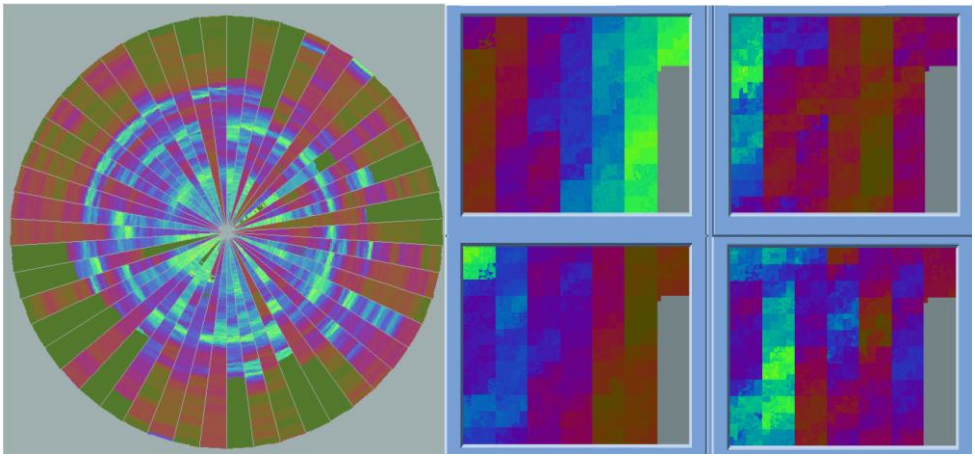


Figura 3.10: Exemplo de Segmentos da Técnica Orientada a Pixels
Fonte: (ANKERST, 2001).

Para ambas as janelas de visualizações é observada a desvantagem para quando se apresenta um número de atributos muito grande. A resolução de tela esta envolvida

diretamente com isto, pois quanto maior a dimensionalidade dos dados maior será o número necessário de janelas para a representação.

3.1.5 Técnicas de Visualização Hierárquicas

Esta classe de técnicas geralmente é aplicada a dados cuja própria natureza apresenta uma formação hierárquica. Aplicada na representação de métodos de agrupamento hierárquicos, em que o espaço é dividido em subespaços que são organizados uns dentro dos outros e exibido de forma hierárquica (SILVA NETO, 2008).

O dendrograma, Figura 3.11, é um método muito conhecido para demonstrar níveis hierárquicos entre dados. É uma árvore que apresenta o arranjo, ou ligação, entre grupos, para o caso de métodos de Agrupamento Hierárquico. Para cada ligação feita entre dois dados é formado um grupo. A formação desta árvore se dá de baixo para cima (*bottom-up*), sendo que o último elemento formará a raiz desta árvore que ao final irá formar somente um grupo.

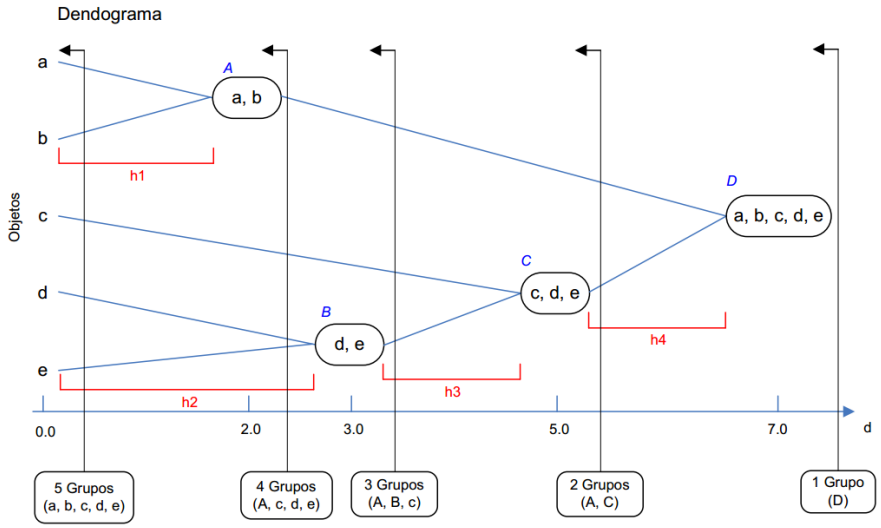


Figura 3.11: Representação Hierárquica de Método de Agrupamento Aglomerativo

Fonte: (VALE, 2005).

A principal vantagem deste método é a facilidade da visualização dos grupos formados, podendo saber, por exemplo, qual dado tem maior correlação com outro, dependendo diretamente do método de agrupamento de dados hierárquico utilizado. A desvantagem é a limitação de representação gráfica, já que, dependendo do número de dados da base de dados a representação hierárquica ficará sobreposta, ou muito próxima uma a outra, não retornando uma visão clara para o usuário.

Uma alternativa de representação 3D para esta árvore 2D (dendrograma) são os métodos *Cone Trees* e *Cam Trees* (ROBERTSON *et al.*, 1991). A Figura 3.12 representa ambas as técnicas que por se utilizarem da representação 3D podem se utilizar de métodos de interação com usuário, como *zoom*, para suprir o problema de visualização de muitos padrões, como já apresentado.

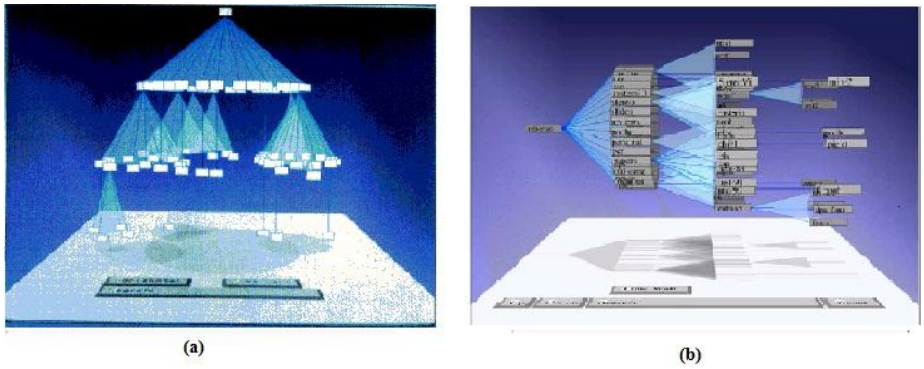


Figura 3.12: Representação de *Cone Tree* (a) e *Cam Tree* (b)
Fonte: Adaptada de (ROBERTSON *et al.*, 1991).

3.2 Técnicas Implementadas na Ferramenta YADMT

Nesta seção são apresentados os métodos implementados na YADMT. Os métodos serão explanados de modo que fique claro o motivo de sua escolha, bem como as vantagens e as desvantagens do seu uso.

A implementação dos métodos foi feita na linguagem de programação Java, seguindo o modelo de implementação inicial proposto por (BENFATTI *et al.*, 2010) para a YADMT, que é dividida em três etapas: Pré-Processamento, Mineração de Dados e Pós-Processamento, pensada para ser construída de forma modular, para que fosse um ambiente de desenvolvimento acadêmico colaborativo evolutivo/expansível, no qual essa proposta está inserida.

3.2.1 Gráfico de Dispersão

Este método faz parte da classe de Técnicas de Visualização em 2D e 3D, sendo que pode ser implementado em ambas dimensões, bidimensional ou tridimensional. Ainda, pode ser dividido em uma representação geral da base de dados ou a partir dos grupos formados por um método de agrupamento (CLEVELAND, 1993).

Como o nome do método propõe, a técnica demonstra a dispersão dos dados baseada em um conjunto de atributos da base de dados, conjunto composto por dois elementos (visualização 2D) ou por três elementos (visualização 3D).

A representação geral da base de dados que será minerada é uma forma auxiliar de entendê-la, pois é feita em um passo anterior à mineração de dados. A sua representação é feita a partir dos atributos selecionados. O par, ou tripla, será mostrado nos eixos cartesianos equivalentes, ilustrando como é a dispersão destes dados.

A Figura 3.13 mostra a dispersão dos dados da base de dados Íris (FRANK, ASUNCION, 2010) que é formada por quatro atributos numéricos, 150 padrões e três grupos, sendo na cor vermelha o grupo Iris-Setosa, na cor azul o grupo Iris-Versicolor e na cor verde Iris-Virgínica.

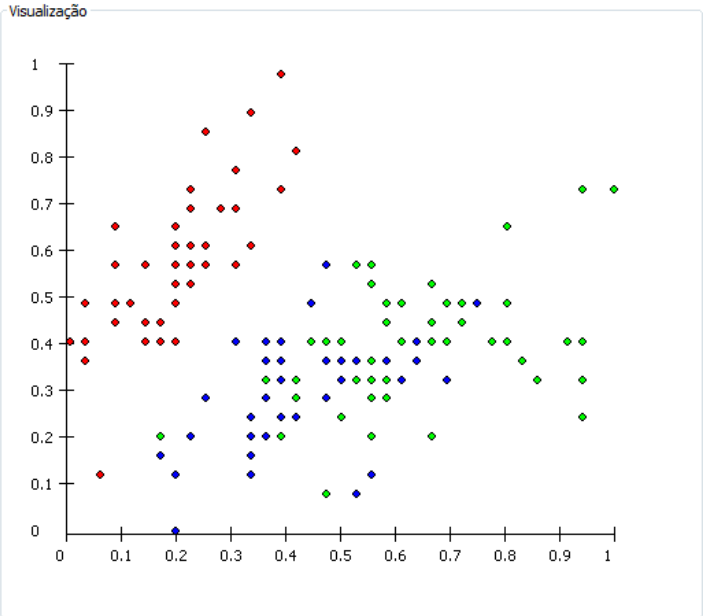


Figura 3.13: Representação de Dispersão geral da base de dados Íris, *Sepallenght* e *Sepalwidth*

A Figura 3.13 ilustra a representação da base de dados Íris considerando as coordenadas x e y com os atributos comprimento de sépala (*Sepallength*) e largura de sépala (*Sepalwidth*), respectivamente. Vale ressaltar que a base de dados foi normalizada para a sua representação, assim como para aplicação nos métodos de agrupamento de dados, justificando a escala apresentada de zero a um.

A Figura 3.14 apresenta a mesma base de dados, porém, diferenciadas as coordenadas projetadas, na qual se tem o comprimento da sépala (*Sepallenght*) e o comprimento da pétala (*Petallenght*).

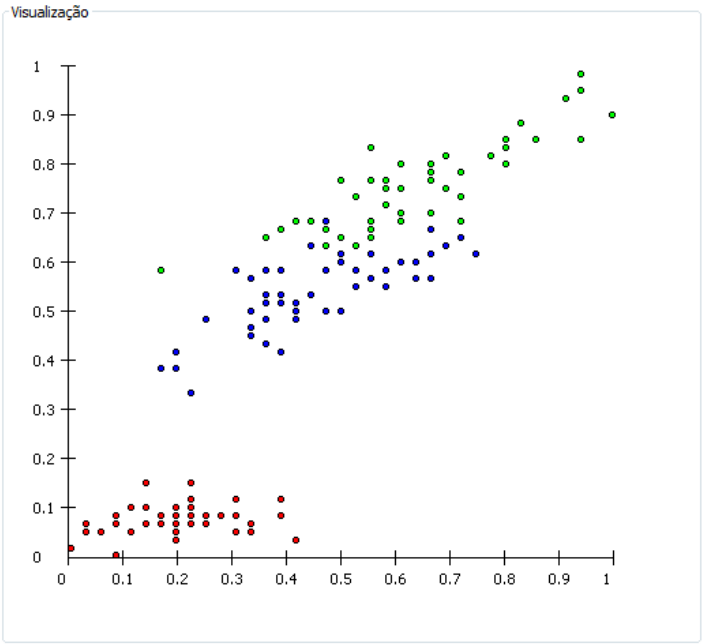


Figura 3.14: Representação de Dispersão geral da base de dados Íris, *Sepallenght* e *Petallenght*

Como dito, a representação geral da base de dados a ser minerada auxilia no seu entendimento. Com o auxílio da técnica de dispersão de dados podemos ver que a base de dados Íris possui o grupo Iris-Setosa linearmente bem separado dos outros dois grupos, o que induz que na mineração este grupo será bem definido no agrupamento.

A representação por grupos é feita pós-mineração de dados, que a partir de uma estrutura de dados tem-se a distribuição espacial dos grupos. Após isto, é possível selecionar os grupos a serem mostrados, assim como os atributos. Esta representação irá demonstrar a dispersão dos atributos de cada grupo, podendo assim ser feita a análise da qualidade do agrupamento.

A Figura 3.15 apresenta cada grupo formado pelo algoritmo de agrupamento baseado em colônia de formigas, utilizando-se da distância euclidiana para sua execução e o método de ligação simples para a recuperação dos grupos. Os grupos são apresentados em função do comprimento e de largura da sépala, eixos x e y, respectivamente.

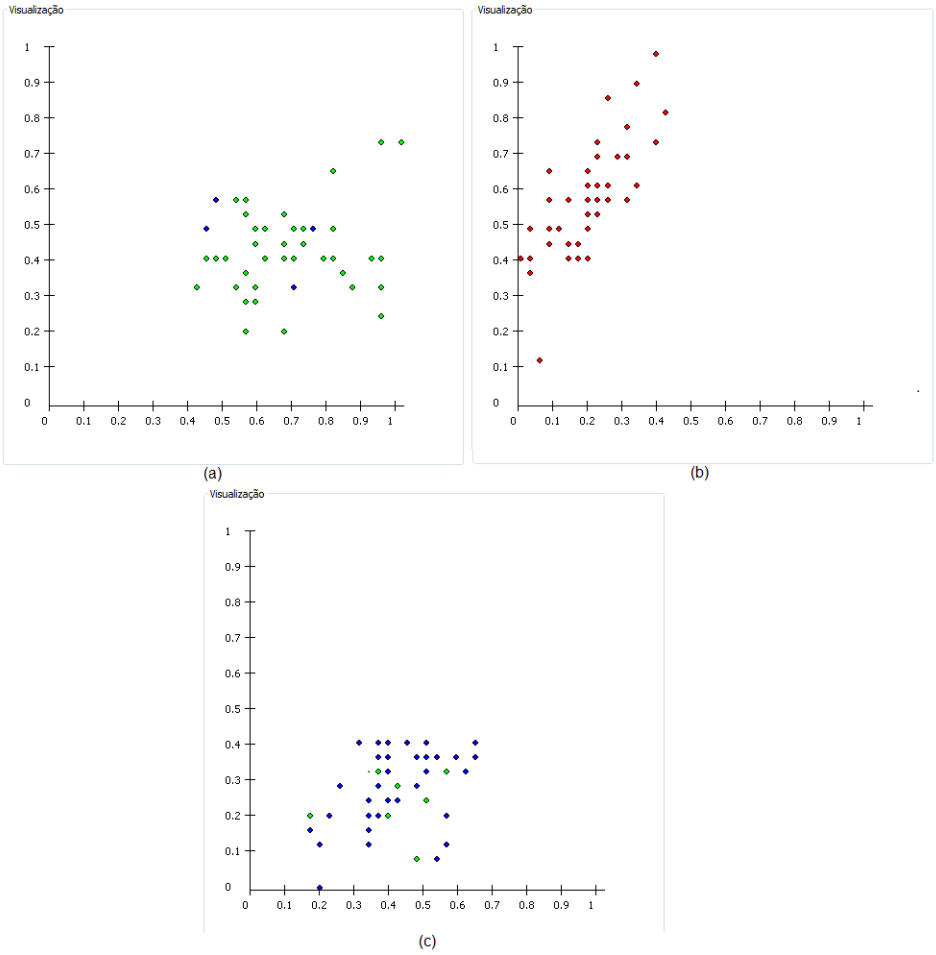


Figura 3.15: Grupos gerados pelo algoritmo de agrupamento baseado em colônia de formigas. (a) *Cluster* 1, (b) *Cluster* 2, (c) *Cluster* 3.

Como visto na representação geral da base de dados, o grupo Iris-Setosa ficou bem agrupado, tendo alguns padrões pertencentes ao grupo Iris-versicolor definido como Iris-Setosa. Para o grupo Iris-versicolor alguns dos seus cinquenta padrões foram agrupados corretamente. Também como visto na representação geral da base de dados, houve a junção do grupo Iris-Virgínica com o grupo Iris-Versicolor.

O principal benefício desta técnica é que se pode ver a distribuição espacial dos atributos de uma determinada base de dados e entre grupos pós-mineração de dados, entendendo assim, inicialmente a base de dados e posteriormente, o resultado do agrupamento de dados, também podendo certificar-se da qualidade do agrupamento gerado pelo método aplicado.

Como um ponto negativo desta técnica tem-se a limitação do espaço visual, ou seja, o número de dimensões que podem ser representadas. O método, assim como qualquer outro tipo de representação visual, tem a limitação de até três dimensões para ser representado.

Caso queira-se visualizar a distribuição espacial da base de dados, ou de um determinado grupo, a visualização dos atributos limitar-se-á a três ficando a cargo de quem esta utilizando o método a escolha destes atributos.

Outra possível limitação desta técnica envolve o dispositivo de saída, monitor, em que será representado. Isto ocorre pelo fato de que o número de padrões (objetos) de uma base de dados a ser representada pode ser muito elevado fazendo com que o número de pixels da área de visualização, em largura ou altura, não seja o suficiente para a representação de todos os atributos de maneira geral.

Um problema desta representação é a sobreposição gráfica de pontos, que se dá pelo fato dos padrões possuírem atributos com valores muito próximos, ou iguais, que na projeção para a tela são mapeados para um mesmo ponto de pixel. A dispersão em 3D dos atributos pode contornar este problema, já que o número de atributos para a representação aumenta e a possibilidade de dois pontos caírem em um mesmo ponto de pixel diminui.

Outro benefício desta representação é que as dimensões para representação aumentam, o que permite uma análise maior sobre a base de dados. A Figura 3.16 a técnica de dispersão geral em 3D, projetando em tela a base de dados Íris, com os atributos *sepallength*, *sepalwidth* e *petallength.*

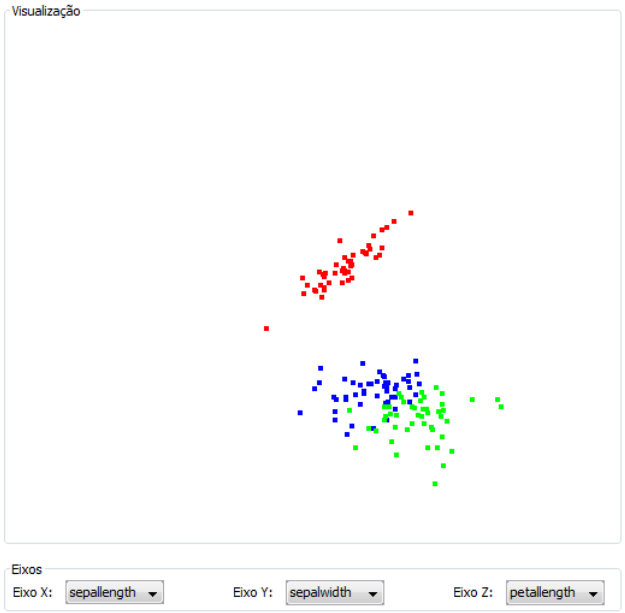


Figura 3.16: Representação base de dados Íris em três dimensões.

3.2.2 Matriz de Correlação

Este método faz parte da classe de Técnicas de Visualização em 2D e 3D e representa a correlação entre os objetos de um determinado grupo, apresentando esta correlação de uma forma bastante intuitiva, (SILVA NETO, 2008).

A representação é feita por meio de uma matriz $n\ x\ n$ em que n é o tamanho do grupo representado, ou seja, o número de objetos presentes neste grupo. Para cada par $M_{(i,j)}$ é calculada a correlação do par, considerando-se todos os atributos do objeto, e de acordo com uma classificação da correlação, em níveis, é apresentada uma cor diferente para este par, o que permite uma visualização bastante clara e intuitiva da qualidade do agrupamento, pois para um conjunto de objetos a disposição das cores apresentadas pela matriz deverá seguir um padrão uniforme, independentemente da qualidade do agrupamento. Para padrões, pontuais, que foram agrupados erroneamente no grupo visualizado, a percepção deste fato é dada de maneira muito intuitiva, pois irá fugir completamente do padrão de cor que foi atribuída ao restante dos objetos.

A Figura 3.17 ilustra o método aplicado a um agrupamento gerado pelo método de agrupamento de dados baseado em colônia de formigas.

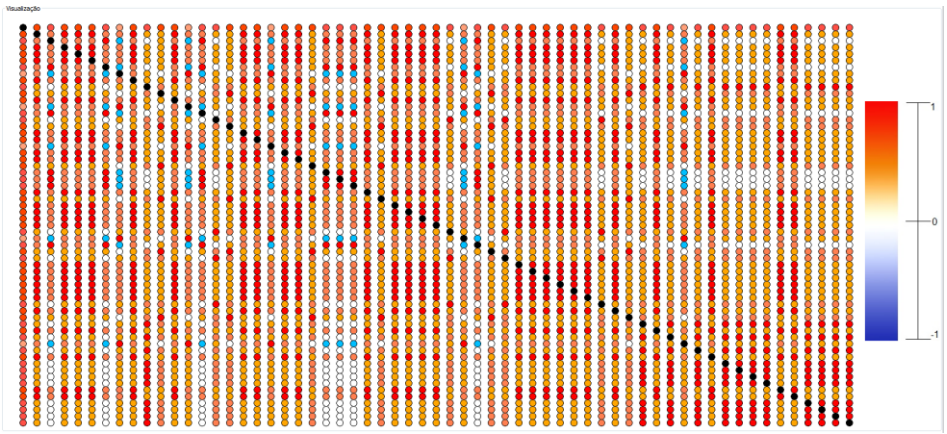


Figura 3.17: Representação de grupos por matriz de correlação

Para as posições de i e j iguais tem-se a analise de um padrão consigo mesmo, sendo assim a correlação igual a um, para este caso foi atribuída a cor preta para toda a diagonal principal da matriz de correlação. O restante da coloração é definido de acordo com uma escala, mostrada pela Figura 3.18.

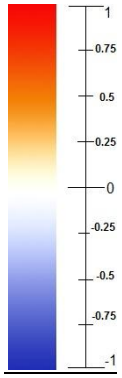


Figura 3.18: Escala de cores para matriz de correlação

A matriz de correlação pode confirmar que padrões que não são pertencentes ao mesmo grupo podem ser agrupados como tal, pois apresentam uma correlação suficientemente alta para tanto. Também a técnica permite identificar aqueles padrões que foram agrupados de forma errada, apresentando correlação negativa (coloração azulada) ou não apresentando nenhuma correlação (coloração branca).

Outro fator positivo desta técnica de visualização é que, ao contrário das demais técnicas apresentadas, esta se utiliza de todos os atributos de um objeto, o que permite que a análise seja feita de uma maneira completa, sem desconsiderar qualquer atributo na análise.

Uma dificuldade encontrada neste método é justamente a de criar essa classificação de cor de acordo com uma faixa de correlação, o que se dá pela dificuldade de saber em quantos níveis de correlação terá que ser dividida a faixa total de correlação. Teoricamente, quanto mais níveis houver, melhor será feita a representação, porém, chega-se em outra limitação que é a representação de cores presente na linguagem abordada.

Como limitação tem-se, assim como as outras técnicas, a tela do dispositivo em que será feita a representação da técnica. Também para esta técnica o número de objetos de um único grupo é um limitador, pois poderá passar dos limites da área de representação. Ainda, considerando-se como é feita a técnica, representação da matriz, não é possível considerar somente um pixel para cada posição i e j da matriz, pois este único pixel não é suficiente para visualizar a posição.

Logo, para cada i e j da matriz será necessária uma pequena área de pixels que seja suficiente para que seja possível visualizar corretamente a posição i e j da matriz. Tendo isso em mente, para ilustrar esse problema basta considerar um grupo com 300 objetos, a matriz de correlação será uma matriz de 300 linhas por 300 colunas. Considerando a área de cada

posição i e j desta matriz representada por um quadrado constituído por 10 linhas de pixels e 10 colunas de pixels a área total que este grupo irá necessitar será de 3000 pixels por 3000 pixels, o que ultrapassa o limite dos dispositivos convencionais que se tem atualmente disponível no mercado.

3.2.3 Coordenadas Paralelas

A técnica de Coordenadas Paralelas foi introduzida por Inselberg e Dimsdale (1990) para representar múltiplas dimensões sem utilizar eixos cartesianos ortogonais. Os dados são apresentados utilizando-se de linhas verticais e horizontais, em que cada linha vertical representa um atributo, sendo o total de linhas verticais igual à dimensionalidade da base de dados. As linhas horizontais são mapeadas para pontos nas linhas verticais, de forma que cada item de dado seja representado como uma linha poligonal que intercepta a linha vertical no seu ponto correspondente ao valor do atributo.

Os extremos de cada linha vertical indicam, de baixo para cima, o valor mínimo e máximo de cada atributo, respectivamente. A Figura 3.19 apresenta a técnica em que é representada a base de dados Íris, as cores apresentadas são as cores de cada classe presente na base de dados.

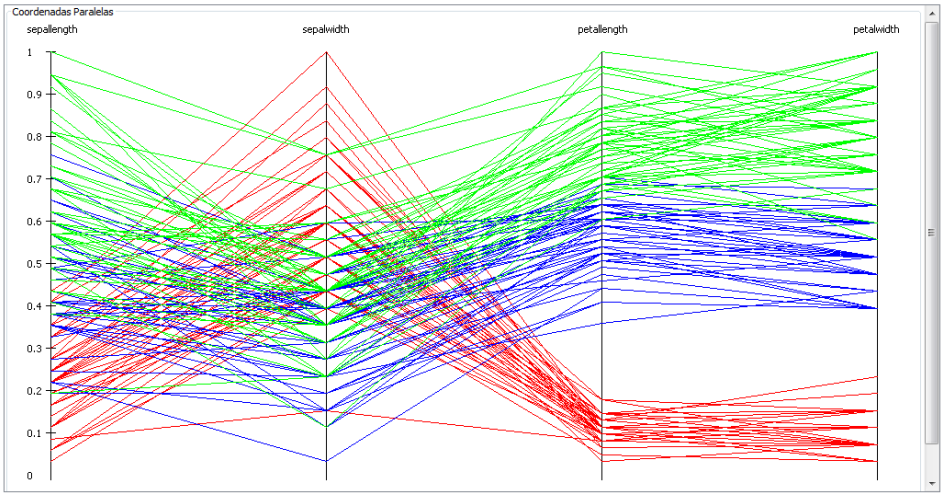


Figura 3.19: Coordenadas Paralelas – base de dados Íris

A representação visual das Coordenadas Paralelas facilita a identificação de características como: distribuição dos dados e correlações entre atributos. Em contrapartida existe a limitação da quantidade de dados, padrões, a ser apresentado. Segundo Keim (2002), o grande número de padrões causa a sobreposição das linhas poligonais assim criando uma limitação

no número de padrões a ser visualizado, que segundo o mesmo é de 1000 padrões. Para resolver este problema técnicas adequadas de iteração podem ser aplicadas para uma melhor visualização, como o método de seleção de pontos. Com esta técnica pode-se selecionar um ponto, em um dos eixos verticais, e visualizar, destacadamente, todas as outras linhas que por ali passam.

3.2.4 Coordenadas Paralelas Circulares

Também denominada Gráfico de Estrelas Sobrepostas, correspondem à versão circular das Coordenadas Paralelas. A ideia de representação dos dados é a mesma da primeira técnica, mudando apenas a disposição das dimensões da base de dados, formando um gráfico em estrela.

Para este formato de representação as linhas na parte externa do círculo são mais longas, indicando que os valores, ali mapeados, são maiores. Para valores menores o mapeamento se concentra na parte central do círculo. As mesmas características de representação e contrapartidas do método anterior são apresentadas neste método. A diferenciação deste método para o anterior é de que a diferença na representação de dados com valores altos e baixos gera assimetrias bastante claras, de forma a ser possível a detecção padrões nessas assimetrias (FAYYAD; GRINSTEIN; WIERSE, 2002).

A Figura 3.20 traz a visualização da técnica Coordenadas Paralelas Circulares, a partir da base de dados Iris, em que as cores utilizadas são as classes presentes na base de dados.

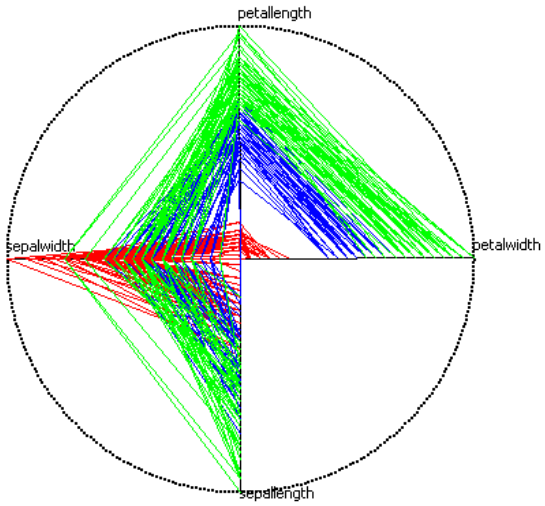


Figura 3.20: Coordenadas Paralelas Circulares – base de dados Íris

3.2.5 *Scatter Matrix*

A técnica *Scatter Matrix* é uma técnica de projeção pertencente à classe de Projeções Geométricas. Esta classificação se dá, uma vez que seus dados são plotados em coordenadas *x* e *y* de forma semelhante a outros gráficos bidimensionais, diferenciando que para este técnica são exibidas simultaneamente as múltiplas dimensões da base de dados. Por exemplo, para uma base de dados com 4 atributos, como a base de dados Íris, será apresentado uma matriz 4x4 e em cada posição *i* e *j* desta matriz terá a apresentação bidimensional dos atributos, par a par, da base de dados.

A Figura 3.21 apresenta a base de dados Íris em que cada atributo é representado par a par. Com esta técnica também é possível visualizar a correlação dos atributos presentes na base de dados.

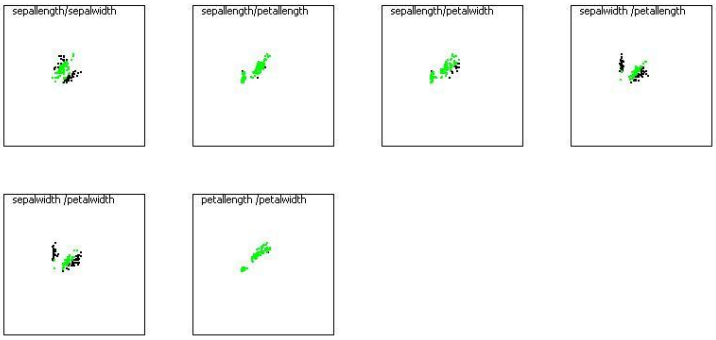


Figura 3.21: *Scatter Matrix* – base de dados Íris

A interpretação da *Scatter Plots* é fácil, pois permite a visualização de todas as possíveis correlações entre os pares de dimensões. Porém, esta técnica apresenta duas principais limitações que são:

- Limitação gráfica, tamanho de tela, para a representação de bases de dados que contenham muitas dimensões. As dimensões não poderão ser visualizadas todas simultaneamente;
- Sobreposição de padrões, quando a base de dados for constituída de muitos padrões, poderá ocasionar a sobreposição dos mesmos na representação em tela.

3.2.6 Dendrograma

O dendrograma é a representação da árvore hierárquica de uma base de dados, ou seja, apresenta a hierarquia entre os dados desta base de dados. Esta hierarquia é obtida por meio de um método de agrupamento de dados, como os apresentados na Seção 2.1.2. A formação da árvore inicia de baixo para cima, em que inicialmente cada folha é um grupo e estas se juntam através de nós formando um novo grupo a cada nó, e encerrando quando ocorre a formação da raiz da árvore.

A formação da árvore hierárquica se dá inicialmente pela obtenção das ordens de formação de grupos pelo método de agrupamento hierárquico. Isto evita que haja o cruzamento de arestas na representação gráfica da árvore, dendrograma. O segundo passo é a formação da árvore propriamente dita, que a partir das ordens de formação de grupos junta-os como nó na árvore.

Ambos os métodos descritos foram implementados neste trabalho, porém para implementação da representação gráfica do dendrograma foi utilizado a ferramenta Prefuse (HEER *et al.*, 2005), que é uma ferramenta livre para a representação gráfica de estruturas de dados como árvores binárias e grafos. A utilização desta ferramenta justifica-se pela complexidade de representar graficamente uma estrutura de dados como a utilizada. Também há métodos de iterações gráficos que facilitam a visualização de todos os dados apresentados.

A Figura 3.22 apresenta a utilização Prefuse representando parte do agrupamento aglomerativo formado pelo método *single-linkage* utilizando a distância euclidiana, onde tem-se o padrão 1 como raiz da árvore, significando que todos os outros padrões pertencem a este padrão.

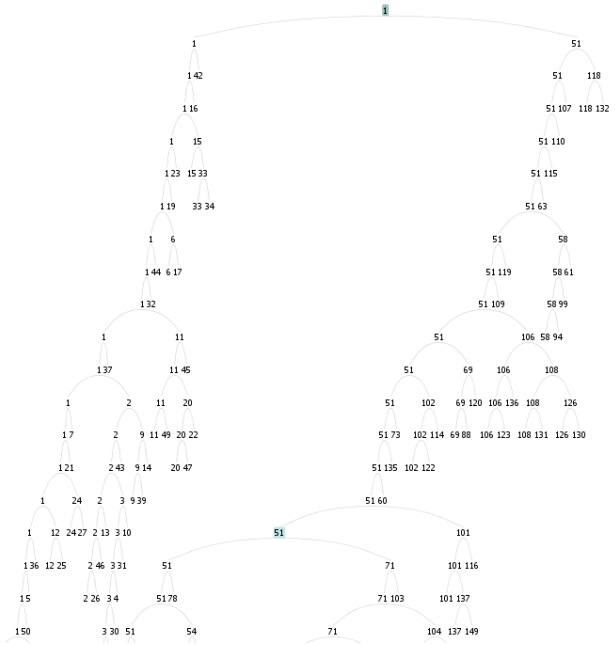
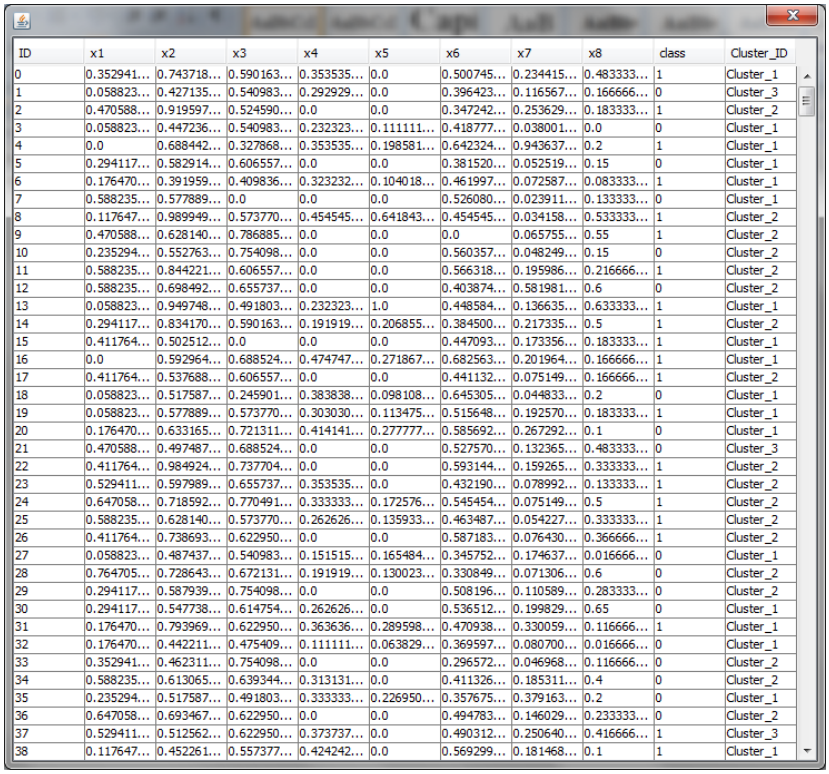


Figura 3.22: Dendrograma gerado pelo método *single-linkage* – base de dados Íris

3.2.7 Tabela de Visualização

O método consiste em apresentar, em forma de tabela, a base de dados selecionada apresentando o grupo em que cada padrão desta base de dados se encontra. Também é possível de visualizar todos os atributos, e seus valores, que constitui a base de dados, assim como a classe em que cada padrão pertence. A Figura 3.23 apresenta a base de dados Pima (FRANK, ASUNCION, 2010) agrupada pelo método *k-means*. No método é possível visualizar os valores dos atributos que constituem um determinado padrão e ordenando uma determinada coluna da tabela entender o agrupamento formado.



ID	x1	x2	x3	x4	x5	x6	x7	x8	class	Cluster_ID
0	0.352941...	0.743718...	0.590163...	0.353535...	0.0	0.500745...	0.234415...	0.483333...	1	Cluster_1
1	0.058823...	0.427135...	0.540983...	0.292929...	0.0	0.396423...	0.116567...	0.166666...	0	Cluster_3
2	0.470588...	0.919597...	0.524590...	0.0	0.0	0.347242...	0.253629...	0.183333...	1	Cluster_2
3	0.058823...	0.447236...	0.540983...	0.232323...	0.111111...	0.418777...	0.038001...	0.0	0	Cluster_1
4	0.0	0.688442...	0.327868...	0.353535...	0.198581...	0.642324...	0.943637...	0.2	1	Cluster_1
5	0.294117...	0.582914...	0.606557...	0.0	0.0	0.381520...	0.052519...	0.15	0	Cluster_1
6	0.176470...	0.391959...	0.409836...	0.323232...	0.104018...	0.461997...	0.072587...	0.083333...	1	Cluster_1
7	0.588235...	0.577889...	0.0	0.0	0.0	0.526080...	0.023911...	0.133333...	0	Cluster_1
8	0.117647...	0.989949...	0.573770...	0.454545...	0.641843...	0.454545...	0.034158...	0.533333...	1	Cluster_2
9	0.470588...	0.628140...	0.786885...	0.0	0.0	0.0	0.065755...	0.55	1	Cluster_2
10	0.235294...	0.552763...	0.754098...	0.0	0.0	0.560357...	0.048249...	0.15	0	Cluster_2
11	0.588235...	0.844221...	0.606557...	0.0	0.0	0.566318...	0.195986...	0.216666...	1	Cluster_2
12	0.588235...	0.698492...	0.655737...	0.0	0.0	0.403874...	0.581981...	0.6	0	Cluster_2
13	0.058823...	0.949748...	0.491803...	0.232323...	1.0	0.448584...	0.136635...	0.633333...	1	Cluster_1
14	0.294117...	0.834170...	0.590163...	0.191919...	0.206855...	0.384500...	0.217335...	0.5	1	Cluster_2
15	0.411764...	0.502512...	0.0	0.0	0.0	0.447093...	0.173356...	0.183333...	1	Cluster_1
16	0.0	0.592964...	0.688524...	0.474747...	0.271867...	0.682563...	0.201964...	0.166666...	1	Cluster_1
17	0.411764...	0.537688...	0.606557...	0.0	0.0	0.441132...	0.075149...	0.166666...	1	Cluster_2
18	0.058823...	0.517587...	0.245901...	0.383838...	0.098108...	0.645305...	0.044833...	0.2	0	Cluster_1
19	0.058823...	0.577889...	0.573770...	0.303030...	0.113475...	0.515648...	0.192570...	0.183333...	1	Cluster_1
20	0.176470...	0.633165...	0.721311...	0.414141...	0.277777...	0.585692...	0.267292...	0.1	0	Cluster_1
21	0.470588...	0.497487...	0.688524...	0.0	0.0	0.527570...	0.132365...	0.483333...	0	Cluster_3
22	0.411764...	0.984924...	0.737704...	0.0	0.0	0.593144...	0.159265...	0.333333...	1	Cluster_2
23	0.529411...	0.597989...	0.655737...	0.353535...	0.0	0.432190...	0.078992...	0.133333...	1	Cluster_2
24	0.647058...	0.718592...	0.770491...	0.333333...	0.172576...	0.545454...	0.075149...	0.5	1	Cluster_2
25	0.588235...	0.628140...	0.573770...	0.262626...	0.135933...	0.463487...	0.054227...	0.333333...	1	Cluster_2
26	0.411764...	0.738693...	0.622950...	0.0	0.0	0.587183...	0.076430...	0.366666...	1	Cluster_2
27	0.058823...	0.487437...	0.540983...	0.151515...	0.165484...	0.345752...	0.174637...	0.016666...	0	Cluster_1
28	0.764705...	0.728643...	0.672131...	0.191919...	0.130023...	0.330849...	0.071306...	0.6	0	Cluster_2
29	0.294117...	0.587939...	0.754098...	0.0	0.0	0.508196...	0.110589...	0.283333...	0	Cluster_2
30	0.294117...	0.547738...	0.614754...	0.262626...	0.0	0.536512...	0.199829...	0.65	0	Cluster_1
31	0.176470...	0.793969...	0.622950...	0.363636...	0.289598...	0.470938...	0.330059...	0.116666...	1	Cluster_1
32	0.176470...	0.442211...	0.475409...	0.111111...	0.063829...	0.369597...	0.080700...	0.016666...	0	Cluster_1
33	0.352941...	0.462311...	0.754098...	0.0	0.0	0.296572...	0.046968...	0.116666...	0	Cluster_2
34	0.588235...	0.613065...	0.639344...	0.313131...	0.0	0.411326...	0.185311...	0.4	0	Cluster_2
35	0.235294...	0.517587...	0.491803...	0.333333...	0.226950...	0.357675...	0.379163...	0.2	0	Cluster_1
36	0.647058...	0.693467...	0.622950...	0.0	0.0	0.494783...	0.146029...	0.233333...	0	Cluster_2
37	0.529411...	0.512562...	0.622950...	0.373737...	0.0	0.490312...	0.250640...	0.416666...	1	Cluster_3
38	0.117647...	0.452261...	0.557377...	0.424242...	0.0	0.569299...	0.181468...	0.1	1	Cluster_1

Figura 3.23: Tabela de visualização da base de dados Íris

3.3 Considerações Finais

As técnicas de visualização permitem uma análise dos dados facilitando a extração do conhecimento de uma determinada base de dados ou uma facilitação no entendimento do resultado de métodos de agrupamento de dados. Para cada tipo de dado é sugerido um determinado método de visualização que melhor se encaixe. A escolha de um método que melhor represente uma base de dados é difícil, porém, a utilização em conjunto de várias técnicas facilita o processo de extração de conhecimento.

Este capítulo apresentou, brevemente, os tipos de dados que podem ser visualizados, e também aplicados a métodos de agrupamento de dados. Também apresentou classes de técnicas de visualização assim como os métodos que constituem estas classes.

Também foram apresentados os métodos de visualização que foram implementados na YADMT. Cada método de agrupamento de dados implementado poderá ter seus resultados visualizados por métodos de visualização, cada qual apresentando suas vantagens e desvantagens, como já discutido.

Capítulo 4

A Implementação da Ferramenta YADMT

Métodos de Visualização de Dados para Agrupamento de Dados, assim como os métodos de Agrupamento de Dados, foram implementados na YADMT – *Yet Another Data Mining Tool*, desenvolvida pela Universidade Estadual do Oeste do Paraná – no Grupo de Pesquisa em Inteligência Aplicada (GIA), seguindo o modelo proposto por Benfatti *et al.* (2010).

Conforme descrição de (Benfatti *et al.*, 2010), a YADMT é uma ferramenta de *Data Mining* que executa todas as etapas do processo *KDD*, conforme Figura 1.1. A construção da ferramenta se da de forma modular, possibilitando assim que técnicas pertencentes a cada etapa do processo *KDD* seja desenvolvida e inserida de forma independente, fazendo com que estes módulos sejam agrupados como pacotes independentes. Em complemento, foram definidas regras para a construção de cada módulo e também, meta informações contidas no próprio código-fonte em Java.

Como primeira etapa do desenvolvimento da YADMT foram implementadas técnicas de aquisição de dados, estes que são essenciais para as etapas seguinte do processo *KDD*. Esta aquisição de dados pode ser feito através de SGBD – Sistema Gerenciador de Banco de Dados, que para a ferramenta é utilizado o PostgreSQL utilizando o serviço de conectividade JDBC (*Java Database Connectivity*). Outra fonte de dados para YADMT tem-se a leitura de arquivos *ARFF* (*Attribute-Relation File Format*) (ARFF, 2002), a escolha deste padrão de arquivo deve-se a utilização deste padrão na ferramenta de mineração de dados Weka (WEKA, 2002).

Após, os métodos de classificação de dados k-NN, C4.5 e *Naive Bayes* (BENFATTI, 2010), e três RNA (Redes Neurais Artificiais) *Multilayer Perceptron* (*MLP*), *Radial Basis Function* (*RBF*) e *Learning Vector Quantization* (*LVQ*), (BONIFÁCIO, 2010) constituíram o módulo de Classificação. Como uma terceira etapa de desenvolvimento foi desenvolvida uma Máquina de Comitê Estática (SIPPERT, 2012) também para a tarefa de Classificação. Ainda, foi desenvolvido o módulo de Extração de Características por (RÖHSING SILVA, 2012).

Como mais uma etapa da ferramenta tem-se o desenvolvimento deste trabalho que faz parte de um novo módulo da YADMT, o módulo de agrupamento de dados, constituído dos

métodos de Agrupamento e de Visualização de Dados apresentados nas seções anteriores. A seguir, será apresentada a parte do módulo de Agrupamento de Dados desenvolvido por este trabalho.

4.1 O Módulo de Agrupamento de Dados

Os métodos de Agrupamento de Dados e Visualização de Dados foram implementados seguindo a filosofia de implementação da ferramenta. Todos os métodos do módulo de Agrupamento são conectados, podendo, a partir de um método de Agrupamento utilizar-se de todos os métodos de recuperação e avaliação do agrupamento, assim como os métodos de visualização, salvo os métodos de agrupamento hierárquico e de visualização hierárquica (dendrograma).

Para iniciar a utilização de qualquer método de agrupamento, ou de visualização de dados, deve-se fazer a aquisição da base de dados a ser utilizada, conforme ilustra a Figura 4.1.

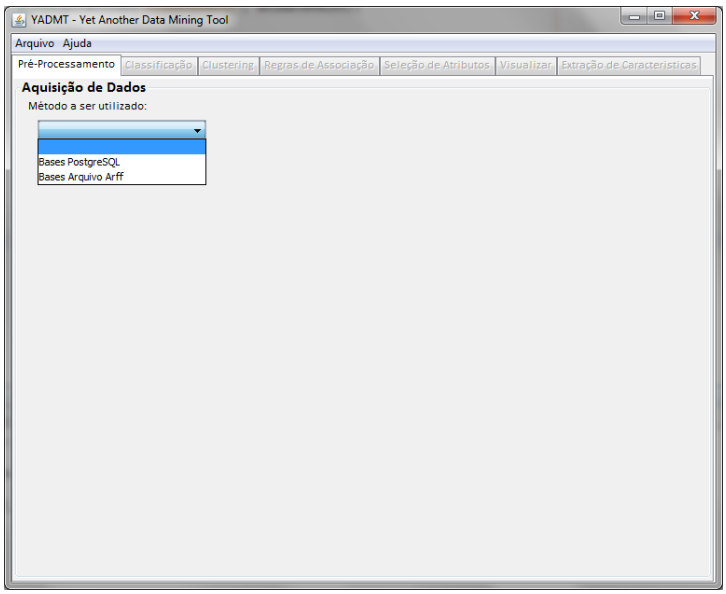


Figura 4.1: Tela aquisição de dados – YADMT

A aquisição da base de dados pode ser feita através do SGBD PostgreSQL, em que na YADMT, deve-se especificar o local do banco de dados, porta de conexão, nome da base de dados, usuário e senha. Após estas informações pode-se criar a conexão entre o banco de dados e a YADMT. A Figura 4.2 apresenta a tela do sistema em questão.

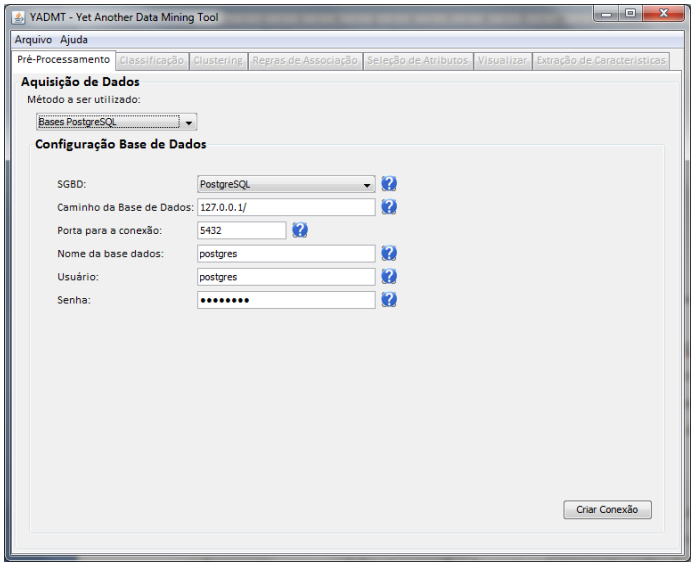


Figura 4.2: Tela aquisição de dados via SGBD – YADMT

A aquisição da base de dados através de arquivos do tipo *ARFF* é feita por meio da tela representada pela Figura 4.3. Nesta tela especifica-se o local do arquivo, que carregado para a ferramenta, e possibilitará a escolha de quais colunas serão utilizadas nos métodos presentes na YADMT, além de poder visualizar a base de dados escolhida.

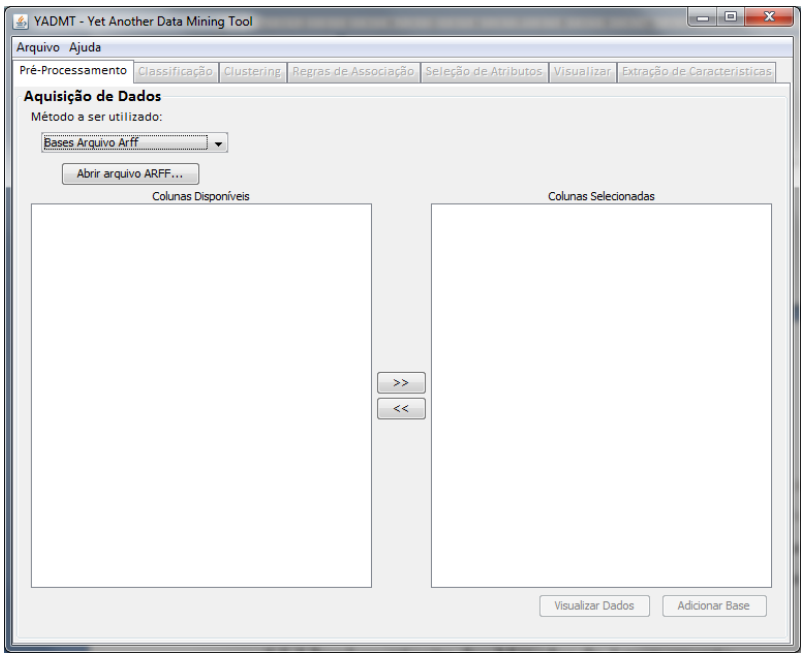


Figura 4.3: Tela aquisição de dados via *ARFF* – YADMT

4.1.1 Implementação dos Métodos de Agrupamento

Os métodos de agrupamento de dados foram implementados de forma com que não tivessem dependências entre si, assim, a implementação se deu de forma individual em que cada método tem sua própria execução sem dependências. Isso é possível de ser visualizado pela Figura 4.4, que mostra a tela de escolha dos métodos.

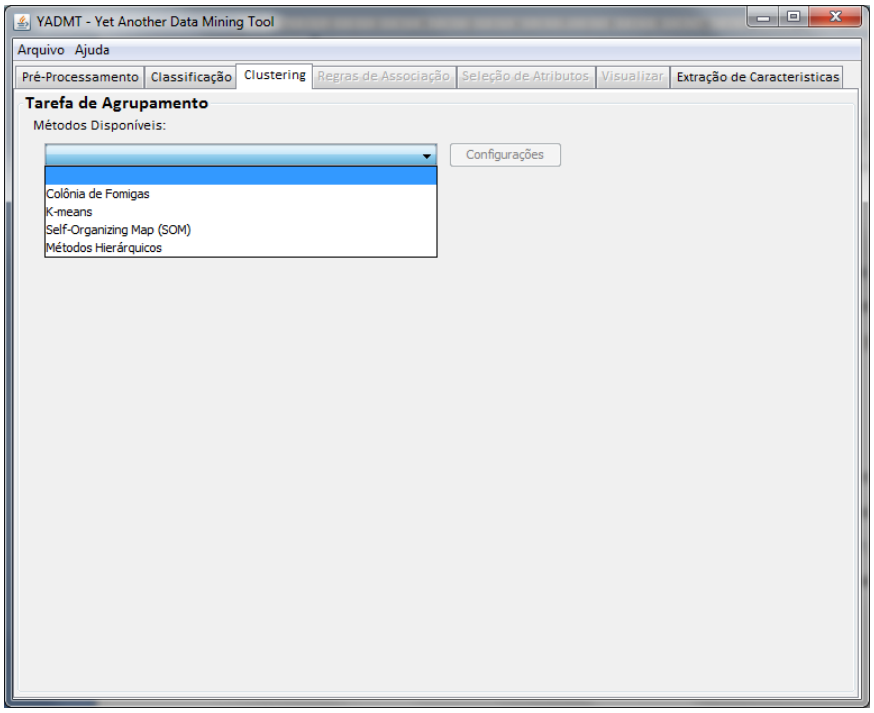


Figura 4.4: Tela escolha de método – YADMT

A Figura 4.5 traz a tela do método de Agrupamento Baseado em Colônia de Formigas, apresentando suas opções de execução, com destaque para as medidas de distância, descritas no Anexo A, e os métodos de recuperação de grupos. Para a escolha de parâmetros do método tem-se o botão “Configurações”, representado pela Figura 4.6. Durante a execução do método é possível acompanhar (aba “Simulação”) a representação da sua execução da movimentação dos padrões da base de dados. Após a execução do método é possível, através do botão “Visualização” ter acesso aos métodos de visualização de dados implementados. Ainda nesta tela também é possível selecionar uma execução anterior do método, podendo assim comparar resultados de diferentes execuções.

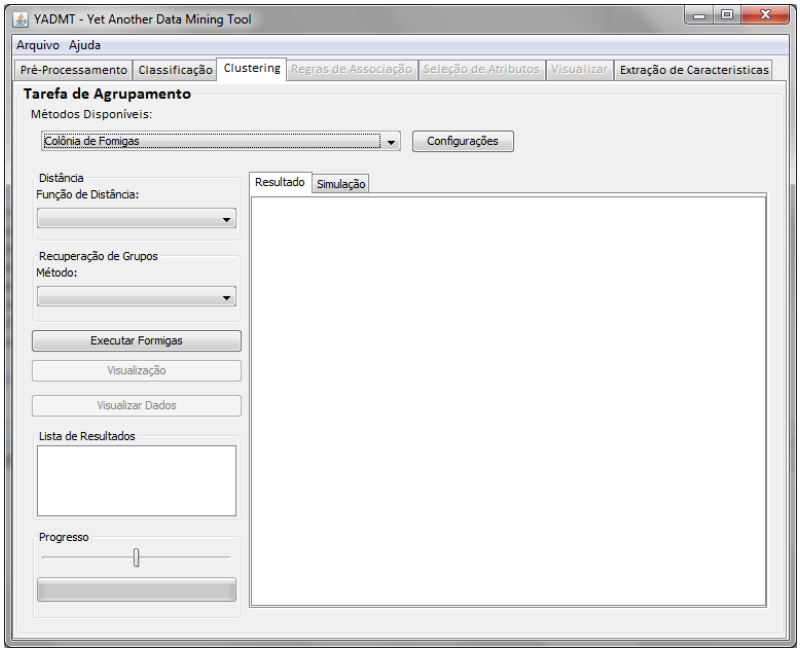


Figura 4.5: Tela método Colônia de Formigas – YADMT

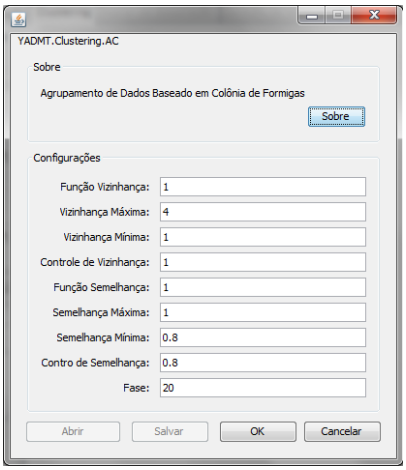


Figura 4.6: Tela de configurações do método Colônia de Formigas – YADMT

A tela de configurações do método de Colônia de Formigas apresenta todos os parâmetros possíveis de serem alterados, como o raio de vizinhança (σ), o grau de semelhança entre padrões (α). Os parâmetros "Vizinhança Máxima", "Vizinhança Mínima", "Controle de Vizinhança", "Semelhança Máxima", "Semelhança Mínima" e "Controle de Semelhança" controlam os valores máximos e mínimos que σ e α podem atingir. O parâmetro "Fase" indica a porcentagem de iteração que os parâmetros de vizinhança e semelhança irão crescer, sendo para o parâmetro σ uma progressão aritmética e para α uma progressão geométrica. Além da

definição dos parâmetros, a tela apresenta opções para abrir e salvar as configurações para posterior utilização.

A Figura 4.7 apresenta a tela de execução do método de agrupamento *k-means*, na qual pode-se escolher a medida de distância para uso no método, todos os outros parâmetros possíveis de serem definidos estão presentes no botão "Configurações", apresentados na Figura 4.8. O botão "Imprimir Histórico" irá apresentar todos os centroides e disposição dos grupos, formados durante a execução do método. O botão "Visualização" irá abrir os métodos de visualização de dados para agrupamento de dados disponíveis.

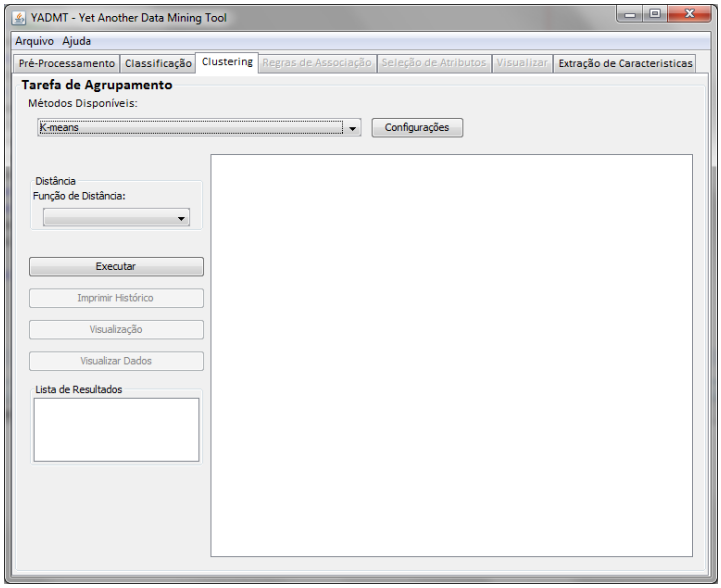


Figura 4.7: Tela método *k-means* – YADMT

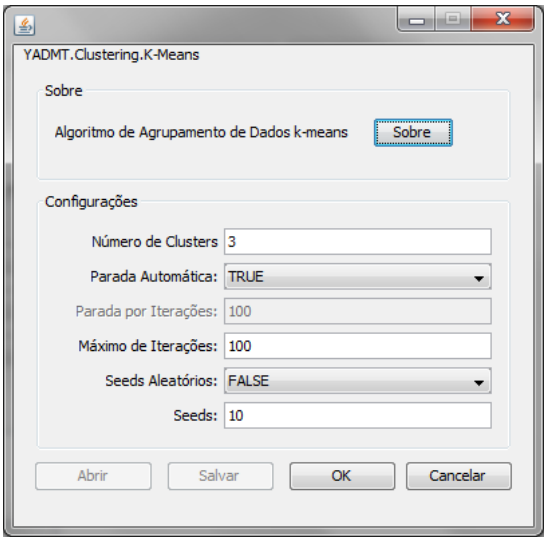


Figura 4.8: Tela de configurações do método *k-means* – YADMT

A tela de configurações do método *k-means* permite ajustar os parâmetros para o método. Nesta tela, é possível especificar o número de grupos a ser formados (*k*), se a parada do método será de maneira automática (quando não houver mudanças nos grupos) ou, por atingir um número limite de iterações, um máximo de iterações que é utilizado para prevenir laços infinitos (caso sempre haja mudança nos grupos) e a determinação dos centroides iniciais (*seeds*), que pode ser feito por meio de centroides iniciais com valores aleatórios (*seeds* aleatórios) ou que contenham a média de uma porcentagem da base de dados.

A Figura 4.9 apresenta a tela dos métodos hierárquicos, na qual é possível escolher qual método será executado, e também é possível escolher a função distância a ser utilizado e o número de grupos a ser formado pelo método de agrupamento hierárquico, isto via botão "Configurações". No botão "Visualização" é possível abrir os métodos de visualização disponíveis. A lista de resultados apresenta todas as execuções feitas durante uma sessão.

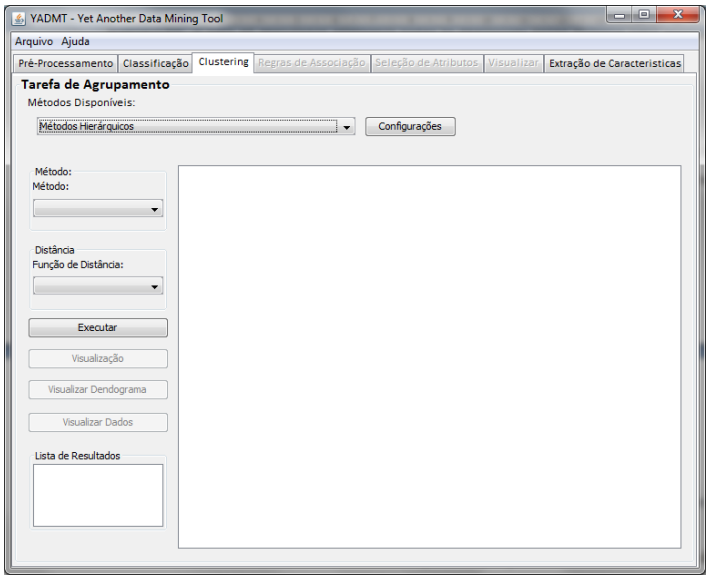


Figura 4.9: Tela métodos hierárquicos – YADMT

Em todas as telas de métodos de agrupamento é possível visualizar o botão "Visualizar Dados" que é responsável por executar o método de visualização "Tabela de Visualização", que somente é possível de ser executado após um método de agrupamento.

4.1.1 Implementação dos Métodos de Visualização de Dados

A implementação dos métodos de visualização de dados para o módulo de agrupamento de dados da YADMT foi feita para que todos os métodos de agrupamento possam ter acesso às visualizações do grupos e da base de dados, com exceção do dendrograma que é exclusivo para os métodos hierárquicos.

Para tanto, foi desenvolvida uma interface padrão, que possuísse todos os métodos de visualização implementados neste trabalho (Seção 3.2). Esta interface é representada pela Figura 4.10. Na aba desta tela pode-se escolher qual método será visualizado; por padrão inicia com a técnica de "Gráfico de Dispersão Geral". Ainda, é possível observar a tela para o método Gráfico de Dispersão, onde é possível fazer a escolha dos eixos a serem plotados em tela, caso os três eixos sejam escolhidos a visualização passará de 2D para 3D, sem necessitar a mudança de aba.

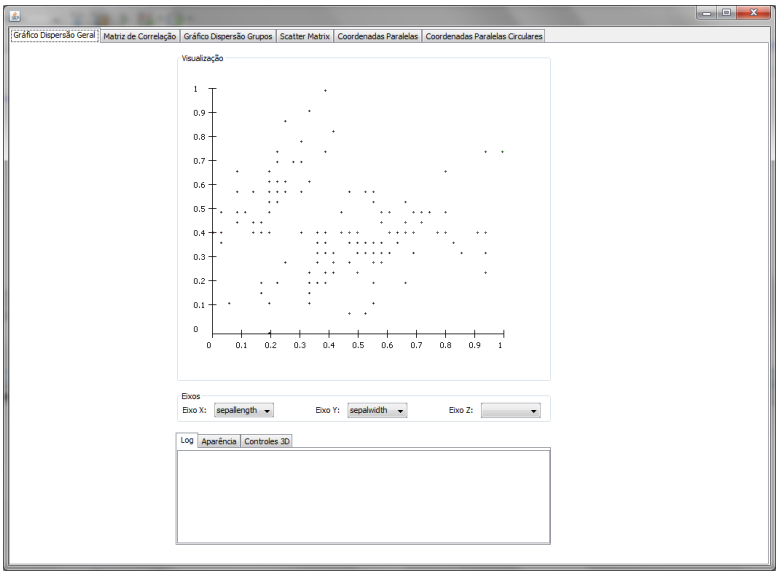


Figura 4.10: Tela gráfico de dispersão geral – YADMT

Ainda nesta primeira aba tem-se as sub abas "*Log*", "Aparência" e "Controles 3D". Na aba "*Log*" há a apresentação dos dados selecionados em tela, identificando qual padrão foi selecionando e também mostrando os valores dos atributos que formam este padrão. Na aba "Aparência" pode-se escolher o tamanho do ponto, em *pixels*, que será apresentado em tela e também habilitar a opção de seleção de pontos. Na aba "Controles 3D" tem-se os controles para a representação tridimensional: rotações no eixo *X*, rotações no eixo *Y* e *zoom In* e *Out.*

Na segunda aba, Figura 4.11, tem-se o método de “Matriz de Correlação” onde é possível escolher qual grupo será representado em tela, o coeficiente de correlação que será utilizado para calcular a correlação entre os padrões do grupo e a quantidade padrões presentes neste grupo.

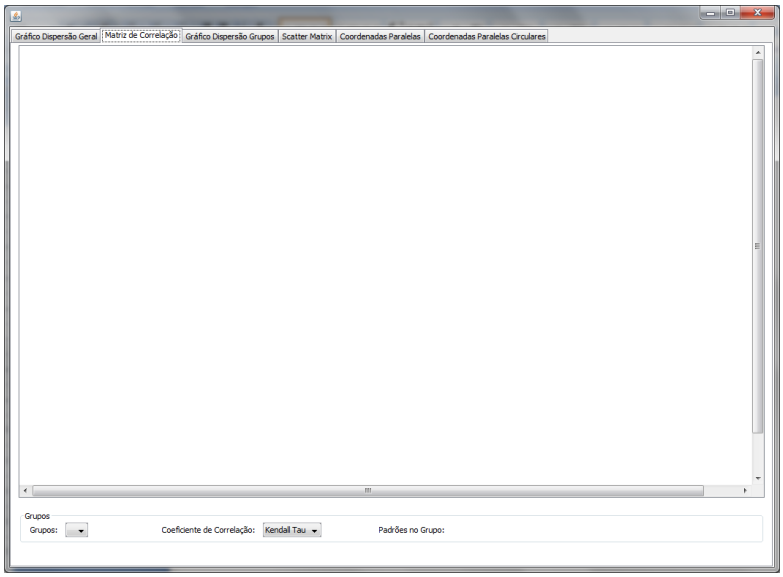


Figura 4.11: Tela matriz de correlação – YADMT

Na terceira aba, Figura 4.12, encontra-se o método de gráfico de dispersão para grupos, nesta tela, assim como a primeira aba, pode-se escolher quais atributos irão compor os eixos de representação, em que o eixo Z preenchido transforma a visualização de 2D para 3D. Possui as mesmas sub abas da primeira aba, Figura 4.10, com adição da aba “Seleção de Grupos” em que são apresentados os grupos formados por um método de agrupamento e estes podem ser adicionados, ou removidos, da visualização.

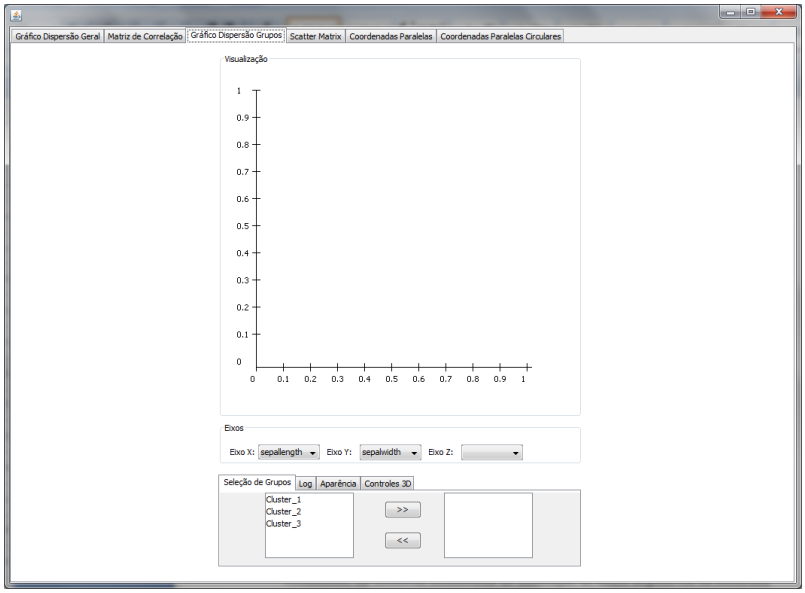


Figura 4.12: Tela gráfico de dispersão de grupos – YADMT

Na quarta aba, Figura 4.13, é apresentado o método de *Scatter Matrix*, além da apresentação simultânea das múltiplas dimensões, par a par, de uma base de dados é apresentado, por meio de uma tabela, a correlação de cada atributo da base de dados com outro par de atributos, também na aba "Medida de Correlação" é possível escolher qual coeficiente de correlação será utilizado para calcular as correlações que serão exibidas na tabela de correlações. Como a matriz formada é uma matriz simétrica somente os atributos não repetidos são apresentados. Os pontos coloridos na cor verde são aqueles pontos que tem alguma significância dentro da base de dados, de acordo com o calculo do *p* valor.

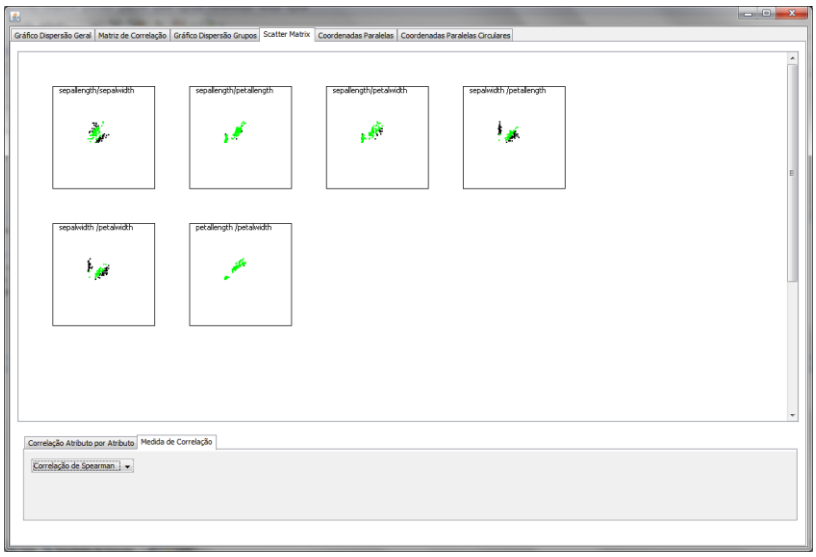


Figura 4.13: Tela *scatter matrix* – YADMT

Na quinta aba, Figura 4.14, encontra-se o método de coordenadas paralelas, nela é possível, além de visualizar o método, determinar a espessura das linhas, o espaçamento entre os eixos que representam os atributos de uma base de dados. Também é possível selecionar um ponto em um dos eixos e, por meio do campo de tela visualizar o padrão selecionado e os atributos que o formam. As mesmas informações contidas na quinta aba encontram-se na sexta aba, com o diferencial que nesta última será apresentado o método de coordenadas paralelas circulares (Figura 4.15).

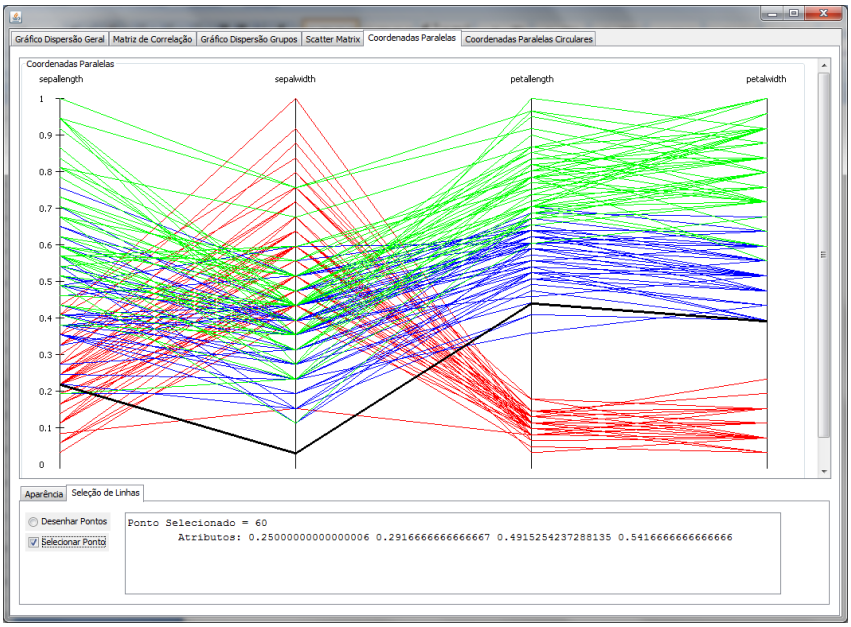


Figura 4.14: Tela coordenadas paralelas – YADMT

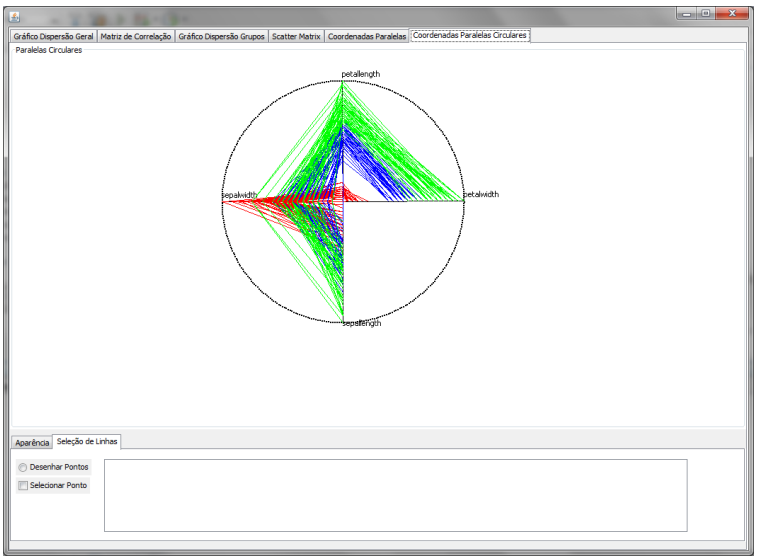


Figura 4.15: Tela coordenadas paralelas circulares – YADMT

O método de visualização para os métodos de agrupamento de dados hierárquico, dendrograma, e também a tabela de visualização de dados são métodos que são acessados fora do escopo dos outros métodos de visualização, portanto o acesso a estes se dá na janela de execução do método, como mostrado na Seção 4.1.1.

4.1.2 Implementação dos Métodos de Iteração

Considerando a linguagem escolhida para desenvolvimento deste trabalho e considerando que os recursos gráficos desta são limitados, os métodos de iteração implementados são básicos. O desenvolvimento de métodos de iteração mais complexos e funcionais somente com os recursos nativos da linguagem de programação escolhida é possível, porém, demandaria de um tempo maior do que o disponível para a realização de trabalho completo. Sendo assim os métodos de iteração presentes neste trabalho são:

- **Escolha do tamanho de ponto**: pode-se escolher o tamanho do ponto a ser apresentado em tela, facilitando assim a sua visualização e também seleção. Presente nos métodos de gráfico de dispersão geral e por grupos.
- **Escolha da espessura de linha**: pode-se escolher a espessura da linha a ser apresentada em tela, facilitando a visualização destas. Isto está presente nos métodos de coordenadas paralelas e coordenadas paralelas circulares.
- **Definição da distância entre eixos das coordenadas paralelas:** pode-se determinar a distância (em pixels) entre os eixos de visualização do método coordenadas paralelas, isto aproxima, ou afasta, os eixos facilitando a visualização da base de dados.
- **Representação de pontos de interseção:** pode-se representar os pontos de interseção dos eixos de atributos com as linhas que representam o valor destes atributos nos métodos de coordenadas paralelas e coordenadas paralelas circulares. Isto facilita a visualização de onde ocorre a interseção, podendo haver várias linhas em um mesmo ponto, e também facilita a seleção deste ponto para apresentação dos atributos que ali passam.
- **Controles 3D:** é possível, para os métodos de gráfico de dispersão geral e por grupos, fazer rotações, *zoom in* e *zoom out* aos dados representados em tela, para melhor compreensão dos mesmos.

- **Seleção de Pontos:** é possível selecionar um ponto em tela para apresentar-se que padrão é exatamente este e apresentar os atributos que este é composto. Este método é capaz de apresentar vários padrões que estão sobre-escritos por outro padrão (deficiência do gráfico de dispersão). Para as coordenadas paralelas e coordenadas paralelas circulares a seleção de um ponto acarreta na apresentação de todos os padrões que ali se intersectam.

A Figura 4.16 apresenta um exemplo com o método de coordenadas paralelas, utilizando a base de dados Íris, com a seleção de um ponto com várias interseções em que as linhas que passam por este ponto são destacadas na cor preta. Para os métodos de dispersão geral e por grupos os pontos escolhidos são destacados na cor preta.

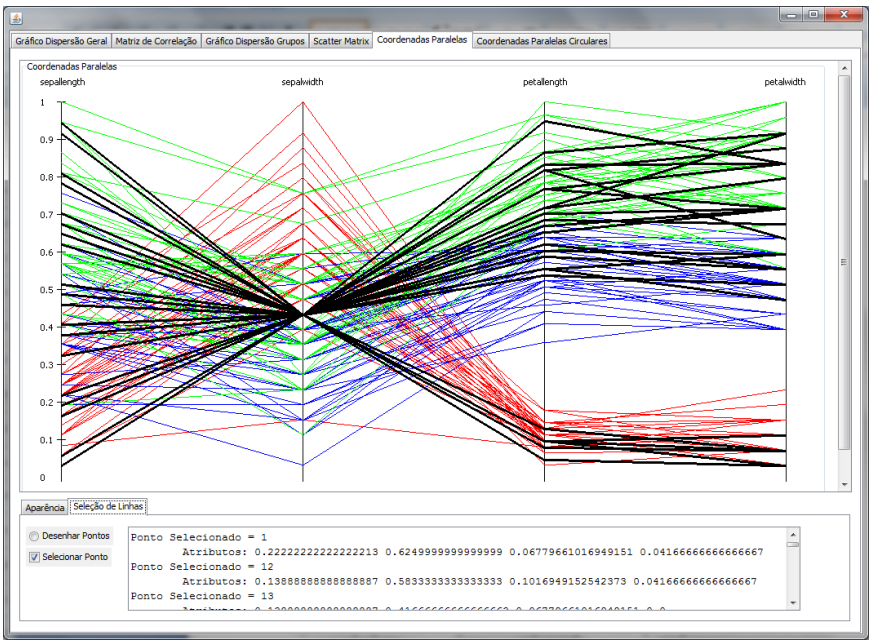


Figura 4.16: Exemplo de utilização de método de interação

Com a utilização da ferramenta Prefuse para a representação gráfica do dendrograma também foi possível utilizar-se dos métodos de iteração nela contida:

- *Zoom In/Out*;
- Mover o dendrograma;
- Selecionar nós, apresentando suas ocorrências nos níveis mais baixos da árvore e;
- Busca por nós.

4.2 Considerações Finais

Este capítulo teve como objetivo apresentar o módulo de agrupamento de dados da YADMT desenvolvido, apresentado de maneira breve e geral a ferramenta. Mais especificamente, apresentou os métodos de agrupamento de dados e de visualização de dados.

Para cada método disponível para execução é necessário a aquisição de uma base de dados, feito isso é possível de se utilizar a ferramenta com todos os seus módulos descritos. O módulo de agrupamento de dados possui a particularidade de todos os métodos nela contida serem mutuamente excludentes, não dependendo um dos outros para sua execução. Também apresenta a característica de ligação dos métodos de agrupamento de dados com os métodos de visualização de dados, em que, com exceção do dendrograma, todos os métodos resultados de agrupamento podem ser visualizados.

Ao término das implementações obteve-se um módulo de agrupamento de dados bastante robusto, oferecendo métodos de agrupamento já consagrados, como o *k-means*, e métodos de visualização de dados, que são uma ferramenta a mais para interpretação e extração de conhecimento sobre o resultado de um agrupamento de dados.

Capítulo 5

Resultados e Discussão

Atualmente há inúmeras ferramentas que possibilitam a execução dos passos do processo de *KDD*. A YADMT vem com uma proposta diferente, em que possibilita além do desenvolvimento das etapas do processo *KDD*, o acoplamento de novos módulos. Este trabalho teve como propósito desenvolver mais um módulo, o módulo de agrupamento de dados, juntamente com métodos de visualização para estes métodos de agrupamento.

A YADMT possui um total de quatro métodos de agrupamento de dados, sendo eles: Colônia de Formigas, *k-means*, Mapas Auto-organizáveis - *SOM* (FAINO, 2013) e métodos Hierárquicos (*single-linkage*, *complete-linkage*, *average-linkage* e *ward*), e um total de 10 métodos de visualização de dados. Possui também métodos para a recuperação do agrupamento gerado e apresentação de índices de qualidade. Esta recuperação de agrupamento na YADMT pode ser feita tanto pelos métodos disponíveis para tanto como também por métodos de visualização de dados, que nas ferramentas avaliadas pode ser feito somente através dos métodos de visualização, com exceção da ferramenta Weka, que apresenta alguns índices de qualidade de agrupamento.

Os possíveis fluxos de execução, desde o início até o final de um método de agrupamento de dados juntamente com um método de visualização de dados é ilustrado pelo fluxograma apresentado na Figura 5.1.

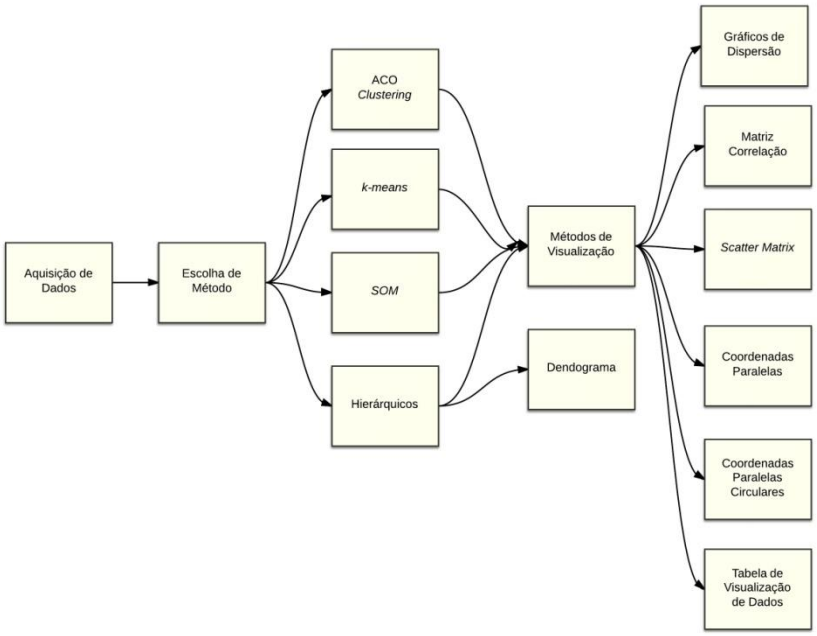


Figura 5.1: Fluxograma de execução do módulo de Agrupamento de Dados da YADMT

Pelo fluxograma é possível ver que todos os métodos de agrupamento tem o mesmo fluxo de execução. Parte da aquisição de dados, a escolha do método de agrupamento de dados e por fim a escolha do método de visualização de dados. Também é possível de se ver que os métodos não hierárquicos não tem acesso ao método que gera a visualização do dendrograma.

Apesar de anteriormente ter sido enumerado 10 métodos de visualização que constituem o módulo de agrupamento de dados, o fluxograma apresenta somente sete, pois métodos como gráfico de dispersão geral e para grupos foram classificados no mesmo grupo.

Atualmente há inúmeras ferramentas de mineração de dados disponíveis, sejam proprietárias ou *Open Source*, porém, para cada tipo de problema e base de dados que deseja-se extrair conhecimento há um método que se faz mais eficiente para tanto. Com isso, esta seção tem como objetivo avaliar ferramentas de mineração de dados, para identificar os métodos de agrupamento de dados e também visualização de dados, para ao final comparar os métodos de visualização de dados presentes nestas ferramentas com os métodos que a YADMT disponibiliza.

Na Tabela 5.1 são apresentadas as ferramentas de mineração de dados escolhidas para a avaliação e também a avaliação preliminar destas para a comparação com a YADMT. A pesquisa foi realizada através de leitura das documentações das ferramentas, execuções sobre

as mesmas e também por meio de tutorias disponibilizados pelos desenvolvedores. As ferramentas analisadas foram:

- Knime (KNIME, 2013);
- Orange Canvas (CANVAS, 2013);
- Tanagra (TANAGRA, 2003);
- RapidMiner (RAPIDMINER, 2013);
- Weka (WEKA, 2002).

É importante ressaltar que na avaliação e contagem dos métodos das ferramentas selecionados foram considerados somente métodos considerados bases, como o *k-means* ou seja, suas variantes, como o *fast k-means*, presente em algumas ferramentas não foram contadas como mais um método de agrupamento de dados presente na ferramenta. Para os métodos hierárquicos não foram considerados cada ligação como um método e sim a presença da função de agrupamento por um método hierárquico, já que as ligações hierárquicas são consideras, em todas as ferramentas, um parâmetro de execução do método.

Com esta avaliação foi possível levantar o número de métodos de agrupamento de dados e visualização de dados das principais ferramentas de mineração de dados disponíveis hoje na literatura. Com isso, é possível comparar a YADMT, em desenvolvimento, com estas ferramentas.

Tabela 5.1: Ferramentas de mineração de dados escolhidas para avaliação

Ferramenta	**Características**	
Knime	**Descrição**: é uma plataforma de trabalho gráfica para a análise de dados, prove acesso a dados, transformação de dados e análises preditivas e visuais.	
	URL: http://www.knime.org/	
	Versão: 2.7.2	**Livre**: (x) sim () não
	Métodos de Agrupamento de Dados:	
	Quantidade: 5	
	Métodos	
	Fuzzy c-Means, Hierárquicos (*single-linkage, complete-linkage e average-linkage*), *SOTA* – (*Self Organizing Tree Algorithm*) *Learner e Predictor e k-means.*	
	Métodos de Visualização de Dados:	
	Quantidade: 8	
	Métodos	
	Box Plot, Histograma, *Lift Chart*, *Line Plot*, Coordenadas Paralelas, Gráfico de Pizza, Matriz de *Scatter Plot* e Gráfico de Dispersão.	

<table>
<tr><td rowspan="10">Orange Canvas</td><td colspan="2">Descrição: aplicativo Open Source de análise de dados e visualização de dados, voltado para analistas novatos e especialistas. Prove mineração de dados através de métodos visuais.</td></tr>
<tr><td colspan="2">URL: http://orange.biolab.si/</td></tr>
<tr><td>Versão: 2.6.1</td><td>Livre: (x) sim () não</td></tr>
<tr><td colspan="2">Métodos de Agrupamento de Dados:</td></tr>
<tr><td colspan="2">Quantidade: 2</td></tr>
<tr><td colspan="2">Métodos</td></tr>
<tr><td colspan="2">k-means, Hierárquicos (single-linkage, complete-linkage, average-linkage e ward)</td></tr>
<tr><td colspan="2">Métodos de Visualização de Dados:</td></tr>
<tr><td colspan="2">Quantidade: 12</td></tr>
<tr><td colspan="2">Métodos</td></tr>
<tr><td colspan="2">Distribuição de Frequências, Box Plot, Dispersão Geral, Projeção Linear, Radviz, Polyviz, Coordenadas Paralelas, Survey Plot, Análise Correspondente, Mosaic Display, Sieve Diagram, Sieve Multigram.</td></tr>
<tr><td rowspan="10">Tanagra</td><td colspan="2">Descrição: é um software Open Source voltado para pesquisas e ensino na área de mineração de dados e estatística. Apresenta suporte a várias tarefas de mineração de dados, como: Visualização, agrupamento e classificação.</td></tr>
<tr><td colspan="2">URL: http://eric.univ-lyon2.fr/~ricco/tanagra/</td></tr>
<tr><td>Versão: 1.4.48</td><td>Livre: (x) sim () não</td></tr>
<tr><td colspan="2">Métodos de Agrupamento de Dados:</td></tr>
<tr><td colspan="2">Quantidade: 7</td></tr>
<tr><td colspan="2">Métodos</td></tr>
<tr><td colspan="2">Clustering Tree, Expectation-Maximization Clustering, Hierárquicos (Ward), k-means, SOM-Kohonem, Kohonem's Learning Vector Quantization, Neighborhood Graph</td></tr>
<tr><td colspan="2">Métodos de Visualização de Dados:</td></tr>
<tr><td colspan="2">Quantidade: 2</td></tr>
<tr><td colspan="2">Métodos</td></tr>
<tr><td colspan="2">Dispersão de Correlação e Dispersão Geral</td></tr>
</table>

<table>
<tr><td rowspan="9">RapidMiner</td><td colspan="2">Descrição: é um ambiente para aprendizado de máquina, mineração de dados, mineração de texto, análise preditiva, e Business Analytics. É utilizado para pesquisa, educação, treinamento, prototipagem rápida, desenvolvimento de aplicativos e aplicações industriais.</td></tr>
<tr><td colspan="2">URL: http://rapidminer.com/products/rapidminer-studio/</td></tr>
<tr><td>Versão: 5.3.015</td><td>Livre: () sim (x) não</td></tr>
<tr><td colspan="2">Métodos de Agrupamento de Dados:</td></tr>
<tr><td colspan="2">Quantidade: 9</td></tr>
<tr><td colspan="2">Métodos</td></tr>
<tr><td colspan="2">k-means, k-medoids, DBSCAN, Expectation Maximization Clustering, Support Vector Clustering, Random Clustering, Hierárquicos (single-linkage, complete-linkage e average-linkage), Top Down Clustering e Flatten Clustering.</td></tr>
<tr><td colspan="2">Métodos de Visualização de Dados:</td></tr>
<tr><td colspan="2">Quantidade: 3</td></tr>
<tr><td colspan="2">Métodos</td></tr>
<tr><td colspan="2">Lift Chart, Curvas ROC – (Receiver Operating Characteristics) e Visualização de Modelo por SOM – (Self Organizing Maps)</td></tr>
<tr><td rowspan="9">Weka</td><td colspan="2">Descrição: é uma das mais famosas ferramentas de mineração de dados, desenvolvida pela Universidade de Waikato, Nova Zelândia, contém uma variedade de métodos para manipulação de dados, visualização e análise.</td></tr>
<tr><td colspan="2">URL: http://www.cs.waikato.ac.nz/ml/weka/</td></tr>
<tr><td>Versão: 3.7.8</td><td>Livre: (x)sim () não</td></tr>
<tr><td colspan="2">Métodos de Agrupamento de Dados:</td></tr>
<tr><td colspan="2">Quantidade: 7</td></tr>
<tr><td colspan="2">Métodos</td></tr>
<tr><td colspan="2">Cobweb, Expectation Maximization, Fatherst First, Filtered Clusterer, Hierárquicos (single-linkage, complete-linkage, average-linkage, mean-linkage, centroid-linkage, ward, adjcomplete e neighbor-joining), Make Density Based Clusterer, k-means.</td></tr>
<tr><td colspan="2">Métodos de Visualização de Dados:</td></tr>
<tr><td colspan="2">Quantidade: 4</td></tr>
<tr><td colspan="2">Métodos</td></tr>
<tr><td colspan="2">Histograma, Gráfico de Dispersão, Matriz Scatter Matrix e Visualização de Árvore.</td></tr>
</table>

YADMT	**Descrição**: é uma ferramenta para aplicação das etapas do processo *KDD*. Atualmente esta sendo desenvolvida na Unioeste e tem como característica ser modular para que qualquer desenvolvedor possa contribuir para o seu desenvolvimento.	
	URL: http://www.inf.unioeste.br/~clodis/yadmt	
	Versão: v0	**Livre**: (x)sim () não
	Métodos de Agrupamento de Dados:	
	Quantidade: 4	
	Algoritmo de Agrupamento de Dados baseado em Colônia de Formigas, *k-means*, Hierárquicos, *SOM.*	
	Métodos de Visualização de Dados:	
	Quantidade: 10	
	Métodos	
	Gráfico de Dispersão Geral 2D/3D, Gráfico de Dispersão para Grupos 2D/3D, Matriz de Correlação, *Scatter Matrix*, Coordenadas Paralelas, Coordenadas Paralelas Circulares, Tabela de Visualização de Dados e Dendrograma	

5.1 Testes de Execução dos Métodos de Agrupamento

Nesta seção serão apresentados os resultados das execuções dos métodos de agrupamento de dados contidos no módulo da Ferramenta YADMT. Serão apresentados os resultados textuais dos métodos para as bases de dados apresentadas na Tabela 5.2, que podem ser obtidas no repositório *UCI* (FRANK, ASUNCION, 2010).

Tabela 5.2: Base de dados escolhidas para testes

Base de Dados	Número de Registros	Número de Atributos	Número de Classes
Dermatology	366	34	6
Íris	150	4	3
Libras Movement	360	90	15
Pima	768	8	2
Vehicle	946	18	4

Estas bases de dados foram escolhidas pela variação do número de atributos, registros e classes, para que os métodos de visualizações possam ser testados e avaliados corretamente. A base de dados *Dermatology* originalmente possui atributos faltantes em sua composição, considerando que os métodos implementados na YADMT não podem ser executados com esse

tipo de atributo foi usado o método de filtragem disponível na WEKA para substituir estes valores, e assim tornar possível sua execução na YADMT.

A YADMT apresenta seus resultados de duas maneiras, textualmente em que apresenta detalhes como nome da base de dados utilizada, o número de instâncias (padrões) que a compõem, suas classes e os parâmetros utilizados na execução do método. Também apresenta os índices de avaliação de agrupamento: Medida (F), Índice Aleatório (R) (índices externos), porcentagem de acerto e a Variância Intra-Grupos (V) (índice interno), para os três primeiros índices quanto mais perto de um o valor melhor é o agrupamento. É possível também visualizar a distribuição dos padrões em cada grupo formado, e também a matriz confusão que apresenta a distribuição dos padrões em cada grupo, corretamente e erroneamente. Tanto a porcentagem de acerto como a matriz confusão são construídos com base no grupo que um determinado padrão ficou e qual a classe real deste padrão. A Figura 5.2 ilustra um exemplo desta tela da YADMT.

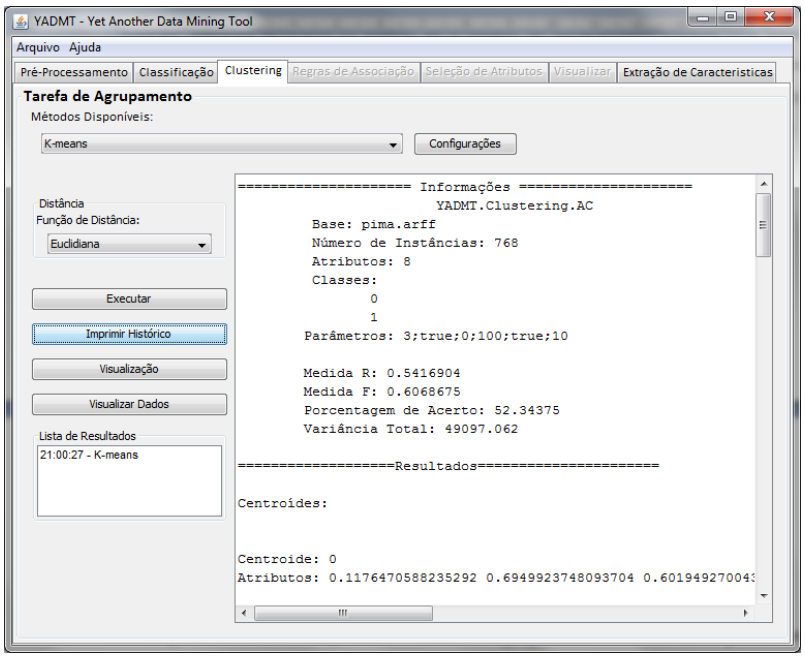


Figura 5.2: Tela de saída de resultado textual da YADMT

5.1.1 Execuções dos Métodos de Agrupamento de Dados com as Bases de Dados Escolhidas

Os parâmetros de execuções dos métodos de agrupamento de dados são os parâmetros padrões já pré-definidos na ferramenta, mostrados no Capítulo 4, utilizando-se como medida de distância a medida euclidiana. Para o método de agrupamento de dados baseado em colônia de formigas foi utilizado o método ligação completa (*complete-linkage*) como recuperação de grupos. Para os métodos hierárquicos foi escolhida a ligação simples (*simple-linkage*) como ligação para efetuar o agrupamento.

- **Base de Dados *Dermatology* (Tabela 5.3)**

Tabela 5.3: Resultados para base de dados *Dermatology* – *k-means*

Dermatology	**Colônia de Formigas**	***k-means***	***single-linkage***
Medida *F*	0,38	**0,68**	0,5
Medida *R*	0,74	**0,8**	0,52
Porcentagem Acerto (%)	32,24	**47,27**	30,87

- **Base de Dados Íris (Tabela 5.4)**

Tabela 5.4: Resultados para base de dados Íris – *k-means*

Íris	**Colônia de Formigas**	***k-means***	***single-linkage***
Medida *F*	0,61	**0,89**	0,5
Medida *R*	0,67	**0,87**	0,34
Porcentagem Acerto (%)	60,67	**88,67**	34

- **Base de Dados Libras *Movement* (Tabela 5.5)**

Tabela 5.5: Resultados para base de dados Libras *Movement* – *k-means*

Libras *Movement*	**Colônia de Formigas**	***k-means***	***single-linkage***
Medida *F*	0,19	**0,23**	0,12
Medida *R*	**0,82**	0,68	0,20
Porcentagem Acerto (%)	**15,28**	4,17	6,94

- **Base de Dados Pima (Tabela 5.6)**

Tabela 5.6: Resultados para base de dados Pima – *k-means*

Pima	**Colônia de Formigas**	***k-means***	***single-linkage***
Medida *F*	0,57	0,61	**0,69**
Medida *R*	0,5	0,54	**0,55**
Porcentagem Acerto (%)	49,9	52,60	**65,1**

- **Base de Dados *Vehicle* (Tabela 5.7)**

Tabela 5.7: Resultados para base de dados *Vehicle* – *k-means*

Vehicle	**Colônia de Formigas**	***k-means***	***single-linkage***
Medida *F*	0,312	**0,462**	0,4
Medida *R*	**0,613**	0,525	0,26
Porcentagem Acerto (%)	28,49	**32,62**	25,53

Como visto pelas tabelas, nenhum dos métodos de agrupamento apresentou resultados satisfatórios no ponto de vista de qualidade de agrupamento, com exceção do método *k-means* para a base de dados Íris. Este fato se dá, inicialmente, os parâmetros usados pelos métodos não são os ideais, podendo melhorar a qualidade do agrupamento com a mudança nos parâmetros. Segundo fato, a medida de distância usada para as bases de dados também pode não ter sido a ideal, também podendo melhor a qualidade utilizando-se de outra medida de distância.

Um fator que dificulta encontrar um resultado satisfatório para um determinado agrupamento e base de dados é a escolha desses parâmetros de execução, e não há uma heurística definida para a escolha dos mesmos sendo que esta escolha é feita em cima de repetidas execuções com diferentes valores de parâmetros para cada base de dados e método escolhido.

5.2 Testes de Execução Comparativa com outras ferramentas

Esta seção tem como principal objetivo expor os testes comparativos feitos com a YADMT, mais especificamente do módulo de agrupamento de dados, com as ferramentas avaliadas no início deste capítulo. Os testes comparativos foram feitos utilizando o método de agrupamento de dados *k-means* e a base de dados Pima, e terão como objetivo apresentar os métodos de visualizações das outras ferramentas avaliadas para posterior avaliação em relação aos constantes na YADMT.

O método *k-means* foi escolhido para estes testes comparativos, pois é o único presente em todas as ferramentas avaliadas. Com exceção do parâmetro "*k*", que define o número de grupos a ser formado pelo método, todos os outros parâmetros serão executados com os valores pré-definidos em todas as ferramentas. O valor de "*k*" será dois, pois este valor retorna o melhor agrupamento formado e foi definido após algumas execuções com diferentes valores deste parâmetro.

As técnicas de visualização mostradas são aquelas que permitem a visualização, interpretação e extração de informação de um grupo formado pelo método *k-means*, sendo também apresentando os métodos mais relevantes de cada ferramenta e que podem ser aplicados a um método de agrupamento não hierárquico.

5.2.1 Testes com a Ferramenta YADMT

A YADMT possui duas formas de retorno de resultado do agrupamento de dados, a primeira é o retorno textual (Figura 5.3), contendo o os índices citados na Seção 5,1 e que para o método *k-means* acrescenta-se os centroides formados. A segunda forma são os métodos de visualização implementados neste trabalho, mais especificamente o método matriz de correlação, que apresenta a correlação de um determinado grupo, o gráfico de dispersão por grupos em 2D e 3D em que se pode escolher qual grupo que será representado em tela, a tabela de visualização que mostra todos os dados da base de dados incluindo a qual grupos cada padrão pertence e o dendrograma, exclusivo para os métodos hierárquicos.

Os métodos de gráfico de dispersão geral 2D e 3D, coordenadas paralelas, coordenadas paralelas circulares e matriz de *scatter matrix* são métodos para a visualização da base de

dados em geral, mas que podem ser adaptados para a visualização dos grupos, como no gráfico de dispersão de grupos.

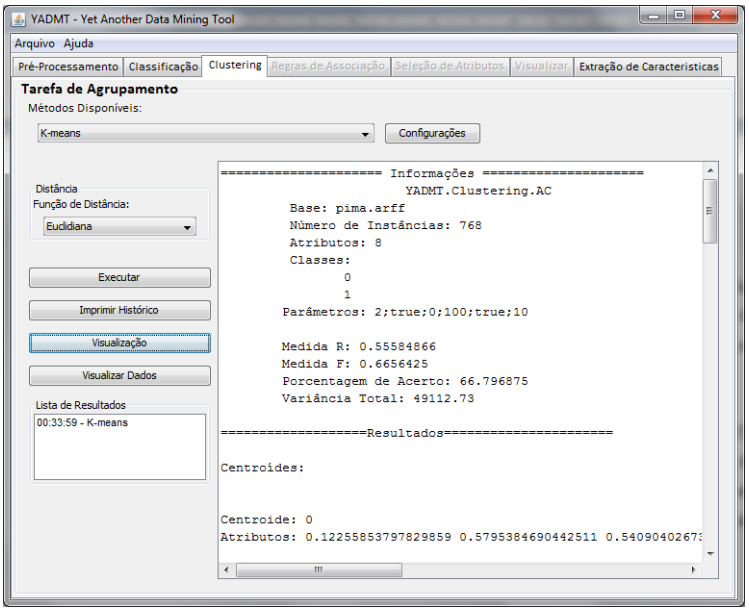


Figura 5.3: Resultado método *k-means* para a base de dados Pima

A seguir, serão apresentados os métodos de visualização para a base de dados. A Figura 5.3 apresenta o método de gráfico de dispersão geral 2D, no qual é possível visualizar os atributos da base de dados em que estes não apresentam uma separação linear entre as classes.

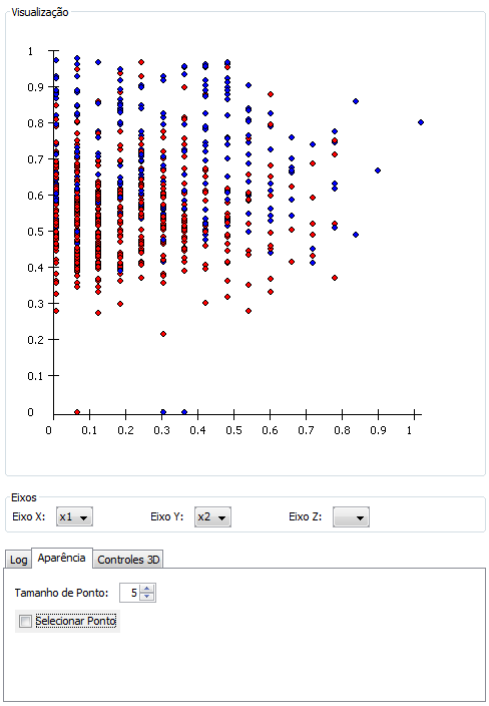


Figura 5. 4: Representação da base de dados Pima pelo método de gráfico de dispersão geral – atributos “x1” e “x2”

A Figura 5.5 ilustra a técnica de matriz de *scatter matrix*, que mostra a correlação entre os atributos da base de dados apresentando uma uniformidade destas e também uma correlação baixa de atributo por atributo. Devido à alta dimensionalidade da base de dados a técnica não pode ser ilustrada por inteiro nessa figura.

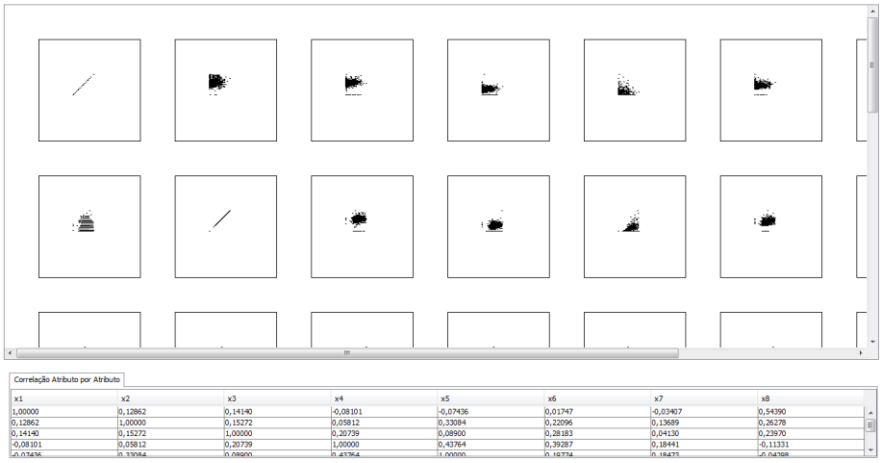


Figura 5.5: Matriz *scatter matrix* para a base de dados Pima

A Figura 5.6 apresenta a técnica de coordenadas paralelas que assim como a de dispersão geral também apresenta os dados não separados linearmente. A base de dados Pima apresenta linhas distantes no eixo X e que apresentam bastante cruzamento entre si, o que demonstra uma baixa relação entre as linhas.

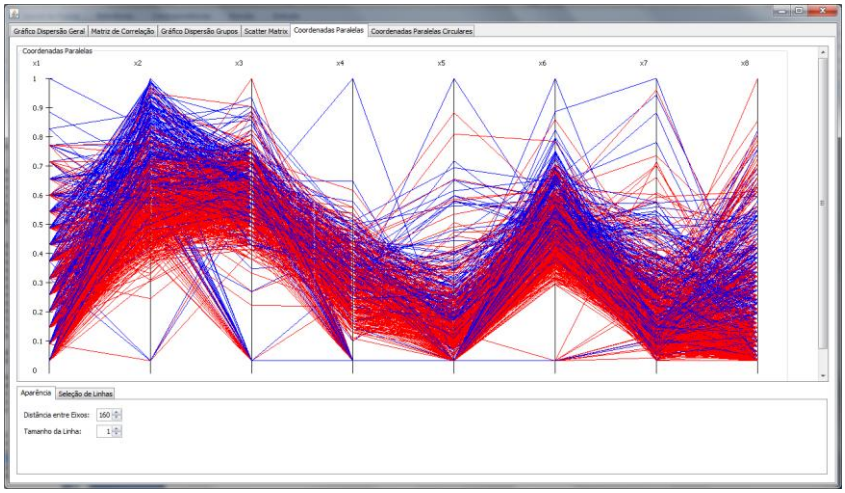


Figura 5.6: Representação da técnica coordenadas paralelas para a base de dados Pima

A Figura 5.7 apresenta a técnica de coordenadas paralelas circulares aplicada à base de dados em questão. A técnica possui as mesmas características da técnica de coordenadas

paralelas circulares com o diferencial de poder identificar valores de atributos maiores, que se encontram nas extremidades do círculo e valores de atributos menores que se encontram no centro do círculo. Com a visualização da Figura 5.7 pode-se perceber que para esta base de dados os valores dos atributos concentram-se no meio do raio do círculo.

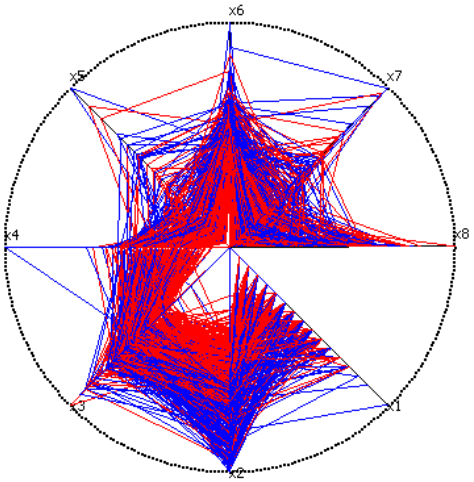


Figura 5.7: Representação da técnica coordenadas paralelas circulares para a base de dados Pima

Para a Figura 5.8 apresenta-se um dos grupos formado pelo método *k-means*, o grupo em questão é formado por 253 padrões da base de dados e a partir da dispersão destes padrões é possível visualizar que seus valores são aproximados, justificando o agrupamento como mesmo grupo.

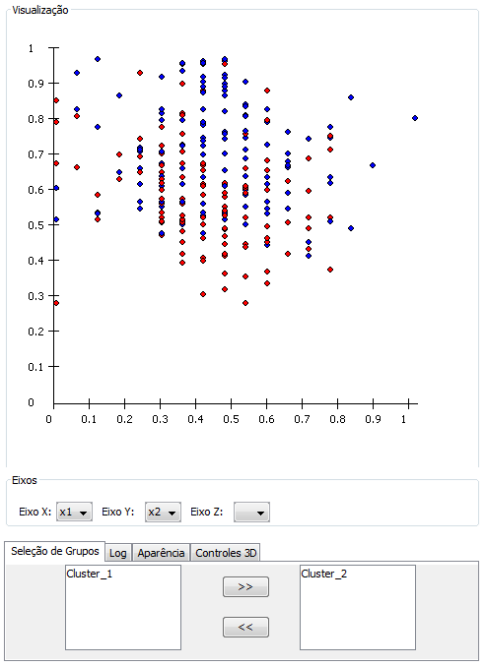


Figura 5.8: Representação da dispersão de grupo gerado pelo método *k-means*

A matriz de correlação (Figura 5.9) representa o mesmo grupo apresentado na Figura 5.8 e, através da cor uniforme apresentada pela matriz de correlação, que é predominantemente vermelha (alta correlação), é possível entender o motivo pelo qual houve o agrupamento de padrões entre as classes reais da base de dados.

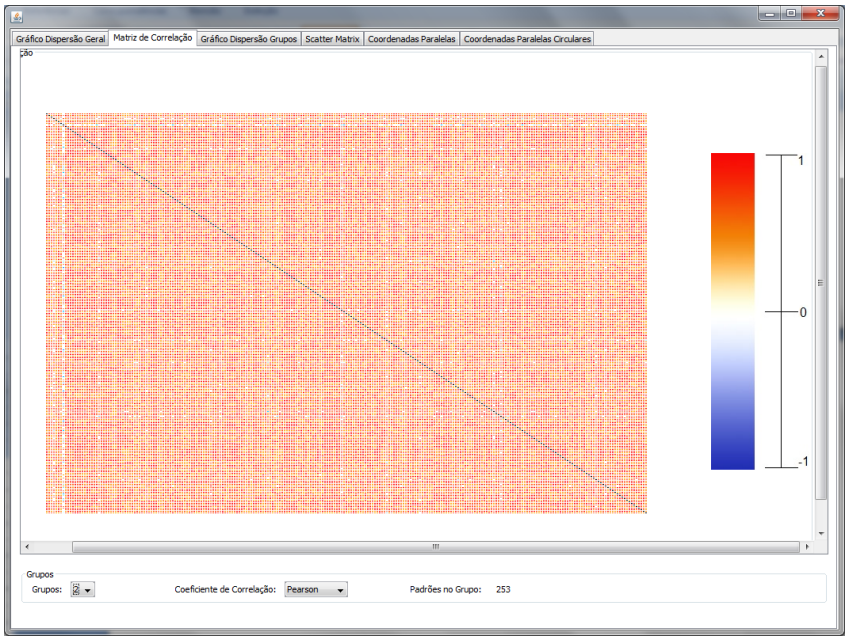


Figura 5.9: Representação de grupo gerado pelo método *k-means* pela matriz de correlação

5.2.2 Testes comparativos com a Ferramenta KNIME

A KNIME é uma ferramenta *Open Source* que possui métodos de agrupamento de dados e visualização como os apresentados na Tabela 5.10. Os parâmetros possíveis de serem ajustados da ferramenta, para o método *k-means* são as colunas que serão consideradas no método, o máximo de iterações e o número *k* de grupos para ser formado.

A KNIME traz como resultado textual somente a formação dos centroides, apresentando os valores que cada atributo foi composto, e por quantos padrões este centroide foi constituído. Este resultado textual é apresentado em uma segunda janela de visualização, ficando a parte do fluxo de execução principal da ferramenta, como ilustrado na Figura 5.10.

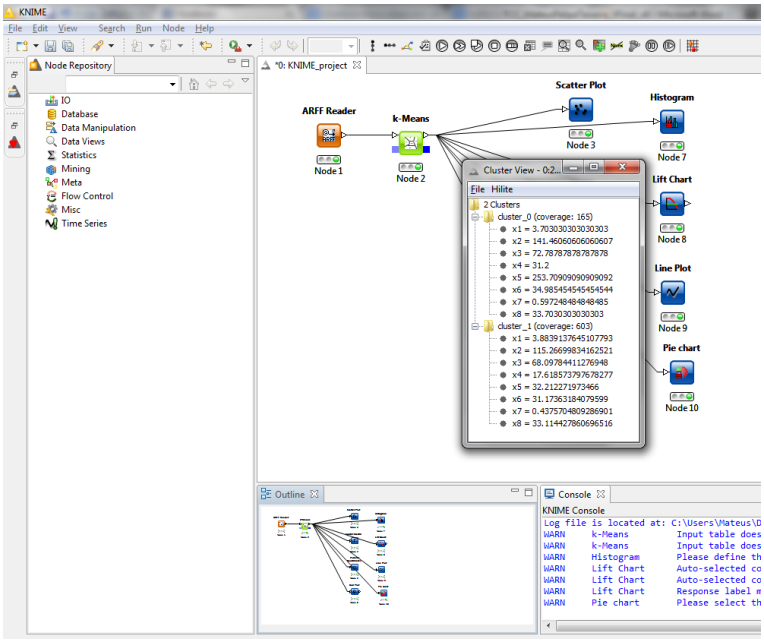


Figura 5.10: Tela principal - KNIME

O funcionamento da KNIME é todo baseado na ideia de adição de nodos de métodos a um fluxo de execução, no quanto esquerdo superior é possível visualizar os métodos disponível na ferramenta, no centro situa-se o fluxo de execução, com o fluxo montado para a execução deste teste e em destaque a tela de resultado textual do método da *k-means* da ferramenta.

O método de gráfico de dispersão da KNIME (Figura 5.11), denominada na ferramenta como *scatter plot*, possui o mesmo funcionamento básico do método da YADMT, possuindo também a mesma seleção de eixos cartesianos, tamanho de ponto, seleção de pontos e outros, como *zoom*.

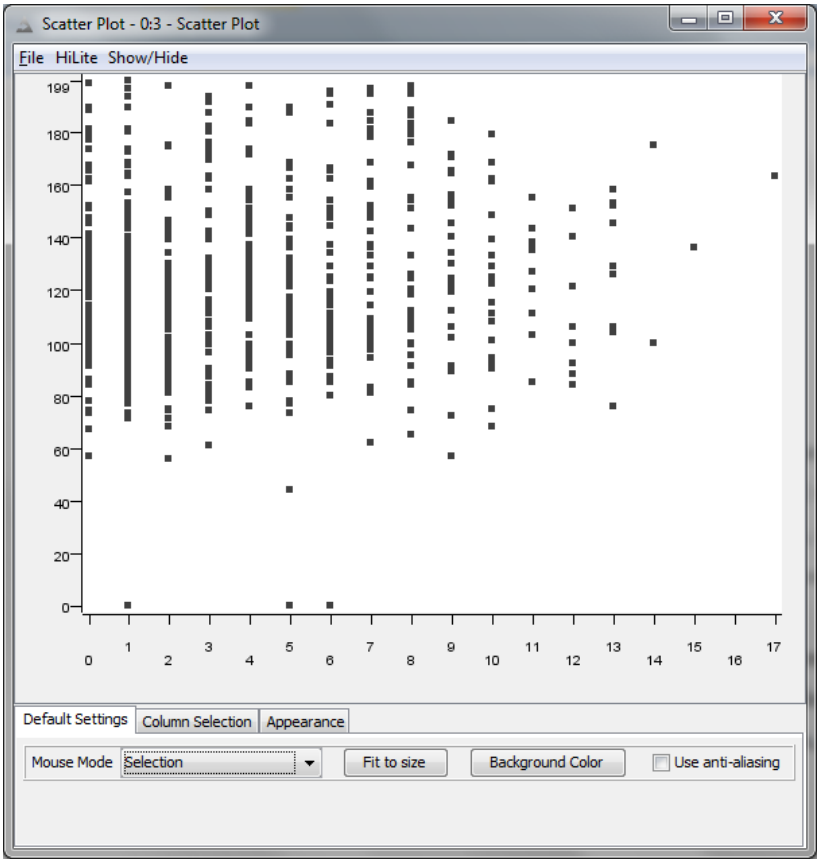


Figura 5.11: Gráfico de dispersão – KNIME

A matriz de *scatter matrix* da KNIME (Figura 5.11) também possui funcionamento semelhante ao da YADMT, com a principal diferença de que para essa ferramenta é possível selecionar as colunas que serão apresentadas par a par, acabando com o problema de dimensionalidade alta. Ainda, é possível redimensionar os dados para que todos possam ser exibidos em tela de uma só vez, porém quando a base de dados tiver muitas dimensões ainda sim terá o problema de dimensionalidade já que o redimensionamento feito será alto.

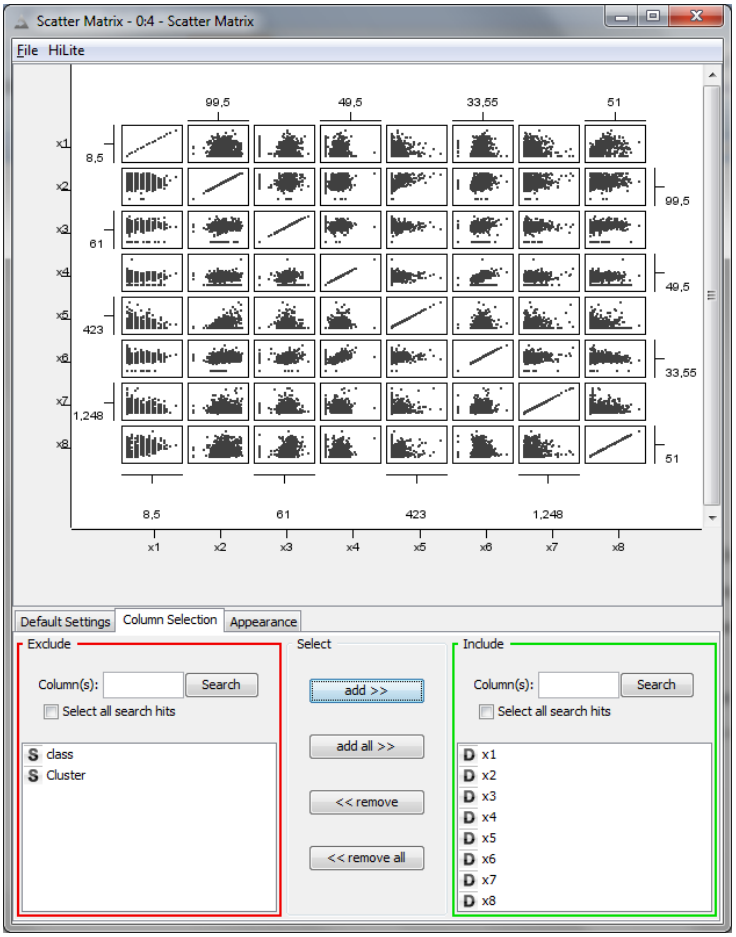


Figura 5.12: Matriz *scatter matrix* – KNIME

A técnica de coordenadas paralelas (Figura 5.13) da KNIME, não traz uma representação clara das classes da base de dados, apesar de que pode-se escolher quais colunas irão ser representadas pela técnicas para uma base de dados grande, muitos padrões, a representação das linhas fica muito densa, assim ficando difícil de distinguir os dados.

Para a KNIME a técnica possui um mecanismo de seleção das colunas, eixos verticais, que serão representados, assim como seleção dos pontos de interseção que também são implementados na YADMT. A principal diferença das duas ferramentas é a escolha dos eixos verticais em que para a KNIME é possível a visualização dos grupos.

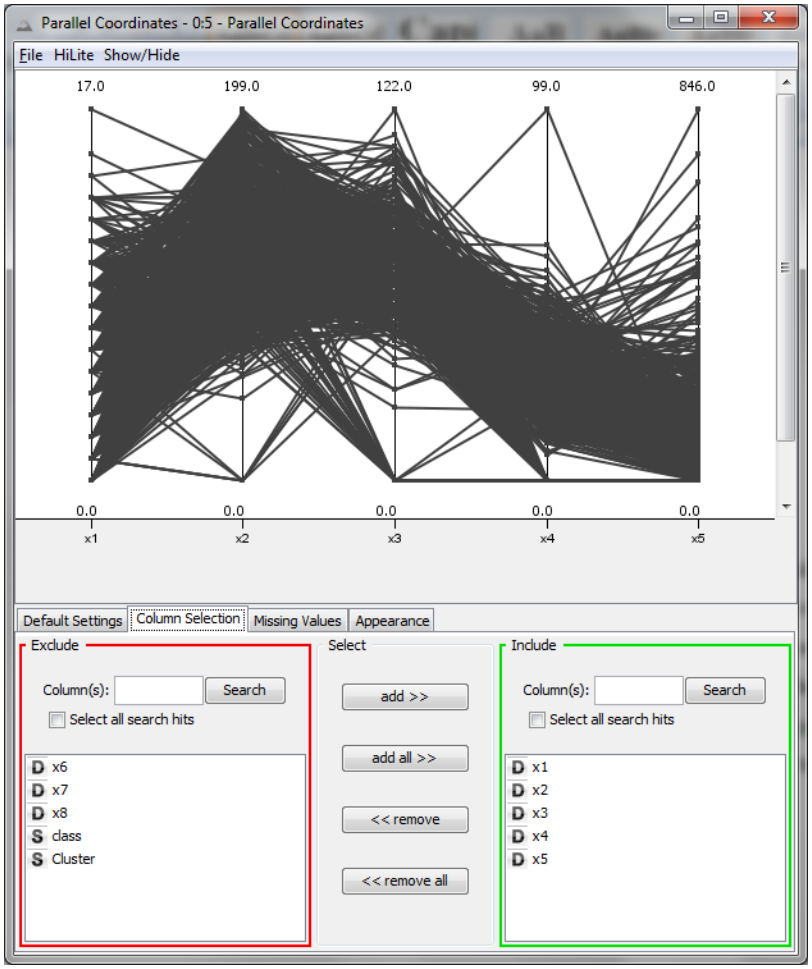


Figura 5.13: Coordenadas paralelas – KNIME

A ferramenta possui a técnica de histograma (Figura 5.14), que agrupa os dados de uma maneira uniforme para representação de acordo com algum critério, na KNIME o histograma representa os grupos formados pelo método de agrupamento e pode representa a média de todos os padrões dentro de um grupo (por atributo), a soma de todos os padrões de um grupo (por atributo), a contagem de padrões completa e também a contagem de padrões que possuem valores de atributos faltantes.

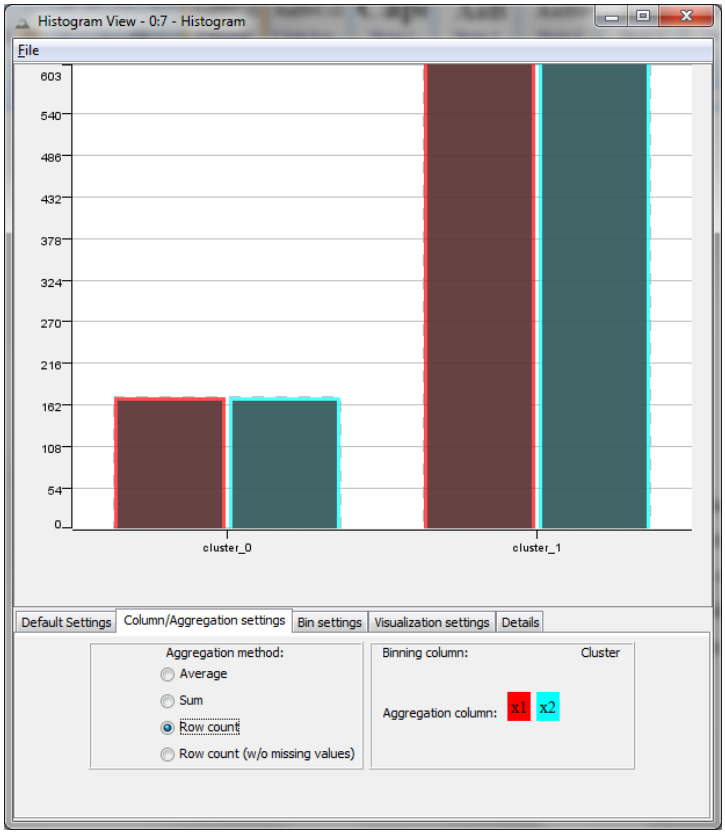


Figura 5.14: Histograma – KNIME

Com funcionalidade parecida, porém com a distribuição gráfica diferente, a técnica de *pie chart*, ou gráfico de pizza, (Figura 5.15) apresenta os mesmos dados que o histograma, diferenciando-se na representação e na coluna de agregação que somente é possível uma por vez. No exemplo da Figura 5.15 são ilustradas as mesmas informações da Figura 5.13, porém, somente para a coluna de atributos “x1” da base de dados.

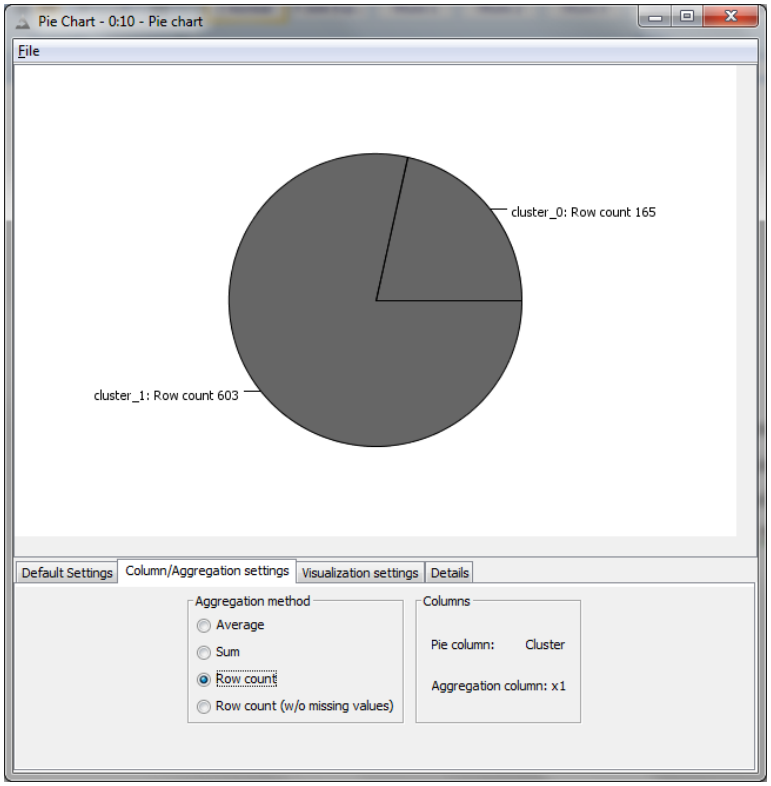


Figura 5.15: Gráfico de Pizza – KNIME

A KNIME é uma proposta de ferramenta para o uso na mineração de dados, estatística e outras áreas. Possui diversos métodos que possibilitam uma extração de conhecimento completa de uma determinada base de dados. O seu fluxo de execução é bastante intuitivo, pelo uso do sistema de nodos, em que cada nodo executa alguma funcionalidade. Porém, ainda possui algumas funcionalidades que não são explicitadas o que pode dificultar o seu uso, ou não tornar o uso da ferramenta ágil.

5.2.3 Testes comparativos com a Ferramenta ORANGE CANVAS

A ORANGE CANVAS é uma ferramenta *Open Source* de mineração de dados com enfoque para classificação de dados, regressão de dados e mineração visual de dados, mas também possuindo métodos de avaliação e associação de dados. Possui apenas dois métodos de agrupamento de dados, métodos hierárquicos e *k-means*, este último usado para as execuções desta seção.

Seus parâmetros de execução do método *k-means* são o número de grupos a serem formados, a medida de distância e a inicialização dos *seeds*. Destes, somente o parâmetro *k* foi modificado para o valor dois. A ferramenta não possui um retorno de resultados textual,

sendo que os resultados podem ser somente visualizados pelos métodos de visualização disponível na ferramenta.

O primeiro método disponível é o método de distribuição de frequência (Figura 5.16) que apresenta a distribuição dos padrões em cada grupo diferenciando por classes através de cores, cor azul para classe “0” e cor vermelha para classe “1”.

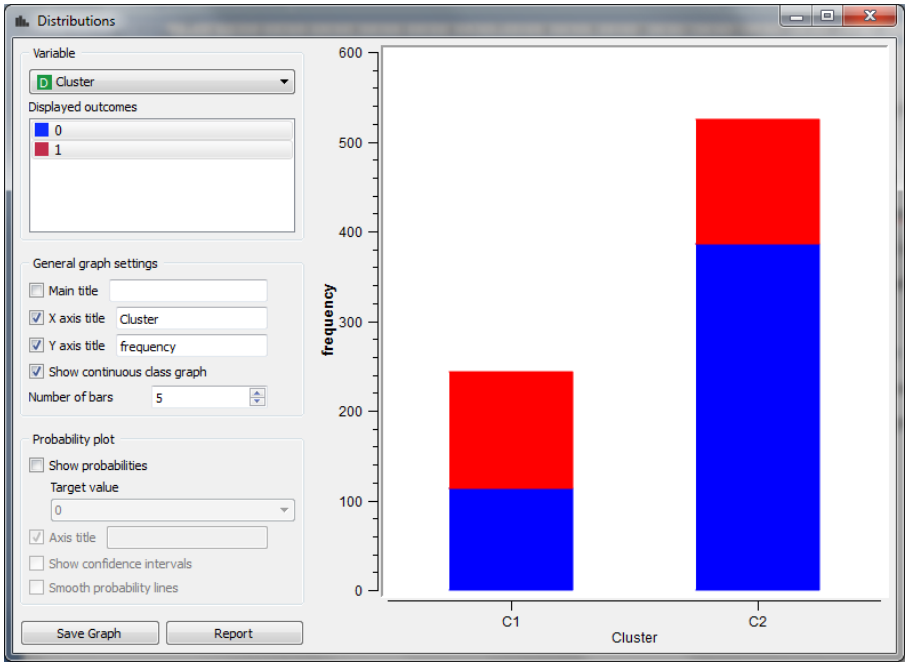


Figura 5.16: Distribuição de frequência – ORANGE CANVAS

A técnica de dispersão (Figura 5.17) da ORANGE CANVAS é bastante parecida a técnica implementada neste trabalho, uma vez que apresenta os padrões em função dos eixos cartesianos, e identifica os grupos aos quais estes padrões pertencem, diferenciando da YADMT onde todos os grupos são apresentados juntos.

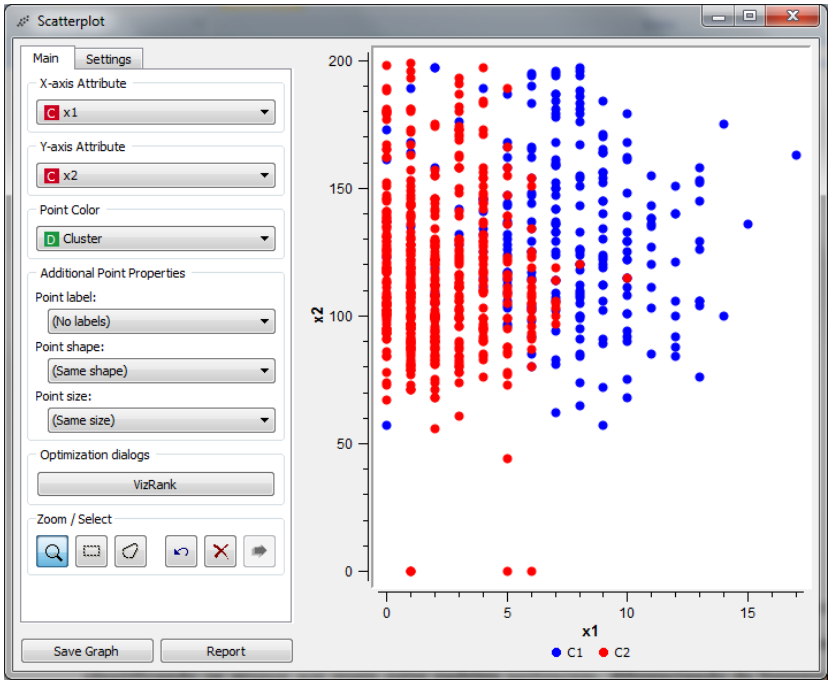


Figura 5.17: Gráfico de dispersão – ORANGE CANVAS

A técnica de coordenadas paralelas da ORANGE CANVAS (Figura 5.18) também é bastante similar à da YADMT e a KNIME, esta apresenta a coloração das linhas conforme seus grupos e aponta a qual classe determinada linha pertence (padrão) e também possui recursos para seleção de eixos verticais.

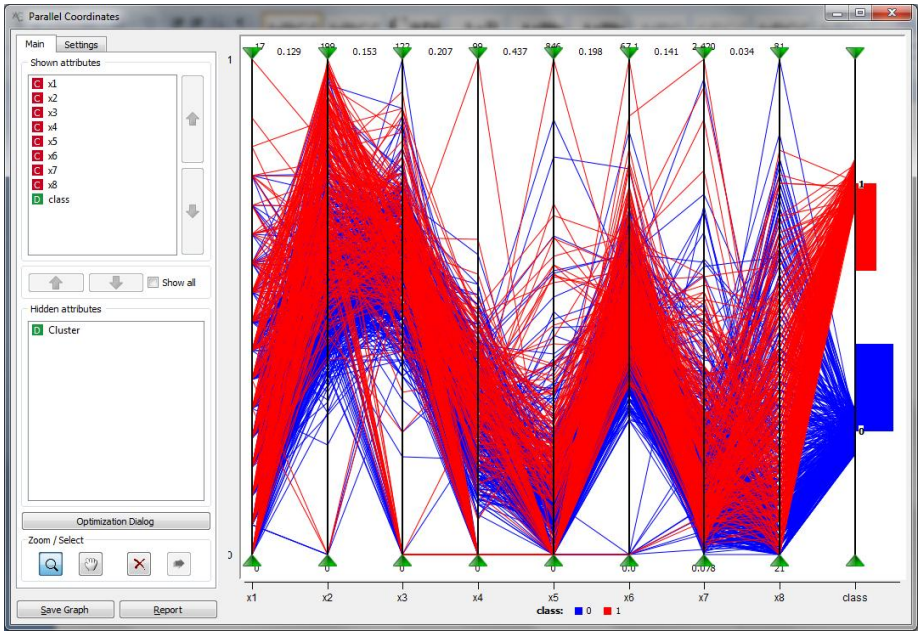


Figura 5.18: Coordenadas Paralelas – ORANGE CANVAS

A ORANGE CANVAS possui outros seis métodos de visualização que podem ser aplicados ao agrupamento, porém, não retornam resultados relevantes na extração de conhecimento sobre esse agrupamento gerado pelo método *k-means* e também sobre a base de dados.

O fluxo de execução da ferramenta também é bastante simples, e assim como a KNIME apresenta a estrutura de nodos em que cada nodo adicionado ao campo de fluxo irá executar uma determinada tarefa. Também apresenta um sistema de *feedback* para cada método, que retorna ao usuário os dados de entrada e de saída de cada método.

A Figura 5.19 apresenta um fluxo de execução da ferramenta em que, para este fluxo, será executado o método *k-means*, a partir da leitura de um arquivo *ARFF*, e após a execução do método será plotado em tela a base de dados pelo método de gráfico de dispersão.

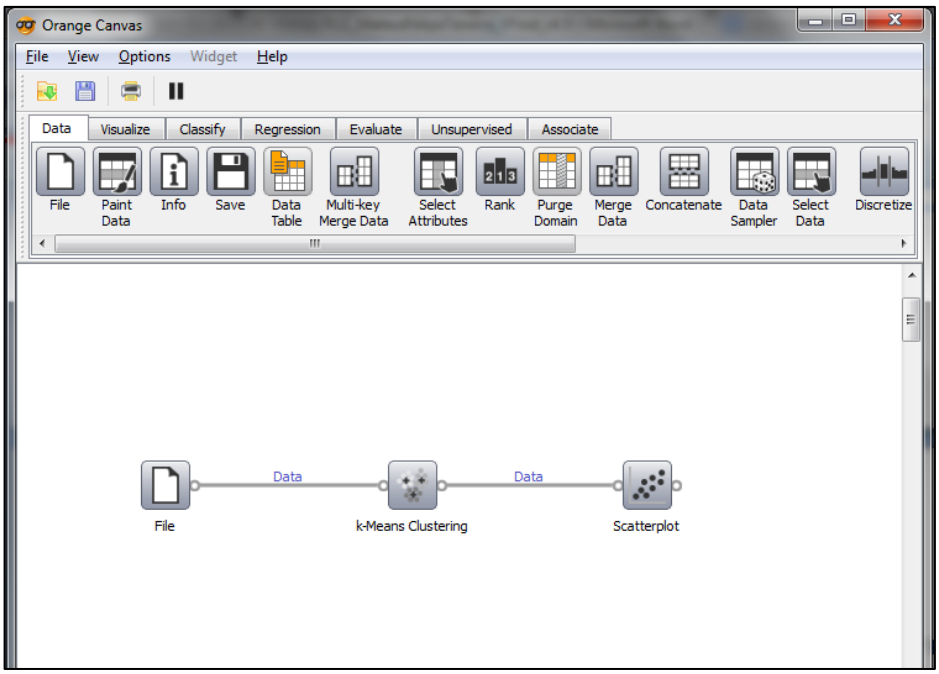


Figura 5.19: Tela principal da ORANGE CANVAS

5.2.3 Testes comparativos com a Ferramenta RAPIDMINER STUDIO

A RAPIDMINER STUDIO é uma ferramenta proprietária, que possui uma versão de testes, de mineração de dados também voltada à estatística, banco de dados e processos de análises em dados. Tem como o principal foco disponibilizar um ambiente de trabalho totalmente gráfico, com a presença de elementos gráficos que significam uma operação em

questão, por exemplo, um método de mineração de dados. Possui nove métodos de agrupamento de dados, sendo o *k-means* o mais conhecido.

O método *k-means* da ferramenta pode ser ajustado de acordo com seis parâmetros sendo eles: número de grupos, máximo de iteração, boa inicialização dos *seeds*, tipo da medida de distância (numérica, nominal e mista), a medida de distância (de acordo com o tipo) e o número máximo de passos de otimização.

O retorno para o método na ferramenta é textual, apresentando o número de grupos formados e por quantos padrões estes são formados, e também uma tabela de dados semelhante a apresentada na Seção 3.2.7. A ferramenta não possui métodos de visualização aplicáveis a agrupamento de dados.

A ferramenta possui um fluxo de execução bastante intuitivo, seguindo a ideia de fluxo de execução baseada em nodos que representam um determinado processo. A Figura 5.20 ilustra este processo para a execução do método *k-means* utilizando-se da base de dados Pima. Para este fluxo tem-se o leitor de *ARFF*, um conversor de valores nominais para numéricos, por exigência do método de agrupamento, e o método de *k-means*.

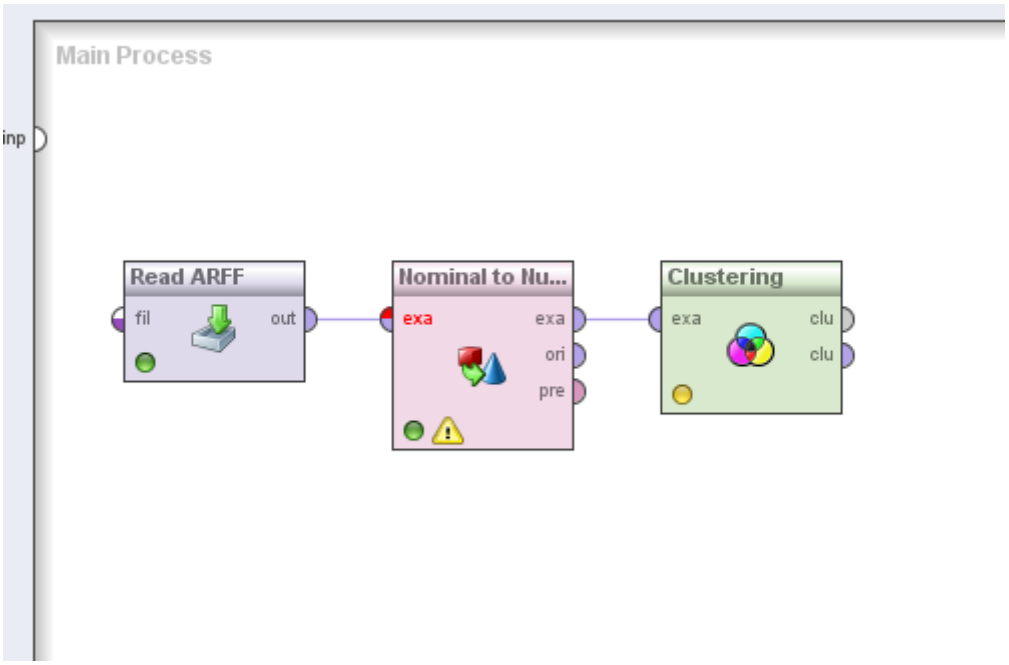


Figura 5.20: Fluxo de Execução RapidMiner Studio

5.2.4 Testes comparativos com a Ferramenta TANAGRA

A TANAGRA é uma ferramenta *Open Source* que possui métodos de mineração de dados, porém com um enfoque maior a métodos de estatística. Possui sete métodos de agrupamentos de dados sendo o *k-means*, *ward* e *SOM* os mais utilizados, também existentes na YADMT.

O método *k-means* da TANAGRA pode ser ajustado por meio de seis parâmetros: número de grupos, máximo de iterações, número de tentativas (a ferramenta não deixa clara o que este parâmetro faz), normalização da distância, computação de médias, e inicialização dos *seeds*. Como já dito, o único parâmetro alterado foi o parâmetro número de grupos.

O retorno do método *k-means* para a TANAGRA é textual apresentando algumas informações sobre o agrupamento realizado. Inicialmente apresenta os parâmetros utilizados para execução do método, a avaliação global da execução, o tamanho de cada grupo formado e os centroides finais para cada grupo.

O único método de visualização aplicável a agrupamento da TANAGRA é o gráfico de dispersão geral. O funcionamento deste é semelhante a todos as outras ferramentas, em especial com a ORANGE CANVAS, porém, os rótulos utilizados para identificação de cada padrão em cada grupo dificulta a visualização apropriada dos grupos formados e da dispersão destes, como pode ser visualizado na Figura 5.20 que apresenta os dois grupos formados pelo método *k-means* e que há certa dificuldade para visualizar a dispersão dos padrões, pois é atributo a cada padrão um formato (círculo e triângulo para o exemplo) que para lugares em que há grande concentração de padrões não há uma distinção clara destes formatos, o que não ocorreria com a utilização de cores.

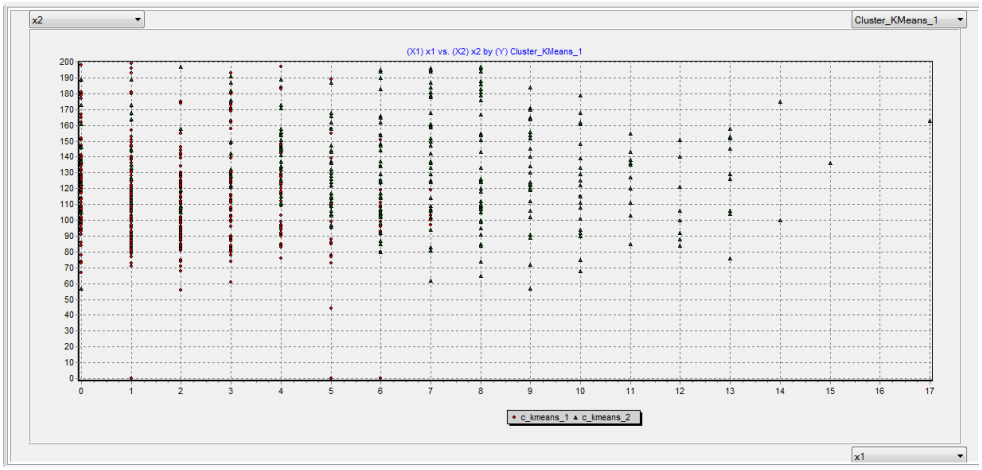


Figura 5.21: Gráfico de dispersão – TANAGRA

Diferente das ferramentas apresentadas, a TANAGRA possui um fluxo de execução em forma de árvore em que cada nodo de execução é adicionado conforme sua hierarquia, ou sequência, pretendida para obter-se um resultado final, isto pode ser visualizado através da Figura 5.21.

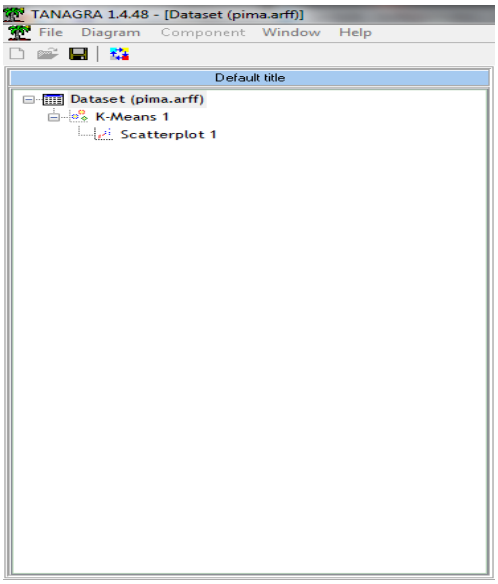


Figura 5.22: Tela principal – TANAGRA

5.2.5 Testes comparativos com a Ferramenta WEKA

A WEKA assim como as apresentadas também é uma ferramenta *Open Source* e é uma das mais utilizadas e completas ferramentas de mineração de dados atualmente. Possui uma grande quantidade de métodos de classificação de dados, agrupamento de dados e associação de dados. Também possui métodos para pré-processamento, como transformações de dados.

Para agrupamento de dados possui sete métodos entre estes o método *k-means* que será utilizado para as execuções de teste. O método possui quatro parâmetros possíveis de serem alterados e que apresentam influência no resultado do agrupamento, que são: número de grupos, número máximo de iterações, medida de distância e inicialização de *seeds*. Como dito o único parâmetro alterado para as execuções foi o número de grupos para dois, mantendo os outros parâmetros *default* da ferramenta.

Como resultado para o método em questão, é apresentado o número de iteração, a soma do quadrado do erro dentro dos grupos e os centroides formados. A Figura 5.22 ilustra a saída da execução do método *k-means*, que apresenta a formação dos grupos com o mesmo número de padrões que a execução do mesmo método nas outras ferramentas.

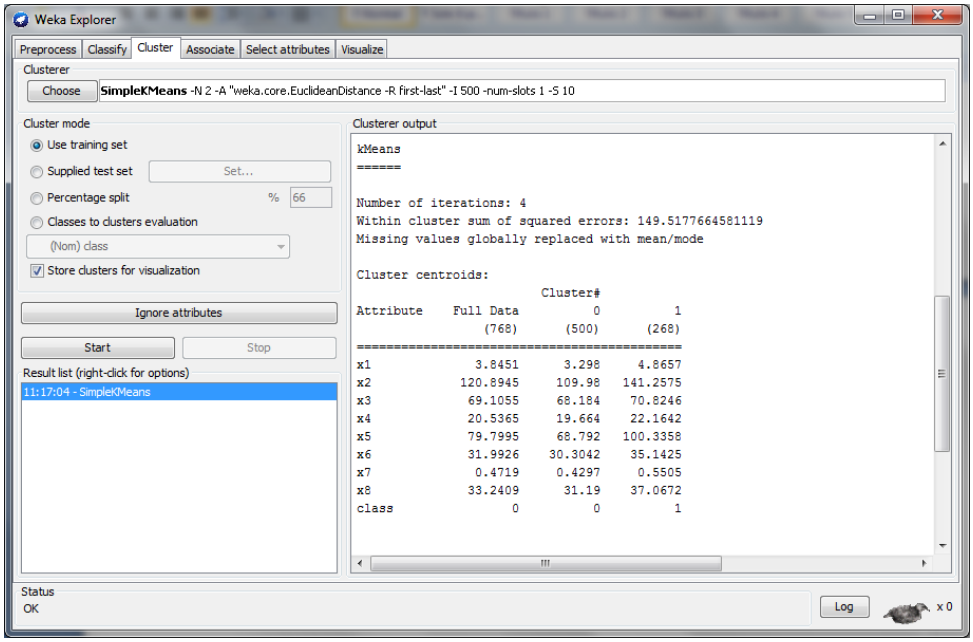


Figura 5.23: Saída de resultado da WEKA

Os métodos de visualização presentes na WEKA são histograma, gráfico de dispersão, matriz *scatter matrix* e visualização de árvore. Destes quatros, o único aplicável ao resultado obtido com o método *k-means* é a técnica de gráfico de dispersão, que pode ser aplicado à base de dados ou ao agrupamento obtido. Porém, a técnica de gráfico de dispersão da WEKA não mostra a dispersão dos dados de um determinado grupo e sim a dispersão dos dados de acordo com um de seus atributos em função do grupo que esta pertence. A Figura 5.23 apresenta a técnica aplicada ao resultado do método *k-means* sendo representados os padrões pertencentes aos dois grupos em função do atributo "*x2*" da base de dados Pima, representado no eixo cartesiano x.

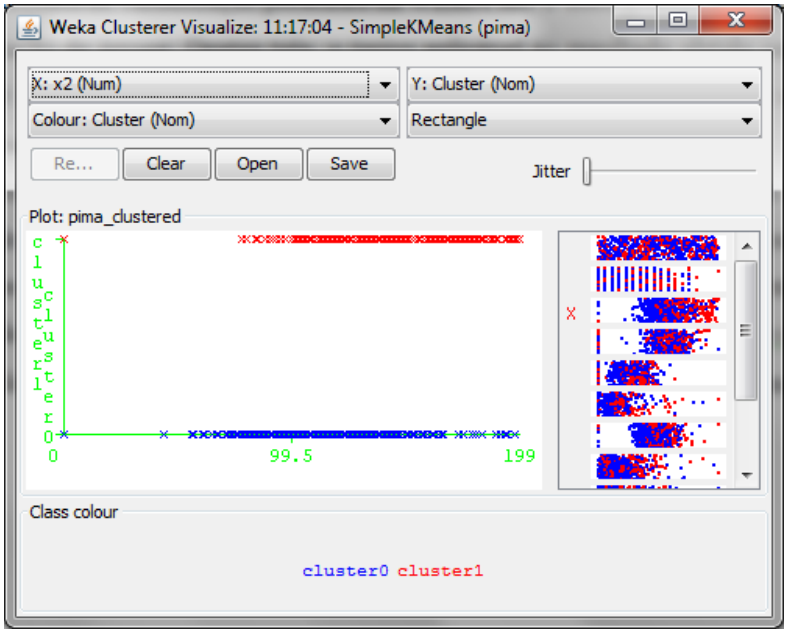


Figura 5.24: Gráfico de dispersão de grupos – WEKA

A utilização da WEKA é a mais simples dentre as ferramentas analisadas e estudadas, tendo o fluxo de execução semelhante ao proposto pela YADMT, e apesar de possuir uma variedade de métodos de mineração de dados, abrangendo praticamente todas as etapas do processo *KDD*, a ferramenta não possui métodos de visualização o que poderia agregar mais ao processo de extração de conhecimento que a ferramenta se propõe a fazer.

5.3 Considerações Finais

As avaliações realizadas sobre as ferramentas propostas tiveram como principal objetivo comparar estas ferramentas com a ferramenta em desenvolvimento YADMT, mais especificamente, em comparar o módulo de agrupamento de dados.

Com a avaliação pode-se perceber que em geral as ferramentas tem uma proposta bastante forte com relação a visualização de dados acoplando os métodos de agrupamento de dados, já que grande parte das ferramentas apresentam em grande parte de seus métodos de visualização a possibilidade de visualizar os grupos gerados pelo método de agrupamento, e a partir destes entender, interpretar e extrair conhecimento.

De modo geral, a usabilidade, facilidade na operação da ferramenta e seus métodos, é boa e pode abranger públicos com qualquer nível de conhecimento sobre a ferramenta, ou até mesmo sobre os métodos que as ferramentas disponibilizam, já que para a execução todas as ferramentas apresentam um fluxo de execução bastante claros, que tem um inicio, meio e fim,

como por exemplo a Figura 5.19 que apresenta a aquisição da base de dados, o método de agrupamento e por fim a visualização dos dados. Para os métodos todas as ferramentas apresentam *feedbacks* sobre o método, como dados de entrada e de saída.

Um ponto negativo é que as ferramentas, com exceção de WEKA e TANAGRA, não apresentam um resultado textual do método de agrupamento, contendo métricas de avaliação para o agrupamento, que aliado aos métodos de visualização podem proporcionar um melhor entendimento dos grupos formado, melhorando assim a extração de conhecimento.

O módulo de agrupamento de dados da YADMT se equivale às ferramentas apresentadas, possuindo métodos de visualização de dados e de grupos que as outras ferramentas apresentam, possuindo como diferencial as representações em 3D da base de dados e dos grupos, possibilitando uma melhor visualização. Também possui a matriz de correlação que não é encontrada nas outras ferramentas.

Em geral a YADMT é uma ferramenta de mineração de dados robusta, podendo atender a qualquer tipo de público, tanto pela sua interface simples de ser usada quanto pelos métodos que nela se encontram, atendendo um público mais específico que procuram por métodos de mineração de dados que traga uma extração de conhecimento sobre uma base de dados rápida e efetiva.

Capítulo 6
Conclusões

Este trabalho discutiu análise de agrupamento de dados e visualização de dados a partir do projeto e implementação de um módulo para tais tarefas na ferramenta YADMT, em desenvolvimento na Unioeste, campus de Cascavel. Este capítulo traz as considerações finais do trabalho desenvolvido, discute alguns dos resultados obtidos e, perspectivas da pesquisa.

6.1 Principais Considerações

Atualmente há grandes volumes de dados, que necessitam de ferramentas de análise para que se chegue a uma tomada de decisão efetiva a partir da compreensão dos dados históricos armazenados. Nesse sentido, este trabalho apresentou uma proposta de desenvolvimento de métodos de agrupamento de dados e visualização de dados para integrar uma ferramenta de mineração de dados, visando agregar valor a esta, deixando-a mais completa para a extração de conhecimento sobre qualquer conjunto de dados.

O primeiro capítulo trouxe uma breve introdução sobre o processo *KDD* como um todo, de forma a contextualizar a pesquisa. O Capítulo 2 objetivou definir e explicar os conceitos e técnicas principais de agrupamento de dados, que podem ser usados nos mais diversos domínios de aplicação.

O Capítulo 3 trouxe os propósitos das visualizações de dados e de agrupamento de dados, assim como as técnicas de visualização, que facilitam a interpretação e entendimento de um agrupamento de dados, possibilitando também a geração de conhecimento. Neste mesmo capítulo, foram discutidos todos os métodos de visualização que constituem o módulo de agrupamento de dados.

No Capítulo 4 os métodos de agrupamento de dados implementados foram introduzidos e, finalmente, o Capítulo 5 relata os testes feitos com os métodos implementados, além de uma avaliação comparativa com outras ferramentas de mineração de dados.

A utilização de métodos de mineração de dados torna-se cada vez mais importante, considerando que os valores contidos em inúmeras bases de dados são altos. Com isso, o

estudo sobre a mineração de dados também se torna importante para que possam surgir métodos e ferramentas que ofereçam diferencias sobre os existentes.

Este trabalho situou-se no estudo de métodos de mineração de dados, mais especificamente no agrupamento de dados e visualização de dados, incorporando-os à ferramenta YADMT.

A principal dificuldade deste trabalho foi justamente o estudo de métodos de visualização de dados para o agrupamento de dados, uma vez que em sua grande maioria, são indicados para a compreensão da base de dados como um todo e não, como forma de visualização dos resultados de uma etapa de agrupamento de dados.

Para atingir o objetivo proposto foi necessário, na maior parte dos métodos, realizar adaptações para visualizar os grupos formados, o que exigiu estudos de qual seria a melhor maneira de realiza-lo. A despeito disso, o trabalho alcançou os objetivos propostos de forma satisfatória.

6.2 Principais Contribuições

Este trabalho teve seu foco na implementação de métodos de visualização de dados para agrupamento de dados, para acoplamento no módulo de agrupamento de dados, porém, também teve como foco secundário a implementação de métodos de agrupamento de dados para constituir o módulo de agrupamento de dados. Sendo assim obteve-se:

- O desenvolvimento de três métodos de agrupamento de dados, acoplados ao módulo de agrupamento de dados da Ferramenta YADMT;
- O desenvolvimento de dez métodos de visualização de dados, acoplados ao módulo de agrupamento de dados da Ferramenta YADMT;
- A realização de uma comparação avaliativa com outras ferramentas de mineração de dados, especificando as diferenças entre os métodos nelas contidos com a YADMT.

6.3 Trabalhos Futuros

Como possíveis trabalhos futuros destacam-se:

- Realização de mudanças na interface gráfica da ferramenta YADMT. Durante os testes comparativos surgiram ideias que podem ser aplicadas para melhorar a

interface gráfica da ferramenta. Estas mudanças criariam uma identidade para a ferramenta, uma vez que o atual *design* é inspirado em ferramentas já existentes;

- Inclusão de novos métodos de agrupamento de dados. Outros métodos de agrupamento de dados poderiam ser estudados e acoplados ao módulo de agrupamento de dados, tornando-o ainda mais robusto;
- Implementação de um módulo de pré-processamento completo para a ferramenta YADMT, para que não seja necessária a utilização de ferramentas externas para este processo, tornando-a uma ferramenta mais completa pela ótica de um processo de KDD;
- Implementação de métodos de interação com as visualizações de dados para prover uma melhor extração de conhecimento destas visualizações;
- Estudo da mineração de dados aliada à Interação Humano-Computador (IHC), de forma a melhorar a experiência do usuário com a ferramenta, além de realização de testes de usabilidade, visando obter insumos para melhorar a extração de conhecimento a partir também da interação com a ferramenta.

Anexo A – Medidas de Distâncias

Os métodos de Agrupamento de Dados descritos no Capítulo 2 deste trabalho se utilizam de medidas de distâncias para expressar a relação entre um objeto e outro. Esta relação, utilizando-se das medidas, é representada por uma matriz de dissimilaridade, ou de proximidade, de acordo com a medida utilizada.

Cada entrada $M_{(i,j)}$ na matriz consiste em um valor numérico que representa o relacionamento entre o objeto i e j. Uma representação desta matriz pode ser dada pela Figura A.1, em que esta matriz é triangular inferior.

$$d = \begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Figura A.1: Representação da matriz de distâncias

A seguir, uma breve descrição das medidas de distâncias utilizadas na implementação do módulo de agrupamento de dados acoplado à YADMT. Estas, segundo (EVERITT, RABEHESKETH, 1997), são as medidas mais utilizadas em agrupamento de dados.

- **Distância de Chebyshev:** Também chamada de distância máxima de valor, examina a magnitude absoluta das diferenças entre as coordenadas de um par de objetos. Pode ser utilizada tanto para variáveis ordinais e quantitativas. A distância é dada por:

$$d_{Ch}(x,y) = \max |x_i - y_i| \tag{A.1}$$

- **Distância de City Block:** A distância de City Block representa a distância entre dois objetos em uma grade de estrada de uma cidade, examinando a diferença absoluta entre as coordenadas do par de objetos. A distância é dada por:

$$d_{CB}(x,y) = \sum_{i=1}^{n} |x_i - y_i| \tag{A.2}$$

- **Coeficiente de Correlação de Pearson:** A correlação mede quão parecidos são dois objetos, quanto mais perto de um maior é a correlação entre os mesmos. Apresenta valores entre [-1, 1]. A correlação é dada por:

$$C_{pearson}(x,y) = \frac{\underline{\sum x * y - \sum x * \sum y}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) * \left((\sum y^2) - \frac{(\sum y)^2}{n}\right)}} \quad \text{(A.3)}$$

- **Coeficiente de Correlação de Spearman:** É uma medida de correlação não-paramétrica. Avalia a relação entre variáveis sem fazer nenhuma suposição sobre a distribuição de frequências das variáveis, não requer que as variáveis sejam lineares. A correlação é dada por:

$$C_{spearman}(x,y) = 1 - \frac{6\sum d_i^2}{(n^3 - n)} \quad \text{(A.4)}$$

- **Coeficiente de Correlação de Kendall Tau:** Assim como o coeficiente de correlação de Spearman também é uma medida de correlação não-paramétrica e é dada por:

$$C_{kendall}(x,y) = \frac{n_x - n_y}{\frac{1}{2}n(n-1)} \quad \text{(A.5)}$$

- **Similaridade de Cosseno:** Representa o cosseno do ângulo entre dois vetores, no caso entre objetos. Apresenta valores entre [-1, 1]. A Similaridade de Cosseno é dada por:

$$S(x,y) = \frac{\sum_{i=1}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2} * \sqrt{\sum_{i=1}^{n} (y_i)^2}} \quad \text{(A.6)}$$

- **Distância Euclidiana:** Mede a dissimilaridade entre dois objetos, em que quanto mais perto de zero à dissimilaridade entre dois objetos mais parecidos estes são. A Distância Euclidiana é dada por:

$$d_e(x,y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} \quad \text{(A.7)}$$

- **Distância de Mahalanobis:** É baseada nas correlações entre variáveis com as quais distintos padrões podem ser identificados e analisados. É uma estatística útil para determinar a similaridade entre uma amostra desconhecida e uma conhecida. É dada por:

$$d_m(x,y) = \sqrt{(x-y)^T * cov^{-1} * (x-y)} \quad \text{(A.8)}$$

Referências

ANDREWS, D. F. **Plots of high-dimensional data**. Biometrics, 1972.

ANKERST, M. **Visual Data Mining with Pixel-oriented Visualization Techniques**. Seatle, WA. 2001.

ANKERST, M.; KEIM, D. A.; KRIEGEL, H. P. **Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets**. Proc Visualization, San Francisco, Ca, 1996.

ARFF. **Especificação do Formato ARFF**. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/arff.html. Acesso: 23 de agosto de 2013.

BENFATTI, E. W.; BONIFACIO, F. N.; GIRARDELLO, A. D.; BOSCARIOLI, C. **Descrição da Arquitetura e Projeto da Ferramenta YADMT - Yet Another Data Mining Tool**. Relatório Técnico nº 01 do Curso de Ciência da Computação, UNIOESTE, Campus de Cascavel, 2010.

BENFATTI, E. W.; **Um estudo sobre a aplicação dos algoritmos KNN, C45 e redes de Bayes na classificação de dados.** Dissertação (Trabalho de Conclusão de Curso) – UNIOESTE – Universidade Estadual do Oeste do Paraná, Cascavel – PR, 2010.

BONIFÁCIO, F. N.; **Comparação entre as redes neurais artificiais MLP, RBF e LVQ na classificação de dados.** Dissertação (Trabalho de Conclusão de Curso) – UNIOESTE – Universidade Estadual do Oeste do Paraná, Cascavel – PR, 2010.

BORYCZKA, U. **Finding groups in data: Cluster analysis with ants**. Applied Soft Computing, v. 9, p. 61-70, 2009.

BOSCARIOLI, C. **Análise de Agrupamentos baseada na Topologia dos Dados e em Mapas Auto-organizáveis**. São Paulo, 2008.

CANVAS. **Site Oficial da Ferramenta ORANGE CANVAS**. Disponível em: http://orange.biolab.si/. Acesso em 25 de setembro de 2013.

CARVALHO, J. G. **Coordenadas Paralelas: Uma Metodologia para Visualização em 3D**. Dissertação de Mestrado, Programa de Pós Graduação em Ciência da Computação, Pontifícia Universidade Católica do Rio Grande do Sul (PUC-RS), Porto Alegre, Brasil, 2001.

CHERNOFF, H. **The use of Faces to Represent Points in K-Dimensional Space Graphically**. Journal of American Statistical Association, vol. 68, p. 361-368, 1973.

CHUAH, M. C.; ROTH, S. F.; MATTIS, J.; KOLOJEJCHICK, J. SDM: **Selective Dynamic Manipulation of Visualizations**. In Proceedings of the ACM Symposium on User Interface Software and Technology, 3D User Interface, pages 61-70. 1995.

CLEVELAND, W. S. **Visualizing Data**. AT&T Bell Laboratories, Murray Hill, NJ, 1993.

CORMACK, R. M. **A Review of Classifications**. JRSS, A. 134-321-367. 1971.

DENEUBOURG, J.-L., GOSS, S., FRANKS, N., SENDOVA-FRANKS, A., DETRAIN, C. CHRÉTIEN, L. **The dynamics of collective sorting: Robot-like ants and ant-like robots**. In Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animals 1 (pp. 356–365). Cambridge, MA: MIT Press, 1991.

DEZA, M. M.; DEZA, E.; **Encyclopedia of Distances**. Springer, 2009.

EVERITT, B. S.; LANDAU, S.; MORVEN, L. **Cluster Analysis**. 4a ed. Londres: Hodder Arnold Publishers, 2001.

EVERITT, B. S.; RABE-HESKETH, S. **The Analysis of Proximity Data**. Londres: Hodder Arnold Publishers, 1997.

FAINO, T. M.; **Agrupamento de Dados a partir de Mapas Auto-Organizáveis na Ferramenta YADMT.** Dissertação (Trabalho de Conclusão de Curso) – UNIOESTE – Universidade Estadual do Oeste do Paraná, Cascavel – PR, 2013.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHRUSAMY, R. Advances **in knowledge Discovery & Data Mining**. California: AAAI/MIT, 1996.

FAYYAD, U.; GRINSTEIN, G. G.; WIERSE, A. **Information Visualization in Data Mining and Knowledge Discovery.** San Francisco: Academic Press, 2002.

FRANK, A.; ASUNCION, A. **UCI Machine Learning Repository**, Irvine, CA: University of California, School of Information and Computer Science, 2010. Disponível em: http://archive.ics.uci.edu/ml. Acesso em 26 de agosto de 2013.

FREITAS, C.M.D.S.; WAGNER, F.R. **Ferramentas de Suporte às Tarefas da Análise Exploratória Visual.** Revista de Informática Teórica e Aplicada, v.2, n. 1, p.5-36, jan. 1995.

HANDL, J.; KNOWLES, J.; DORIGO, M. **Ant-Based Clustering and Topographic Mapping.** Artificial Life, v. 12, n. 1, p. 35-61, 2006.

HANDL, J.; MEYER, B. Ant-**based and swarm-based clustering**. Swarm Intell, v. 1, p. 95-113, 2007.

HEER, J.; CARD, K, S.; LANDAY, J. A. **Prefuse: a toolkit for interactive information visualization.** In: Proccedings of the SIGCHI conference on Human factors in computing systems: 421-430, Portland, Oregon, USA, 2005.

HERNÁNDEZ, L.; BALADRÓN, C.; AGUIAR, J. M.; CARRO, B. SÁNCHES-ESGUEVILLAS, A. **Classification and Clustering of Electricity Demand Patterns in Industrial Parks**. 2012.

HOFFMAN, P. E. **DNA Visual and Analytic Data Mining.** In: IEEE VISUALIZATION, 1997.

INSELBERG, A,; DIMSDALE, B. **Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry.** In: IEEE VISUALIZATION, 1990.

JAIN, A. K.; DUBES, R. C. **Algorithms for Clustering Data**. Nova Jersey, USA: Prentice Hall, 1988.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. **Data Clustering: A Review**. ACM Computing Surveys. v. 31, n. 3, 1999.

JOHNSON, R.A.; WICHERN, D.W. **Applied Multivariate Statistical Analysis**. Fourth Edition. New Jersey: Prentice Hall, 1998.

KEIM, D. A. **Information Visualization and Visual Data Mining**. IEEE Transactions on Visualization And Computers Graphics, vol. 8, n. 1, p. 1-8, 2002.

KEIM, D. A.; KRIEGEL, H, P.; VisDB: **Database Exploration using Multidimensional Visualization**. IEEE Computer Graphics and Applications, 1994.

KEIM, D. A.; KRIEGEL, H.P. **Visualization Techniques for Mining Large Databases: A Comparison**. IEEE Trans. Knowledge & Data Engineering, p. 923-936, 1996.

KEIM, D.; WARD, M. Visual Data Mining Techniques. **Intelligent Data Analysis: An Introduction**. University of Konstanz, Germany. And Worcester Polytechnic Institute, USA. 2002.

KNIME. **Site Oficial da Ferramenta KNIME**. Disponível em: http://www.knime.org/. Acesso em 25 de setembro de 2013.

KOSARA. R. **Parallel Coordinates**, 2010. Disponível em: <http://eagereyes.org/techniques/parallel-coordinates>. Acesso em: 09 junho de 2013.

MACQUEEN, J. B. **Some Methods for classification and Analysis of Multivariate Observations**. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297. 1967.

RABELO, E. **Avaliação de Técnicas de Visualização para Mineração de Dados**. Dissertação (Mestrado em Ciência da Computação), Universidade Estadual de Maringá, Paraná, 2007.

RAPIDMINER. **Site Oficial da Ferramenta RAPIDMINER STUDIO**. Disponível em: http://rapidminer.com/products/rapidminer-studio/. Acesso em 17 de novembro de 2013.

ROBERTSON, G. G.; MACKINLAY, J. D.; CARD, S. K.; **Cone Trees: Animated 3D Visualizations of Hierarchical Information**. In Robertson, S. P., Olson, G. M., and Olson, J.

S., editors, Proc. ACM Conf. Human Factors in Computing Systems, CHI, pages 189–194. ACM Press. 1991.

RÖHSING SILVA, V. H; **Um estudo comparativo entre métodos de extração de seleçõ de características.** Dissertação (Trabalho de Conclusão de Curso) – UNIOESTE – Universidade Estadual do Oeste do Paraná, Cascavel – PR, 2012.

SHNEIDERMAN, B. **The eye have it: A task by data type taxonomy for information visualization**. In Visual Languages, 1996.

SILVA NETO, M. A. **Mineração Visual de Dados: Extração do Conhecimento a partir das Técnicas de Visualização da Informação e Mineração de Dados**. Dissertação (Mestrado em Métodos Numéricos em Engenharia) – Setor de Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2008.

SIPPERT, T. A. S.; **Desenvolvimento de uma máquina de comitê estática para a tarefa de classificação na ferramenta YADMT.** Dissertação (Trabalho de Conclusão de Curso) – UNIOESTE – Universidade Estadual do Oeste do Paraná, Cascavel – PR, 2012.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Inc. Boston, MA, USA: Addison-Wesley Longman Publishing Co. 2005.

TAN, S. C.; TING, K. M.; TENG, S. W.; **Simplifying and improving ant-based clustering**. International Conference on Computacional Science, 2011.

TANAGRA**. Site Oficial da Ferramenta TANAGRA**. Disponível em: http://eric.univ-lyon2.fr/~ricco/tanagra/. Acesso em 25 de setembro de 2013.

VALE, M. N.; **Agrupamento de Dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos**. Dissertação de Mestrado (Mestre em Ciência em Engenharia Elétrica). Pontifícia Universidade Católica do Rio de Janeiro (PUC-RJ), Rio de Janeiro, RJ, 2005.

VALIATI, E. **Avaliação de Técnicas de Visualização de Informações Multidimensionais.** Trabalho Individual (Mestrado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre, 2004.

VILLWOCK, R. **Técnicas de Agrupamento e de Hierarquização no Contexto de Kdd – Aplicação a Dados Temporais de Instrumentação Geotécnica-Estrutural da Usina Hidrelétrica de Itaipu**. 125 f. Tese (Doutorado em Métodos Numéricos em Engenharia) – Setor de Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2009.

WARD, M. O.; **Xmdvtool: Integrating Multiple Methods for Visualizing Multivariate Data**. Washington, DC, 1994.

WEKA. **Waikato Environment for Knowledge Analysis**. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/. Acesso em 25 de setembro de 2013.

WONG, P. C. **Visual Data Mining**. IEEE Computer Graphics and Applicatinos, Los Alamitos, v.19, no. 5, p. 20-21, 1999.