MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
CURSO DE MESTRADO EM SISTEMAS E COMPUTAÇÃO

RAPHAELA BAPTISTA FONSECA

UMA ESTRATÉGIA DE APOIO À SELEÇÃO DE ALGORITMOS DE CLUSTERIZAÇÃO DE DADOS

Rio de Janeiro
2008

INSTITUTO MILITAR DE ENGENHARIA

RAPHAELA BAPTISTA FONSECA

UMA ESTRATÉGIA DE APOIO À SELEÇÃO DE ALGORITMOS DE CLUSTERIZAÇÃO DE DADOS

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Mestre em Sistemas e Computação.

Orientadora: Prof[a]. Maria Claudia Cavalcanti - D.Sc.
Co-orientador: Prof. Ronaldo R. Goldschmidt - D.Sc.

Rio de Janeiro
2008

INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80-Praia Vermelha
Rio de Janeiro-RJ CEP 22290-270

F676u Fonseca, R. B.
Uma Estratégia de Apoio à Seleção de Algoritmos de Clusterização de Dados/ Raphaela Baptista Fonseca. – Rio de Janeiro: Instituto Militar de Engenharia, 2008.
84 p.: il.

Dissertação (mestrado) – Instituto Militar de Engenharia – Rio de Janeiro, 2008.

1. KDD. 2. Clusterização. 3. Mineração de Dados. I. Título. II. Instituto Militar de Engenharia.

CDD 006.312

INSTITUTO MILITAR DE ENGENHARIA

RAPHAELA BAPTISTA FONSECA

UMA ESTRATÉGIA DE APOIO À SELEÇÃO DE ALGORITMOS DE CLUSTERIZAÇÃO DE DADOS

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Mestre em Sistemas e Computação.

Orientadora: Prof[a]. Maria Claudia Cavalcanti - D.Sc.
Co-orientador: Prof. Ronaldo R. Goldschmidt - D.Sc.

Aprovada em 30 de julho de 2008 pela seguinte Banca Examinadora:

Prof[a]. Maria Claudia Cavalcanti - D.Sc. do IME - Presidente

---

Prof. Ronaldo R. Goldschmidt - D.Sc. do IST-Rio/FAETEC

Prof. Ricardo Choren Noya - D.Sc., do IME

Prof. Eduardo Bezerra da Silva - D.Sc., do CEFET-Rio

Rio de Janeiro
2008

Dedico esta à memória de Claudio: um batalhador, sincero, companheiro, amigo e dedicado pai. E à Rosemery, uma inigualável mãe.

AGRADECIMENTOS

A Deus, em primeiro lugar, que me permitiu concluir com êxito mais esta importante etapa em minha vida.

Ao meu orientador, professor Ronaldo Ribeiro Goldschmidt, que me apoiou durante todo curso. Seus conselhos, opiniões, compreensão e amizade foram fundamentais para o êxito do trabalho.

Aos Professores da Seção de Engenharia de Sistemas e Computação (SE/8) do Instituto Militar de Engenharia pelos valiosos ensinamentos.

À Petrobras e ao Tecgraf pela oportunidade e pelas condições oferecidas, sem as quais este trabalho não poderia ter sido realizado.

Aos amigos de turma do IME que muito me incentivaram durante o curso.

Ao meu namorado Felipe, por seu amor, compreensão e apoio incondicionais em todos os momentos de minha vida.

A todos os meus familiares pelos momentos que não pude compartilhar.

A todas as pessoas que contribuíram com o desenvolvimento desta dissertação de mestrado, tenha sido por meio de críticas, idéias, apoio, incentivo ou qualquer outra forma de auxílio.

Por fim, a todos os professores e funcionários da Seção de Engenharia de Sistemas (SE/8) do Instituto Militar de Engenharia.

Raphaela Baptista Fonseca

"Tudo posso naquele que me fortalece"
Filipenses 4:13

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SÍMBOLOS

ABREVIATURAS

ACO	-	*Ant Colony Optimization*
AR	-	*Average Ranking*
ARD	-	*Adjusted Ratio of Distances*
AG	-	*Algoritmos Genéticos*
IKDD	-	*Intelligent Knowledge Discovery in Databases*
IME	-	*Instituto Militar de Engenharia*
KDD	-	*Knowledge Discovery in Databases*
NFL	-	*No Free Lunch Theorems*
PSO	-	*Particle Swarm Optimization*
TI	-	*Tecnologia da Informação*

RESUMO

A clusterização de dados, foco da presente dissertação, é uma tarefa de KDD que se caracteriza por ser um processo de otimização que pode apresentar uma diversidade de soluções possíveis. A busca por boas soluções nesse espaço caracteriza-se como um problema NP-completo. Diante disso, uma pergunta natural em uma aplicação envolvendo a tarefa de clusterização de dados refere-se à escolha entre inúmeros métodos de clusterização de dados disponíveis, de qual ou quais métodos seriam os mais recomendados para o problema que esteja sendo analisado. Uma alternativa para a escolha de métodos de mineração de dados seria a experimentação individual dos métodos disponíveis. Tal abordagem mostra-se, muitas vezes, inviável na prática, considerando o grande número de métodos a serem experimentados. Uma alternativa de cunho prático mais viável sugere a ordenação dos métodos de mineração de dados com base no desempenho destes métodos em experiências similares realizadas anteriormente. Assim sendo, a abordagem proposta pela presente dissertação utiliza conhecimento experimental sobre o desempenho dos métodos de clusterização de dados em situações anteriores de forma a propor ordenações entre estes métodos segundo seu potencial de utilização em novas situações.

Palavras-Chaves: Descoberta de Conhecimento em Bases de Dados; Mineração de Dados; Clusterização de Dados; Algoritmos de Clusterização de Dados.

ABSTRACT

The clustering data, focus of this dissertation, is a task of KDD that is characterized as a process of optimization that can present a lot of possible solutions. The search for good solutions in this space is characterized as an NP-complete problem. Given this, a natural question in an application involving the task of clustering data refers to choose among many data clustering methods available, which methods would be the most recommended to the problem that is being analyzed. An alternative to the choice of data mining methods would be the testing of individual available methods. This approach is shown, a lot of times, unviable in practice, considering the large number of methods to be tested. An alternative more viable, suggests that the sort of data mining methods are based on the performance of these methods in similar experiments accomplished previously. So, the approach proposed by this dissertation uses experimental knowledge about the performance of data clustering methods in previous situations in order to propose ordinances between these methods by their potential for use in new situations.

Keywords: Knowledge Discovery in Databases; Data Mining; KDD Assistance; Data Clustering; Data Clustering Methods.

1 INTRODUÇÃO

1.1 POSICIONAMENTO E MOTIVAÇÃO

Os freqüentes avanços na área de Tecnologia da Informação (TI) têm proporcionado recursos voltados à captação e ao armazenamento de grandes volumes de dados. Diversas tecnologias, tais como a Internet, leitores de códigos de barras, sistemas gerenciadores de banco de dados e sistemas de informação, são alguns exemplos de recursos que têm tornado viável o surgimento de inúmeras bases de dados de natureza comercial, administrativa, governamental e científica, tais como informações espaciais da NASA, e empresas como Banco do Brasil e Caixa Econômica Federal.

Assim sendo, na medida em que as bases crescem cada vez mais, torna-se imprescindível o desenvolvimento de ferramentas que auxiliem o homem na tarefa de análise e interpretação dos dados, em busca de conhecimentos que possam ser úteis no contexto de cada aplicação (GOLDSCHMIDT, 2005). Neste momento, é importante destacar as diferenças e a hierarquia entre dado, informação e conhecimento conforme mostra a Figura 1.1.

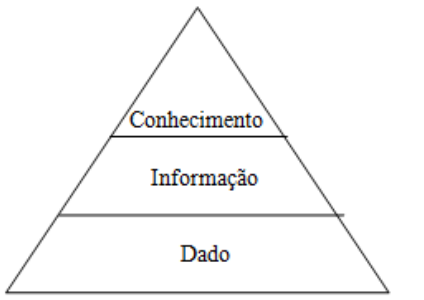


FIG. 1.1: Hierarquia entre Dado, Informação e Conhecimento

Os dados, representados pela base da pirâmide, podem ser interpretados como itens elementares, captados e armazenados através de recursos de TI. A informação, representada pelo meio da pirâmide, representa os dados processados, com significado e contexto bem definidos. O conhecimento, padrão ou conjunto de padrões, representado pelo topo da pirâmide, pode envolver e relacionar dados e informação, não podendo ser abstraído de bases de dados por meio de recursos tradicionais de TI.

Diante deste cenário, surge uma nova área denominada *Descoberta de Conhecimento em Bases de Dados* (*Knowledge Discovery in Databases − KDD*) que busca conhecimento a partir de bases de dados. (FAYYAD, 1996) define KDD como *"Um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, validos, novos e potencialmente úteis a partir de grandes conjuntos de dados"*. Desta forma, o processo de Descoberta do Conhecimento em Bases de Dados é constituído de várias etapas (fig. 1.2) e tem como objetivos encontrar, representar e interpretar, a partir de grandes bases de dados, conhecimentos úteis, através da aplicação de algoritmos e da análise de resultados (AGRAWAL, 1993).

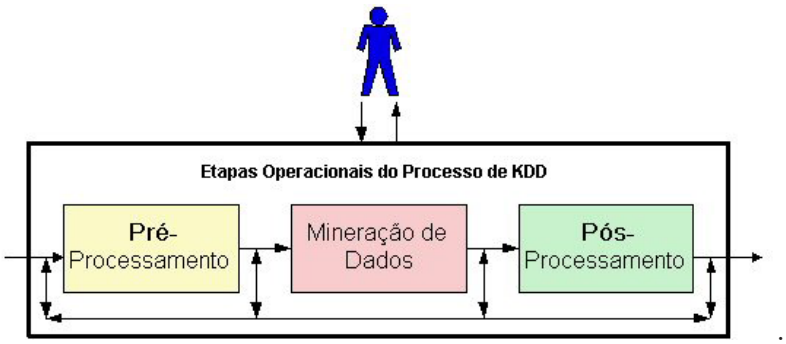


FIG. 1.2: Etapas Operacionais do Processo de KDD

Segundo (GOLDSCHMIDT, 2005), a etapa de pré-processamento compreende as funções relacionadas à captação, à organização e ao tratamento dos dados, tendo como principal objetivo preparar os dados para os algoritmos da etapa seguinte, a Mineração de Dados. Durante a etapa de Mineração de Dados, é realizada a busca efetiva por conhecimento no contexto da aplicação de KDD. A etapa de pós-processamento compreende o tratamento do conhecimento obtido pela Mineração de Dados, tendo como objetivo viabilizar a avaliação da utilidade do conhecimento descoberto.

A Mineração de Dados envolve a aplicação de algoritmos sobre os dados, em busca de conhecimentos implícitos e úteis. Nela são definidos as técnicas e os algoritmos a serem utilizados no problema em questão. Redes Neurais (HALKIDI, 2001), Algoritmos Genéticos (DAVIS, 1991), Modelos Estatísticos e Probabilísticos (MICHIE, 1995) são exemplos de técnicas que podem ser utilizadas na etapa de Mineração de Dados.

Existem diversas tarefas de KDD. A seguir encontram-se citadas e comentadas algumas delas: (a) Descoberta de Associação − Tarefa que abrange a busca por itens que freqüentemente ocorram de forma simultânea em transações do banco de dados (ZAKI,

2000); (b) Classificação − Consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de rótulos categóricos pré-definidos, denominados classes (MICHIE, 1995); (c) Sumarização − Tarefa utilizada para identificar e indicar características comuns entre conjuntos de dados (WEISS, 1998); (d) Detecção de Desvios − Tarefa voltada a identificar registros do banco de dados cujas características não atendam aos padrões considerados normais no contexto (WEISS, 1998); (e) Clusterização de Dados − Foco da presente dissertação, utilizada para separar, em função da similaridade dos dados, os registros de uma base de dados em subconjuntos ou *clusters* (FAYYAD, 1996).

De uma forma geral, a complexidade do processo de KDD decorre da dificuldade de percepção e interpretação de fatos observáveis durante o processo, e da dificuldade em conjugar dinamicamente tais interpretações de forma a decidir quais ações devem ser realizadas em cada caso (GOLDSCHMIDT, 2003). De acordo com (FAYYAD, 1996; HELLERSTEIN, 1999), a complexidade origina-se de diversos fatores que podem ser subdivididos em operacionais e de controle. Os fatores operacionais podem ser identificados, por exemplo, como a dificuldade de integração de diversos algoritmos, e a necessidade de manipulação de grandes e heterogêneos volumes de dados. Os fatores de controle envolvem considerações sobre como conduzir o processo de KDD. Dentre eles podem ser citados:

- A dificuldade na escolha de um algoritmo de mineração de dados, intensificada a partir do surgimento de novos algoritmos com abordagens diferentes. Geralmente a escolha fica a cargo do analista de KDD, que pode deixar de considerar alternativas promissoras (BRAZDIL, 2003);

- A dificuldade na escolha da parametrização adequada dos algoritmos de mineração de dados a cada novo problema, aumentando o número de experimentos na medida em que os algoritmos são testados com diferentes parâmetros (WEISS, 1998);

- A incapacidade humana de memorização de resultados e alternativas processadas conforme o tempo passa e o número de experimentos aumenta, comprometendo a comparação entre alternativas e resultados (HELLERSTEIN, 1999);

Assim sendo, cabe ao analista humano a difícil tarefa de orientar a execução do processo de KDD, utilizando sua experiência, conhecimento e intuição para interpretar e combinar os fatos, decidindo sobre qual a melhor estratégia a ser adotada (FAYYAD, 1996).

A Descoberta de Conhecimento em Bases de Dados surge como uma área de grande interesse pela comunidade cientifica e comercial, cuja busca por resultados vem crescendo nos últimos anos (AGRAWAL, 1993). Diante do exposto acima, a criação de ferramentas inteligentes que auxiliem o homem no controle do processo de KDD torna-se bastante oportuna (MITCHELL, 1997).

A clusterização, foco da presente dissertação, faz parte da etapa de Mineração de Dados, sendo uma tarefa utilizada para particionar os registros de uma base de dados em subconjuntos ou *clusters*, que compartilhem características comuns que os diferenciem dos demais *clusters*. O principal objetivo da clusterização é a maximização da similaridade intra-*cluster* e a minimização da similaridade inter-*cluster*.

(HRUSCHKA, 2003) destacam que a clusterização de dados é um problema que envolve otimização, em que a busca por uma melhor clusterização é NP-Completo, e computacionalmente inviável, a não ser que n (número de objetos) e k (número de *clusters*) sejam extremamente pequenos, visto que o número de partições distintas em que podemos dividir n objetos em k *clusters* é dada pela fórmula:

$$N\,(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{i}\,\binom{k}{i}\,(k\,-\,i)^{n}$$

Segundo (COLE, 1998), para clusterizar 25 objetos em 5 grupos, existem mais de 2.43 quatrilhões de maneiras possíveis. E se o número de *clusters* é desconhecido, precisa-se somar todas as partições possíveis para cada número de *clusters* entre 1 e 5 que fornece um valor superior a 4 x 10^{18} partições possíveis.

Devido a grande diversidade de algoritmos, um mesmo algoritmo pode obter bons resultados em uma determinada base de dados, e ruins em outras. Segundo (DICARLANTONIO, 2001), a qualidade dos resultados obtidos pelos algoritmos de clusterização pode variar de uma base de dados para outra, tornando o processo de clusterização ainda mais complexo, devido a fatores como:

- Natureza dos dados;
- Parametrização utilizada por cada algoritmo;
- Forma como a abordagem se baseia para orientar a busca;
- Formato dos *clusters*.

Sendo assim, procurou-se na literatura trabalhos que buscassem auxiliar o homem no processo de seleção de algoritmos de clusterização de dados a cada novo problema. Como não foram encontrados trabalhos similares na referida pesquisa, o presente trabalho procura preencher uma lacuna relevante, possibilitando o desenvolvimento de trabalhos futuros na área de clusterização de dados.

1.2 OBJETIVOS

Inspirada em (BRAZDIL, 2003) e (GOLDSCHMIDT, 2003), a presente dissertação tem como objetivos pesquisar, formalizar, implementar e avaliar um ambiente de apoio que auxilie na seleção de algoritmos de clusterização de dados no processo de KDD, de forma a propor uma ordenação que apresente recomendações de bons algoritmos de clusterização a serem aplicadas em novas bases de dados. Tal ambiente, ao receber uma nova base de dados, irá realizar os seguintes passos:

- Calcular os *descritores quantitativos* do novo conjunto de dados em questão definidos para a tarefa de clusterização. O novo conjunto passará a ser representado de forma resumida pelos descritores calculados;
- Aplicar alguma *técnica de filtragem de conjuntos de dados*, procurando obter os *conjuntos de referência* mais similares ao novo conjunto. A técnica de filtragem de conjuntos de dados utilizará os descritores quantitativos de conjunto de dados no cálculo de similaridade;
- Selecionar as *medidas de desempenho local* de cada método de clusterização nos conjuntos de referência mais similares ao novo conjunto;
- Combinar as *medidas de desempenho local* dos métodos, calculando uma *medida de desempenho global* para cada método;
- Ordenar os métodos de clusterização de dados em função das *medidas de desempenho global* calculadas anteriormente, produzindo uma lista ordenada de métodos como saída do *ambiente de apoio ao processo de seleção de algoritmos de clusterização de dados*.

Além disso, a abordagem proposta pela presente dissertação possibilitará a incorporação de algoritmos de clusterização, descritores quantitativos, filtros, conjuntos de

dados, e a construção de históricos de desempenho que viabilizem uma análise em busca de novos conhecimentos, além da sua utilização como ferramenta de apoio ao controle do processo de KDD.

1.3 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação possui mais cinco capítulos, conforme descrito abaixo.

O Capítulo 2 fornece definições e conceitos básicos sobre a tarefa de clusterização de dados, além de descrever em detalhes cada algoritmo de clusterização utilizado no trabalho.

No Capítulo 3, encontra-se um resumo com os principais trabalhos relacionados à área de Assistência ao controle do processo de KDD.

No Capítulo 4, encontra-se uma especificação conceitual detalhada da abordagem proposta para o Ambiente de Apoio à Seleção de Algoritmos de Clusterização de dados, abrangendo aspectos conceituais e funcionais de sua formulação.

Detalhes envolvendo a metodologia de testes e a especificação dos experimentos realizados com o protótipo do ambiente de assistência, além de uma análise dos resultados obtidos são fornecidos no Capítulo 5.

O Capítulo 6 descreve as conclusões e as principais contribuições proporcionadas pelo presente trabalho. Alternativas de trabalhos futuros também encontram-se indicadas.

2 TAREFA DE CLUSTERIZAÇÃO

Este capítulo está estruturado da seguinte forma: a seção 2.1 apresenta conceitualmente a tarefa de clusterização de dados, e a seção 2.2 apresenta detalhadamente uma descrição do funcionamento dos métodos utilizados nesta dissertação.

2.1 DEFINIÇÃO E CONCEITOS BÁSICOS

Segundo (GOLDSCHMIDT, 2005), a clusterização de dados, também chamada de agrupamento, é uma tarefa utilizada para criar partições dos registros de uma base de dados em subconjuntos ou *clusters*. Diferente da classificação, que tem grupos pré-definidos, a clusterização precisa automaticamente identificar os grupos. Por esta razão, a clusterização é também denominada indução não supervisionada, sendo definida como uma das tarefas básicas da Mineração de Dados, auxiliando os usuários na realização de agrupamentos naturais de registros em conjuntos de dados.

A análise de *clusters* envolve a organização de um conjunto de padrões (usualmente representados na forma de vetores de atributos ou pontos em um espaço multidimensional - espaço de atributos) em *clusters*, de acordo com alguma medida de similaridade. Intuitivamente, padrões pertencentes a um dado *cluster* devem ser mais similares entre si (pois devem compartilhar um conjunto de propriedades comuns) do que em relação a padrões pertencentes a outros *clusters*.

Segundo (DICARLANTONIO, 2001), os *clusters* são conjuntos de dados que podem apresentar diversos formatos. Os métodos de clusterização baseados em densidade têm maior facilidade para descobrir *clusters* que tenham um formato arbitrário, tais como elíptica, cilíndrica, espiralada. Facilidade que para os métodos de clusterização baseados em particionamento, praticamente, não existe, pois os mesmos só trabalham com *clusters* de formas esféricas. A figura 2.1 mostra o que podem ser *clusters* de formatos arbitrários.

(AGRAWAL, 1993) destacam que, para o método de clusterização baseado em densidade, um *cluster* é uma região que tem uma densidade mais alta de objetos do que sua região vizinha, ou, *clusters* são regiões densas de objetos que são separados por regiões de baixa densidade.

Formalmente, supõe-se a existência de n pontos de dados x_1, x_2, ... x_n tais que cada

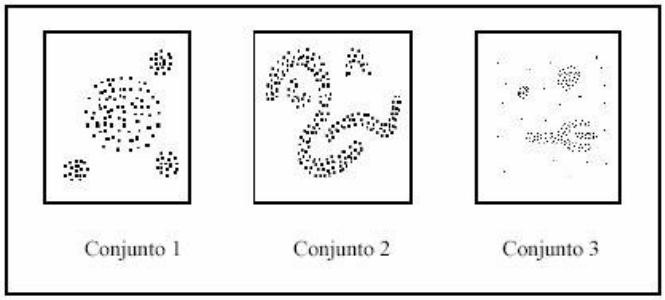


FIG. 2.1: *Clusters* de formatos diferentes

ponto pertençam a um espaço d dimensional R^d. O problema de clusterização destes pontos de dados, separando-os em k *clusters* consiste em encontrar k pontos m_j em R_d de tal forma que a expressão:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\min_j d^2(x_i, m_j)\right)$$

seja minimizada, onde $d(x_i, m_j)$ denota uma distância, normalmente a distância euclidiana, entre x_i e m_j e os pontos m_j são denominados centróides ou médias dos *clusters*.

Informalmente, deseja-se encontrar k centróides de *clusters*, de tal forma que a distância entre cada ponto de dado e o centróide do *cluster* mais próximo a ele seja minimizada (GOLDSCHMIDT, 2005). A tarefa de clusterização consiste de um processo de otimização que pode apresentar uma diversidade de soluções possíveis. A busca por boas soluções nesse espaço caracteriza-se como um problema NP-completo. (AGRAWAL, 1998; HAN, 2000; DICARLANTONIO, 2001) ressaltam que o método ideal de clusterização deveria atender aos seguintes requisitos:

- Descoberta de *clusters* com formato arbitrário: os métodos de clusterização baseados em densidade têm maior facilidade para descobrir *clusters* que tenham um formato arbitrário, tais como elíptica, cilíndrica e espiralada, enquanto métodos baseados em medidas de distância Euclidiana ou Manhattan tendem a encontrar *clusters* de formatos esféricos e densidades similares;

- Aceitação de diversos tipos de variáveis: os métodos têm de ser capazes de lidar com variáveis escaladas em intervalos ou proporções, variáveis binárias, categóricas, ordinais, ou combinações entre elas;

- Indiferença na ordem de apresentação dos objetos: Se um mesmo conjunto de objetos é apresentado em ordens diferentes, devem-se fornecer resultados iguais;

- Variedade no número de atributos/dimensões: os métodos devem trabalhar eficientemente com objetos em altas dimensões, proporcionando resultados inteligíveis;

- Fornecimento de resultados interpretáveis e utilizáveis: os resultados das clusterizações devem possuir representações simples, de forma que os usuários compreendam facilmente;

- Robusto na presença de ruídos: a maioria das bases de dados possui dados faltosos, desconhecidos ou errados, mas sua existência não deve afetar o mínimo possível a qualidade dos resultados obtidos;

- Conhecimento mínimo para determinação dos parâmetros de entrada: os valores adequados dos parâmetros são muito difíceis de determinar, principalmente em conjuntos de objetos de alta dimensionalidade. Em alguns métodos, uma pequena mudança nos parâmetros de entrada reflete nos resultados da clusterização;

- Definição do número adequado de *clusters*: muitos métodos necessitam de um valor de referência como parâmetro de entrada. Encontrar o número ideal de *clusters* é uma tarefa bastante complicada.

Embora o método ideal devesse atender a todos esses requisitos, (AGRAWAL, 1998) sinaliza que nenhum método de clusterização conseguiria atender todos estes pontos de forma adequada, visto que cada método possui uma abordagem diferente. Sendo assim, existem vários algoritmos voltados para a tarefa de clusterização, desde os mais tradicionais até os mais modernos, cada um deles baseado em um paradigma. Entre os mais tradicionais, pode-se citar o K-Means (MACQUEEN, 1967), que é baseado em métodos estatísticos; o Kohonen (KOHONEN, 1997), baseado em redes neurais; e os algoritmos genéticos (AG) (GOLDSCHMIDT, 2005). Já entre os mais recentes, destacam-se o PSO (*Particle Swarm Optimization*), que é baseado no comportamento de grupos de animais (VAN DERMERWE, 2003); e o ACO (*Ant Colony Optimization*), um método multiagente baseado em colônias de formigas (TSAI, 2002).

Segundo (HAN, 2000), uma classificação dos algoritmos de clusterização divide-os da seguinte forma:

- Métodos por particionamento: segundo (DASILVA, 2006), métodos partitivos consideram o número de grupos/*clusters* a serem identificados como um parâmetro

de entrada do algoritmo. Sua tarefa é otimizar (maximizar ou minimizar) uma função-objetivo que envolve cálculos de distância entre os objetos e seus respectivos centros. Os dois maiores representantes da família de métodos partitivos são o K-Means (DUDA, 2000; MACQUEEN, 1967) e o K-Medoids (JAIN, 1988);

- Métodos hierárquicos: segundo (DASILVA, 2006), os algoritmos hierárquicos podem ser subdivididos em divisivos (*top-down*) e aglomerativos (*bottom-up*). Os divisivos começam por considerar a coleção de objetos como um único grupo. A seguir, dividem (particionam) os grupos iterativamente. Já os algoritmos aglomerativos começam por considerar cada objeto da coleção como sendo um grupo. A seguir, iterativamente, unem os grupos menores em grupos cada vez maiores, até que um único grupo que contém todos os objetos seja formado. Um dos representantes desta família é o método Cobweb (FISHER, 1987);

- Métodos baseados em densidade: (DICARLANTONIO, 2001) cita que métodos baseados em densidade são adequados para descobrir *clusters* com forma arbitrária, tais como espiralada, elíptica, cilíndrica etc. (HAN, 2000). Nesta abordagem, um *cluster* é uma região que tem uma densidade mais alta de objetos do que sua região vizinha, ou, *clusters* são regiões densas de objetos que são separados por regiões de baixa densidade (AGRAWAL, 1998). Dois representantes desta família são DBSCAN e OPTICS (HAN, 2000; ANKERST, 1999);

- Métodos baseados em grades: segundo (DICARLANTONIO, 2001), os métodos baseados em grades utilizam uma estrutura de dados em grade de multiresolução, discretizando o espaço de objetos em um número finito de células, que formam uma estrutura de grade na qual as operações de clusterização são efetuadas. Um dos representantes desta família é o STING (SHEIKHOLESLAMI, 1998);

- Métodos baseados em modelos: segundo (DICARLANTONIO, 2001), métodos baseados em modelos são fundamentados na suposição de que os dados são gerados por uma mistura de distribuições de probabilidades. Os métodos baseados em modelos seguem uma das seguintes abordagens (HAN, 2000):

  - Abordagem Estatística: dois representantes desta família são o Cobweb (FISHER, 1987) e o K-Means (DUDA, 2000; MACQUEEN, 1967);

– Abordagem por Rede Neural: um dos principais representantes desta família é o Kohonen (KOHONEN, 1997).

A qualidade dos resultados produzidos pelos algoritmos de clusterização pode variar de uma base de dados para outra. Diversos fatores influenciam nessa variação. São exemplos desses fatores: natureza dos dados, o formato dos conjuntos, a parametrização utilizada e a abordagem que cada algoritmo se baseia de forma a orientar a busca pelo espaço de soluções.

2.2 MÉTODOS DE CLUSTERIZAÇÃO

Esta seção descreve o funcionamento de cada algoritmo de clusterização utilizado na dissertação.

2.2.1 MÉTODO K-MEANS

O método K-Means idealizado por (MACQUEEN, 1967), é um algoritmo guloso cujo objetivo principal é definir k centros (centróides) para cada grupo de *clusters*, através da otimização local de uma função-objetivo. O K-Means procura minimizar a função-objetivo correspondente à distância total entre os objetos e os centróides dos grupos associados a eles. Ou seja, o algoritmo acha um mínimo local para o problema de minimizar a função de soma dos erros médios quadrados, cuja fórmula é apresentada a seguir:

$$\text{MSE} = \frac{1}{|X|}\sum_{i=1}^{k}\sum_{x \in G_i} \|x - \mu_i\|^2$$

Na equação, $|X|$ é a cardinalidade do conjunto de objetos, k é a quantidade de grupos e i é o centróide do grupo G_i. Essa função é também comumente chamada de erro médio quadrático (*mean squared error*) ou dispersão (JAIN, 1988).

O método K-Means pode ser dividido nos seguintes passos:

- Selecionar aleatoriamente k objetos, cada qual representará inicialmente a média de um *cluster* (centróide);
- Para cada objeto remanescente x, x é atribuído a um *cluster* k com o qual x tem maior similaridade. A mesma é baseada na distância de x ao centróide do *cluster* k;

- Este processo é repetido até que algum dos pontos de parada, descritos a seguir, ocorram:
  - Quando os centróides dos *clusters* param de se modificar; ou,
  - Quando não há mudança de objetos para novos *clusters*; ou,
  - Quando o limite máximo de iterações for alcançado.

A função de erro que deve ser minimizada pelo método é representada pela fórmula:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

onde p é o ponto que corresponde a um objeto, m_i é o centróide do *cluster* C_i. Sua complexidade é da ordem de $O(knt)$, onde k é o número de *clusters*, n é o número de objetos e t é o número de iterações.

Algumas questões importantes sobre o método K-Means são que ele necessita de ter como entrada o número de *clusters* desejado, só trabalha com dados numéricos, é sensível a ruídos e elementos estranhos (*outliers*) e é ineficiente para a descoberta de *clusters* que tenham formato não-convexo (HAN, 2000; DICARLANTONIO, 2001).

2.2.2 MÉTODO COBWEB

O método Cobweb, idealizado por (FISHER, 1987), é um algoritmo conceitual incremental que tem como principal objetivo organizar os objetos de um banco de dados, de forma a maximizar a capacidade de prever valores de atributos de um objeto, a partir de informações sobre o conceito associado ao mesmo. O algoritmo organiza incrementalmente os objetos em uma árvore de conceitos (estrutura onde cada um dos nós da árvore representa um conceito, que é resumido pelas distribuições dos valores dos atributos dos objetos pertencentes à sub-árvore do nó). A raiz representa o conceito mais amplo, que resume todo o conjunto de objetos.

A qualidade da distribuição dos valores de atributos em um nó ou sub-agrupamento é calculada através de uma medida de avaliação estatística conhecida como *Utilidade Categórica* (*Category Utility*), que indica o ganho em associar um objeto a um grupo, e é definida como (FISHER, 1987):

$$UC(C_1, C_2, ..., C_k) = \frac{\sum_{l=1}^{k} \Pr[C_l]\left[\left(\sum_i \sum_j P\left(A_i = V_{\text{ij}} \mid C_i\right)^2\right) - \left(\sum_i \sum_j P\left(A_i = V_{\text{ij}}\right)^2\right)\right]}{k}$$

onde k é o número de nós, conceitos ou categorias, A_i indica o atributo i dos objetos e V_{ij} o *j-ésimo* valor desse mesmo atributo. É importante ressaltar que a probabilidade de um grupo é obtida através do total de objetos que pertencem a esse grupo, e que, as folhas da árvore de conceitos são consideradas como um grupo de um único objeto. Particularmente, a *Utilidade Categórica* é uma troca entre a similaridade intra-*cluster* e a dissimilaridade inter-*cluster*, onde os objetos são descritos por pares do tipo *atributo-valor*. A similaridade intra-*cluster* é representada por uma probabilidade condicional na forma $P(A_i = V_{\text{ij}} \mid C_k)$, onde $A_i = V_{ij}$ é um par *atributo-valor* e C_k é o *cluster*. Quanto maior a probablilidade, mais próximos os objetos encontram-se do *cluster*. A similaridade inter-*cluster* é representada pela função $P(C_k \mid A_i = V_{ij})$. Quanto maior esta probabilidade, mais afastados encontram-se os objetos entre os *clusters*.

A listagem abaixo detalha o pseudo-código do algoritmo (FISHER, 1987):

```
Para cada exemplo
    Se a raiz é uma folha
        Adicionar o objeto à folha;
    Senão
        Inserir o objeto no melhor nó descendente, e escolher um dos seguintes passos
            Adicionar um novo objeto ao nó;
            Criar um novo nó com o objeto;
            Juntar os dois melhores nós e inserir no melhor nó descendente (MERGE)
            ou
            Separar o nó (SPLIT)
```

O algoritmo Cobweb possui como parâmetro de entrada um conjunto de registros, em que cada registro é um conjunto de atributo = valor, e como parâmetro de saída uma hierarquia de conceitos. As operações de *MERGE* e *SPLIT* são exemplificadas a seguir através das figuras 2.2 e 2.3:

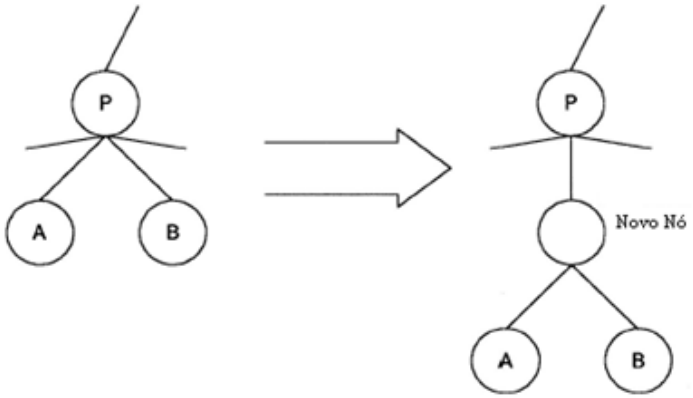


FIG. 2.2: Função de *MERGE*

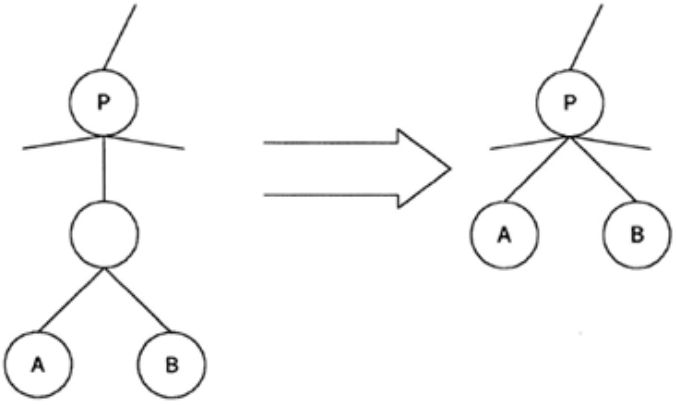


FIG. 2.3: Função de *SPLIT*

Vale ressaltar que a fórmula para o cálculo da *Utilidade Categórica* mencionada anteriormente é utilizada apenas para o cálculo de atributos discretos, visto que ela não considera a distância entre dois valores numéricos. Desta forma, o método CLASSIT (GENNARI, 1989; FISHER, 1987) (uma variação do Cobweb), supõe que os valores dos atributos contínuos estão distribuídos de forma normal e utiliza a curva de Gauss para determinar a probabilidade de ocorrência de um determinado valor. Nesse caso, a soma do quadrado das probabilidades de um atributo discreto da fórmula anterior, torna-se o quadrado da integral da distribuição normal do atributo contínuo, conforme a *Category Utility* abaixo (FERREIRA, 2005):

$$\text{CU numeric } = \frac{1}{n}\sum_{k=1}^{n} P(C_k)\left[\sum_i \frac{1}{2\sqrt{\pi\sigma_{\text{ik}}}} - \sum_i \frac{1}{2\sqrt{\pi\sigma_{\text{ip}}}}\right]$$

onde k é a quantidade de classes da partição, i é a quantidade de atributos, σ_{ik} é o desvio-padrão do atributo i na classe k, e σ_{ip} é o desvio-padrão do atributo i no topo da hierarquia. Quando o desvio-padrão se torna zero, CLASSIT usa um parâmetro chamado *acuidade*, que representa a menor diferença perceptível entre dois valores numéricos. Experimentos mostraram (LI, 1995; REICH, 1991; YOO, 1995) que a qualidade da hierarquia gerada depende fortemente da escolha do valor para acuidade (FERREIRA, 2005).

2.2.3 MÉTODO EXPECTATION-MAXIMIZATION (EM)

O método Expectation-Maximization (EM) idealizado por (DEMPSTER, 1977), é um algoritmo iterativo para estimativa da máxima verossimilhança dos dados. Segundo (GIBSON, 2007), este método não associa cada instância a um *cluster*, mas sim calcula a probabilidade de cada instância pertencer ou não ao *cluster*. Ou seja, cada instância é associada ao *cluster* de maior probabilidade, sendo este representado por uma distribuição de Gauss ou Normal (outras também podem ser utilizadas) com diferentes médias e variâncias.

Cada iteração do algoritmo EM é composta por dois passos: o *expectation* (mais conhecido como passo E) e o *maximization* (passo M). No passo E, a probabilidade dos *clusters* é calculada para cada instância, enquanto que no passo M a distribuição dos parâmetros é calculada de forma a maximizar a distribuição das probabilidades dos mesmos. No final de cada iteração é calculada uma probabilidade global, multiplicando-se as probabilidades individuais de cada instância.

O algoritmo consiste na formalização da idéia intuitiva de lidar com dados incompletos. No caso desta dissertação, podem-se considerar dados incompletos os atributos de classificação dos conjuntos de dados que não foram considerados na utilização do EM. Assim sendo, o algoritmo repete os seguintes passos, até que um critério de convergência seja alcançado (LITTLE, 1986):

- Substitui os valores incompletos por valores estimados;
- Estima os parâmetros (Passo E);
- Reestima os valores incompletos considerando que os novos parâmetros são corretos;
- Reestima os parâmetros (Passo M).

Segundo (GIBSON, 2007), assim como o K-Means, o algoritmo EM garante apenas convergir a um máximo local e não global. Desta forma, o algoritmo deveria ser repetido

várias vezes com suposições iniciais diferentes para os valores dos parâmetros. Neste caso, a verossimilhança pode ser utilizada para comparar diretamente as configurações finais resultantes obtendo o maior dos máximos locais.

2.2.4 MÉTODO FARTHEST-FIRST TRAVERSAL

(HOCHBAUM, 1985) introduziram o método *Farthest-First Traversal* como um algoritmo de aproximação para o que chamamos de *k-center problem*, objetivando encontrar uma clusterização com k grupos a partir de uma função de custo, maximizando o raio do cluster. Sua idéia é escolher um ponto qualquer no conjunto de dados (entenda ponto como uma tupla do conjunto de dados) para iniciar o processamento e, a partir daí, escolher o ponto mais afastado dele, depois o ponto mais afastado dos dois primeiros (a distância de um ponto x a partir de um conjunto S é dada por $min\{d(x, y) : y \in S\}$), até que k pontos sejam obtidos. Esses pontos são tidos como os centróides dos clusters e cada ponto restante é atribuído ao centróide mais próximo (DASGUPTA, 2005).

O algoritmo *Farthest-First Traversal* difere do K-Means apenas no cálculo dos seus centróides. Enquanto no K-Means os centróides são calculados de forma a minimizar a distância dos vários elementos que compõe o *cluster* a seu respectivo centróide através da fórmula:

$$\min |x_i - u_i|^2$$

onde x_i é um elemento do cluster e u_i o seu respectivo centróide, o cálculo do Farthest-First Traversal determina o centróide como o elemento do cluster com o valor máximo das distâncias mínimas aos centros atuais (DASGUPTA, 2005), através da fórmula:

$$\max_x \min_c d(x, c)$$

onde x é um elemento do cluster e c o seu centróide. As figuras 2.4 e 2.5 ilustram um exemplo dos centróides obtidos pelo K-Means e Farthest-First Traversal respectivamente.

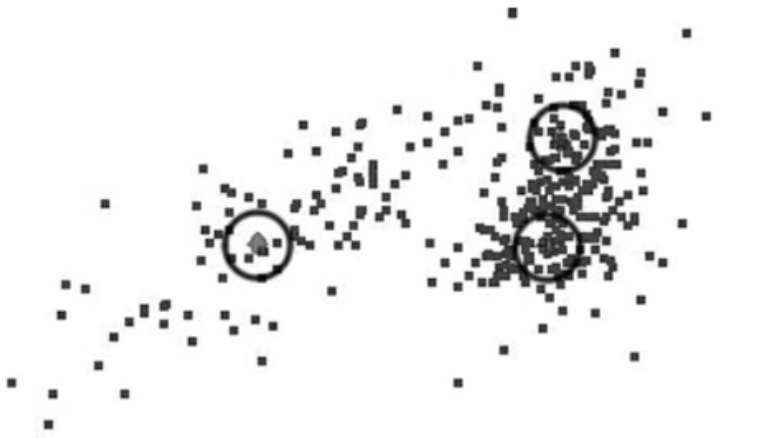


FIG. 2.4: Centróides do algoritmo K-Means (PEIXOTO)

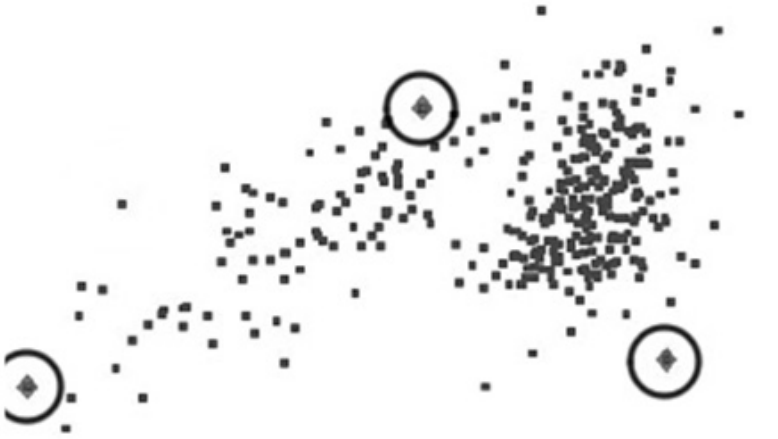


FIG. 2.5: Centróides do algoritmo Farthest-First Traversal (PEIXOTO)

2.3 CONSIDERAÇÕES FINAIS

A escolha destes métodos deveu-se aos seguintes fatores:

- A diversificação da abordagem em que cada algoritmo se baseia, orientando a busca pelo espaço de soluções. O K-Means é um método partitivo, o Cobweb é hierárquico, Farthest-First é um método baseado em modelos, e o EM possui princípios estatísticos;
- A utilização destes métodos em experimentos similares realizados na comunidade científica (BRAZDIL, 2003; SOARES, 2001; BENSUSAN, 2000; GONÇALVES, 2001);

- A disponibilidade dos métodos por meio de uma ferramenta *opensource*, denominada *WEKA*, o que permitiu construir uma ferramenta de apoio (*Automated Weka*) que auxiliasse na grande e diversificada massa de experimentos a serem realizados.

3 TRABALHOS RELACIONADOS

Este capítulo tem como objetivo reunir os principais trabalhos relacionados à área de KDD e à presente dissertação.

3.1 CONSIDERAÇÕES INICIAIS

A grande inspiração para esta dissertação foi o trabalho descrito em (GOLDSCHMIDT, 2003), onde foi proposta, desenvolvida e avaliada uma *Máquina de Assistência Inteligente à Orientação do Processo de KDD*, denominada *Máquina de IKDD (Intelligent Knowledge Discovery in Databases)*, utilizada como ferramenta didática voltada à formação de profissionais técnicos na área da Descoberta de Conhecimento em Bases de Dados. Essa máquina foi formalizada com base na *Teoria do Planejamento para Resolução de Problemas* (RUSSELL, 1995) da *Inteligência Artificial*, e implementada a partir da integração de funções de assistência utilizadas em diferentes níveis de controle do processo de KDD: *Definição de Objetivos*, *Planejamento de Ações de KDD*, *Execução dos Planos de Ações de KDD* e *Aquisição e Formalização do Conhecimento*. O nível *Definição de Objetivos* auxilia o homem na identificação de tarefas de KDD cuja execução seja potencialmente viável na base de dados em análise. O *Planejamento de Ações de KDD* auxilia o homem na escolha de qual algoritmo de mineração de dados é o mais apropriado para determinado problema, utilizando uma metodologia de ordenação dos algoritmos de mineração baseada no desempenho prévio destes algoritmos em problemas similares (SOARES, 2001; BRAZDIL, 2003). Esta função permite gerar ainda os chamados *Planos de Ação em KDD*. Um *Plano de Ação em KDD* é uma seqüência de métodos que tem como finalidade atingir os objetivos estabelecidos a partir do estado em que se encontra a base de dados em análise. A *Execução dos Planos de Ações de KDD* é responsável por executar os algoritmos de KDD previstos no plano, analisando os resultados obtidos. Por último, a *Aquisição e Formalização do Conhecimento* permite adquirir e disponibilizar os conhecimentos sobre KDD em uma representação e organização que viabilizem o processamento das funções de assistência mencionada nos níveis anteriores. Convém mencionar que a *Máquina de IKDD*, no nível de *Planejamento de Ações de KDD*, foi estruturada de forma flexível, permitindo a incorporação de qualquer algoritmo de mineração de dados.

Apesar disso, apenas algoritmos de classificação de dados foram utilizados neste nível. A presente dissertação procurar "estender" esta estrutura, trabalhando com algoritmos de clusterização de dados neste nível.

O trabalho descrito em (GOLDSCHMIDT, 2003) ressalta ainda que, o conhecimento sobre como realizar ações de KDD pode se apresentar como:

- Meta Conhecimento Teórico;
- Meta Conhecimento Experimental;
- Meta Conhecimento Híbrido.

O meta conhecimento teórico consiste de formas de representação de conhecimento que retratam a experiência de especialistas humanos sobre um determinado assunto. Portanto, a formulação do meta conhecimento teórico requer a participação de um especialista em KDD na aquisição e formalização do referido conhecimento. Na alternativa baseada em meta conhecimento experimental é necessária a existência de um histórico contendo exemplos de aplicações de KDD previamente realizadas, com o objetivo de traçar considerações sobre como utilizar este histórico na extração do referido conhecimento. O meta conhecimento experimental não necessita da participação de um especialista de KDD para sua operacionalização. Por último, a alternativa baseada em meta conhecimento híbrido, combina as duas alternativas anteriores, tirando proveito dos benefícios proporcionados por cada uma delas. O meta conhecimento híbrido requer a participação de um especialista em KDD, e de um histórico de dados sobre o assunto em questão.

Levando em consideração a diversidade de trabalhos na área de IKDD, foram criadas subseções em função das dimensões de assistência abrangidas por cada trabalho como *Assistência ao Planejamento de Ações de KDD*, *Definição de Objetivos* e *Execução de Ações de KDD*.

3.2 ASSISTÊNCIA AO PLANEJAMENTO DE AÇÕES EM APLICAÇÕES DE KDD

Em (BERNSTEIN, 2002), foi implementado um protótipo denominado *IDEA*, que, produz planos de KDD baseados em ontologias de KDD. Este protótipo permite que os usuários ordenem conjuntos de planos produzidos, facilitando sua escolha. Apenas as tarefas de classificação, sumarização e regressão foram consideradas neste trabalho. Em (GOLDSCHMIDT, 2002b), os autores propuseram uma extensão ao trabalho (BERNSTEIN,

2002), incorporando heurísticas para filtragem dos algoritmos em função de restrições especificadas pelos usuários.

Em (GAMA, 1995; SPILIOPOULOU, 1998; BRAZDIL, 1998; KALOUSIS, 1999; KELLER, 2000; BRAZDIL, 2000; BENSUSAN, 2000; SOARES, 2001; FÜRNKRANZ, 2001; PFAHRINGER, 2000; BRAZDIL, 2003) foram propostos critérios para ordenação de algoritmos de classificação de dados, baseados no desempenho dos mesmos em experiências anteriores. Assim sendo, ao receber um novo conjunto de dados, buscam-se conjuntos de dados similares e apresentam-se ordenadamente, os algoritmos que obtiveram os melhores resultados em tais conjuntos. Estas abordagens encontram-se restritas à tarefa de classificação.

Em (BRODLEY, 1995), foram apresentados exemplos de meta conhecimento teórico, expressos na forma de regras, guiando a seleção de algoritmos de classificação. Ao incluir novos algoritmos, é necessário reavaliar as regras previamente incorporadas, viabilizando a inclusão de novas regras consistentes com o conjunto de regras original.

Em (MORIK, 2000), a autora analisa a influência de ações de pré-processamento de dados no desempenho dos métodos de mineração de dados.

(MICHIE, 1995) buscou caracterizar sobre quais circunstâncias determinados algoritmos devem ser utilizados em detrimento de outros. Neste trabalho, foram realizados vários experimentos em aproximadamente 20 bases de dados distintas, aplicando-se algoritmos de várias abordagens (estatística, conexionista e aprendizado de máquina). Este trabalho foi um dos subprodutos do projeto *StatLog* e contribuiu significativamente em diversos estudos sobre classificação.

A tabela 3.1 e 3.2 abaixo resumem as principais características dos trabalhos referentes à *Assistência ao Planejamento de Ações de KDD*.

TAB. 3.1: Trabalhos em Assistência ao Planejamento de Ações de KDD (GOLDSCHMIDT, 2003)

Características da Assistência em KDD	**(BERNSTEIN, 2002)**	**(SUYAMA, 1998)**	**(BRAZDIL, 2003)**
Etapas do Processo de KDD	Todas	Todas	Mineração de Dados
Algoritmos e Técnicas de KDD	Diversos	Diversos	Diversos
Tarefas de KDD	Clusterização, Regressão, Sumarização	Classificação	Classificação
Mecanismo de Assistência	Planejamento	Programação Genética	Critérios de ordenação por desempenho
Acoplamento Assistência - Execução	Baixo	Baixo	Baixo
Recursos de Paralelismo e Distribuição	Não	Não	Não
Suporte a iterações no Processo de KDD	Não	Não	Não
Conhecimento do Domínio da Aplicação	Não	Não	Sim
Conhecimento de KDD	Sim	Sim	Sim
Representação do Conhecimento	Ontologias	Ontologias	Histórico de desempenho
Meta Conhecimento	Teórico	Experimental	Experimental
Independência Assistência - Aplicações	Sim	Sim	Sim
Incorporação de Novos Conhecimentos	Sim	Sim	Sim
Capacidade de Aprendizado	Não	Não	Não
IHM na Definição de Objetivos	Não se aplica	Não se aplica	Não se aplica
IHM no Planejamento de Ações	Possível	Possível	Possível
IHM na Execução de Ações	Não se aplica	Não se aplica	Não se aplica

TAB. 3.2: Trabalhos em Assistência ao Planejamento de Ações de KDD - continuação

Características da Assistência em KDD	**(MICHIE, 1995)**	**(BRODLEY, 1995)**
Etapas do Processo de KDD	Mineração de Dados	Mineração de Dados
Algoritmos e Técnicas de KDD	Diversos	Diversos
Tarefas de KDD	Classificação, Regressão	Classificação
Mecanismo de Assistência	Seleção de Métodos baseada em experiências prévias	Seleção de Métodos baseada em heurísticas
Acoplamento Assistência - Execução	Baixo	Baixo
Recursos de Paralelismo e Distribuição	Não	Não
Suporte a iterações no Processo de KDD	Não	Não
Conhecimento do Domínio da Aplicação	Sim	Sim
Conhecimento de KDD	Sim	Sim
Representação do Conhecimento	Diversas	Regras de produção
Meta Conhecimento	Híbrido	Teórico
Independência Assistência - Aplicações	Sim	Sim
Incorporação de Novos Conhecimentos	Sim	Sim
Capacidade de Aprendizado	Não	Não
IHM na Definição de Objetivos	Não se aplica	Não se aplica
IHM no Planejamento de Ações	Possível	Possível
IHM na Execução de Ações	Não se aplica	Não se aplica

3.3 DEFINIÇÃO DE OBJETIVOS

Em (ENGELS, 1997; VERDENIUS, 1998), os autores utilizam um modelo de processos que permite a decomposição hierárquica de tarefas a partir de objetivos definidos pelo usuário humano em uma etapa não automatizada.

Na referência (KERBER, 1998) encontra-se descrito o conceito de *áreas de trabalho ativas*. Uma área de trabalho ativa é uma estrutura de dados utilizada para armazenar seqüências de métodos de KDD que tenham sido bem sucedidas. Neste mesmo trabalho, os autores utilizam hipertextos e recursos gráficos para apresentação de informações sobre as áreas de trabalho ativas, procurando orientar a realização de novos processos de KDD.

Em (JENSEN, 1999), encontra-se proposta uma linguagem de programação e um ambiente interativo para programação nesta linguagem. A idéia é que o especialista em KDD utilize o ambiente e a linguagem para programar os passos a serem realizados em

cada aplicação. No programa, a responsabilidade pela execução de cada passo é atribuída a um agente (MAES, 1994; HAYES-ROTH, 1995; WOOLDRIDGE, 1995), que pode ser computacional ou humano.

Nas referências (GOLDSCHMIDT, 2002c,a), os autores propuseram um modelo computacional de auxílio à prospecção de objetivos baseada em conceitos da *Teoria da Equivalência entre Atributos de Bancos de Dados* (LARSON, 1989) e de *Espaços Topológicos* (LIPSCHUTZ, 1973), representando as bases de dados como a união de padrões definidos a partir dos metadados dos atributos destas bases. Uma vez que uma base de dados tenha sido mapeada na nova representação, o mecanismo de assistência instancia operações, sugerindo alternativas de tarefas de KDD potencialmente executáveis nesta base.

A tabela 3.3 abaixo resume as principais características dos trabalhos referentes à *Assistência à Definição de Objetivos*.

TAB. 3.3: Trabalhos em *Assistência à Definição de Objetivos* (GOLDSCHMIDT, 2003)

Características da Assistência em KDD	**(ENGELS, 1997)**	**(JENSEN, 1999)**	**(KERBER, 1998)**
Etapas do Processo de KDD	Todas	Todas	Todas
Algoritmos e Técnicas de KDD	Diversos	Diversos	Diversos
Tarefas de KDD	Diversas	Diversas	Diversas
Mecanismo de Assistência	Decomposição hierárquica de tarefas	Agenda de Tarefas	Áreas de trabalho ativas
Acoplamento Assistência - Execução	Baixo	Baixo	Baixo
Recursos de Paralelismo e Distribuição	Não	Sim	Não
Suporte a iterações no Processo de KDD	Não	Sim	Sim
Conhecimento do Domínio da Aplicação	Não	Não	Sim
Conhecimento de KDD	Sim	Sim	Sim
Representação do Conhecimento	Codificação em métodos	Associações entre tarefas e agentes	Hipertextos, áreas de trabalho ativas
Meta Conhecimento	Teórico	Teórico	Teórico
Independência Assistência - Aplicações	Sim	Sim	Sim
Incorporação de Novos Conhecimentos	Sim	Sim	Sim
Capacidade de Aprendizado	Não	Não	Não
IHM na Definição de Objetivos	Necessária	Necessária	Necessária
IHM no Planejamento de Ações	Possível	Necessária	Necessária
IHM na Execução de Ações	Inexistente	Inexistente	Inexistente

3.4 EXECUÇÃO DE AÇÕES DE KDD

Grande parte destes trabalhos envolve a otimização dos parâmetros de algoritmos de mineração de dados (GOLDBERG, 1989; FAYYAD, 1996; CURRAN, 2002), em que o espaço de busca se restringe aos valores possíveis para os parâmetros dos algoritmos.

Na abordagem proposta em (SHEN, 1996), o processo de descoberta do conhecimento baseia-se em *meta-consultas* que são aplicadas em um algoritmo de aprendizado relacional, produzindo todas as especializações possíveis das meta-consultas fornecidas. Uma

meta-consulta, também denominada *meta-padrão*, é uma expressão da *Lógica de Segunda Ordem* (SHOENFIELD, 1967; SMULLYAN, 1971; RUSSELL, 1995), que descreve o tipo de padrão a ser descoberto.

Em (LIVINGSTON, 2001b, 2000, 2001a), os autores propuseram uma forma de assistência baseada nos conceitos de *Agenda e Justificativa* utilizados em (LENAT, 1984). Uma Agenda é um repositório onde são armazenadas tarefas a serem executadas pelo sistema. Tarefas são operações voltadas à descrição de atributos do conjunto de dados. A identificação, a inclusão e a ordenação de tarefas na estrutura de *Agenda* são realizadas com base em heurísticas previamente formuladas pelo homem.

A tabela 3.4 abaixo resume as principais características dos trabalhos referentes à *Assistência à Execução de Ações de KDD*.

TAB. 3.4: Trabalhos em *Assistência à Execução de Ações de KDD* (GOLDSCHMIDT, 2003)

Características da Assistência em KDD	**(LIVINGSTON, 2001b)**	**(SHEN, 1996)**
Etapas do Processo de KDD	Mineração de Dados	Mineração de Dados, Pós-processamento
Algoritmos e Técnicas de KDD	RL	Aprendizado Relacional e Lógica de 2a Ordem
Tarefas de KDD	Sumarização	Sumarização
Mecanismo de Assistência	Agenda e Justificativa	Meta-Consultas
Acoplamento Assistência - Execução	Alto	Alto
Recursos de Paralelismo e Distribuição	Não	Não
Suporte a iterações no Processo de KDD	Sim	Não
Conhecimento do Domínio da Aplicação	Sim	Não
Conhecimento de KDD	Sim	Sim
Representação do Conhecimento	Regras de Produção	Regras de produção
Meta Conhecimento	Teórico	Teórico
Independência Assistência - Aplicações	Sim	Sim
Incorporação de Novos Conhecimentos	Sim	Não
Capacidade de Aprendizado	Não	Não
IHM na Definição de Objetivos	Não se aplica	Não se aplica
IHM no Planejamento de Ações	Não se aplica	Não se aplica
IHM na Execução de Ações	Inexistente	Possível

3.5 COMPARAÇÕES COM TRABALHOS RELACIONADOS

A tabela 3.5 resume as principais características do *Ambiente ASAC*, facilitando sua comparação com os principais trabalhos da área.

TAB. 3.5: Resumo das Características do *Ambiente ASAC*

Características da Assistência em KDD	**Ambiente ASAC**
Etapas do Processo de KDD	Mineração de Dados
Algoritmos e Técnicas de KDD	Diversos
Tarefas de KDD	Clusterização
Dimensões da Assistência em KDD	Planejamento das Ações de KDD
Mecanismo de Assistência	Critérios de ordenação por desempenho
Acoplamento Assistência - Execução	Baixo
Recursos de Paralelismo e Distribuição	Não
Suporte a iterações no Processo de KDD	Não
Conhecimento do Domínio da Aplicação	Sim
Conhecimento de KDD	Sim
Representação do Conhecimento	Histórico de desempenho
Meta Conhecimento	Experimental
Independência Assistência - Aplicações	Sim
Incorporação de Novos Conhecimentos	Sim
Capacidade de Aprendizado	Não
IHM na Definição de Objetivos	Não se aplica
IHM no Planejamento de Ações	Possível
IHM na Execução de Ações	Não se aplica

Em uma comparação direta com os trabalhos relacionados, convém destacar os seguintes comentários:

- O trabalho de (BERNSTEIN, 2002) está restrito à tarefa de classificação, enquanto que o *Ambiente ASAC* propõe uma expansão para trabalhar não só com a tarefa de clusterização, mas também com outras tarefas de mineração de dados;

- (BRAZDIL, 2003) propôs critérios para ordenação de métodos de classificação. O *Ambiente ASAC* incorpora este critérios e propõe a expansão do mecanismo de ordenação para outros métodos de mineração de dados;

- (BRODLEY, 1995) e (MICHIE, 1995) propõem meta-conhecimento teórico para seleção de algoritmos de classificação. São heurísticas cujo teor é fortemente dependente do conjunto de métodos de classificação disponíveis (BRODLEY, 1995;

MICHIE, 1995). Tal limitação não ocorre no *Ambiente ASAC*, visto que ele propõe a utilização de conhecimento experimental, podendo portanto incorporar novos métodos, demandando apenas a configuração dos mesmos no ambiente.

4 ABORDAGEM PROPOSTA

Este capítulo está estruturado da seguinte forma: a seção 4.1 apresenta a complexidade da tarefa de clusterização de dados e a dificuldade em se determinar o desempenho dos algoritmos de clusterização de dados, e a seção 4.2 e subseções descrevem detalhadamente a abordagem proposta pela presente dissertação.

4.1 CONSIDERAÇÕES INICIAIS

Conforme comentado anteriormente, a complexidade inerente ao processo de KDD decorre, sobretudo, de fatores relacionados ao controle do processo. Estes fatores envolvem considerações sobre como conduzir processos de KDD. Entre tais fatores pode ser destacada a dificuldade na escolha de um algoritmo de mineração de dados com potencial para geração de resultados satisfatórios. Tal dificuldade é intensificada na medida em que surjam novos algoritmos com o mesmo propósito, aumentando a diversidade de alternativas. Em geral, a escolha dos algoritmos se restringe às opções conhecidas pelo analista de KDD, deixando muitas vezes de considerar alternativas promissoras (BRAZDIL, 2003).

Uma mesma tarefa de KDD pode ser executada por vários algoritmos distintos. Esses algoritmos são concebidos a partir de diferentes técnicas e buscam obter bons resultados no contexto da tarefa de KDD a que se propõem. Como a natureza de bases de dados a serem analisadas pelo processo de KDD varia, costuma variar também a qualidade dos resultados obtidos pelos algoritmos nos mais variados contextos.

A clusterização de dados, objeto da presente dissertação, é uma tarefa utilizada para particionar os registros de uma base de dados em subconjuntos ou *clusters*. Esse particionamento deve ocorrer de tal forma que elementos em um *cluster* compartilhem um conjunto de propriedades comuns que os distingam dos elementos de outros *clusters*. O objetivo da clusterização, é maximizar a similaridade intra-*cluster* e minimizar a similaridade inter-*cluster*. Diferente da classificação, que tem rótulos pré-definidos, a clusterização precisa automaticamente identificar os rótulos. Por esta razão, a clusterização é também denominada indução não supervisionada (GOLDSCHMIDT, 2005), sendo definida como uma das tarefas básicas da Mineração de Dados, auxiliando os usuários na realização de agrupamentos naturais de registros em um conjunto de dados.

A clusterização de dados é uma tarefa de KDD que se caracteriza por ser um processo de otimização que pode apresentar uma diversidade de soluções possíveis. A busca por boas soluções nesse espaço caracteriza-se como um problema NP-completo. Diante disso, uma pergunta natural em uma aplicação envolvendo a tarefa de clusterização de dados refere-se à escolha entre inúmeros métodos de clusterização de dados disponíveis, de qual ou quais métodos seriam os mais recomendados para o problema que esteja sendo analisado.

Considerando a crescente diversidade de métodos de mineração de dados, a definição de quais destes métodos possuem melhor desempenho em determinados problemas tem sido uma questão de grande relevância e interesse na comunidade científica (BRODLEY, 1995; MICHIE, 1995; GAMA, 1995; WOLPERT, 1996; BRAZDIL, 1998; SPILIOPOULOU, 1998; BRAZDIL, 2000; BENSUSAN, 2000; SOARES, 2000a,b, 2001).

A ausência de comprovação formal quanto à existência de métodos de mineração de dados cujo desempenho seja superior ao dos demais em qualquer problema torna a escolha incondicional por determinados métodos uma questão de mera preferência pessoal. Convém destacar que os teoremas NFL (*No Free Lunch Theorems*) comprovam a inexistência de métodos que sejam universalmente superiores aos demais em qualquer problema (WOLPERT, 1996).

Uma alternativa para a escolha de métodos de mineração de dados seria a experimentação individual dos métodos disponíveis. Tal abordagem mostra-se, muitas vezes, inviável na prática, considerando o grande número de métodos a serem experimentados.

Uma outra alternativa de cunho prático mais viável sugere a ordenação dos métodos de mineração de dados com base no desempenho destes métodos em experiências similares realizadas anteriormente (BRAZDIL, 2003; SOARES, 2001; BENSUSAN, 2000).

Assim sendo, inspirada em (BRAZDIL, 2003) e (GOLDSCHMIDT, 2003), a abordagem proposta pela presente dissertação utiliza conhecimento experimental sobre o desempenho dos métodos de clusterização de dados em situações anteriores de forma a propor ordenações entre estes métodos segundo seu potencial de utilização em novas situações.

4.2 AMBIENTE DE APOIO À TAREFA DE CLUSTERIZAÇÃO DE DADOS

A Figura 4.1 apresenta um diagrama conceitual do *Ambiente de Apoio à Seleção de Algoritmos de Clusterização de Dados* (*Ambiente ASAC*).

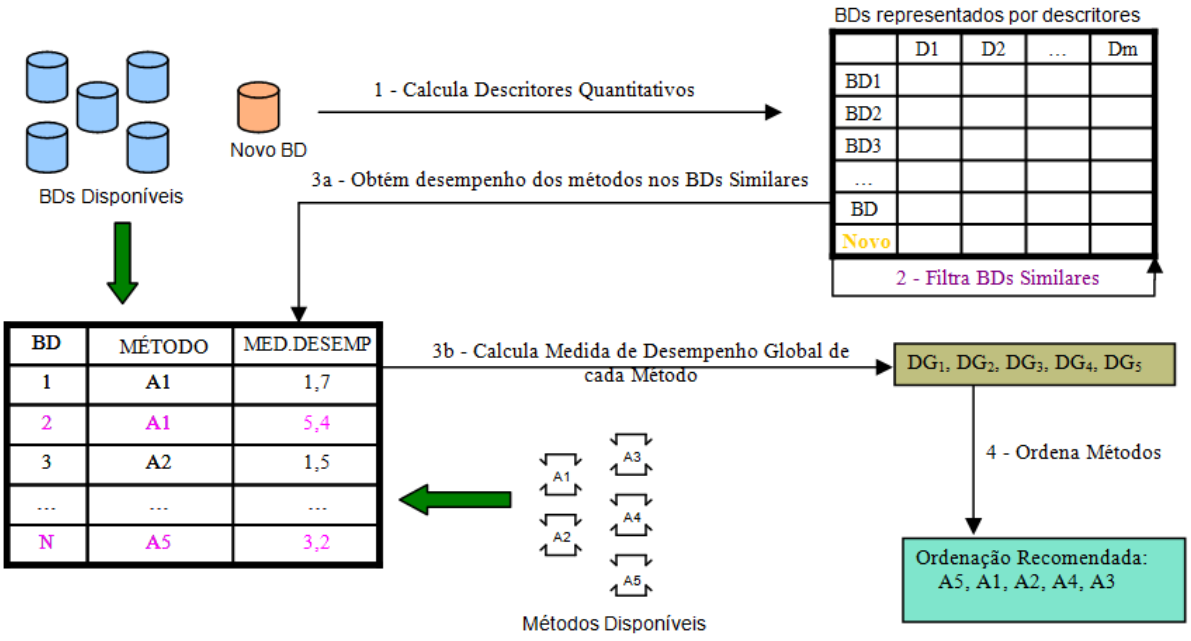


FIG. 4.1: Ambiente ASAC

O *Ambiente ASAC*, ao receber um novo conjunto de dados, realiza os seguintes passos:

- Calcula os *descritores quantitativos de conjunto de dados*, representando o novo conjunto de forma resumida pelos descritores calculados;
- Aplica alguma *técnica de filtragem de conjuntos de dados*, obtendo os *conjuntos de referência* mais similares ao novo conjunto. Esta técnica é um dos elementos configuráveis do ambiente, e utiliza os descritores quantitativos dos conjuntos de dados para o cálculo da similaridade;
- Seleciona as *medidas de desempenho local* de cada método de clusterização de dados nos *conjuntos de referência* mais similares ao novo conjunto;
- Combina as *medidas de desempenho local* dos métodos calculando uma *medida de desempenho global* para cada método;
- Ordena os métodos de clusterização de dados em função das *medidas de desempenho global* calculadas anteriormente, produzindo uma lista ordenada *ranking* dos métodos de clusterização de dados sendo interpretada pelo especialista de KDD como uma sugestão que poderá ou não ser adotada.

A operação do ambiente proposto pressupõe que os seguintes elementos tenham sido definidos e estejam disponíveis:

- *Conjuntos de Dados de Referência* (representado na figura como Bases de Dados Disponíveis) − são bases de dados onde a tarefa de clusterização de dados tenha sido previamente realizada;

- *Métodos de Mineração de Dados* − algoritmos de clusterização de dados disponíveis que tenham sido aplicados nas bases de dados selecionadas;

- *Histórico de Desempenho* − estrutura de dados que contém o desempenho de cada método de clusterização em cada um dos conjuntos de dados de referência;

- *Conjuntos de Dados Representados por Descritores Quantitativos* − estrutura de dados que contém uma representação resumida de cada conjunto de dados por meio de *descritores quantitativos*.

Conforme representado na Figura 4.1, o processamento do ambiente proposto encontra-se dividido em quatro fases distintas: 1 − Calcula descritores quantitativos, 2 − Filtra BDs Similares, 3a − Obtém desempenho dos métodos nos BDs Similares, 3b − Calcula Medida de Desempenho Global de cada Método e 4 − Ordena Métodos.

Ao receber um novo conjunto de dados sobre o qual se deseja realizar a tarefa de clusterização, o *Ambiente ASAC* executa em ordem cada uma destas fases. A descrição detalhada de cada uma delas encontra-se nas subseções.

4.2.1 CALCULA DESCRITORES QUANTITATIVOS

Segundo (GOLDSCHMIDT, 2003; SOARES, 2001; BRAZDIL, 2003; GAMA, 1995; BENSUSAN, 2000; KALOUSIS, 1999), descritores quantitativos são medidas que expressam relacionamentos existentes entre os dados de um conjunto de dados, sendo aplicados em problemas em que se deseja aferir graus de similaridade entre conjuntos. (MICHIE, 1995; ENGELS, 1998; BRAZDIL, 2003) classificam os descritores quantitativos em:

- Medidas Simples − fornecem medidas sobre a complexidade ou tamanho do problema. Exemplos: número de casos, número de atributos, número de classes etc.;

- Medidas Estatísticas − descrevem relações entre atributos quantitativos. Exemplos: assimetria, curtose, desvio-padrão etc.;

- Medidas de Teoria da Informação − utilizadas para descrever atributos categóricos. Exemplos: entropia de atributo, entropia de classe, taxa de ruído etc.

O presente trabalho, embora concebido para considerar um número arbitrário de descritores, terá seu funcionamento ilustrado com descritores de medidas simples e estatísticas na fase 1 − Calcula Descritores Quantitativos do conjunto de dados em questão, de forma que o novo conjunto passe a ser representado de maneira resumida pelos descritores calculados. Os descritores escolhidos e utilizados neste trabalho foram:

- Número de casos / exemplos: representado pela quantidade de registros / casos / linhas no conjunto de dados em questão;
- Número de atributos: representado pela quantidade de campos / características / colunas no conjunto de dados em questão;
- Número de classes: representado pela quantidade de atributos do tipo categórico no conjunto de dados em questão;
- Assimetria: segundo (fre), é o grau de desvio ou afastamento da simetria de uma distribuição, sendo calculada da seguinte forma:

$$\text{Assimetria} = \frac{X - moda}{s}$$

  onde X representa a média da distribuição e s o desvio-padrão.

- Desvio-padrão: segundo (fre), o desvio-padrão entre X_1, X_2, ..., X_n é dado por:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \overline{X}\right)^2}{N}}$$

  onde x_i é o valor corrente, $\overline{X}$ é a média da distribuição e N a quantidade de valores da distribuição.

É importante salientar que os cálculos de assimetria e desvio-padrão são obtidos através da média das assimetrias e desvios-padrão de todas as colunas do conjunto de dados em questão.

Ao final desta fase, cada conjunto de dados passa a ser representado pelos valores dos descritores quantitativos calculados.

4.2.2 FILTRA BDS SIMILARES

Nesta fase, o objetivo é aplicar uma técnica de filtragem de padrões, procurando obter os conjuntos de dados mais similares ao novo conjunto. Para isso, esta técnica utiliza o conceito de *distância* entre os padrões. Desta forma, o cálculo da distância entre dois conjuntos de dados envolve os valores dos descritores quantitativos utilizados na representação dos respectivos conjuntos.

Neste trabalho, optou-se por utilizar o algoritmo *K-NN*, dos *K-vizinhos mais próximos* (COVER, 1967) para implementar a técnica de filtragem. A função de distância utilizada foi a *Distância Euclidiana* definida por:

$$\text{dist}(d_i, d_j) = \sqrt{\sum_x \left(V_{x,d_i} - V_{x,d_j}\right)^2}$$

onde d_i e d_j são conjuntos de dados, V_{x,d_i} é o valor do descritor quantitativo x no conjunto d_i. Cada parcela do cálculo da distância refere-se a um *descritor quantitativo*. Assim sendo, quando um novo conjunto de dados é submetido à técnica de filtragem, sua distância é calculada em relação aos conjuntos de dados de referência, mas apenas os K conjuntos com menor distância são selecionados e considerados nas fases subseqüentes.

4.2.3 OBTÉM DESEMPENHO DOS MÉTODOS NOS BDS SIMILARES

O objetivo desta fase é avaliar o desempenho dos métodos de clusterização de dados de acordo com duas medidas: *medidas de desempenho local* e *medidas de desempenho global*. Segundo (GOLDSCHMIDT, 2003), a *medida de desempenho local* expressa a adequação de um método em relação a algum conjunto de dados, enquanto que a *medida de desempenho global* fornece uma avaliação comparativa entre os *desempenhos locais* de um método em relação aos demais (BRAZDIL, 2003). São exemplos de *medidas de desempenho global* a *Taxa de Distâncias Ajustada* (*ARD - Adjusted Ratio of Distances*) e a *Média das Posições* (*AR - Average Ranking*).

Neste trabalho, foram considerados como medidas de desempenho local:

- O tempo de processamento gasto para clusterizar um conjunto de dados utilizando um determinado método;

- A qualidade do resultado da clusterização. Esta qualidade será medida de acordo com os seguintes passos para cada resultado de uma clusterização:

  - Calculam-se os centróides de cada *cluster*;
  - Calcula-se a distância euclidiana de cada tupla ao seu respectivo centróide;
  - Somam-se estas distâncias, obtendo-se uma média parcial do *cluster*;
  - Calcula-se a distância média global dos *clusters*.

Conforme foi citado anteriormente, para o cálculo das medidas de desempenho global, usamos a *Taxa de Distâncias Ajustada*, adaptada da fórmula *Taxa de Taxas Ajustada* (BRAZDIL, 2003) e descrita abaixo:

$$ARD^{d_i}_{a_p,a_q} = \frac{\frac{DM^{d_i}_{a_q}}{DM^{d_i}_{a_p}}}{1 + AccD \times \log\left(\frac{T^{d_i}_{a_p}}{T^{d_i}_{a_q}}\right)}$$

onde:

- $DM^{d_i}_{a_p}$ representa o desempenho local do método a_p no conjunto d_i, calculado através da distância média global dos *clusters*. Convém mencionar que quanto maior forem as medidas de desempenho local, maior será a medida de desempenho global do método;
- $T^{d_i}_{a_p}$ representa o tempo de processamento do método a_p no conjunto de dados d_i;
- $AccD$ representa a importância relativa da qualidade do resultado em relação ao tempo (configurável). Quanto maior o valor desse parâmetro, mais significativa é a influência do tempo na comparação. Para $AccD = 0$ apenas a qualidade do resultado da clusterização é considerada. O tempo do processo, nesse caso, é desprezado.

Segundo (BRAZDIL, 2003), o numerador acima pode ser considerado como uma medida da vantagem do método a_p em relação ao método a_q, enquanto que o denominador expressa uma medida do custo do método a_p em relação ao método a_q. Como a taxa entre os tempos pode variar, utiliza-se o logaritmo para prover uma medida da ordem de magnitude da taxa.

Conforme podemos observar, esta fórmula calcula apenas a taxa de distâncias ajustada entre dos métodos. A fim de comparar o desempenho de um método em relação aos demais é necessário usar a seguinte agregação, adaptada de (BRAZDIL, 2003):

$$ARD_{a_p} = \frac{\sum_{a_q} \sqrt[n]{\pi_{d_i} ARD^{d_i}_{a_p,a_q}}}{m}$$

onde n e m representam o número de conjuntos de dados selecionados e métodos de clusterização, respectivamente. Note que, quanto menor o valor de $DM^{d_i}_{a_p}$ maior o valor de $ARD^{d_i}_{a_p,a_q}$, e que quanto maior o valor ARD de um método, melhor a posição dele na ordenação dos métodos.

4.2.4 ORDENA MÉTODOS

O processo de ordenação requer a existência de um histórico que contenha o desempenho de cada método de clusterização em cada um dos conjuntos de dados de referência. Assim sendo, a inclusão de um novo método de clusterização demanda a experimentação e a avaliação do novo método em todos os conjuntos de dados de referência. De forma análoga, a inclusão de um novo conjunto de dados requer a sua experimentação por todos os métodos de clusterização associados a tarefas de KDD viáveis no novo conjunto.

Esta fase tem como objetivo sugerir uma ordenação entre os métodos de clusterização de dados, a partir das medidas de *desempenho global* calculadas na seção anterior. Como saída do *Ambiente ASAC*, é apresentada uma lista ordenada decrescentemente (pela medida de desempenho) com os métodos de clusterização, sendo interpretada pelo especialista de KDD como uma sugestão que poderá ser adotada ou não.

4.2.5 CONSIDERAÇÕES COMPLEMENTARES

Nas seções 4.2.2 e 4.2.3 foram apresentados os elementos do *Ambiente ASAC* cujo conteúdo pode ser configurado: escolha de qual ou quais descritores quantitativos, métodos de clusterização e conjunto de dados utilizar, número k de vizinhos e $AccD$. Assim sendo, denomina-se *critério de ordenação*, qualquer configuração de elementos utilizada na implementação do processo de ordenação de métodos de clusterização de dados. Por exemplo, uma configuração para $k = 3$, $AccD = 1$, utilizando-se os descritores assimetria e desvio-padrão descreve um *critério de ordenação*. Desta forma, cada combinação de valores dos parâmetros K (do K-NN), $AccD$ e combinação entre descritores quantitativos foi consid-

erada um *critério de ordenação*, conforme será comentado no Capítulo 5. A escolha de um determinado *critério de ordenação* em detrimento de outros pode ou não conduzir a uma ordenação de métodos mais próxima da *ordenação ideal* (descrita abaixo pelo passo 4) (BRAZDIL, 2003). Desta forma, é necessário avaliar dentre os *critérios* disponíveis, qual deve ser utilizado. Esta avaliação foi realizada da seguinte forma:

1. Para cada *critério de ordenação*
2. Para cada conjunto de dados de referência
3. Aplica ambiente e constrói *ordenação recomendada*
4. Consulta *ordenação ideal*
5. Cálculo do *Coeficiente de Correlação de Spearman Local* entre as ordenações
6. Cálculo do *Coeficiente de Correlação de Spearman Global* do *critério de ordenação*

Cabe destacar as seguintes observações:

a) No laço compreendido entre os passos 2 e 5, deve ser utilizado o procedimento de testes denominado *leave-one-out* (MICHIE, 1995). Em nossa aplicação, este procedimento consistiu em considerar cada conjunto de dados de referência como um novo conjunto e os restantes como conjuntos de referência. Assim sendo, para cada critério de ordenação, tal processo se repete tantas vezes quanto for o número de conjuntos de dados de referência;

b) No passo 3, a construção da *ordenação recomendada* para um conjunto de testes d_i consiste em:

   1) Obter os K conjuntos de referência mais similares a d_i, utilizando a técnica de filtragem *K-NN*;

   2) Para cada método de mineração de dados a_p, calcular a Taxa de Distâncias Ajustada ARD_{a_p}, definida na seção 4.2.3;

   3) Ordenar os valores ARD obtidos para cada método. O maior valor de ARD confere ao respectivo método a primeira posição na *ordenação recomendada*, o segundo valor de ARD a segunda posição e assim sucessivamente.

c) No passo 4, a construção da *ordenação ideal* para um conjunto de testes d_i consiste na ordenação dos experimentos previamente realizados para construção do histórico de desempenho, que obtiveram o menor *desempenho local*;

d) No passo 5, o cálculo da *correlação* entre a *ordenação recomendada* e a *ordenação ideal* utiliza o coeficiente de *Spearman* para medir a similaridade entre as ordenações (NEAVE, 1992). O *Coeficiente de Correlação Spearman* é calculado segundo a fórmula:

$$r_s = 1 - \frac{6 \sum_{i=1}^{m} (\mathrm{rr}_i - \mathrm{ir}_i)^2}{m^3 - m}$$

onde:

- r_s é o *coeficiente de correlação de Spearman*;
- rr_i e ir_i correspondem às posições do *i-ésimo* método nas ordenações *recomendada* (rr) e *ideal* (ir);
- m é o número de métodos de clusterização considerados.

e) No passo 6, uma vez calculados todos os coeficientes de correlação entre as ordenações *recomendada* e *ideal* em todos os conjuntos, procede-se o cálculo do *coeficiente de correlação global*. O *coeficiente de correlação global* é expresso pela média dos coeficientes de *Spearman* obtidos em todos os conjuntos de referência, conforme demonstra a fórmula abaixo:

$$\mathrm{rg}_s = \frac{\sum_{i=1}^{n} \mathrm{rs}_i}{n}$$

onde n representa o número de conjunto de dados de referência considerados e rs_i representa o coeficiente de Spearman no conjunto de dados d_i.

f) A melhor correlação será a que estiver mais próxima de 1 (um). O capítulo a seguir apresenta todos os resultados obtidos e sua respectiva análise.

5 PROTÓTIPO, EXPERIMENTOS E RESULTADOS

Este capítulo está estruturado da seguinte forma: a seção 5.1 apresenta uma introdução ao *Ambiente ASAC*, citando suas fases e os processos de *configuração* e *avaliação* do ambiente. A seção 5.2 detalha o protótipo construído, e a seção 5.3 e subseções descrevem os experimentos e resultados obtidos com pelo *Ambiente ASAC*, assim como a metodologia de testes adotada.

5.1 CONSIDERAÇÕES INICIAIS

Conforme descrito no Capítulo 4, o *Ambiente ASAC* é dividido em quatro fases distintas. A primeira fase procura representar os conjuntos de dados através de descritores quantitativos. A segunda fase filtra os conjuntos de dados mais similares ao novo conjunto que se deseja analisar, através da escolha de um número k de vizinhos. A terceira fase calcula os desempenhos locais e globais dos métodos de clusterização nos conjuntos de dados similares ao novo conjunto de dados. A quarta e última fase, sugere uma ordenação entre os métodos de clusterização baseando-se nas medidas de desempenho calculadas.

Para que estas quatro fases ocorram, é necessário que o *Ambiente ASAC* passe antes por um processo de *configuração*, responsável pela definição e disponibilização de todos os elementos configuráveis do ambiente:

- Escolha dos conjuntos de dados de referência. Cabe ressaltar que, embora a presente abordagem possa ser futuramente estendida para trabalhar com atributos categóricos, na versão ora descrita somente bases de dados numéricas foram utilizadas;
- Escolha dos métodos de clusterização de dados;
- Criação do histórico de desempenho. É importante ressaltar que para criar o histórico de desempenho, é necessário realizar clusterizações com todos os métodos escolhidos em todos os conjuntos de dados de referência, avaliando os desempenhos dos métodos por meio das medidas de desempenho *locais* e *globais*;

- Escolha dos descritores quantitativos a serem utilizados na representação dos conjuntos de dados por estes descritores.

Após a *configuração* e o processamento das quatro fases, os resultados obtidos pelo *Ambiente ASAC* são interpretados por um especialista em KDD que irá avaliar, dentre os *critérios de ordenação* escolhidos, qual ou quais podem ou não conduzir a uma ordenação de métodos mais próxima da *ordenação real*. As próximas seções apresentam detalhadamente, o protótipo e os experimentos e resultados do *Ambiente ASAC*.

5.2 PROTÓTIPO

Conforme apresentado no Capítulo 4, um dos elementos que faz parte do *Ambiente ASAC* é o histórico de desempenho. Para que tal histórico fosse construído, duas ferramentas de apoio foram desenvolvidas para auxiliar no processo. A primeira delas procurou automatizar as clusterizações realizadas nos conjuntos de dados de referência, enquanto que a segunda era responsável por armazenar os resultados das clusterizações no histórico de desempenho a ser utilizado pelo ambiente.

A primeira ferramenta foi desenvolvida a partir do *WEKA* (WITTEN, 2005), um *software* de KDD *opensource* desenvolvido em Java pela Universidade de Waikato, que possui implementados os algoritmos de clusterização utilizados nesta dissertação. A utilização do WEKA em sua implementação original conduziria a um processo demorado e custoso, visto que um elevado número de experimentos teria de ser realizado manualmente. Desta forma, optou-se pelo desenvolvimento de uma ferramenta de apoio que acessasse as bibliotecas do WEKA e realizasse de forma automatizada todos os experimentos. Esta primeira ferramenta de apoio, denominada *Automated Weka*, foi desenvolvida em C# no Microsoft Visual Studio 2005®. Maiores detalhes sobre tal ferramenta encontram-se descritos no Apêndice 8.1.

A segunda ferramenta de apoio desenvolvida ficou responsável pela leitura de todos os arquivos de experimentos gerados pelo *Automated Weka*, inserindo-os no histórico de desempenho. O histórico de desempenho é uma tabela que faz parte de um modelo de banco de dados criado para o *Ambiente ASAC*. A segunda ferramenta de apoio, denominada *Banco de Experimentos*, também foi desenvolvida em C# no Microsoft Visual Studio 2005®, assim como todo o protótipo desenvolvido na dissertação. O banco de dados utilizado foi o MySQL 5.1®. Maiores detalhes sobre esta segunda ferramenta também podem

ser encontrados no Apêndice 8.1.

O *Ambiente ASAC* foi dividido em classes, sendo cada uma delas representada pelas fases descritas no Capítulo 4 desta dissertação. A classe "DescritoresQuantitativos" construída é responsável pelo cálculo dos descritores quantitativos de um novo conjunto de dados de referência conforme foi descrito na seção 4.2.1 desta dissertação. A classe "filtros" é responsável pela execução do filtro K-NN descrito na seção 4.2.2 desta dissertação. A classe "MedidasDesempenho" calcula os desempenhos locais e globais dos métodos conforme foi apresentado na seção 4.2.3, e por último a classe "spearman" ordena os métodos de clusterização e avalia este critério de ordenação conforme foi explicitado na subseção 4.2.5. Mais detalhes sobre o funcionamento do *Ambiente ASAC* encontram-se descritos no Apêndice 8.2.

5.3 EXPERIMENTOS E RESULTADOS

5.3.1 METODOLOGIA DE TESTES

Conforme descrito na seção 5.1, é necessário que o *Ambiente ASAC* passe por um processo de *configuração*, que consiste na definição dos seguintes elementos essenciais:

- Conjunto de dados de referência;
- Métodos de clusterização de dados;
- Medidas de desempenho dos métodos;
- Representação dos dados;
- Técnicas de filtragem dos conjuntos de dados.

As seções a seguir detalham a *configuração* de cada um desses elementos para os experimentos realizados.

5.3.1.1 CONJUNTO DE DADOS DE REFERÊNCIA

Foram utilizados quinze conjuntos de dados como referência, obtidos por meio de sites da comunidade científica na Internet. A tabela 5.1 descreve as propriedades de cada um dos conjuntos de dados, obtidos através de *sites* da comunidade científica na *Internet* (DEDADOS, 2008).

TAB. 5.1: Propriedades dos Conjuntos de Dados

Conjuntos de Dados	**Quantidade de Atributos**	**Quantidade de Tuplas**	**Descrição**
Balance scale	4	625	Conjunto de dados contendo características e resultados de experimentos psicológicos envolvendo crianças.
Bupa	6	345	Conjunto de dados contendo amostras de teste sanguíneo realizadas em pessoas do sexo masculino, que possuem tendência a ter doença hepática, após o consumo excessivo de álcool.
Cmc	9	1473	Conjunto de dados contendo informações sobre uma pesquisa nacional realizada na Indonésia sobre métodos anticoncepcionais. As amostras são de mulheres casadas que nunca engravidaram ou não sabiam se estavam grávidas no momento da entrevista.
Ecoli	7	336	Conjunto de dados com informações sobre classificação de proteínas.
Glass	10	214	Conjunto de dados extraídos de análises físico-químicas de amostras de vidro obtidas em um estudo sobre investigações de crimes.
Haberman	3	306	Conjunto de dados com informações sobre um estudo realizado entre 1958 e 1970 no Hospital da Universidade de Chicago, sobre pacientes que sobreviveram após uma cirurgia de câncer de mama.
Hayes roth	5	132	Conjunto de dados contendo informações pessoais sobre indivíduos entrevistados assim como sua opinião sobre um determinado assunto.
Heart Disease	13	270	Conjunto de dados contendo informações sobre características físicas observadas em pacientes incluindo diagnóstico quanto à presença ou ausência de doença cardíaca em cada paciente.
Housing	13	506	Conjunto de dados com informações sobre os valores de habitação nos subúrbios de Boston.
Iris	4	150	Conjunto de dados contendo medições de comprimento e largura de caules e pétalas e a classificação do tipo de planta.
Lenses	4	24	Conjunto de dados contendo informações sobre diversos pacientes e sobre o tipo de lentes de contato indicado para cada caso.
Pima Indians Diabetes	8	768	Conjunto de dados contendo informações sobre pacientes do sexo feminino de descendência indígena, incluindo a classificação em portadora ou não portadora de diabetes.
Tae	5	151	Conjunto de dados contendo avaliações de desempenho do Corpo Docente do Departamento de Estatística da Universidade de Wisconsin-Madison durante 3 semestres seguidos e 2 cursos de verão.
Tic-tac-toe	9	958	Conjunto de dados contendo informações sobre possíveis estados de conclusão do jogo da velha, indicando o vencedor em cada caso.
Wine	13	178	Conjunto de dados resultantes de análises químicas de vinhos cultivados numa mesma região da Itália, porém provenientes de diferentes cultivadores de videiras.

5.3.1.2 MÉTODOS DE CLUSTERIZAÇÃO DE DADOS

Os métodos de clusterização de dados adotados para os experimentos foram: K-Means (DUDA, 2000; MACQUEEN, 1967), Cobweb (FISHER, 1987), Farthest-First (HOCHBAUM, 1985) e Expectation-Maximization (EM) (DEMPSTER, 1977). A partir da escolha dos métodos, cada um deles foi estudado separadamente, analisando-se possíveis formas de parametrização. O principal objetivo era atingir o maior número possível de experimentos, variando os parâmetros dos métodos escolhidos.

Os métodos K-Means, Farthest First e EM possuem como parâmetro de entrada o número de *clusters*. Desta forma, para cada conjunto de dados de referência procurou-se variar este parâmetro de dois até o número total de tuplas do conjunto menos um. Por exemplo, para um conjunto de dados de referência com 150 tuplas, 148 experimentos foram realizados, onde o primeiro experimento consiste de 2 *clusters*, o segundo 3 *clusters*, o terceiro 4 *clusters*, e assim sucessivamente até que os algoritmos fossem experimentados com 149 *clusters*.

O método Cobweb não possui como parâmetro de entrada o número de *clusters*, mas sim um parâmetro muito específico denominado *acuity* e que foi descrito na seção 2.2.2. Este sofreu variações de 0,025 até 0,35 com intervalos de 0,025. Valores acima de 0,35 tornaram-se irrelevantes visto que o algoritmo inseria todo o conjunto de dados de referência num único *cluster*.

Outras formas de parametrização poderiam ter sido realizadas, mas por limitações de tempo não foi possível executá-las. Desta forma, foram realizados 19.428 experimentos com os conjuntos de dados normalizados (para tal foi utilizada a normalização linear), e mais 19.428 experimentos com os conjuntos de dados não-normalizados, conforme ilustra a tabela 5.2, armazenando-se os resultados no histórico de desempenho.

TAB. 5.2: Total de experimentos de clusterização realizados

Conj. de Dados de referência	**K-Means**	**Cobweb**	**Farthest-First**	**EM**	**Quantidade de Experimentos**	**Tempo Clusterização (hs)**	**Tempo histórico desempenho (hs)**
Balance scale	1246	28	1246	1246	3766	20,5	28,2
Bupa	686	28	686	686	2086	11,35	15,62
Cmc	2942	28	2942	2942	8854	48,15	66,3
Ecoli	668	28	668	668	2032	11,04	15,22
Glass	424	28	424	424	1300	7,1	9,73
Haberman	608	28	608	608	1852	10,1	13,87
Hayes-roth	260	28	260	260	808	4,4	6,05
Heart-disease	536	28	536	536	1636	8,9	12,25
Housing	1008	28	1008	1008	3052	16,6	22,85
Iris	296	28	296	296	916	5	6,86
Lenses	44	28	44	44	160	0,9	1,2
Pima-indians-diabetes	1532	28	1532	1532	4624	25,14	34,63
Tae	298	28	298	298	922	5,01	6,9
Tic-tac-toe	1912	28	1912	1912	5764	31,34	43,16
Wine	352	28	352	352	1084	5,9	8,12
Total de Experimentos					**38.856**	**211,43**	**290,96**

Note que, para atingir o número de experimentos indicados na tabela 5.2 foi necessário variar o parâmetro *número de clusters* dos métodos K-Means, Farthest-First e EM de dois até o número de tuplas de cada conjunto de dados de referência menos um. Desta forma, cada variação no parâmetro *número de clusters* foi considerada um experimento. O conjunto de dados de referência *Balance Scale* por exemplo, possui 625 *tuplas* onde foram realizados experimentos variando-se o parâmetro *número de clusters* de 2 à 624 , totalizando 623 experimentos em conjuntos de dados normalizados e 623 em conjuntos de dados não-normalizados. A tabela 5.3 abaixo ilustra mais detalhadamente, a quantidade de experimentos realizados pelos métodos K-Means, Farthest-First e EM nos conjuntos de dados de referência.

TAB. 5.3: Total de experimentos realizados pelos métodos K-Means, Farthest-First e EM em conjuntos de dados normalizados e não-normalizados

Conjunto de dados de referência	**Quantidade de Tuplas**	**Quantidade de Experimentos Normalizados**	**Quantidade de Experimentos não-normalizados**	**Total**
Balance scale	625	623	623	1246
Bupa	345	343	343	686
CMC	1473	1471	1471	2942
Ecoli	336	334	334	668
Glass	214	212	212	424
Haberman	306	304	304	608
Hayes roth	132	130	130	260
Heart Disease	270	268	268	536
Housing	506	504	504	1008
Iris	150	148	148	296
Lenses	24	22	22	44
Pima-Indians-Diabetes	768	766	766	1532
Tae	151	149	149	298
Tic-tac-toe	958	956	956	1912
Wine	178	176	176	352

Diferentemente dos outros métodos, o Cobweb (conforme descrito no Capítulo 2) possui o *acuity* como parâmetro de entrada. Para cada conjunto de dados de referência normalizado e não-normalizado foram realizados 14 experimentos onde o parâmetro *acuity* assumiu os valores 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225, 0.25, 0.275, 0.3, 0.325 e 0.35.

5.3.1.3 MEDIDAS DE DESEMPENHO DOS MÉTODOS

Posteriormente à criação do histórico de desempenho, é necessário escolher quais medidas de *desempenho local* e *global* devam ser adotadas.

As medidas de desempenho *local* utilizadas foram a distância euclidiana média entre os registros e os centróides dos respectivos *clusters*, além do tempo total necessário para a realização de cada experimento, expressando a adequação de um método em relação à um conjunto de dados de referência.

A medida de *desempenho global* adotada, a *ARD* (*Taxa de Distâncias Ajustada*), fornece uma avaliação comparativa entre os *desempenhos locais* de um método em relação aos demais, conforme foi descrito detalhadamente no Capítulo 4. De forma a avaliar a

influência do tempo na ordenação dos métodos, foram consideradas as seguintes situações:

- $AccD = 0\%$ − Neste caso não atribui-se importância ao tempo de processamento, privilegiando a média da distância euclidiana gerada pelos métodos. Útil em situações em que a qualidade e o grau de compactação dos *clusters* são prioritários;
- $AccD = 1\%$ − Neste caso busca-se o equilíbrio entre a distância euclidiana e o tempo de processamento gerado pelos métodos.

5.3.1.4 REPRESENTAÇÃO DOS DADOS

Para a representação dos dados foram escolhidos os descritores quantitativos de número de casos, número de atributos, número de classes, assimetria e desvio-padrão, descritos na seção 4.2.1 do Capítulo anterior. Ao todo foram realizadas sete combinações diferentes entre os descritores para representar cada conjunto de dados. A tabela 5.4 abaixo ilustra as combinações utilizadas entre os descritores

TAB. 5.4: Combinações entre os descritores quantitativos

	Descritores Quantitativos				
Número da Combinação	Classes	Atributos	Tuplas	Assimetria	Desvio-Padrão
1	x	x	x		
2	x	x	x		x
3		x	x	x	
4		x	x		x
5		x	x	x	x
6				x	x
7	x	x	x	x	x

A primeira combinação teve como objetivo avaliar a influência dos descritores que fornecem medidas simples sobre a complexidade ou tamanho de um problema. A segunda e quarta combinação procurou combinar os descritores de medidas simples, com um descritor de medida estatística, avaliando a relação entre os atributos de um conjunto de dados. A terceira combinação desconsiderou o descritor "número de classes" pelo fato de ser um descritor cujo cálculo só é possível em conjuntos de dados que contenham atributos de classificação, e avaliou a relação entre medidas simples e estatísticas. A quinta combinação avaliou descritores de medidas simples e estatísticas. A sexta combinação procurou avaliar apenas a influência de descritores de medidas estatísticas, e a sétima e

última combinação teve como objetivo avaliar a influência de todos os descritores sobre o conjunto de dados.

5.3.1.5 TÉCNICA DE FILTRAGEM DOS CONJUNTOS DE DADOS

Conforme comentado no Capítulo 4, o algoritmo *K-NN* (*K-vizinhos* mais próximos) foi adotado nos testes do *Ambiente ASAC* como uma técnica de filtragem dos conjuntos de dados mais similares. Os seguintes valores de K foram utilizados:

- $K = 1$ − Este valor foi escolhido por ser aproximadamente 6% do total de conjuntos de referência (GOLDSCHMIDT, 2003);
- $K = 3$ − Este valor foi escolhido pelos bons resultados obtidos com este número de vizinhos em outros experimentos de natureza similar (BRAZDIL, 2003);
- $K = 7$ − Este valor foi escolhido por ser aproximadamente 50% do total de conjuntos de referência;
- $K = 14$ − Como são ao todo 15 conjuntos de referência, o máximo de conjuntos similares a qualquer um deles exceto ele mesmo, são 14 conjuntos.

5.3.1.6 CRITÉRIOS DE ORDENAÇÃO AVALIADOS

Ao todo foram avaliados cento e dois *critérios de ordenação*, cinqüenta e seis em conjuntos de dados normalizados e cinqüenta e seis em conjuntos de dados não-normalizados. A tabela 5.5 mostra a relação de *critérios de ordenação* avaliados. Cabe destacar que, por limitações de espaço, a terceira coluna da tabela 5.5 não foi replicada para cada combinação de valores de K e *AccD*.

K	AccD	Descritores Quantitativos
1	0	- Número de Classes, Número de Atributos, Número de Tuplas - Número de Classes, Número de Atributos, Número de Tuplas, Desvio-Padrão - Número de Atributos, Número de Tuplas, Desvio-Padrão - Número de Atributos, Número de Tuplas, Assimetria, Desvio-Padrão - Número de Atributos, Número de Tuplas, Assimetria - Assimetria, Desvio-Padrão - Número de Classes, Número de Atributos, Número de Tuplas, Assimetria, Desvio-Padrão
1	1	
3	0	
3	1	
7	0	
7	1	
14	0	
14	1	

TAB. 5.5: Critérios de Ordenação

5.3.1.7 EXPERIMENTOS E RESULTADOS

Conforme exposto anteriormente, os testes realizados avaliaram os *critérios de ordenação* descritos na tabela 5.5. As tabelas 5.6 e 5.7 apresentam o coeficiente de Spearman Global obtido em cada um dos testes em conjuntos de dados normalizados e não-normalizados, respectivamente.

TAB. 5.6: Coeficientes de Spearman Global em conjuntos normalizados

	(K-vizinhos, AccD)							
Combinação de Descritores	**(1, 0)**	**(1, 1)**	**(3, 0)**	**(3, 1)**	**(7, 0)**	**(7, 1)**	**(14, 0)**	**(14, 1)**
Classes, Atributos, Tuplas	-0,067	0,067	-0,067	0,24	-0,067	0,16	-0,067	0,093
Classes, Atributos, Tuplas, Desvio-Padrão	-0,067	0,067	-0,067	0,24	-0,067	0,16	-0,067	0,093
Atributos, Tuplas, Desvio-Padrão	-0,067	0,067	-0,067	0,24	-0,067	0,16	-0,067	0,093
Atributos, Tuplas, Assimetria, Desvio-Padrão	-0,067	0,067	-0,067	0,24	-0,067	0,16	-0,067	0,093
Atributos, Tuplas, Assimetria	-0,067	0,067	-0,067	0,24	-0,067	0,16	-0,067	0,093
Assimetria, Desvio-Padrão	-0,067	0,12	-0,067	0,267	-0,067	0,13	-0,067	0,093
Classes, Atributos, Tuplas, Desvio-Padrão, Assimetria	-0,067	0,067	-0,067	0,24	-0,067	0,16	-0,067	0,093

TAB. 5.7: Coeficientes de Spearman Global em conjuntos não-normalizados

	(K-vizinhos, AccD)							
Combinação de Descritores	**(1, 0)**	**(1, 1)**	**(3, 0)**	**(3, 1)**	**(7, 0)**	**(7, 1)**	**(14, 0)**	**(14, 1)**
Classes, Atributos, Tuplas	0,722	0,741	0,722	0,787	0,722	0,817	0,722	0,833
Classes, Atributos, Tuplas, Desvio-Padrão	0,722	0,741	0,722	0,779	0,722	0,771	0,722	0,833
Atributos, Tuplas, Desvio-Padrão	0,722	0,733	0,722	0,779	0,722	0,794	0,722	0,833
Atributos, Tuplas, Assimetria, Desvio-Padrão	0,722	0,733	0,722	0,779	0,722	0,794	0,722	0,833
Atributos, Tuplas, Assimetria	0,722	0,733	0,722	0,787	0,722	0,817	0,722	0,833
Assimetria, Desvio-Padrão	0,722	0,779	0,722	0,814	0,722	0,817	0,722	0,833
Classes, Atributos, Tuplas, Desvio-Padrão, Assimetria	0,722	0,741	0,722	0,779	0,722	0,771	0,722	0,833

Primeiramente podemos observar que os resultados obtidos a partir do conjunto de dados normalizado não foram satisfatórios, visto que o coeficiente de *Spearman* global obteve aproveitamento máximo de 0,26. Isto se deve a proximidade dos valores nos conjuntos normalizados. Em certos casos, a diferença entre valores era mínima, ocorrendo em casas decimais pouco significativas, ocasionando ordenações completamente opostas, em que, por exemplo, um método que ficava em primeiro lugar na *ordenação recomendada* ficava em último na *ordenação real*. Apesar disso, a tabela 5.6 mostra que o melhor resultado obtido foi com $k = 3$ e $AccD = 1$. Outra observação importante está relacionada ao valor de *AccD*. A influência do tempo nos testes foi significativa, visto que foram obtidos coeficientes de *Spearman* positivos para os pares (1, 1), (3, 1) e (7, 1). A combinação de descritores que obteve melhor desempenho foi (Assimetria, Desvio-Padrão). Por último, um fato importante observado foi que não houve variação no coeficiente de *Spearman* quando foram utilizados os pares (14, 0) e (14, 1). Isso porque neste caso, foram considerados todos os outros conjuntos de dados como vizinhos.

Por outro lado, analisando a tabela 5.7 na qual foram utilizados conjuntos de dados não normalizados, foi obtido um aproveitamento de 0,83. O melhor resultado obtido foi a partir de 14 vizinhos com a influência no tempo de processamento ($AccD = 1$), apesar da configuração com 7 vizinhos e $AccD = 1$ também ter ficado muito próxima

(0,81). Assim como nos conjuntos normalizados, a influência do tempo nos testes foi significativa, visto que foram obtidos coeficientes de *Spearman* maiores para os pares (1, 1), (3, 1), (7, 1) e (14, 1). Igualmente nos conjuntos normalizados, não houve variação no coeficiente de *Spearman* quando foram utilizados os pares (14, 0) e (14, 1), devido ao fato de terem sido considerados todos os outros conjuntos de dados como vizinhos. Por último, diferentemente dos conjuntos de dados normalizados que tiveram melhor desempenho na combinação de descritores (Assimetria e Desvio-Padrão), os conjuntos não-normalizados tiveram desempenho igual para todas as combinações de descritores quantitativos, quando utilizados os pares (14, 1).

A título ilustrativo seguem algumas situações envolvendo a comparação entre as ordenações *real* e *recomendada* pelo *Ambiente ASAC*.

- *Critério de Ordenação A*:
  - Descritores Quantitativos: número de atributos, número de tuplas, assimetria e desvio-padrão;
  - Conjunto de dados: wine;
  - $K = 14$;
  - $AccD = 1$;

TAB. 5.8: Resultado do *Critério de Ordenação A*

Posição	**Ordenação Real**	**Ordenação Recomendada**
1	EM	EM
2	Farthest-First	Farthest-First
3	Cobweb	Cobweb
4	K-Means	K-Means

- *Critério de Ordenação B*:
  - Descritores Quantitativos: assimetria e desvio-padrão;
  - Conjunto de dados: pima-indians-diabetes;
  - $K = 3$;
  - $AccD = 0$;

TAB. 5.9: Resultado do *Critério de Ordenação B*

Posição	**Ordenação Real**	**Ordenação Recomendada**
1	EM	Farthest-First
2	K-means	EM
3	Cobweb	Cobweb
4	Farthest-First	K-Means

- *Critério de Ordenação C*:
  - Descritores Quantitativos: número de atributos, número de tuplas e desvio-padrão;
  - Conjunto de dados: glass;
  - $K = 7$;
  - $AccD = 1$;

TAB. 5.10: Resultado do *Critério de Ordenação C*

Posição	**Ordenação Real**	**Ordenação Recomendada**
1	EM	EM
2	Cobweb	Farthest-First
3	Farthest-First	Cobweb
4	K-Means	K-Means

Analisando os *critérios de ordenação* exemplificados acima, podemos observar que na tabela 5.8 houve a ocorrência de uma *correlação* igual à um, ou seja, as duas ordenações *real* e *recomendada* tiveram o mesmo resultado. No caso da tabela 5.9 é possível observar a ocorrência de uma baixa *correlação*, já que o método Farthest-First se apresentou em posições opostas (1 e 4). Por último, a tabela 5.10 apresenta uma *correlação* média, visto que os algoritmos Cobweb, K-Means e Farthest-First se apresentaram em posições misturadas.

6 CONSIDERAÇÕES FINAIS

Este capítulo tem como objetivo descrever um breve retrospecto de todo o trabalho, assim como as contribuições e trabalhos futuros proporcionados.

6.1 RETROSPECTO

A Descoberta de Conhecimento em Bases de Dados, KDD (*Knowledge Discovery in Databases*) busca encontrar e interpretar, a partir de grandes bases de dados, conhecimentos úteis através da aplicação de algoritmos e da análise de resultados. Nos últimos anos, a crescente demanda por aplicações de KDD, a quantidade insuficiente de especialistas em KDD para atender à necessidade, a existência de vários métodos de mineração de dados e a impossibilidade de experimentação de todos eles, motivou a realização desta dissertação, que teve como principais objetivos pesquisar, formalizar, implementar, configurar e avaliar um ambiente de apoio que auxiliasse na seleção de algoritmos de clusterização de dados no processo de KDD, propondo ordenações que apresentem recomendações de boas estratégias de clusterização a serem aplicadas em novas bases de dados.

A fase de pesquisa e formalização foi estruturada a partir da reunião de conceitos e idéias identificados ao longo de extensos estudos e pesquisas. Um desses conceitos é baseado na utilização de conhecimento experimental sobre o desempenho dos métodos de classificação de dados em situações anteriores, propondo ordenações destes métodos segundo seu potencial de utilização em novas situações (BRAZDIL, 2003; GOLDSCHMIDT, 2003). Outro conceito estudado foi a representação dos conjuntos de dados por meio de descritores quantitativos, que expressam os relacionamentos existentes entre os dados de um conjunto de dados. Uma vez que os conjuntos de dados sejam representados por vetores de características (compostos pelos valores dos descritores), é possível aferir graus de similaridade entre os conjuntos, a partir do cálculo das distâncias entre os respectivos vetores de características. A partir da escolha dos descritores, foi possível aplicar uma técnica de filtragem, procurando obter os conjuntos de dados mais similares a cada novo conjunto apresentado. Por último foram estudados e definidos critérios de ordenação dos métodos de clusterização de dados, em função das suas medidas de desempenho.

A fase de implementação foi marcada pelo desenvolvimento de duas ferramentas de

apoio e do *Ambiente ASAC*. A primeira ferramenta de apoio, denominada *Automated WEKA*, automatizou as clusterizações realizadas nos conjuntos de dados. A segunda ferramenta de apoio, denominada *Banco Experimentos*, foi responsável por guardar os resultados das clusterizações no histórico de desempenho que é utilizado pelo *Ambiente ASAC*. Por último, o *Ambiente ASAC* possibilitou uma estrutura que permitisse sua expansão através da incorporação de novos métodos de clusterização de dados, novos descritores quantitativos para representação de conjuntos, novos conjuntos de dados e novas técnicas de filtragem. O protótipo do *Ambiente ASAC* foi desenvolvido de forma a atender aos requisitos estabelecidos pelo diagrama conceitual fornecido no Capítulo 4.

A fase de configuração do *Ambiente ASAC* foi marcada pela realização de vários experimentos de clusterização, em que se aplicaram cada método nos quinze conjuntos de dados existentes, variando-se a parametrização de cada método em particular.

Na fase de avaliação iniciaram-se os testes no *Ambiente ASAC*, a partir dos valores propostos para os parâmetros K e $AccD$. A partir da combinação dos valores dos parâmetros e dos descritores quantitativos foi possível avaliar sua influência no processo de ordenação proposto pela presente dissertação.

6.2 CONTRIBUIÇÕES

Entre as principais contribuições proporcionadas por esta dissertação, podemos destacar:

- Concepção e implementação de um *Ambiente de Apoio à Seleção de Algoritmos de Clusterização de Dados* cuja estrutura possui as seguintes características:
  - Potencialidade para ser utilizada como ferramenta de apoio ao controle do processo de KDD em aplicações práticas;
  - Possibilidade de incorporação de novos algoritmos de clusterização de dados. A inclusão de um novo algoritmo de clusterização requer um estudo aprofundado do funcionamento do algoritmo, uma busca por implementações do mesmo, e sua adaptação ao código-fonte existente;
  - Possibilidade de incorporação de novos descritores quantitativos de conjuntos de dados. A inclusão descritores quantitativos necessita do estudo e implementação do novo descritor, assim como o seu cálculo nos conjuntos de dados de referência existentes;

  - Possibilidade de incorporação de novos filtros de conjuntos de dados similares. A inclusão de filtros requer seu estudo e sua implementação no *Ambiente ASAC*;
  - Possibilidade de incorporação de novos conjuntos de dados para estudo e análise. Para inclusão de um conjunto de dados é necessária a criação de duas tabelas, uma com os dados normalizados linearmente, e outra com os dados não-normalizados no modelo de banco de dados existente, além de sua inclusão em tabelas de dados relacionadas;

- Coleta de quinze conjuntos de dados utilizados no processo de experimentação;
- Reunião de quatro métodos de clusterização de dados em um mesmo ambiente;
- Análise crítica do ambiente;
- Identificação de uma configuração promissora para o ambiente desenvolvido, a partir da avaliação dos *critérios de ordenação* utilizados;
- Construção de históricos de desempenho que viabilizem, posteriormente, uma análise em busca de novos conhecimentos.

6.3 TRABALHOS FUTUROS

O processo de KDD é bastante abrangente e complexo, cujo foco da linha de pesquisa desta dissertação foi a criação de recursos de assistência para condução deste processo. Desta forma, diversos trabalhos futuros dentro desta linha podem ser vislumbrados. Abaixo, encontram-se alguns deles:

- O cálculo do desempenho dos métodos poderia ser aprimorado, considerando uma estimativa da complexidade computacional dos métodos envolvidos. Desta forma, um novo critério de ordenação poderia ser introduzido, priorizando métodos cuja complexidade computacional estimada fosse a menor possível;
- De forma a garantir o bom desempenho na fase de *configuração* do *Ambiente ASAC* em conjuntos de dados muito extensos, seria interessante incluir recursos e técnicas de paralelismo e distribuição de tarefas e dados;

- Utilização do histórico de desempenho como fonte de dados para buscas posteriores de novos conhecimentos sobre KDD;
- Experimentação do ambiente desenvolvido com novos conjuntos de dados, novos descritores e novos algoritmos de clusterização de dados;
- Extensão do ambiente proposto para considerar outras tarefas de KDD tais como regressão, previsão de séries temporais, detecção de desvios, entre outras;
- Investigação de estratégias de combinação de métodos de clusterização de dados na formação de Comitês de Aprendizado (PRODROMIDIS, 2000; STOLFO, 1997) em busca da melhoria do desempenho individual de cada algoritmo;
- Experimentação de descritores quantitativos normalizados;
- Variação do parâmetro $AccD$ nos experimentos.

7 REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R., GEHRKE, J., GUNOPULOS, D. e RAGHAVAN, P. Automatic subspace clustering of high dimensional data for data mining applications. *SIGMOD Rec.*, 27(2): 94–105, 1998. ISSN 0163-5808.

AGRAWAL, R., IMIELINSKI, T. e SWAMI, A. Mining association rules between sets of items in large databases. Em *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, págs. 207–216, New York, NY, USA, 1993. ACM. ISBN 0-89791-592-5.

ANKERST, M., BREUNIG, M. M., KRIEGEL, H.-P. e SANDER, J. Optics: ordering points to identify the clustering structure. Em *SIGMOD '99: Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, págs. 49–60, New York, NY, USA, 1999. ACM. ISBN 1-58113-084-8.

BENSUSAN, H., GIRAUD-CARRIER, C. e PFAHRINGER, B. What works well tells us what works better. Em *Proceedings of ICML'2000 workshop on What Works Well Where*, págs. 1–8. ICML'2000, June 2000. URL `http://www.cs.bris.ac.uk/Publications/Papers/1000469.pdf`.

BERNSTEIN, A., HILL, S. e PROVOST, F. Intelligent assistance for the data mining process: An ontology-based approach. IS-02-02. Stern School of Business, New York University, 2002. URL `http://hdl.handle.net/2451/14156`.

BRAZDIL, P. Data transformation and model selection by experimentation and meta-learning. Em C. GIRAUD-CARRIER, M. H. E., editor, *ECML'98 Workshop Notes, Upgrading Learning to the Meta-Level: Model Selection and Data Transformation*. Techische Universitaet Chemnitz, Chemnitzer Informatik-Berichte CSR-98-02, 1998. URL `http://www.liaad.up.pt/pub/1998/Bra98`.

BRAZDIL, P. e SOARES, C. A comparison of ranking methods for classification algorithm selection. Em *ECML '00: Proceedings of the 11th European Conference on Machine Learning*, págs. 63–74, London, UK, 2000. Springer-Verlag. ISBN 3-540-67602-3.

BRAZDIL, P. B., SOARES, C. e COSTA, J. P. D. Ranking learning algorithms: Using ibl and meta-learning on accuracy and time results. *Machine Learning*, 50(3):251–277, 2003. ISSN 0885-6125.

BRODLEY, C. E. Addressing the selective superiority problem: Automatic algorithm/model class selection. Em *Proceedings of the Tenth International Conference on Machine Learning*, págs. 17–24, São Francisco, Califórnia, 1995. Morgan Kaufmann.

COLE, R. M. Clustering with genetic algorithms. Dissertação de Mestrado, University of Western Australia, Nedlands 6907, Australia, 1998. URL http://citeseer.ist.psu.edu/435115.html.

COVER, T. M. e HART, P. E. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1053964.

CURRAN, D. e O'RIORDAN, C. Applying evolutionary computation to designing neural networks: A study of the state of the art. Em *Report number nuig-it-220202. Technical Report, Dept. of IT*. NUI, Galway, 2002.

DA SILVA, E. B. *Agrupamento Semi-Supervisionado de Documentos XML*. Tese de Doutorado, COPPE/UFRJ, Janeiro 2006.

DASGUPTA, S. e LONG, P. M. Performance guarantees for hierarchical clustering. *J. Comput. Syst. Sci.*, 70(4):555–569, 2005. ISSN 0022-0000.

DAVIS, L. D. e MITCHELL, M. Handbook of genetic algorithms. *Van Nostrand Reinhold*, 1991.

DE DADOS, C. Conjuntos de dados utilizados nos experimentos, Junho 2008. URL ftp://ftp.ics.uci.edu/pub/machine-learning-databases.

DEMPSTER, A. P., LAIRD, N. M. e RUBIN, D. B. Maximum likelihood from incomplete data via the algorithm em. *Journal of the Royal Statistical Society*, Série B, v. 39:1–38, 1977.

DI CARLANTONIO, L. M. Novas metodologias para clusterização de dados. Dissertação de Mestrado, COPPE/UFRJ, 2001.

DUDA, R. O., HART, P. E. e STORK, D. G. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000. ISBN 0471056693.

ENGELS, R., LINDNER, G. e STUDER, R. A guided tour through the data mining jungle. 3rd International Conference on Knowledge Discovery in Databases, KDD'97, Newport Beach, 1997. URL http://digbib.ubka.uni-karlsruhe.de/volltexte/62397.

ENGELS, R. e THEUSINGER, C. Using a data metric for preprocessing advice for data mining applications. Em *In Proceedings of the European Conference on Artificial Intelligence (ECAI-98*, págs. 430–434. John Wiley & Sons, 1998.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G. e SMYTH, P. From data mining to knowledge discovery: an overview. págs. 1–34, 1996.

FERREIRA, G., ARAUJO, R., ORAIR, G., GONÇALVES, L., GUEDES, D., FERREIRA, R., FURTADO, V. e JR., W. M. Paralelização eficiente de um algoritmo de agrupamento hierárquico. 2005. URL http://www.lbd.dcc.ufmg.br/wamd2005/WAMD_8.pdf.

FISHER, D. H. Knowledge acquisition via incremental conceptual clustering. *Mach. Learn.*, 2(2):139–172, 1987. ISSN 0885-6125.

FÜRNKRANZ, J. e PETRAK, J. An evaluation of landmarking variants. Em *Proceedings of the ECML/PKDD Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning (IDDM-2001*, págs. 57–68, 2001.

GAMA, J. e BRAZDIL, P. Characterization of classification algorithms. Em *EPIA '95: Proceedings of the 7th Portuguese Conference on Artificial Intelligence*, págs. 189–200, London, UK, 1995. Springer-Verlag. ISBN 3-540-60428-6.

GENNARI, J. H., LANGLEY, P. e FISHER, D. Models of incremental concept formation. *Artif. Intell.*, 40(1-3):11–61, 1989. ISSN 0004-3702.

GIBSON, J., TEKINER, F., HALFPENNY, P., NAZROO, J., FAGAN, C., PROCTER, R. e LIN, Y. Ncess project: Data mining for social scientists. Em *Proceedings of e-Social Science '07*, Michigan, US, 2007.

GOLDBERG, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989. ISBN 0201157675.

GOLDSCHMIDT, R. R., PASSOS, E. L. P. e VELLASCO, M. B. R. A goal definition assistant applied to kdd process. Em ARABNIA, H. R. E MUN, Y., editores, *IC-AI*, págs. 631–637. CSREA Press, 2002a. ISBN 1-892512-26-2. URL `http://dblp.uni-trier.de/db/conf/icai/icai2002-2.html%GoldschmidtPV02`.

GOLDSCHMIDT, R. R. *Assistência Inteligente à Orientação do Processo de Descoberta de Conhecimento em Bases de Dados*. Tese de Doutorado, Engenharia Elétrica - PUC-Rio, 2003.

GOLDSCHMIDT, R. R. e PASSOS, E. *Data Mining: Um Guia Prático - Conceitos, técnicas, ferramentas, orientações e aplicações*. Editora Campus, 2005. ISBN 8535218777.

GOLDSCHMIDT, R. R., PASSOS, E. P. L. e VELLASCO, M. B. R. An action plan definition assistant in kdd process. Em *Proceedings of the Second International Conference on Artificial Intelligence and Applications*, Málaga, Espanha, 2002b.

GOLDSCHMIDT, R. R., PASSOS, E. P. L. e VELLASCO, M. B. R. Assistance in kdd goal definition process. Em *Proceedings of the International Conference on Control and Automation*, págs. 234–235, Xiamen, China, 2002c. ISBN 0-7803-7412-6.

GONÇALVES, B. L. Modelos neuro-fuzzy hierárquicos bsp para classificação de padrões e extração de regras fuzzy em bancos de dados. Dissertação de Mestrado, Pontifícia Universidade Católica do Rio de Janeiro, 2001.

HALKIDI, M., BATISTAKIS, Y. e VAZIRGIANNIS, M. On clustering validation techniques. *J. Intell. Inf. Syst.*, 17:107–145, 2001. ISSN 0925-9902.

HAN, J. e KAMBER, M. *Data mining: concepts and techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000. ISBN 1-55860-489-8.

HAYES-ROTH, B. An architecture for adaptive intelligent systems. *Artif. Intell.*, 72(1-2): 329–365, 1995. ISSN 0004-3702.

HELLERSTEIN, J. M., AVNUR, R., CHOU, A., HIDBER, C., OLSTON, C., RAMAN, V., ROTH, T. e HAAS, P. J. Interactive data analysis: The control project. *Computer*, 32(8):51–59, 1999. ISSN 0018-9162.

HOCHBAUM, D. S. e SHMOYS, D. B. A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, 10(2):180–184, 1985. ISSN 0364-765X.

HRUSCHKA, E. R. e F. EBECKEN, N. F. A genetic algorithm for cluster analysis. *Intell. Data Anal.*, 7(1):15–25, 2003. ISSN 1088-467X.

JAIN, A. K. e DUBES, R. C. *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988. ISBN 0-13-022278-X.

JENSEN, D., DONG, Y., LEGNER, B. S., MCCALL, E. K., OSTERWEIL, L. J., STANLEY M. SUTTON, J. e WISE, A. Coordinating agent activities in knowledge discovery processes. Em *WACC '99: Proceedings of the international joint conference on Work activities coordination and collaboration*, págs. 137–146, New York, NY, USA, 1999. ACM. ISBN 1-58113-070-8.

KALOUSIS, A. e THEOHARIS, T. Noemon: Design, implementation and performance results of an intelligent assistant for classifier selection. *Intelligent Data Analysis*, 3: 319–337, 1999.

KELLER, J., PATERSON, I. e BERRER, H. An integrated concept for multi-criteria-ranking of data-mining algorithms. Em *Proceedings of the ECML-2000 Workshop on Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination*, 2000.

KERBER, R., BECK, H., ANAND, T. e SMART, B. Active templates: Comprehensive support for the knowledge discovery process. Em *KDD*, págs. 244–248, 1998.

KOHONEN, T., editor. *Self-organizing maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1997. ISBN 3-540-62017-6.

LARSON, J. A., NAVATHE, S. B. e ELMASRI, R. A theory of attributed equivalence in databases with application to schema integration. *IEEE Trans. Softw. Eng.*, 15(4): 449–463, 1989. ISSN 0098-5589.

LENAT, D. B. e BROWN, J. S. Why am an euisko appear to work. *Artif. Intell.*, 23(3): 269–294, 1984. ISSN 0004-3702.

LI, C. Extending iterate conceptual clustering scheme in dealing with numeric data. Dissertação de Mestrado, 1995. URL `http://citeseer.ist.psu.edu/168775.html`.

LIPSCHUTZ, S. *Topologia Geral*. McGraw-Hill, Rio de Janeiro, 1973.

LITTLE, R. J. A. e RUBIN, D. B. *Statistical analysis with missing data*. John Wiley & Sons, Inc., New York, NY, USA, 1986. ISBN 0-471-80254-9.

LIVINGSTON, G., ROSENBERG, J. M. e BUCHANAN, B. G. Closing the loop: An agenda - and justification-based framework for selecting the next discovery task to perform. Em *ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining*, págs. 385–392, Washington, DC, USA, 2001a. IEEE Computer Society. ISBN 0-7695-1119-8.

LIVINGSTON, G. R., ROSENBERG, J. M. e BUCHANAN, B. G. A framework for autonomously performing knowledge discovery in databases. Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000. URL `http://citeseer.ist.psu.edu/305603.html`.

LIVINGSTON, G. R. *A framework for autonomous knowledge discovery from databases*. Tese de Doutorado, Pittsburgh, PA, USA, 2001b. Adviser-Bruce G. Buchanan.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. Em CAM, L. M. L. E NEYMAN, J., editores, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, págs. 281–297. University of California Press, 1967.

MAES, P. Agents that reduce work and information overload. *Commun. ACM*, 37(7): 30–40, 1994. ISSN 0001-0782.

MICHIE, D., SPIEGELHALTER, D. J., TAYLOR, C. C. e CAMPBELL, J. *Machine Learning, Neural and Statistical Classifications*. Ellis Horwood, 1995.

MITCHELL, T. M. *Machine Learning*. McGraw-Hill Science/Engineering/Math, March 1997. ISBN 0070428077.

MORIK, K. The representation race - preprocessing for handling time phenomena. Em *ECML '00: Proceedings of the 11th European Conference on Machine Learning*, págs. 4–19, London, UK, 2000. Springer-Verlag. ISBN 3-540-67602-3.

NEAVE, H. R. e WORTHINGTON, P. L. *Distribution Free Tests*. Routledge, 1992.

PEIXOTO, J. J. L. Algoritmos de aprendizagem de contexto em computação ubíqua: Avaliação para o caso de estudo em pdas. Dissertação de Mestrado, Universidade de Coimbra.

PFAHRINGER, B., BENSUSAN, H. e GIRAUD-CARRIER, C. G. Meta-learning by landmarking various learning algorithms. Em *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, págs. 743–750, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-707-2.

PRODROMIDIS, A. L., CHAN, P. K. e ET AL. Meta-learning in distributed data mining systems: Issues and approaches. *Advances in Distributed and Parallel Knowledge Discovery, AAAI Press, Menlo Park*, págs. 81–114, 2000.

REICH, Y. *Building and improving design systems: a machine learning approach.* Tese de Doutorado, Pittsburgh, PA, USA, 1991.

RUSSELL, S. J. e NORVIG, P. *Artificial intelligence: a modern approach.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1995. ISBN 0-13-103805-2.

SHEIKHOLESLAMI, G., CHATTERJEE, S. e ZHANG, A. Wavecluster: A multi-resolution clustering approach for very large spatial databases. Em *VLDB '98: Proceedings of the 24rd International Conference on Very Large Data Bases*, págs. 428–439, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-566-5.

SHEN, W.-M. e LENG, B. A metapattern-based automated discovery loop for integrated data mining-unsupervised learning of relational patterns. *IEEE Trans. on Knowl. and Data Eng.*, 8(6):898–910, 1996. ISSN 1041-4347.

SHOENFIELD, J. R. *Mathematical Logic.* Addison-Wesley, Reading, MA, 1967.

SMULLYAN, R. M. *First-order logic.* Springer-Verlag, 1971.

SOARES, C. e BRAZDIL, P. Zoomed ranking: Selection of classification algorithms based on relevant performance information. Em *PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, págs. 126–135, London, UK, 2000a. Springer-Verlag. ISBN 3-540-41066-X.

SOARES, C., BRAZDIL, P. e COSTA, J. P. Measures to evaluate rankings of classification algorithms. Em KIERS HAL, RASSON JP, G. P. S. M., editor, *Proceedings of the 7th Conference of the International-Federation-of-Classification-Societies*, Studies in Classification, data analysis, and knowledge organization, págs. 119–124, Namur, Belgium, July 2000b. Springer. URL `http://www.liaad.up.pt/pub/2000/SBC00`. ISIProc.

SOARES, C., COSTA, J. e BRAZDIL, P. Improved statistical support to matchmaking: Rank correlation taking rank importance into account. Em *VII Jornadas de Classificação de Dados*, págs. 72–75, Fevereiro 2001.

SPILIOPOULOU, M., FAULSTICH, L. C., KALOUSIS, A. e THEOHARIS, T. An intelligent assistant for classifier selection. págs. 90–97, 1998.

STOLFO, S., PRODROMIDIS, A. L., TSELEPIS, S. e LEE, W. Jam: Java agents for meta-learning over distributed databases. *Proceedings of the 3rd International Conference on Knowledge*, 1997.

SUYAMA, A. e YAMAGUCHI, T. Specifying and learning inductive learning systems using ontologies. Em *AAAI Workshop Methodology of Applying Machine Learning: Problem Definition, Task Decomposition and Technique Selection*, págs. 29–36, 1998.

TSAI, C.-F., WU, H.-C. e TSAI, C.-W. A new data clustering approach for data mining in large databases. Em *ISPAN '02: Proceedings of the 2002 International Symposium on Parallel Architectures, Algorithms and Networks*, pág. 315, Washington, DC, USA, 2002. IEEE Computer Society.

VAN DER MERWE, D. W. e ENGELBRECHT, A. P. Data clustering using particle swarm optimization. Em *Evolutionary Computation, 2003. CEC '03. The 2003 Congress on*, volume 1, págs. 215–220, August-December Dec. 2003.

VERDENIUS, F. e ENGELS, R. A process model for developing inductive applications. Em *Proceedings of the Seventh Belgian-Dutch Conference on Machine Learning*, págs. 119–128, 1998. URL `http://citeseer.ist.psu.edu/220305.html`.

WEISS, S. M. e INDURKHYA, N. *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998. ISBN 1-55860-403-0.

WITTEN, I. H. e FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005. ISBN 0120884070.

WOLPERT, D. H. The lack of a priori distinctions between learning algorithms and the existence of a priori distinctions between learning algorithms. *Neural Computation*, 8 (7):1341–1420, 1996.

WOOLDRIDGE, M. e JENNINGS, N. R. Agent theories, architectures, and languages: a survey. Em *ECAI-94: Proceedings of the workshop on agent theories, architectures, and languages on Intelligent agents*, págs. 1–39, New York, NY, USA, 1995. Springer-Verlag New York, Inc. ISBN 3-540-58855-8.

YOO, J. e YOO, S. Concept formation in numeric domains. Em *CSC '95: Proceedings of the 1995 ACM 23rd annual conference on Computer science*, págs. 36–41, New York, NY, USA, 1995. ACM. ISBN 0-89791-737-5.

ZAKI, M. J. Parallel and distributed data mining: An introduction. Em *Revised Papers from Large-Scale Parallel Data Mining, Workshop on Large-Scale Parallel KDD Systems, SIGKDD*, págs. 1–23, London, UK, 2000. Springer-Verlag. ISBN 3-540-67194-3.

8 APÊNDICES

8.1 APÊNDICE 1: DESCRIÇÃO DAS FERRAMENTAS DE APOIO

8.1.1 AUTOMATED WEKA

A primeira ferramenta de apoio desenvolvida para auxiliar a abordagem proposta nesta dissertação chama-se *Automated WEKA*. Seu principal objetivo é automatizar as clusterizações de todos os métodos a todos os conjuntos de dados existentes, variando-se os parâmetros de cada método a cada novo experimento. Esta ferramenta foi desenvolvida em C#, e acessa as bibliotecas do software *WEKA* para rodar os experimentos. A figura 8.1 mostra a ferramenta sendo executada.

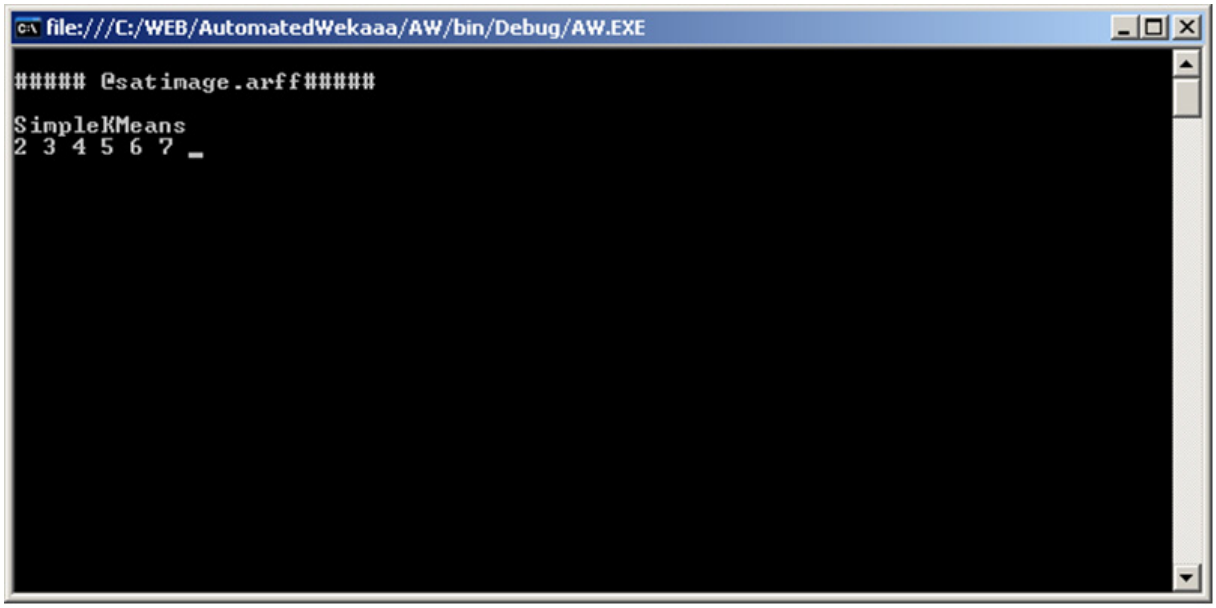


FIG. 8.1: Ferramenta de apoio *Automated Weka* em execução

Inicialmente, a ferramenta busca por arquivos de conjuntos de dados no formato ".arff", aceito pelo software *WEKA* e, para cada arquivo, executa os métodos de clusterização de acordo com a parametrização definida no Capítulo 5. Ao final, são geradas pastas para cada conjunto de dados, contendo todos os experimentos realizados no mesmo em formato ".txt". As figuras 8.2 e 8.3 mostram as pastas e os arquivos ".txt" gerados respectivamente.

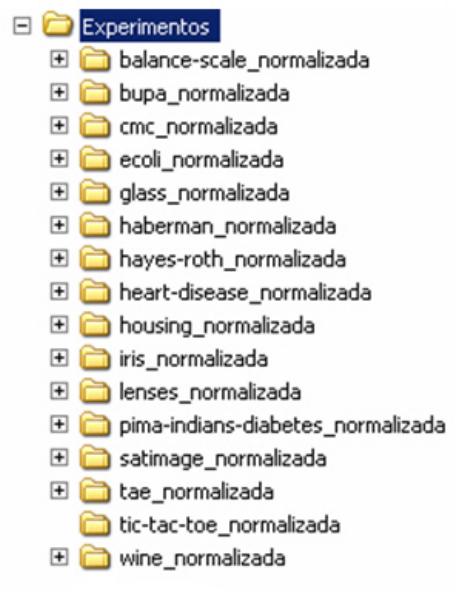


FIG. 8.2: Pastas geradas após a realização dos experimentos de clusterização

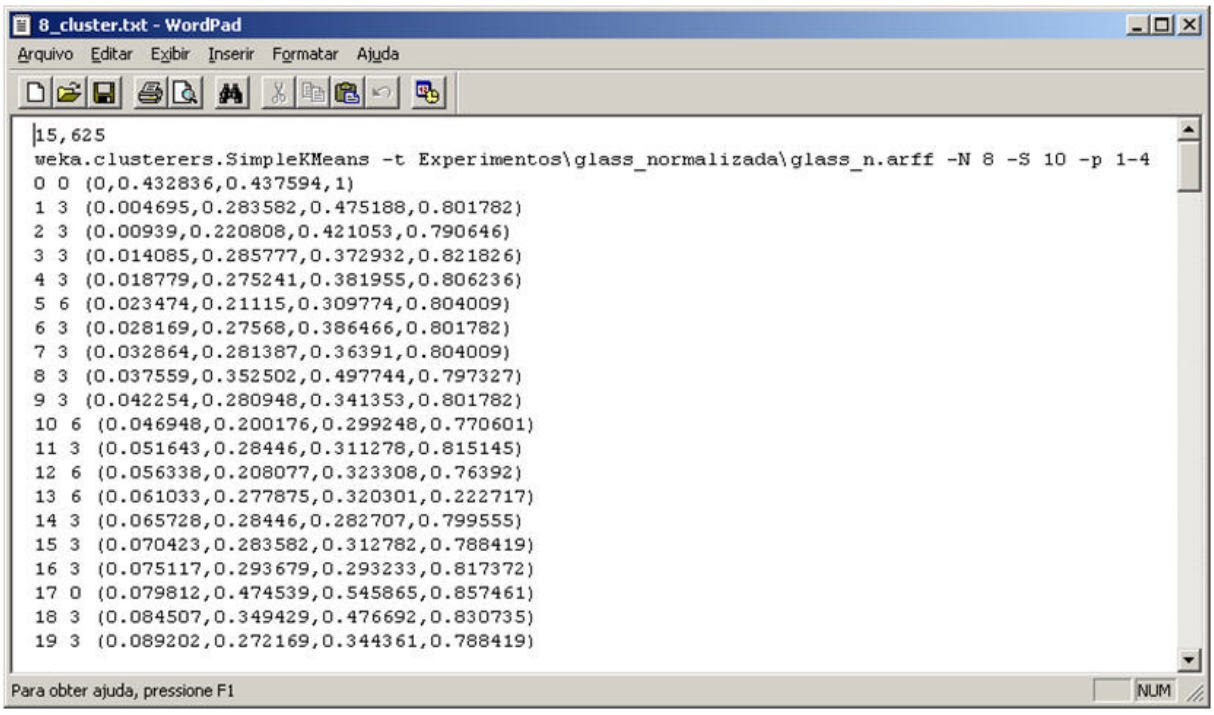


FIG. 8.3: Arquivos no formato ".txt" gerados pela clusterização

O arquivo ".txt" gerado possui algumas características relevantes. A primeira linha contém a informação do tempo gasto pelo algoritmo para realizar a clusterização. A segunda linha contém informações sobre a parametrização utilizada num determinado experimento e, a terceira linha em diante apresenta: na primeira coluna o número da

tupla do conjunto de dados em questão; na segunda coluna, o número de identificação do cluster onde a tupla foi alocada e, a terceira e última coluna, os valores que representam a tupla em questão.

8.1.2 BANCO DE EXPERIMENTOS

A segunda ferramenta de apoio desenvolvida para auxiliar a abordagem proposta nesta dissertação chama-se *Banco de Experimentos*. Ela é responsável pela leitura de todos os arquivos de experimentos gerados pela ferramenta *Automated WEKA*, inserindo-os no histórico de desempenho. O histórico de desempenho é uma tabela que faz parte de um modelo de banco de dados (ilustrado na seção 8.1.3) criado para o *Ambiente ASAC*. O *Banco de Experimentos* também foi desenvolvido em C# no Microsoft Visual Studio 2005, utilizando o Banco de Dados MySQL 5.1. A figura 8.4 ilustra o *Banco de Experimentos* em execução.

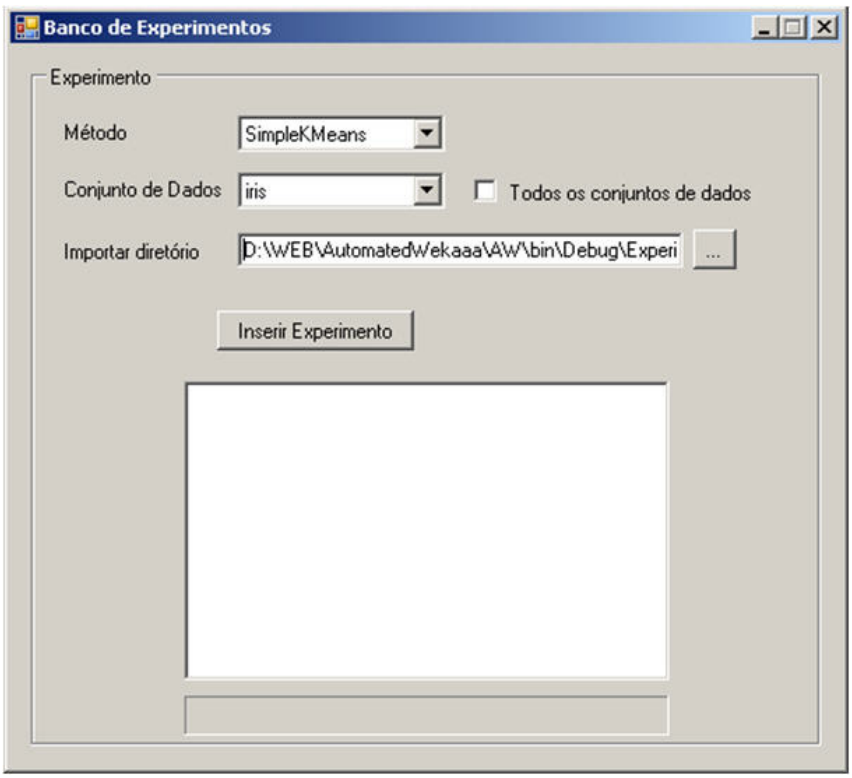
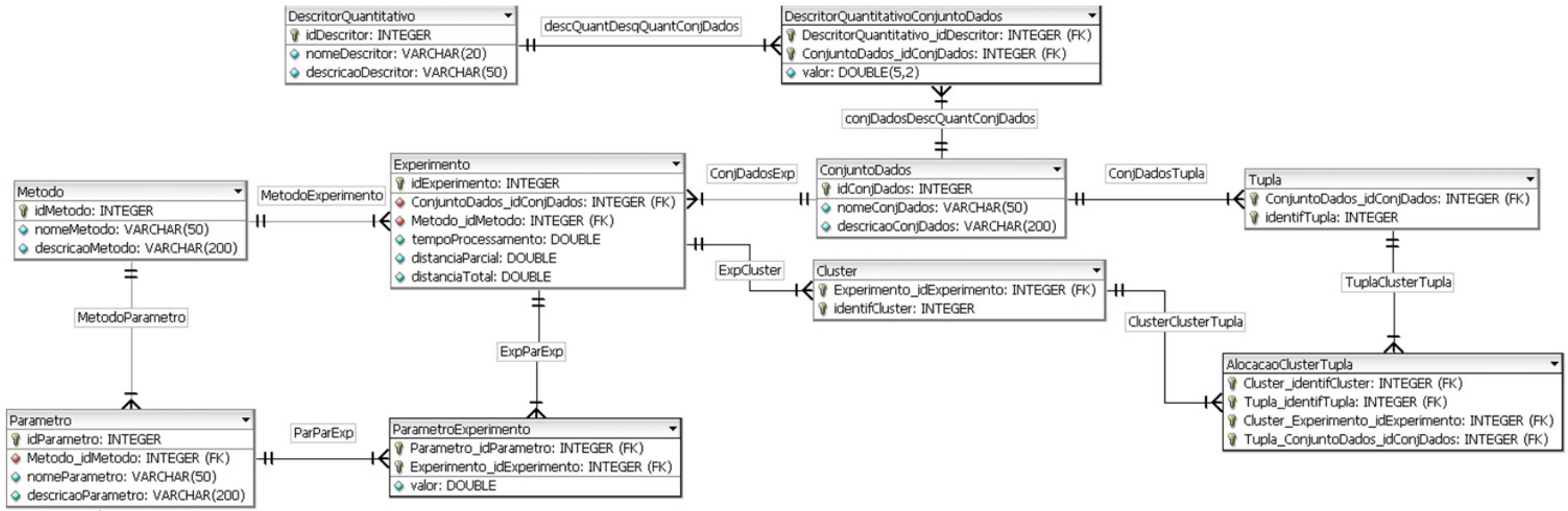


FIG. 8.4: Ferramenta de Apoio Banco de Experimentos em execução

O *Banco de Experimentos* pode inserir experimentos de um determinado método sobre um conjunto de dados ou, inserir todos os experimentos realizados por um determinado método.

8.1.3 MODELAGEM DO BANCO DE DADOS



Conforme detalhado no Capítulo 4 desta dissertação, a abordagem conceitual proposta neste trabalho está dividida em quatro fases distintas e essenciais: 1 − Calcula descritores quantitativos, 2 − Filtra BDs Similares, 3a − Obtém desempenho dos métodos nos BDs Similares, 3b − Calcula Medida de Desempenho Global de cada Método e 4 − Ordena Métodos. Ao implementar o ambiente, cada fase foi representada por uma tela, conforme será ilustrado nas imagens abaixo.

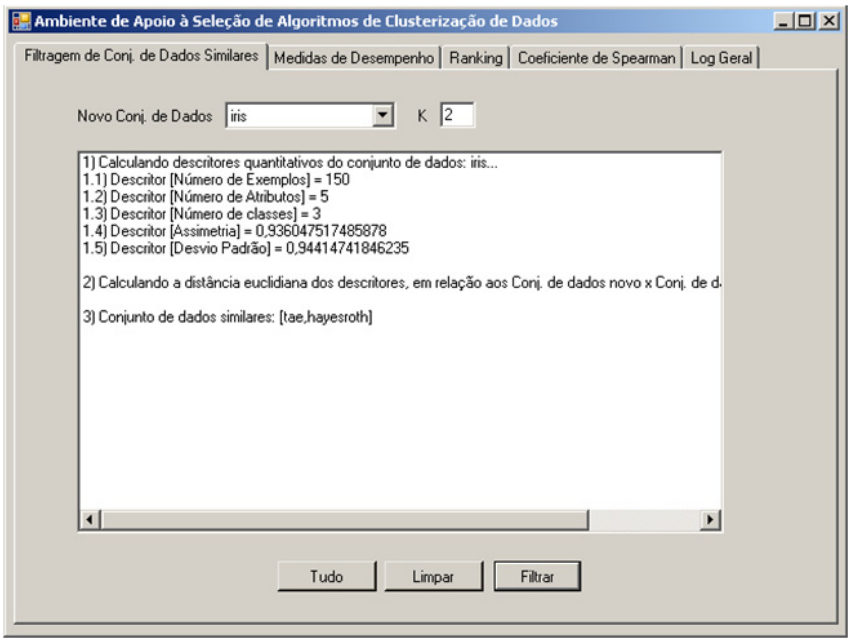


FIG. 8.5: Fase de cálculo dos descritores quantitativos e filtragem dos conjuntos de dados similares

Na figura 8.5 é possível identificar as fases 1 e 2 do *Ambiente ASAC* onde, após a seleção do novo conjunto de dados, e do número k de vizinhos, ao pressionar sobre o botão "Filtrar", obtém-se o cálculo de todos os descritores quantitativos do novo conjunto de dados e dos respectivos conjuntos de dados similares. Um detalhe importante que deve ser mencionado é se ao invés da opção "Filtrar" optar-se por "Tudo", todas as quatro fases do *Ambiente ASAC* são executadas sucessivamente (facilidade implementada para economia de tempo). A opção "Limpar" prepara o Ambiente para uma nova execução.

A figura 8.6 ilustra a fase 3 em que, após a escolha do valor de $AccD$, calculam-se as

medidas de *desempenho local* e *desempenho global* dos métodos de clusterização.

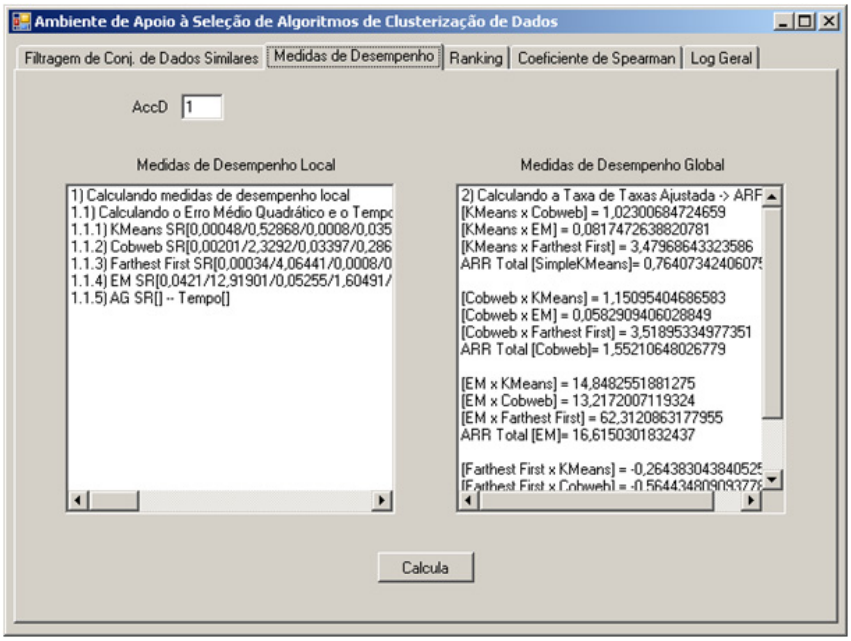


FIG. 8.6: Fase de cálculo das medidas de *desempenho local* e *desempenho global*

A figura 8.7 ilustra a fase 4 onde o *Ambiente ASAC* sugere uma ordenação (*Ranking Recomendado*) e a compara com a *ordenação real* (*Ranking Real*).

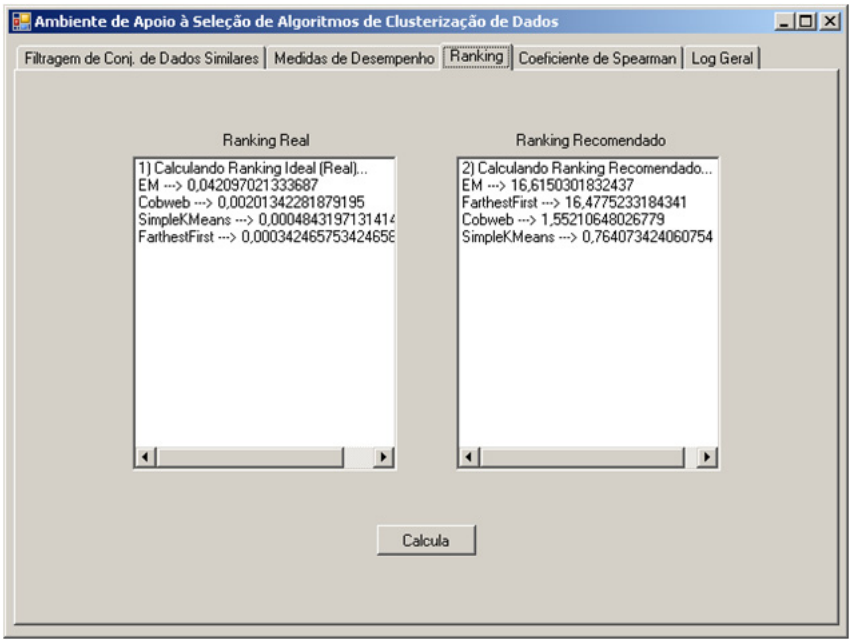


FIG. 8.7: Ordenação dos Métodos de Clusterização

A figura 8.8 ilustra o cálculo do *Coeficiente de Spearman Global*, que irá avaliar a

similaridade entre as ordenações.

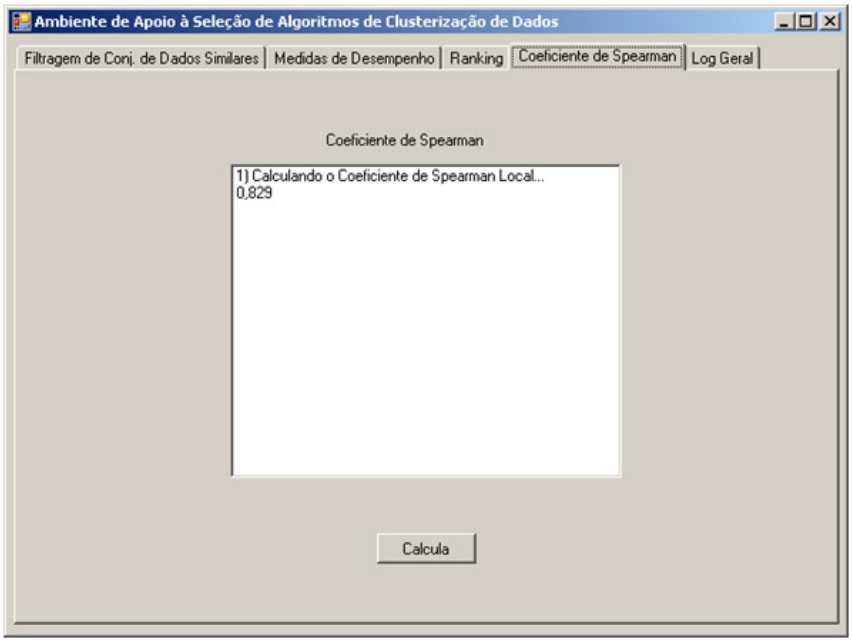


FIG. 8.8: Cálculo do Coeficiente de Spearman

Finalmente, a Figura 8.9 ilustra um *log* com os valores calculados por todas as fases do *Ambiente ASAC*, registrando todos os resultados em um arquivo ".txt".

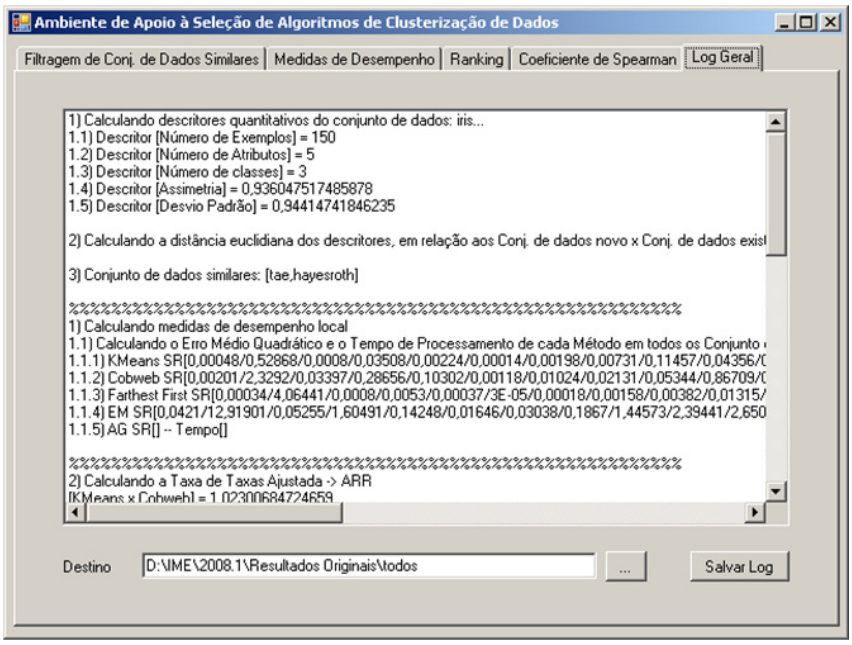


FIG. 8.9: *Log* das quatro fases do *Ambiente ASAC*

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo