Instituto de Ciências Matemáticas e de Computação

ISSN - 0103-2569

MineSet™, Ferramenta de *Data Mining*

Solange Oliveira Rezende
Marcos Ferreira de Paula
Luiz Fernando Figueiredo

Nº 80

RELATÓRIOS TÉCNICOS DO ICMC
Versão 1.0

São Carlos
Dez/1998

Conteúdo

Índice de Figuras

1. Introdução

O estágio atingido nos últimos anos pelas tecnologias de hardware e software aplicadas ao armazenamento, à manutenção e ao compartilhamento de dados permitiu aos usuários de bancos de dados (empresas, pesquisadores e órgãos governamentais) manter quantidades cada vez maiores de informação em suas bases de dados. Esse alto volume de dados acabou por exceder dramaticamente a capacidade humana de análise e compreensão, ainda que utilizando métodos como planilhas eletrônicas e ambientes de consultas ad hoc.

Com o crescimento da capacidade de geração e armazenamento de dados, o processo de extração de conhecimento tem se tornado um problema. Para solucionar este problema surgiu uma tecnologia recente para extrair conhecimento dos grandes volumes de informação, chamada *Knowledge Discovery in Database* (KDD). *Data Mining* (DM) é uma das principais etapas que fazem parte do processo KDD.

Este trabalho se apresenta como um guia para utilização dos utilitários de *Data Mining* da ferramenta MineSet™ versão 2.6, explicando as principais opções disponíveis na sua interface. É importante ressaltar que este relatório foi realizado visando a utilização da interface gráfica, não se preocupando com a possibilidade de utilização dos utilitários através de linha de comando, nem com o modo que os algoritmos de *Data Mining* e as ferramentas de visualização trabalham internamente.

A organização do trabalho está feita da seguinte forma: A seção 2 apresenta uma visão geral do MineSet™ e sua estrutura, mostrando seus componentes e as relações existentes entre eles. A seção 3 apresenta o Módulo de Controle Centralizado, o qual é a interface do usuário com o sistema. Na seção 4 são descritas detalhadamente os utilitários de *Data Mining* que compõem o MineSet™ versão 2.6 Beta 3. A seção 5 cita a importância das ferramentas de visualização. A seção 6 apresenta as considerações finais deste trabalho.

2. MineSet™ e sua Estrutura

O MineSet™ da Silicon Graphics é uma ferramenta que possui vários utilitários para efetuar DM e ferramentas gráficas que juntos proporcionam uma maneira relativamente fácil de extrair e compreender conhecimento embutido nos dados.

MineSet™ suporta a geração e análise de regras de associação e modelos de classificação, usados para previsão, evolução, segmentação e tendências, combinados com visualização animada e interativa. Existem ainda outras funções adicionais no MineSet™, tais como: discretização (por intervalos, por distribuição de freqüências, sendo geradas automaticamente ou declarada pelo usuário), seleção, criação ou eliminação de coluna, filtragem de instâncias no conjunto de dados e troca de tipo de dados.

Desta forma o MineSet™ provê funcionalidades para construir e aplicar modelos de classificação, regressão, clusterização, associação, possibilitando aos usuários a vantagem de extrair padrões de um conjunto de dados. MineSet™ suporta seis modelos de predição:

Regras de Associação, Árvores de Decisão, Árvores de Opção, Árvores de Regressão, Tabelas de Opção, e Classificadores de Evidência (Simples-Bayes). Ainda com os visualizadores especiais, os usuários podem explorar os modelos, determinar precisão e fazer perguntas do tipo *what-if* .

MineSet™ é constituído de três componentes básicos, que podem ser observados na Figura 1, que são:

Figura 1: Componentes básicos do MineSet™

- Módulo de controle centralizado, que consiste de uma ferramenta com interface gráfica chamada *Tool Manager* e um processo chamado de *DataMover*, que é executado no servidor.
- Utilitários de *Data Mining*, constituídos de nove ferramentas:
  - Gerador de Regras de Associação.
  - Classificador e Indutor de Árvore de Decisão.
  - Classificador e Indutor de Árvore de Opção.
  - Classificador e Indutor de Evidência.
  - *Column Importance*.
  - Discretizador Automático.
  - Classificador e Indutor de Tabelas de Decisão.
  - Classificador e Gerador de Árvore de Regressão.
  - Gerador de Cluster.
- Ferramentas de visualização, as quais permitem a visualização dos dados bem como dos resultados dos utilitários. São ao todo dez ferramentas:
  - *Tree Visualizer*: visualizador das árvores de decisão, opção e regressão.
  - *Scatter Visualizer*: visualizador bidimensional ou tridimensional.
  - *Map Visualizer*: visualizador de mapas.
  - *Rule Visualizer*: visualizador de regras.
  - *Evidence Visualizer*: visualizador de evidencias da classe dentro de cada atributo.
  - *Record Visualizer*: visualizador da base ou conjunto de dados.
  - *Statistic Visualizer*: visualizador de estatísticas geradas pelo MineSet™.
  - *Splat Visualizer*: visualizador difuso bidimensional ou tridimensional.
  - *Cluster Visualizer*: visualizador de estatísticas de clusters.

- *Decision Table Visualizer*: visualizador de tabela de decisão.

MineSet™ usa como suporte principal a biblioteca MLC++ (uma biblioteca de classes em C^{++} para Algoritmos de Aprendizado de Máquina [Félix 98, Kohavi 94, Kohavi 97]) que é responsável pela busca de padrões.

Os resultados obtidos com os utilitários de *Data Mining* podem ser visualizados e compreendidos através das ferramentas de visualização que provêem uma interface tridimensional interativa permitindo ao usuário manipular objetos visuais, efetuar buscas, filtragens e animações. O *Tool Manager* provê uma interface entre o usuário e todas as ferramentas do MineSet™ permitindo que através dele sejam configuradas as diversas opções das ferramentas.

Nas próximas seções serão mostrados os principais utilitários do MineSet™.

3. Módulo de Controle Centralizado

O Módulo de Controle Centralizado é constituído do *Tool Manager*, que é a interface gráfica inicial usada para a maioria das interações com os componentes do MineSet™, e do *DataMover* que é um processo executado no servidor cuja função é prover acesso à bases ou arquivos de dados, e transformar estes dados para serem usados pelas ferramentas de *Data Mining* e visualização. Com o *Tool Manager* é possível selecionar, transformar ou analisar a base de dados, configurar e executar os utilitários de *Data Mining* e visualizar os resultados usando as ferramentas individuais do MineSet™.

Como pode ser observado na Figura 2, o *Tool Manager* consiste de dois painéis relacionados ao conjunto de dados e a ferramenta escolhida, e duas seções de informação, que são:

- *Data Transformations*: permite fazer modificações nos dados, tais como discretizar, filtrar, modificar o tipo, entre outras.
- *Data Destination*: permite que sejam aplicadas as ferramentas de *Data Mining* aos dados e que sejam criadas visualizações para os resultados ou para os dados.
- Painel superior: contém informações da base de dados atualmente selecionada.
- Painel inferior: fornece informações sobre o *status* de algumas operações.

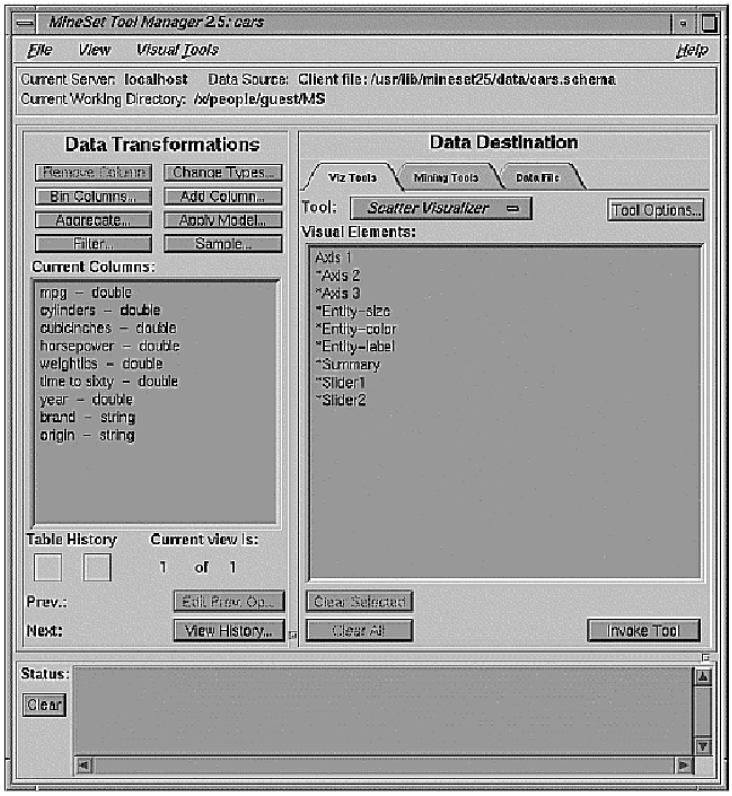

Figura 2: Interface principal do *Tool Manager*

4. Ferramentas de Visualização

As ferramentas de *Data Mining* descobrem padrões construindo modelos que podem ser visualizados através das ferramentas de visualização. Algumas destas ferramentas de visualização, tais como *Scatter Visualizer*, *Record Visualizer*, *Statistic Visualizer* e *Splat Visualizer*, podem ser aplicadas diretamente aos dados e, com isso, uma melhor observação dos mesmos pode ser conseguida. Estas ferramentas permitem ao usuário adquirir um entendimento mais profundo e intuitivo dos dados, permitindo que o próprio usuário possa descobrir padrões escondidos e tendências importantes.

Todas estas ferramentas de visualização possuem interfaces visuais tridimensionais e interativas, que permitem ao usuário a manipulação de objetos na tela, assim como animações. Tal habilidade para visualizar e pesquisar padrões complexos nos dados é um excelente auxílio ao processo de tomada de decisão.

Maiores detalhes sobre as ferramentas de visualização, que fogem ao escopo deste trabalho, podem ser encontrados em [Oliveira 98].

5. Utilitários de *Data Mining*

Como citado anteriormente, o MineSet™ possui várias ferramentas de *Data Mining*. Estas serão descritas a seguir.

5.1. Gerador de Regras de Associação

O Gerador de Regras de Associação gera um arquivo de regras a partir de um arquivo de dados de entrada. Uma regra é uma expressão da forma X=>Y, onde X e Y são conjuntos de atributos. Essas regras indicam as relações entre dois ou mais atributos escolhidos pelo usuário. Essas relações (regras) são quantificadas em três valores:

- *confidence*: indica a freqüência com que um item LHS (*Left Hand Side*. Corresponde ao eixo y do gráfico apresentado pelo Visualizador de Regras) e um item RHS (*Right Hand Side*. Corresponde ao eixo x do gráfico) ocorrem juntos em relação ao número de registros dos quais LSH ocorre. É equivalente à probabilidade condicional P(RHS|LHS). Por exemplo, este valor indica a probabilidade de, dado que uma pessoa comprou leite, qual a probabilidade dela comprar pão.
- *support*: indica a freqüência com que um item RHS e um item LHS ocorrem juntos em relação ao número total de registros. É equivalente à probabilidade P(RHS,LHS). Por exemplo, este valor indica a probabilidade de uma pessoa qualquer comprar leite e pão.
- *expected confidence*: indica a probabilidade de ocorrer um item RHS. É equivalente à probabilidade simples P(RHS). Por exemplo, indica a probabilidade de uma pessoa qualquer comprar leite.
- *lift*: indica a taxa de confidência da confidência esperada. Quanto maior o valor, mais inesperada é a regra.

As regras extraídas pelo gerador de regras de associação são representadas pelos três valores acima citados e podem ser visualizadas usando o *Rules Visualizer*.

O Gerador de Regras de Associação pode ser ativado através do *Tool Manager*, Figura 3.

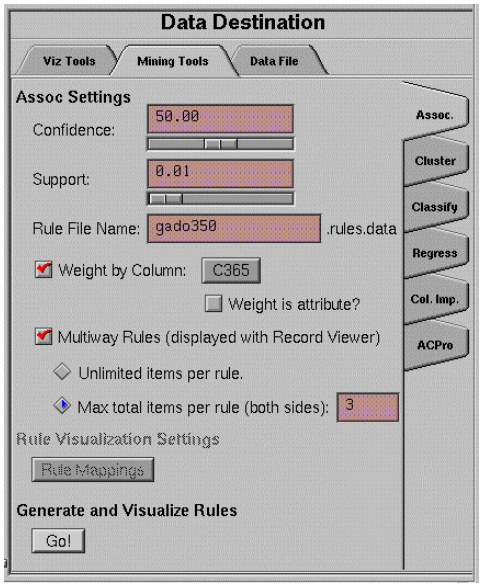

Figura 3: Gerador de Regras de Associação

Os valores *confidence*, *support* e *expected confidence* são representados em forma de barras, discos ou cor no gráfico do Visualizador de Regras, como pode ser visto na Figura 4. A definição de como esses valores serão mostrados no gráfico deve ser realizada pelo usuário. Ao pressionar o botão *Rule Mappings*, será mostrado uma caixa de diálogo onde o usuário pode mapear esses três valores.

Nos campos *Confidence* e *Support*, o usuário pode especificar os valores mínimos para *confidence* e *support* a serem exibidos no visualizador de regras. As regras que possuem valores de *confidence* e *support* menores que os especificados não são criadas. Os valores fornecidos devem estar entre 0 e 100.

Selecionando-se *Weight by Column* é possível definir pesos diferentes para os registros. Esse parâmetro deve ser utilizado quando se conhece que alguns registros são mais importantes que os outros, para isso deve ser selecionado um atributo que contém o peso de cada registro. Quando esta opção não está selecionada, todos os registros têm peso 1. Se a caixa *Weight is attribute?* é selecionada, a coluna *Weight* será incluída nas regras encontradas pelo *Gerador de Regras de Associação*, caso contrário esta coluna será excluída das regras encontradas.

Em alguns casos é útil ter regras mais complexas com múltiplos itens no LHS ou no RHS. Isto é possível selecionando-se a opção *Multiway Rules*. Regras *Multiway* são exibidas pelo *Record Viewer* ao invés do *Rules Visualizer* que é a ferramenta de visualização padrão para os resultados desse utilitário. Elas são exibidas uma por linha, sendo que as primeiras duas colunas da tabela contém o número de itens no LHS e no RHS, as próximas quatro colunas contem os valores de *support*, *confidence*, *expected confidence*, e *lift*. As duas últimas colunas contêm os itens LHS e RHS, que são separados pela palavra *and*.

É possível limitar o tamanho das regras geradas colocando-se um número máximo de itens no campo *Max total items per rule*. O número de itens em uma regra é a soma do número de itens LHS com RHS.

O Gerador de Regras de Associação é acionado ao pressionar o botão *Go!*. O respectivo visualizador também é executado automaticamente. Uma janela semelhante à Figura 4 é apresentada.

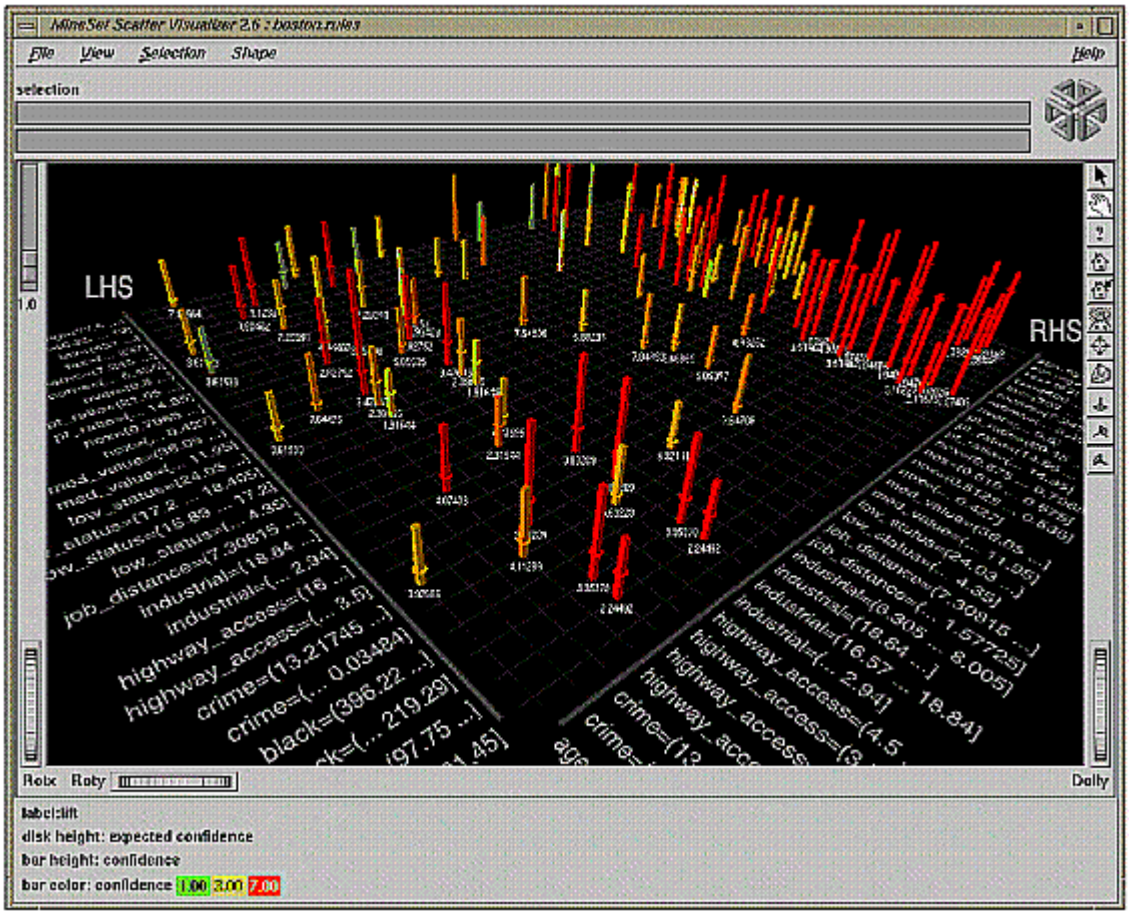

Figura 4: *Rule Visualizer*

5.2. Classificadores

Os classificadores podem ser chamados através do *Tool Manager*. A Figura 5 mostra a janela do *Tool Manager* relacionada a essas ferramentas.

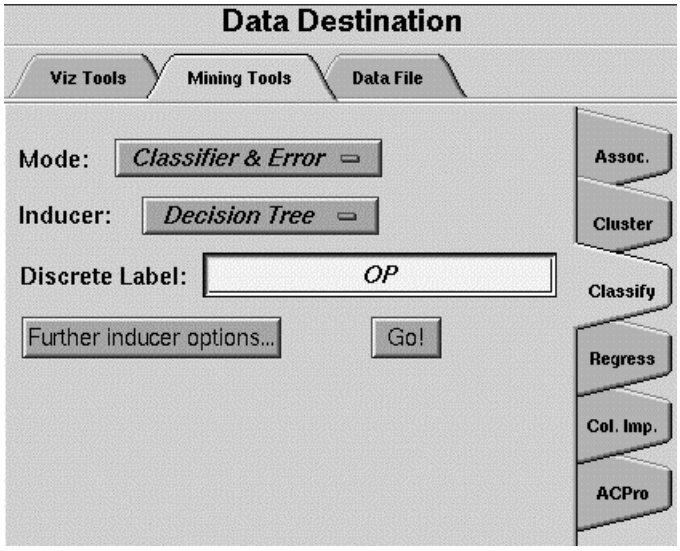

Figura 5: Classificadores

No campo *Inducer* deve ser escolhido o classificador que será utilizado, as possíveis opções são: *Decision Tree* (Classificador de Árvore de Decisão), *Option Tree* (Classificador de Árvore de Opção), *Decision Table* (Classificador de Tabelas de Decisão) e *Evidence* (Classificador de Evidência). Cada indutor e seu respectivo classificador serão explicados detalhadamente nas seções subsequentes.

No campo *Mode* deve ser especificado como o erro deve ser calculado e exibido. Esse campo apresenta as seguintes opções:

- *Classifier & Error*: com esta opção, o indutor calcula o erro do classificador utilizando *Holdout*. O erro é mostrado no painel inferior do *Tool Manager*.
- *Classifier Only*: o indutor não calcula o erro.
- *Error Only*: o erro é calculado utilizando o algoritmo *Cross Validation*, sendo exibido no painel inferior do *Tool Manager*.
- *Learning Curve*: é gerada uma curva de aprendizado para que o comportamento do erro seja melhor compreendido em função do número de registros utilizados para a classificação.

No campo *Discrete Label* deve ser selecionado um atributo discreto da base de dados que será o atributo-meta a ser predito pelo classificador.

5.2.1. Classificador e Indutor de Árvore de Decisão e de Opção

O Classificador de Árvores de Decisão classifica dados de acordo com um conjunto de atributos para uma série de decisões baseadas nesses atributos. O Indutor de Árvores de Decisão gera um classificador de árvore de decisão, a estrutura que vai ser exibida no *Tree Visualizer*, com cada decisão sendo representada por um nó da árvore.

O Classificador de Árvores de Opção funciona de maneira semelhante ao classificador de Árvore de Decisão, mas ao contrário deste, árvores de opção podem conter nós de opção que permitem a escolha de mais de um atributo para ser considerado em cada nível da árvore. As árvores de opção são mais precisas que árvores de decisão, mas geralmente são muito maiores.

Pode-se executar o Classificador de Árvore de Decisão a partir do *Tool Manager* devendo ser selecionada a opção *Decision Tree* no campo *Inducer* da janela mostrada na Figura 5.

Ao pressionar o botão *Further inducer options* é mostrada uma caixa de diálogo, semelhante à Figura 6, que fornece ao usuário opções avançadas deste utilitário.

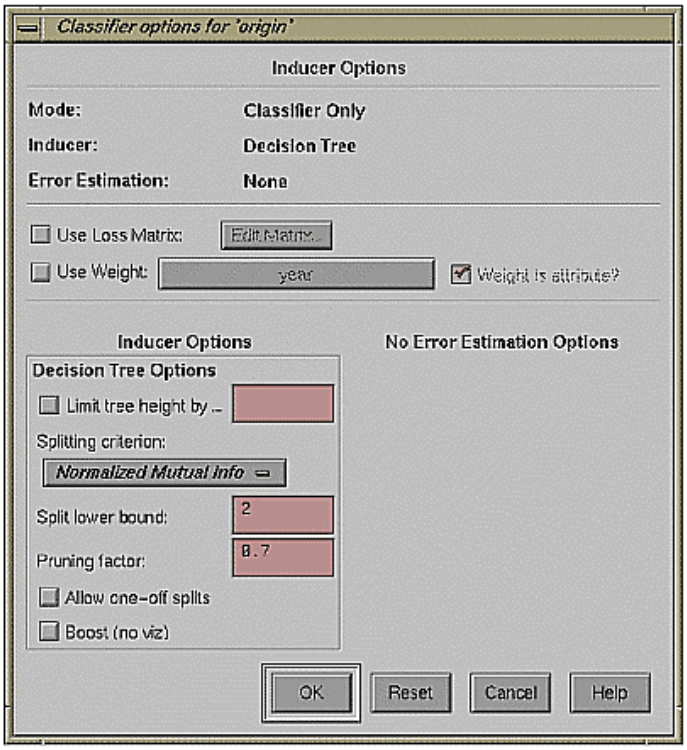

Figura 6: Opções do Indutor de Árvore de Decisão

O Indutor de Árvore de Decisão possui as seguintes opções:

- *Limit tree height by*: por default, não existe limite para o tamanho da árvore. Ela pode ser limitada habilitando esta opção e o tamanho deve então ser especificado. Ao limitar o tamanho da árvore, o tempo de execução do indutor é menor, mas pode diminuir a precisão do classificador;
- *Splitting criterion*: esta opção permite especificar qual critério será utilizado para selecionar entre os atributos divididos durante a indução da árvore. O MineSet™ suporta os seguintes critérios: *Variance*, *Absolute Deviation*, *Normalized Variance*, *Normalized Absolute Deviation*;
- *Split lower bound*: esta opção especifica o número mínimo de registros que pelo

menos dois filhos de um nó devem conter. Isto prove um método de limitar o tamanho da árvore. Ao aumentar o valor dessa opção, a confiança da probabilidade estimada tende a aumentar e também diminui o tamanho das árvores;

- *Pruning factor*: ao selecionar esta opção, o indutor tenta criar árvores otimizadas, diminuindo o número de folhas. O parâmetro determina a relação entre o tamanho da árvore o erro aceitável. Quando o valor é zero, é buscado o menor erro, com 0.5 é buscada a menor árvore em que no máximo 50% dos erros sejam maiores que os erros de uma árvore de erro mínimo. Com valor 1, a menor árvore é obtida. Ao utilizar *Pruning*, a indução fica mais lenta do que ao utilizar *Split lower bound*, mas o erro gerado é normalmente menor.
- *Allow one-off splits*: selecionando esta opção, o indutor isolará exatamente um dos possíveis valores de um atributo. Por exemplo, se um atributo pode ter os valores “vermelho”, “amarelo”, “verde” ou “azul”, e esta opção estiver marcada, o indutor exibirá este atributo com os valores “vermelho” ou “não-vermelho”.
- *Boosting*: é empregado para aumentar a precisão da classificação, mas consome um tempo muito grande na classificação. O visualizador não é chamado quando esta opção está selecionada.

Após o usuário configurar todas as opções desejadas e clicar no botão *Go!*, o Indutor é executado, e em seguida o *Tree Visualizer* também é chamado, mostrando uma Árvore de Decisão como à da Figura 7.

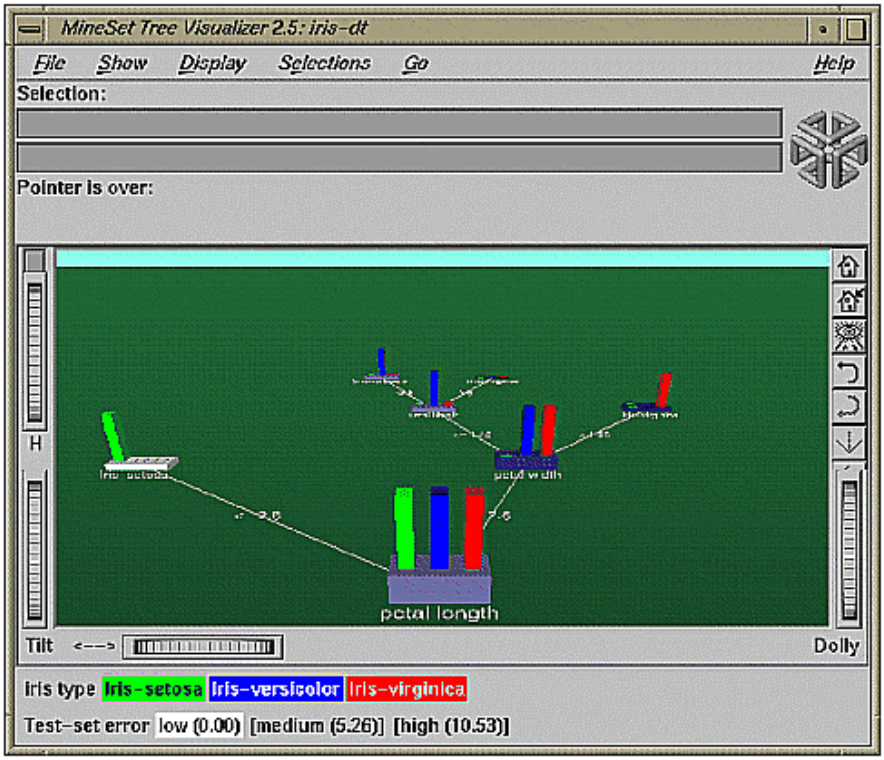

Figura 7: *Tree Visualizer*

5.2.2. Classificador e Indutor de Tabelas de Decisão

O Classificador de Tabelas de Decisão classifica dados baseado em atributos de um registro. O Indutor de Tabelas de Decisão cria um classificador dos dados. O resultado do Classificador de Tabelas de Decisão pode ser visualizado utilizando-se o *Decision Table Visualizer* como mostra a Figura 8.

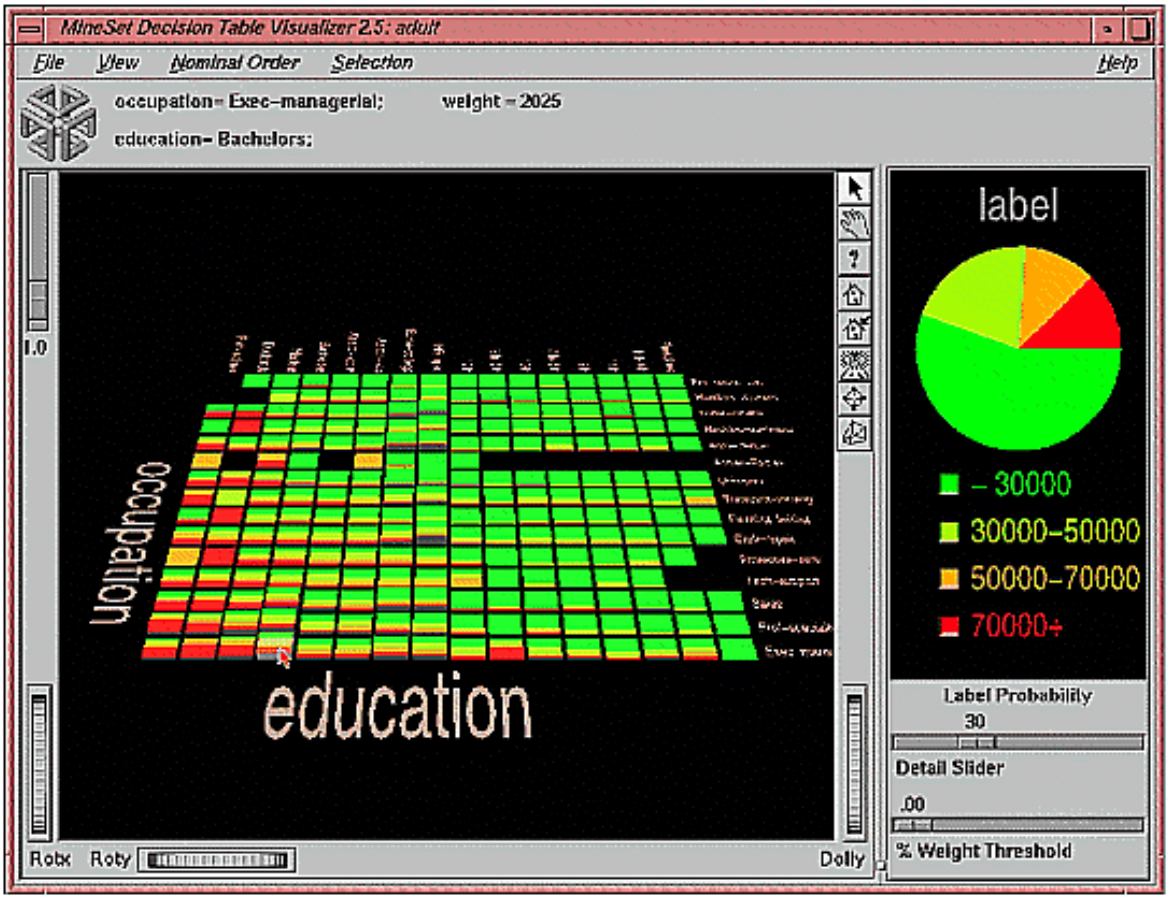

Figura 8: *Decision Table Visualizer*

O Indutor Automático de Tabelas de Decisão é um processo cujos pesos dos registros são usados para calcular as probabilidades, para isso, todos os atributos contínuos são discretizados e as probabilidades são as proporções de cada classe no conjunto de treinamento. O número de elementos em qualquer linha ao longo de um dos pares de eixos mostra o número de conjuntos discretos produzidos pelo indutor. Se existir somente um conjunto, este atributo não é útil para predizer o valor da classe. O Classificador pode ter uma matriz da perda, que pode ser usada para ajustar a distribuição de probabilidade.

Pode-se executar o Classificador de Tabela de Decisão a partir do *Tool Manager* devendo ser selecionada a opção *Decision Table* no campo *Inducer* da janela mostrada na Figura 5.

Após selecionar o modo, o usuário deve clicar em *Go!* para chamar o indutor

utilizando os valores padrão para as opções. Se o usuário não tiver idéia de qual atributo mapear em qual eixo, ele pode clicar em *Suggest* para que o mapeamento seja feito automaticamente. Para ter acesso às opções avançadas, deve-se clicar em *Further Inducer Options* para ter acesso á uma caixa de diálogo semelhante à Figura 9.

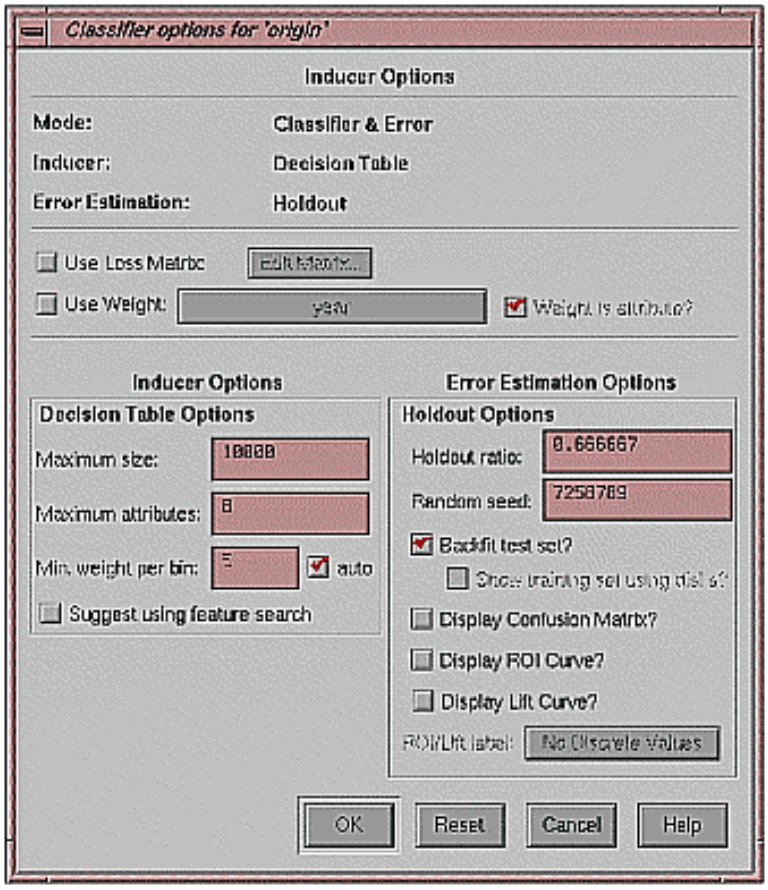

Figura 9: Opções Avançadas do Indutor de Tabela de Decisão

As opções avançadas do Indutor de Tabela de Decisão são:

- *Maximum size*: permite ao usuário limitar o número de nós (os nós correspondem aos gráficos tipo torta na visualização). Esse recurso é bastante útil quando a base de dados é muito grande, pois com a limitação, além da indução ser mais rápida, diminui o espaço de memória necessário para o armazenamento, embora possa aumentar a taxa de erro.
- *Maximum Attributes*: esta opção determina quantas colunas diferentes devem ser examinadas. Este limite afeta somente colunas adicionadas pelo modo *Suggest*. Colunas adicionadas manualmente não são afetadas.
- *Minimum Weight per Bin*: este opção permite ao usuário determinar o número mínimo de elementos por conjunto discreto quando o indutor discretiza os atributos contínuos. Para reduzir o número de conjuntos discretos, deve-se aumentar o valor desta opção.

5.2.3. Classificador e Indutor de Evidência.

O Classificador de Evidência examina a probabilidade de ocorrência de um resultado específico baseada em um atributo dado. O modelo por ele gerado é exibido pelo *Evidence Visualizer*, que mostra gráficos ilustrando as diferentes probabilidades. O *Evidence Visualizer* ajuda a entender a importância do valor de cada atributo para a classificação. Pode ser utilizado para responder questões do tipo “what if”.

Pode-se executar o classificador de Evidência a partir do *Tool Manager*, selecionando a opção *Evidence* no campo *Inducer* da janela na Figura 5.

Ao clicar o botão *Further inducer options* é exibida uma caixa de diálogo semelhante à Figura 10.

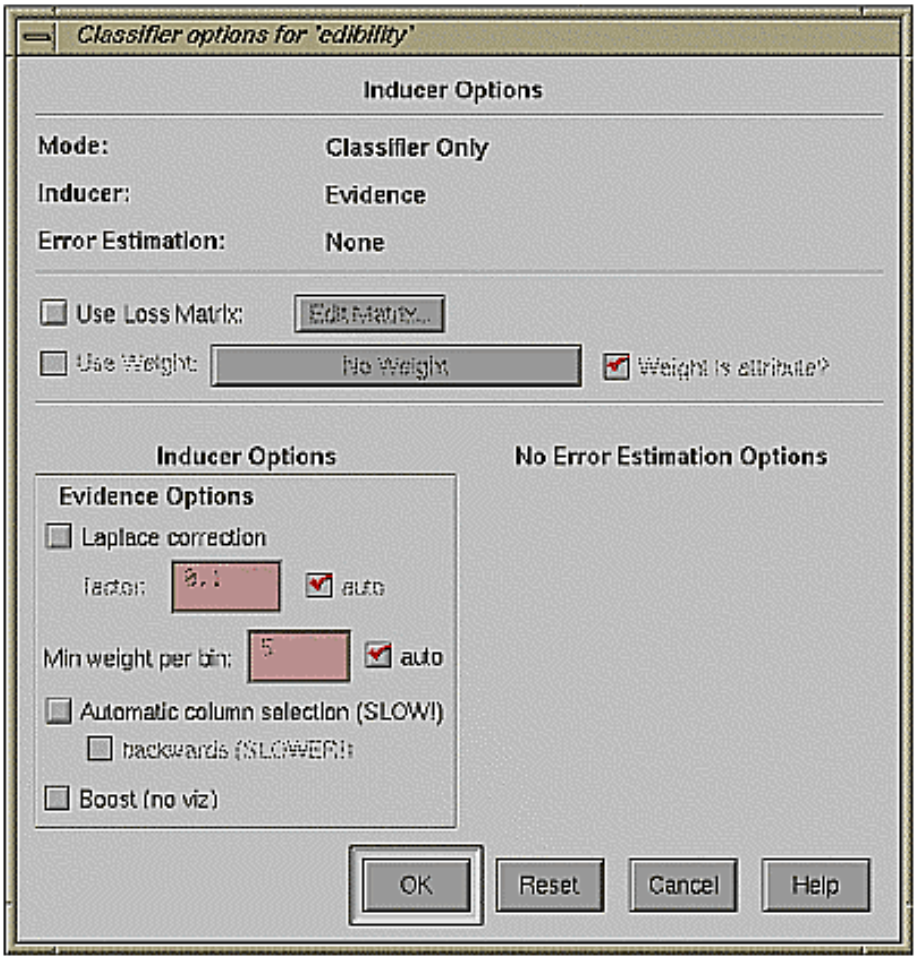

Figura 10: Opções do Indutor de Evidência

As opções para esse indutor são as seguintes:

- *Laplace Correction*: caso esta opção não esteja habilitada, quando o usuário seleciona dois ou mais atributos no painel da esquerda, o gráfico de proporção

exibido pelo *Evidence Visualizer* no painel da direita não vai refletir exatamente a verdadeira proporção dos dados originais, apesar de ser uma boa estimativa.

- *Min weight per bin*: o indutor de evidência discretiza todos os atributos contínuos. Esta opção permite que o usuário defina o número mínimo de instâncias por intervalo. Neste caso, a discretização é por freqüência. A opção *auto* habilita o algoritmo a definir automaticamente este número baseado no tamanho do conjunto de dados. Quanto maior o tamanho, maior será o número mínimo de registros em cada intervalo, e menor será a largura dos intervalos.
- *Automatic column selection*: esta opção permite que o indutor escolha automaticamente os atributos mais importantes. Apesar deste processo gastar muito tempo, é importante para eliminar atributos altamente correlacionados que poderiam diminuir a precisão do classificador. A seleção automática de colunas busca pelo melhor subconjunto de atributos.
- *Boost*: esta opção faz o indutor utilizar um algoritmo que busca dentre diversas combinações de classificadores aquela que apresenta a máxima precisão. Quando esta opção é habilitada, a classificação fica muito lenta.
- *Use Loss Matrix*: permite que a matriz de perda seja editada. Para editá-la, pressione o botão *Edit Matrix*. A matriz de perda é usada para garantir que um dos valores da classe seja predito corretamente, embora possa aumentar o erro nos outros valores.

Após todas as configurações serem feitas, ao clicar no botão *Go!* o indutor é executado e a janela do *Evidence Visualizer*, semelhante à Figura 11, é mostrada automaticamente.

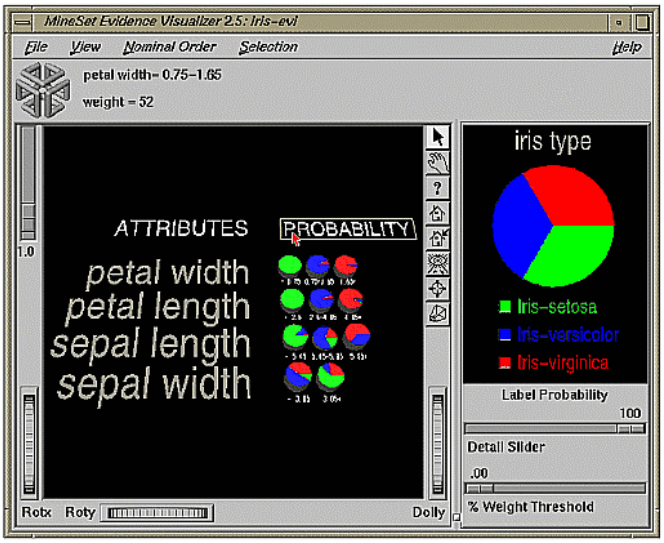

Figura 11: *Evidence Visualizer*

Mais detalhes sobre esse visualizador podem ser obtidos em [Oliveira 98].

5.3. Column Importance

A ferramenta *Column Importance* determina quão importantes os atributos são para determinar o valor do atributo meta.

Esta ferramenta conta ainda com um modo avançado que permite ao usuário entre outras coisas, determinar quais outros atributos são importantes dado que o usuário selecionou explicitamente um atributo.

Essa ferramenta pode ser executada pelo *Tool Manager* na guia correspondente à Figura 12.

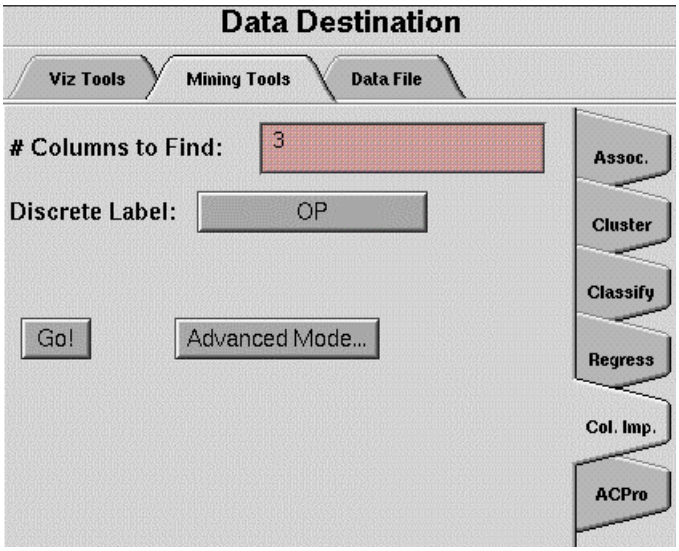

Figura 12: *Column Importance*

Esta ferramenta pode ser executada no modo simples, determinado-se o número de colunas, ou atributos, a serem encontradas, escolhendo um atributo discreto e clicando no botão *Go!*.

Para utilizar o modo avançado uma caixa de diálogo semelhante à Figura 13 deve ser usada. Nesta caixa de diálogo aparecem duas listas, a da esquerda com todos os atributos disponíveis, e a da direita que contém os atributos escolhidos como importantes (pode ser usada pelo usuário ou pelo algoritmo *Column Importance*).

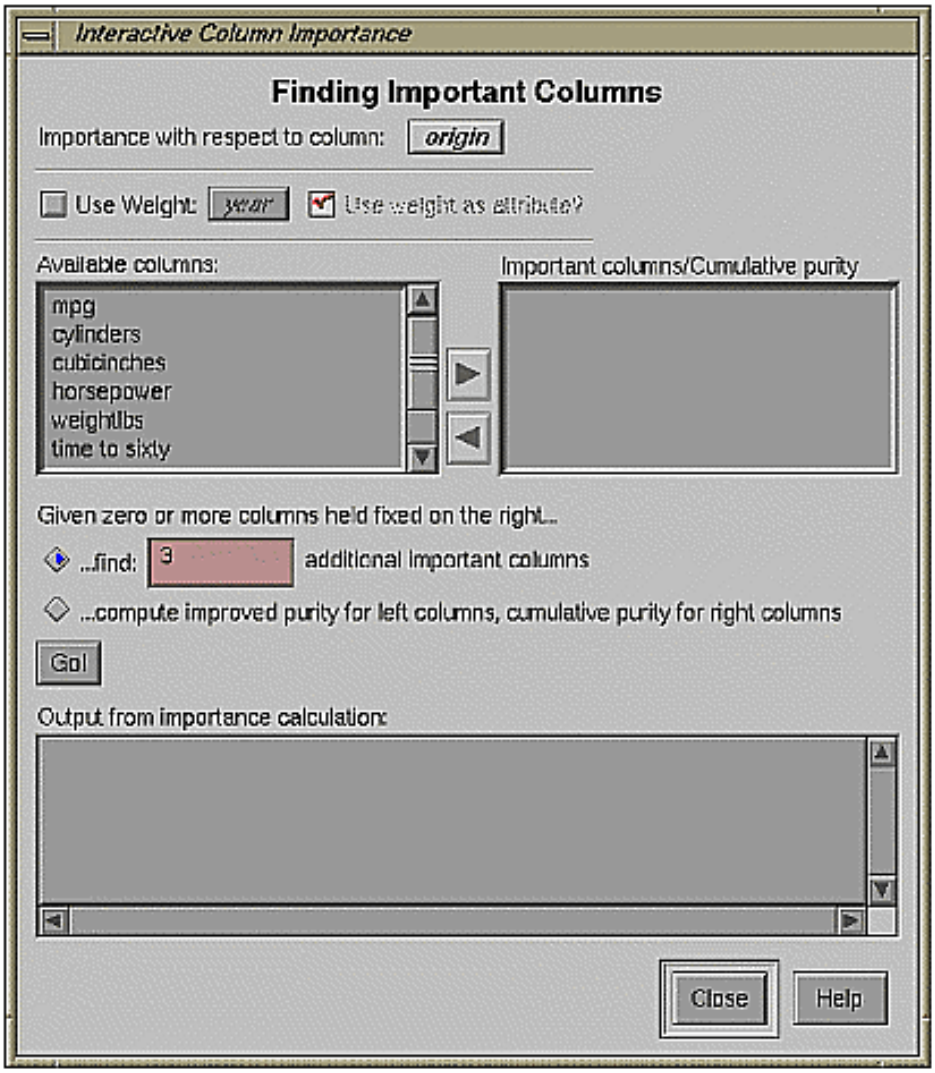

Figura 13: Opções avançadas do utilitário *Column Importance*

No modo avançado, o utilitário pode ser executado de duas formas: encontrar novos atributos importantes ou fazer um "ranking" dos atributos disponíveis.

- Encontrar atributos importantes: para isso, o usuário deve selecionar a caixa *...find n additional important attributes*. Se o botão *Go!* for pressionado sem que outras mudanças sejam feitas, o algoritmo trabalhará como se estivesse no *Simple mode*, procurando o número especificado de atributos importantes e movendo-os para a lista da direita automaticamente. Para cada atributo, a pureza acumulada é exibida.

  É possível também mover atributos da lista da esquerda para a da direita, neste caso o usuário predetermina alguns atributos que considera importantes e deixa o algoritmo encontrar outros. Após clicar em *Go!* será exibida a pureza acumulada de cada atributo. Uma pureza acumulada igual a 100 significa que utilizando os atributos selecionados, é possível determinar perfeitamente os diferentes valores

do atributo meta.

- "ranking" dos atributos disponíveis: Neste modo, a ferramenta permite ao usuário verificar o aumento da pureza que ocorria com a adição de outros atributos à lista de atributos importantes. Ao pressionar o botão *Go!* é exibido para cada atributo da lista da esquerda, o valor do aumento da pureza que este atributo promoveria caso fosse movido para a lista de atributos importantes (lista da direita). Também é exibida na lista da direita a pureza acumulada.

  Para utilizar esta opção da ferramenta, o usuário deve selecionar a caixa *...compute improved purity for left columns, cumulative purity for right columns*.

5.4. Gerador de Clusters

O Gerador de Clusters do MineSet™ particiona os dados em grupos semelhantes (também chamados de *clusters*). MineSet™ pode sugerir a segmentação de um atributo em um número qualquer de grupos distintos. Uma vez utilizada esta ferramenta, é possível ver os resultados através da ferramenta de visualização *Cluster Visualizer* ou aplicar o modelo de segmentação aos dados. É possível também visualizar os resultados no *Scatter Visualizer*.

Um modelo é gerado e armazenado em um conjunto de registros protótipos, um por *cluster*. Eles representam uma média ponderada de todos os dados no *cluster* e é conhecido como Centro do Cluster ou Centróide. Ao contrário dos registros das bases de dados comuns, os Centros de Cluster mantêm uma distribuição para cada coluna.

MineSet™ usa um algoritmo combinatorial baseado no método *k-means*. O algoritmo cria clusters agrupando registros semelhantes com o objetivo de maximizar a similaridade desses registros em cada grupo. MineSet™ prove dois métodos de geração de cluster: simples e interativo.

O utilitário pode ser executado a partir do *Tool Manager* no painel *Data Destination* como pode ser visto na Figura 14.

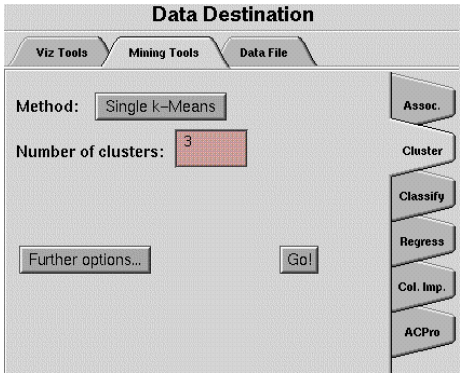

Figura 14: Gerador de cluster

O utilitário pode ser executado de duas maneiras, *Single k-Means* ou *Iterative k-*

Means, bastando selecionar a opção no campo *Method*.

O método *Single k_Means* é a forma mais simples de geração de cluster no MineSet™ . O usuário especifica o número desejado de clusters, e o algoritmo tenta agrupar os registros.

O algoritmo trabalha da seguinte forma:

1. O usuário deve especificar o número de clusters;
2. Os centros dos clusters são inicializados em posições aleatórias;
3. Cada registro é agrupado ao cluster cujo o centro está mais próximo. Então os centros são recalculados baseados nos novos registros existentes em cada cluster;
4. Se existe algum registro que está mais próximo do centro de um cluster diferente, então esse registro é movido para este outro cluster. Os centros são recalculados baseados nos novos registros existentes em cada cluster. Este passo é repetido até que nenhuma melhoria possa ser feita. O número default de interações desse passo é 20.

Para utilizar o método *Iterative k-Means* o usuário deve especificar três valores: número mínimo de clusters, número máximo de clusters e o ponto de escolha (*Choice point*).

O número final de clusters será próximo do valor da soma do mínimo com o número máximo clusters multiplicado pelo valor do *Choice point*. Assim, o método *Iterative k-Means* não requer a especificação do número exato de clusters.

O algoritmo trabalha da seguinte forma:

1. Executa o algoritmo *single k-means* com o número mínimo de clusters para formar os clusters iniciais;
2. Encontra o cluster com a maior dispersão e o divide ao meio. Cria dois novos clusters com a metade dos registros do cluster original cada um. Recalcula os centros desses dois novos clusters baseados nos novos dados;
3. Se existe algum registro que está mais próximo do centro de um cluster diferente, esse registro é movido para este cluster diferente. Os centros são recalculados baseados nos novos registros existentes em cada cluster. Este passo é repetido até que nenhuma melhoria possa ser feita (idem ao passo 4 do *single k-means*).

Os passos 2 e 3 são repetidos até que o número máximo de clusters seja atingido.

Uma vez utilizada esta ferramenta, é possível ver os resultados através da ferramenta de visualização *Cluster Visualizer* ou aplicar o modelo de segmentação aos dados. É possível também visualizar os resultados no *Scatter Visualizer*.

5.7. Classificador e Gerador de Árvore de Regressão

O Gerador de Árvore de Regressão prediz atributos contínuos da mesma forma que os Classificadores de Árvores de Decisão e Opção predizem atributos discretos. Enquanto que

um classificador prediz um evento, um regressor prediz um valor numérico específico.

O Indutor de Árvore de Regressão constrói um modelo de Árvore de Regressão. Este modelo pode ser visualizado e analisado através da ferramenta *Tree Visualizer*.

Quando um regressor é gerado, MineSet™ também gera a visualização. Esta visualização ajuda o usuário a entender como a regressor faz a predição.

Uma árvore de regressão pode ser vista na Figura 15.

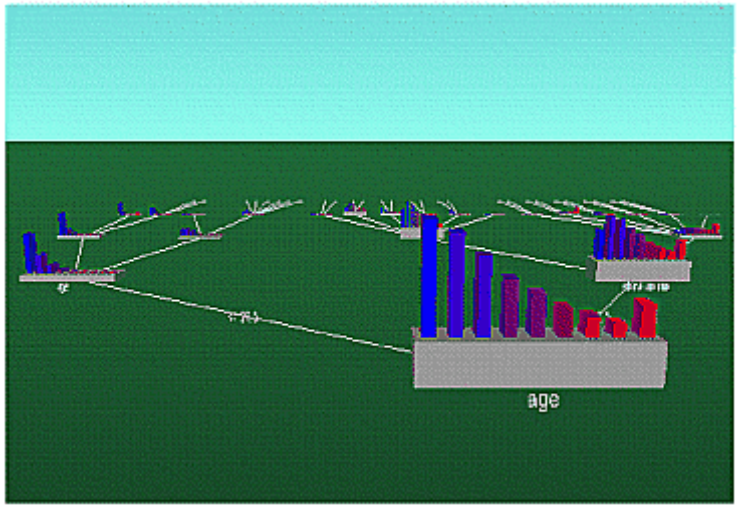

Figura 15: Árvore de regressão

Para acessar o Regressor de Árvore de Decisão, a guia *Regress* do *Tool Manager* deve ser selecionada, como mostra a Figura 16.

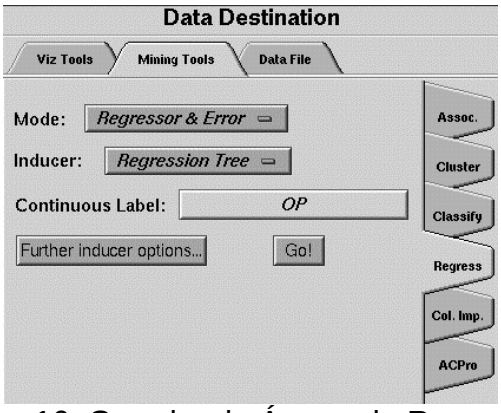

Figura 16: Gerador de Árvore de Regressão

Para executar o indutor de árvore de regressão, deve haver pelo menos um atributo contínuo na base da dados, que será o valor a ser predito. No menu *Continuous Label* deve ser escolhido o atributo contínuo que será predito pelo regressor, se não houver atributo contínuo, *No Continuous Label* é exibido, e o botão *Go!* fica desabilitado.

Ao clicar o botão *Further inducer options...* uma caixa de diálogo semelhante à da Figura 17 será apresentada, onde o usuário pode editar as opções avançadas dessa ferramenta.

Essas opções são semelhantes às do Indutor de Árvore de Decisão, que já foram detalhadas na página 8.

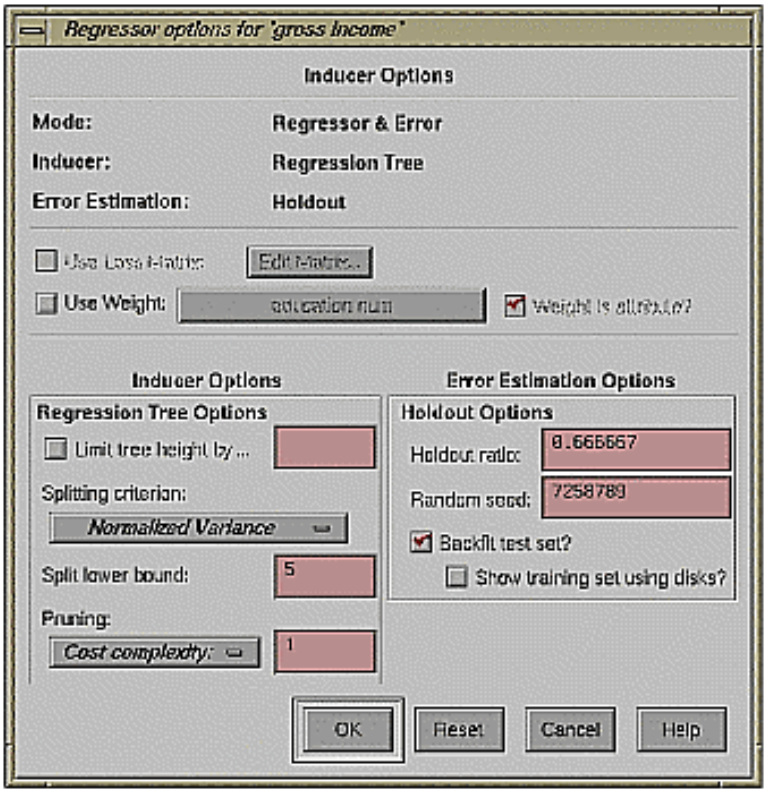

Figura 17: Opções avançadas do Regressor de Árvore de Regressão

5.8. ACPro

O algoritmo *AutoClass Pro* é um *plugin API* baseado nos algoritmos *AutoClass* e *it Snob*, e foi desenvolvido pela empresa privada *baxter-97*. Esse algoritmo também está disponível no software Mineset™.

ACPro tem como objetivo agrupar vários registros em *clusters*. Os exemplos possuem determinadas probabilidades de pertencer aos *clusters* encontrados. O *ACPro* então procura um conjunto de classes que sejam altamente prováveis com os dados, número de *clusters* e os modelos especificados.

Para acessar a ferramenta ACPro, a guia ACPro deve ser selecionada, como mostra a Figura 18.

No Mineset™ versão 2.6 estão disponíveis dois métodos distintos de *clustering*, que devem ser escolhidos no campo *Method*:

- Método *Manual*: Nesse método o usuário define apenas o número de *clusters* desejado,

- Método *Automatic*: Nesse método o usuário define 2 parâmetros, sendo um limite inferior e um limite superior de *clusters*.

Após a execução de um desses métodos, os *clusters* encontrados podem ser visualizados automaticamente na ferramenta *Cluster Visualizer*. Um exemplo dessa ferramenta pode ser visto na Figura 19. A ordem em que os atributos são mostrados no *Cluster Vizualizer* é significante, por default, os atributos são mostrados na ordem de importância para a discriminação entre os *clusters*. A ordem de importância dos atributos em cada um dos *clusters* identificados pode ser diferente.

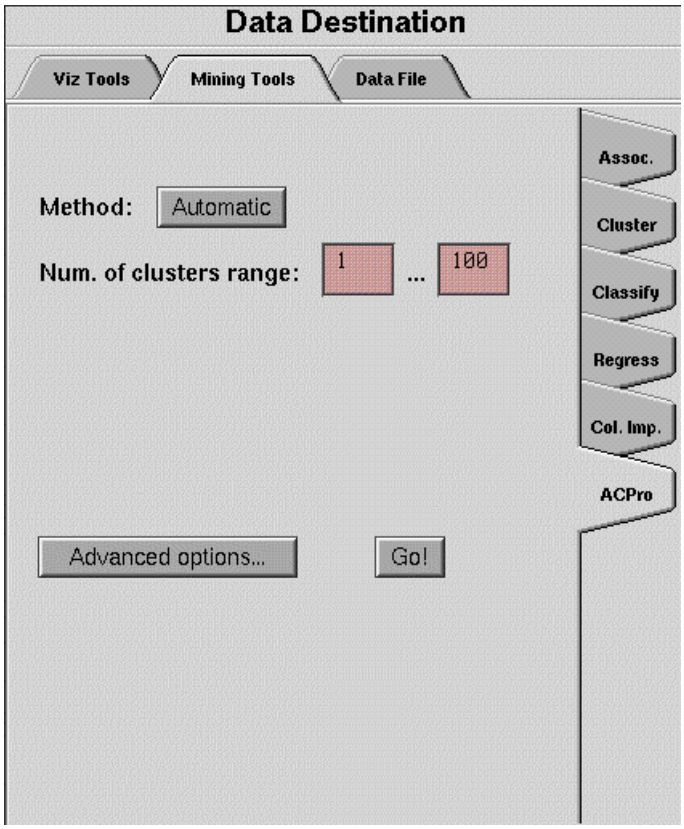

Figura 18: Utilitário ACPro

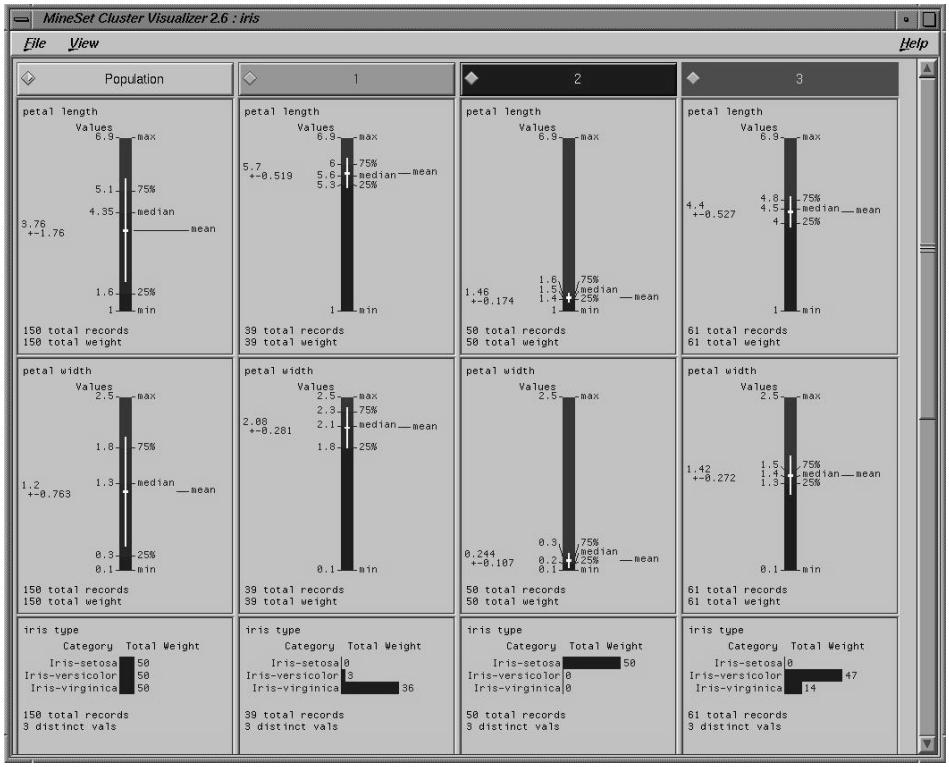

Figura 19: Cluster Visualizer

5.9. Discretizador Automático

O Discretizador Automático agrupa dados numéricos pouco espaçados em categorias discretas. Alguns algoritmos de *Data Mining*, tais como Indutor de Árvore de Decisão, precisam de atributos discretos; o mesmo pode ocorrer com algumas ferramentas de visualização.

Para chamar o Discretizador Automático, deve-se clicar no botão *Bin Columns* no painel *Data Transformations* do *Tool Manager* como mostra a Figura 20. Assim uma caixa de diálogo semelhante a Figura 21 será apresentada. Existem dois modos de utilização do Discretizador Automático: *Specifying Thresholds* ou *Automatically Computed Thresholds*.

Na primeira forma de utilização, o usuário pode escolher uma entre duas formas de especificar os limites dos conjuntos discretos. Em *Use custom thresholds* o usuário deve especificar cada faixa de valores que formarão os conjuntos discretos, por exemplo, se a entrada for 18, 30, 50, 60 os conjuntos criados corresponderão aos intervalos -18, 18-30, 30-50, 50-60, 60+. Em *Use evenly spaced thresholds* o usuário especifica conjuntos discretos igualmente espaçados. Se o usuário especificar, por exemplo, os seguintes valores: *Range start*: 4, *Range end*: 49, *Bin size*: 9, os conjuntos criados corresponderão aos intervalos –4, 4-13, 13-22, 22-31, 31-40, 40-49, 49+.

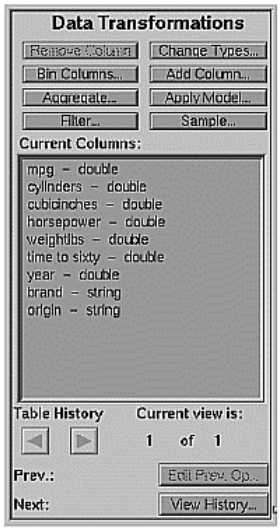

Figura 20: Painel *Data Transformations*

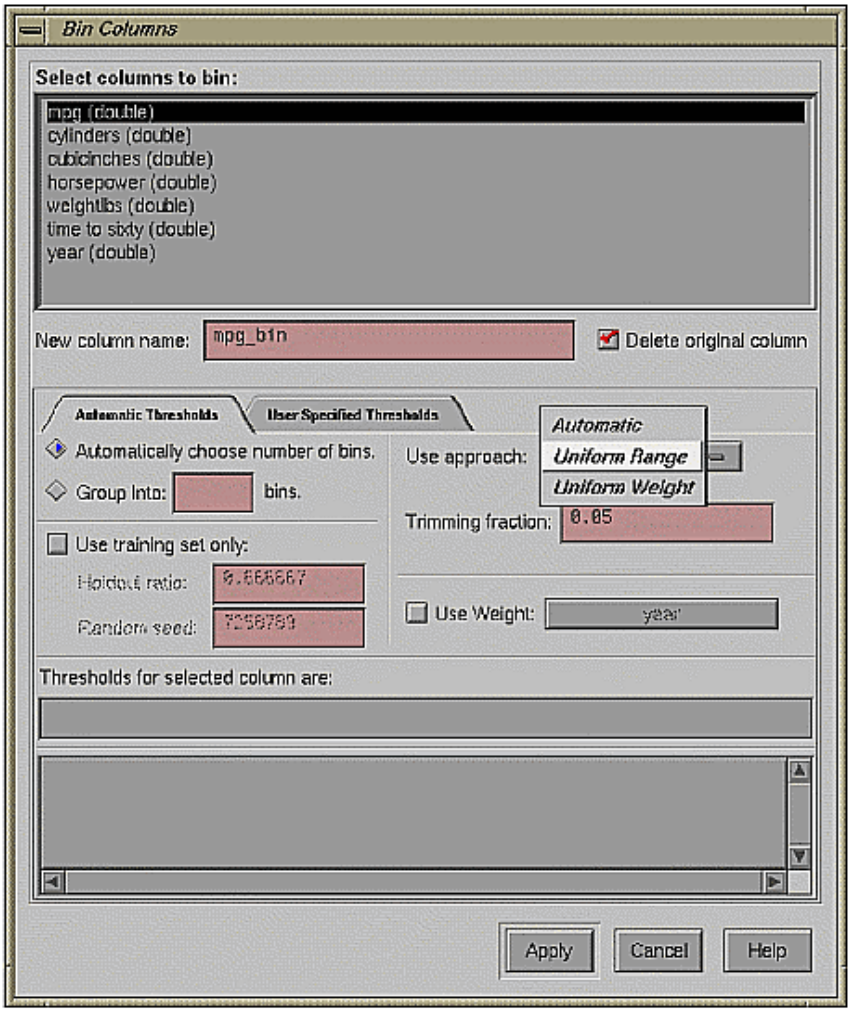

Figura 21: Utilitário *Bin Columns*

Caso a forma de utilização selecionada seja *Automatically Computed Thresholds* o MineSet™ sugere os conjuntos discretos. A primeira escolha a ser feita é entre *Automatically choose number of bins* ou *Group into: ___ bins*. Caso o usuário desejar especificar o número de grupos discretos, este numero dever ser especificado no campo *Group into:*, caso contrário o MineSet™ vai procurar o melhor número de grupos. Neste caso, deve ser selecionado um dos seguintes métodos de discretização no campo *Use approach*:

- *Automatic*: os limites são escolhidos de modo que a distribuição dos dados dentro de diferentes conjuntos discretos seja a mais diferente possível. Este método continua a criar limites que dividem os intervalos até que nenhum intervalo adicional seja significante. No campo *Min weight per bin* deve ser especificado o menor peso para cada conjunto discreto, o que se refere ao número mínimo de instâncias, uma vez que o padrão é cada instância ter peso unitário. É possível deixar o MineSet™ definir este valor automaticamente, para isso, o check box *Auto* deve ser selecionado.
- *Uniform Range*: os intervalos são criados com tamanho uniforme, ou seja, o intervalo de variação de todos os valores é dividido em subconjuntos de mesmo tamanho.
- *Uniform Weight*: os intervalos são criados com peso uniforme. Deve ser especificado um peso, caso contrário o peso assumido será unitário. Neste caso, os conjuntos discretos terão, aproximadamente, o mesmo número de instâncias.

6. Considerações Finais

Na última década, houve um grande crescimento da capacidade de gerar e colecionar dados. A transformação desses grandes volumes de dados em conhecimento se transformou impraticável. Isso tem direcionado várias pesquisas para o estudo do processo de transformação desses dados em conhecimento. Como forma de solucionar esse problema, a Extração de Conhecimento em Bases de Dados desponta como uma tecnologia capaz de cooperar amplamente na busca do conhecimento embutido nos dados.

Várias ferramentas tem sido utilizadas para apoiar as etapas do processo KDD. A ferramenta MineSet™ possui vários algoritmos de AM e Estatísticos implementados além de ferramentas gráficas para visualização que apoia principalmente a etapa de DM.

Este relatório técnico visou dar ao usuário uma noção de como utilizar as ferramentas de DM embutidas no software, além de uma visão geral da utilidade do MineSet™. Neste trabalho foram descritas as ferramentas de classificação: Gerador de Regras de Associação, Classificadores e Indutores de Árvore de Decisão e Opção, de Tabelas de Decisão, e de Evidência. Foram descritas também as ferramentas: Classificador e Gerador de Árvore de Regressão, Gerador de Clusters, ACPro (*AutoClass Pro*), Discretizador Automático e *Column Importance*.

7. Referências Bibliográficas

[Félix 98] Félix, L.C.M., Rezende, S.O., Doi, C.Y., de Paula, M.F., & Romanato, M.J. (1998). *MLC++ Biblioteca de Aprendizado de Máquina em C++*. Technical Report 72, ICMC-USP.

[Kohavi 94] Kohavi, R., John, G., Long, R., Manley, D., & Pfleger, K. (1994). *MLC++: A machine learning library in C++*. Tools with Artificial Inteligence, IEEE Computer Society Press, pages 704-743. http://www.sgi.com/Technology/mlc.

[Kohavi 97] Kohavi, R., Sommerfield, D., & Dougberty, J. (1997). *Data Mining using MLC++: A machine learning library in C++*. International Journal on Artificial Inteligence Tools.

[Oliveira 98] Oliveira, R.B.T., Rezende, S.O., (1998). *Ferramentas de Visualização de Dados do MineSet™*. Relatório Técnico No. 71, ICMC–USP

[Manuais 98] http://www.sgi.com/software/mineset/resources.html