UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

MINERAÇÃO DE DADOS:

UMA APLICAÇÃO NA BASE DE DADOS DE ARTIGOS DE PERIÓDICOS CIENTÍFICOS DAS ÁREAS DE INFORMAÇÃO (ABCDM)

HELD BARBOSA DE SOUZA

DANIELA LEITE NAGLIS

ORIENTADOR: PROF. ROMUALDO ALVES PEREIRA JÚNIOR
CO-ORIENTADOR: PROF. JAYME LEIRO VILAN FILHO

MONOGRAFIA DE ESPECIALIZAÇÃO EM ENGENHARIA ELÉTRICA
ÁREA DE GESTÃO DA TECNOLOGIA DA INFORMAÇÃO

PUBLICAÇÃO: UnBLabRedes.MFE.058/2008

BRASÍLIA / DF: NOVEMBRO/2008

HELD BARBOSA DE SOUZA

DANIELA LEITE NAGLIS

MINERAÇÃO DE DADOS:

UMA APLICAÇÃO NA BASE DE DADOS DE ARTIGOS DE PERIÓDICOS CIENTÍFICOS DAS ÁREAS DE INFORMAÇÃO (ABCDM)

Monografia de Especialização submetida ao Departamento de Engenharia Elétrica da Faculdade de Tecnologia da Universidade de Brasília, como parte dos requisitos necessários para a obtenção do grau de Especialista.

Orientador: Prof. Romualdo Alves Pereira Júnior
Co-Orientador: Prof. Jayme Leiro Vilan Filho

PUBLICAÇÃO: UnBLabRedes.MFE.058/2008

BRASÍLIA / DF: NOVEMBRO/2008

SOUZA, Held Barbosa de; NAGLIS, Daniela Leite.

Mineração de Dados: uma aplicação na base de dados de artigos de periódicos científicos das áreas de informação (ABCDM) [Distrito Federal] 2008.
152 p., 297 mm (ENE/FT/UnB, Especialista, Engenharia Elétrica, 2008).

Monografia de Especialização – Universidade de Brasília, Faculdade de Tecnologia. Departamento de Engenharia Elétrica.

1. Mineração de Dados 2. Mineração de Texto
3. áreas de informação 4. Brasil

I. ENE/FT/UnB. II. Mineração de Dados: uma aplicação na base de dados de artigos de periódicos científicos das áreas de informação (ABCDM)

SOUZA, Held Barbosa de; NAGLIS, Daniela Leite. **Mineração de dados**: uma aplicação na base de dados de artigos de periódicos científicos das áreas de informação (ABCDM). Brasília, Faculdade de Tecnologia, Universidade de Brasília, Monografia de Especialização em Gestão de Tecnologia da Informação, novembro 2008.

CESSÃO DE DIREITOS

NOME DO AUTOR: Held Barbosa de Souza; Daniela Leite Naglis
TÍTULO DA MONOGRAFIA: Mineração de Dados: uma aplicação na base de dados de artigos de periódicos científicos das áreas de informação (ABCDM)
GRAU/ANO: Especialista/2008.

Held Barbosa de Souza
Quadra 6, conjunto C, casa 11, Sobradinho DF

Daniela Leite Naglis
QRSW 5, bloco A5, ap 103, Setor Sudoeste DF

HELD BARBOSA DE SOUZA

DANIELA LEITE NAGLIS

MINERAÇÃO DE DADOS:

UMA APLICAÇÃO NA BASE DE DADOS DE ARTIGOS DE PERIÓDICOS CIENTÍFICOS DAS ÁREAS DE INFORMAÇÃO (ABCDM)

Monografia de Especialização submetida ao Departamento de Engenharia Elétrica da Faculdade de Tecnologia da Universidade de Brasília, como parte dos requisitos necessários para a obtenção do grau de Especialista.

APROVADA POR:

ROMUALDO ALVES PEREIRA JÚNIOR, Mestre, UnB
(ORIENTADOR)

JAYME LEIRO VILAN FILHO, Mestre, UnB
(CO-ORIENTADOR)

EDGARD COSTA OLIVEIRA, Doutor, UnB
(EXAMINADOR EXTERNO)

BRASÍLIA/DF, 04 DE NOVEMBRO DE 2008.

AGRADECIMENTOS

Em primeiro lugar agradeço ao Prof. Tim e ao Prof. Rafael pela oportunidade única de participar desse curso com bolsa integral, contribuindo grandemente para a minha formação.

Agradeço aos meus pais por todo carinho e apoio de sempre; ao meu noivo pelo amor, companheirismo e incentivo especialmente no decorrer deste curso.

Agradeço ao Prof. Romualdo pela disposição e interesse demonstrados não só durante a elaboração deste trabalho, mas em todos os momentos. Sem dúvida foi um ótimo orientador e motivador.

Ao Prof. Jayme, em primeiro lugar agradeço pela amizade. Excelente tutor e exemplo de compromisso com a academia. Só tenho a agradecer por todo o aprendizado decorrente desses anos de convivência.

Agradeço à Profa. Sônia Báo, que na época do início do curso flexibilizou o meu horário de trabalho para que eu pudesse comparecer às aulas. Além disso sempre permitiu e incentivou a participação dos funcionários em cursos diversos.

Agradeço também à uma grande companheira durante o curso e durante o trabalho final: Daniela. Responsável, dedicada e compreensiva. Obrigada pela amizade e carinho!

Held Barbosa de Souza

AGRADECIMENTOS

Agradeço a todos os Professores pela dedicação e atenção demonstradas durante o curso, principalmente aos Professores Tim, Rafael e Edgard.

Agradeço ao nosso orientador Romualdo, pela disponibilidade, interesse e dedicação de seu tempo para nos ajudar com seus conhecimentos.

Agradeço a Held, por ser essa pessoa maravilhosa, esforçada, dedicada e inteligente e que me acompanhou durante todo o curso e acabou se tornando uma grande amiga.

Agradeço aos meus pais pelo apoio, orientação, incentivo e esforço que fizeram para que eu chegasse onde estou. Meu pai que além de ser meu amigo, sempre fez muito mais que sua obrigação e à minha mãe por ser maravilhosa, amiga e por ter dedicado sua vida inteira para realizar meus sonhos e de minha irmã a ponto de desistir de seus próprios sonhos.

À minha irmã agradeço pelo apoio, companheirismo, amizade, paciência e cumplicidade, além do incentivo para que eu ingressasse no curso. Eu não teria conseguido sem a sua ajuda.

Ao meu namorado por todas as vezes que me ajudou e por ser super compreensivo com minha falta de tempo, demonstrando seu carinho e apoio.

Agradeço ao meu chefe por todos os dias que me liberou para que eu pudesse ir aos encontros para elaboração da monografia, sempre muito compreensivo. Não queria deixar também de mencionar o espírito colaborativo de muitos colegas, com quem tive o prazer de trocar experiências.

Agradeço principalmente a Deus, que está sempre ao meu lado proporcionando tudo de bom que aconteceu em minha vida.

Daniela Leite Naglis

RESUMO

Apresenta um estudo do processo de Mineração de Dados na ferramenta Rapid Miner com a base de dados ABCDM, que contém os artigos de periódicos científicos das áreas de informação publicados no Brasil. Com o foco na Mineração de Texto, o processo analisa os títulos dos artigos da base de dados e identifica os assuntos mais relevantes das décadas de 70, 80, 90 e dos anos 2000, até 2007, com base no índice TF/IDF. Os principais assuntos dos títulos dos artigos publicados nos anos 2000 são identificados com maior especificidade através do algoritmo K-Means. Conclui que os resultados confirmaram alguns comportamentos já percebidos pelos pesquisadores das áreas de informação, que o processo de Mineração de Dados também é eficiente na análise de dados bibliográficos e que estudos mais aprofundados poderão ser realizados posteriormente.

Palavras-chave: Mineração de Dados, Mineração de Texto, áreas de informação, Brasil

ABSTRACT

Mining using the tool Rapid Miner over the database ABCDM, that contains the scientific periodic articles of the information science field published in Brazil. Focusing the Text Mining, the process analyzes the headings of articles in the database and identifies the most relevant subjects of the 70s, 80s, 90s and between 2000 and 2007, on the basis of the index TF/IDF. The main subjects of the articles headings published between 2000 and 2007 are identified with the greater specificity through the operator K-Means. The results had confirmed some behaviors already perceived by the researchers of the information areas. The process of Data Mining also is efficient in the analysis of bibliographical data and that deepened studies should be carried out later.

Key-words: Data Mining, Text Mining, information field, Brazil.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Os artigos científicos são de grande importância para a comunidade acadêmica. Além de serem um dos meios mais utilizados para a comunicação de pesquisas e um dos mais lidos, eles revelam indicadores da produção científica. A produção desses indicadores tem sido incentivada por órgãos internacionais e nacionais de fomento à pesquisa como meio para se obter compreensão mais acurada da orientação e da dinâmica da ciência, de forma a subsidiar o planejamento de políticas científicas e avaliar seus resultados.

Para o levantamento desses indicadores, este trabalho se apóia na Bibliometria. – campo disciplinar que estuda aspectos quantitativos da produção bibliográfica. E como fonte de informações da produção bibliográfica este trabalho usa uma base de dados desenvolvida no Departamento de Ciência da Informação e Documentação da Universidade de Brasília, a ABCDM.

A ABCDM é uma base de dados que contém referências bibliográficas dos artigos de periódicos científicos publicados no Brasil e em Portugal, das áreas de informação, aqui entendidas por Arquivologia, Biblioteconomia, Ciência da Informação, Documentação e Museologia.

Algumas análises quantitativas na ABCDM tem sido feitas e publicadas (SOUZA, 2006; VILAN FILHO; SOUZA, 2007; VILAN FILHO; SOUZA; MULLER 2008). Porém. este trabalho é o primeiro a realizar estudos nessa base de dados usando técnicas de Mineração de Textos.

Com mais de 4.000 registros de artigos dentro das 5 áreas de informação, torna-se difícil conhecer a coleção de artigos brasileiros, saber como esses documentos estão relacionados, qual a relação entre os assuntos que esses documentos tratam. Através da aplicação de técnicas de Mineração de Textos é possível, dentro de uma coleção, dividi-la em grupos de documentos que tratam do mesmo assunto, utilizando para isso medidas de similaridades, que aplicam funções de distância sobre as freqüências dos termos dos documentos. E assim é possível conhecer melhor os assuntos que a comunidade científica tem abordado em sua produção.

2 JUSTIFICATIVA

O periódico científico é o meio mais utilizado para a difusão de resultados de pesquisa e para a comunicação entre pares da comunidade científica (MUELLER, 1994, p. 312).

Os artigos de periódicos podem ser colecionados, classificados, catalogados e reproduzidos infinitamente, com isto eles atingem mais rápido que teses e dissertações, um público maior, servindo como fonte de bibliografia e contribuindo para a atualização dos que os lêem (MATOS, 2003).

O propósito da leitura dos artigos de periódicos deve-se à atualização e desenvolvimento profissional, pesquisa, consultorias ou mesmo escrever ou fazer apresentações. Uma pesquisa realizada por King e Tenopir (1998, p. 176) apontou que quase em sua totalidade os leitores indicam que muitos dos resultados positivos provieram da leitura dos artigos, pois melhora a qualidade da pesquisa, ajuda a desempenhar melhor suas atividades e emprega menos tempo.

A quantidade de conhecimento registrado vem aumentando e não é diferente com o número de periódicos e de artigos. Várias causas são apontadas, como o advento das tecnologias da informação, que favorecem a comunicação entre pares, o apoio financeiro, o aumento do número de cursos de pós-graduação, entre outros.

Analisar a evolução desse crescimento pode revelar importantes indicadores a respeito da produção bibliográfica. Esses indicadores vêm ganhando importância crescente como instrumentos para análise da atividade científica e das suas relações com o desenvolvimento econômico e social. A construção de indicadores quantitativos tem sido incentivada por órgãos internacionais e nacionais de fomento à pesquisa como meio para se obter compreensão mais acurada da orientação e da dinâmica da ciência, de forma a subsidiar o planejamento de políticas científicas e avaliar seus resultados (FAPESP, 2005).

A construção de indicadores de produção científica utiliza-se de informações contidas em bases de dados bibliográficas, concebidas fundamentalmente para o armazenamento e a recuperação da informação ou do conteúdo das publicações (FAPESP, 2005). Porém:

> "uma grande dificuldade que existe no Brasil para se estabelecer estratégias de política científica é exatamente à falta de bases de dados que permitam perceber a produção científica em um contexto amplo, que permitam também avaliar o impacto dessa produção local e internacionalmente e que possibilitem perceber, enfim, a dinâmica da circulação de informações." (MENEGHINI, 1998, p. 219)

O acesso à ABCDM, que é uma base de dados com um conteúdo muito precioso, juntamente com o potencial de uma ferramenta de Mineração de Dados, é possível atingir resultados inesperados e importantes, trazendo grandes contribuições para a comunidade científica.

3 CONCEITOS BÁSICOS

3.1 BIBLIOMETRIA

Para realizar qualquer tipo de análise de informações, seja manual ou automatizada, para qualquer finalidade específica, é necessário fazer uso de informações que representem, de forma concisa, a realidade que se pretende analisar. Isto pode ser viabilizado pelo uso de indicadores (ROMÃO, 2002).

Os indicadores científicos surgem da medição dos insumos (recursos humanos, financiamento público e privado, etc.) e dos resultados (produção bibliográficas, patentes, etc.) das instituições científicas (ROMÃO, 2002).

Para o levantamento desses indicadores, alguns conceitos quantitativos foram criados no contexto da atividade científica, conforme segue (SUTCLIFFE apud CHAPULA, 1998):

- **Bibliometria**: é o estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada. Seus resultados são usados para elaborar previsões e apoiar a tomada de decisões.
- **Cientometria**: é o estudo dos aspectos quantitativos da ciência enquanto uma disciplina ou atividade econômica. É aplicada no desenvolvimento de políticas científicas. Sobrepõe-se a bibliometria.
- **Informetria**: é o estudo dos aspectos quantitativos da informação em qualquer formato, não apenas dos cientistas.

Através destes conceitos é possível abstrair a realidade e estabelecer parâmetros numéricos capazes de resumir informações generalizadas sobre investimentos, produção e tendências no campo da ciência e tecnologia (CHAPULA, 1998).

A bibliometria é um meio de situar a produção de uma instituição em relação a seu país e cientistas em relação às suas próprias comunidades (CHAPULA, 1998). Os parâmetros envolvidos pelos indicadores bibliométricos são empregados como medidas indiretas da atividade

da pesquisa científica e contribuem para a compreensão dos objetivos da pesquisa, das estruturas da comunidade científica, do seu impacto social, político e econômico (FAPESP, 2005).

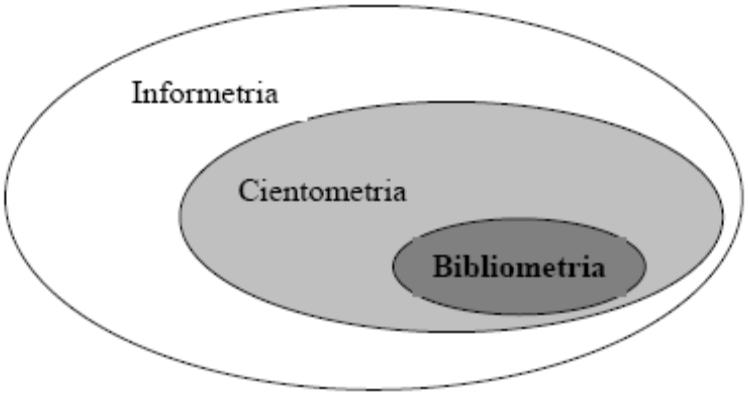


Figura 1. Abrangência dos conceitos de indicadores (ROMÃO, 2002)

Segundo Spinak (1998), a bibliometria é uma disciplina multidisciplinar que analisa um dos aspectos mais relevantes e objetivos da comunidade científica, a comunicação impressa, compreendendo:

- aplicação de análises estatísticas para estudar as características do uso e criação de documentos;
- estudo quantitativo da produção de artigos;
- aplicação de métodos matemáticos e estatísticos no estudo do uso de livros nas bibliotecas;
- estudo quantitativo das unidades físicas publicadas.

Spinak faz comparações concluindo que a bibliometria trata com as várias medições da literatura, dos artigos e outros meios de comunicação, enquanto que a cientometria trata com a produtividade e utilidade científica, mas ambas podem ser aplicadas a:

- identificar as tendências e o crescimento do conhecimento nas diferentes disciplinas;
- estimar a cobertura das revistas secundárias;
- identificar usuários, autores e tendências em diferentes disciplinas;
- prever as tendências de publicação;
- identificar as revistas do núcleo de cada disciplina;

- formular políticas de aquisição baseadas em previsões;
- estabelecer normas para padronização;
- prever a produtividade de editores, autores, organizações, países, etc. (SENGUPTA *apud* SPINAK, 1998).

Wormell (1998) identifica um instrumento que tem sido aplicado no levantamento de informações para estudos bibliométricos: as bases de dados. A autora afirma que é preciso aprender e explorar bases de dados não somente para ter acesso a documentos ou fatos, mas também para traçar as tendências e o desenvolvimento da sociedade, das disciplinas científicas e das áreas de produção e consumo.

Os mecanismos avançados de busca on-line e as técnicas de recuperação da informação proporcionadas pelas bases de dados aumentaram de forma considerável as potencialidades da metodologia de estudos bibliométricos para recuperar informação analisada a partir de grandes coleções de dados bibliográficos (WORMELL, 1998).

3.2 DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS (KDD)

O aumento no volume de dados, associado à crescente demanda por conhecimento novo voltado para decisões estratégicas, tem provocado o interesse crescente em descobrir novos conhecimentos em bancos de dados (ROMÃO, 2002, p. 1).

O desafio que se apresenta pode ser simplificado como a resolução de duas questões básicas: como extrair conhecimento dos dados, e como obter conhecimento que seja estratégico para a tomada de decisões.

O processo de extração de conhecimento a partir de dados é ilustrado pelo triângulo da Figura 2.

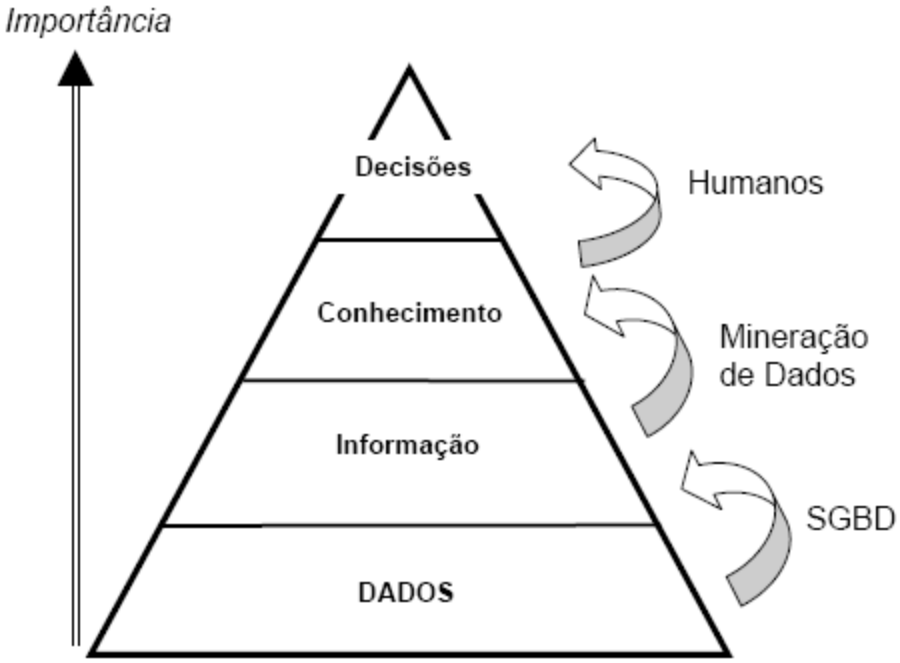


Figura 2. Obtenção de conhecimento para tomada de decisões (ROMÃO, 2002, p. 2)

Na base do triângulo estão os dados, os quais tomam o maior volume da memória do computador, e oferecem pouca utilidade estratégica na hora de se tomar decisões. A partir dos dados é possível obter muita informação através de aplicativos desenvolvidos para fins específicos ou através das ferramentas dos Sistemas Gerenciadores de Banco de Dados (SGBD) que exigem conhecimento das mesmas por parte do analista para se obter o máximo proveito da montanha de dados disponíveis e em crescimento.

A partir das informações ou dos próprios dados é possível extrair um tipo de informação mais completa, o conhecimento, normalmente mais resumido e em menor quantidade, mas de maior inteligibilidade para se tomar decisões.

Finalmente, no topo do triângulo da Figura 2, aparecem as decisões realizadas pelo homem com base no conhecimento obtido pelas ferramentas de Mineração de Dados (MD). A aplicação de algoritmos específicos deve garantir que o tipo e forma do conhecimento obtido estejam adequados ao processo de tomada de decisões rápidas e inteligentes. Estas ferramentas normalmente utilizam métodos baseados na verificação, isto é, o usuário (analista de negócio)

constrói hipóteses sobre relações entre os dados para extrair algum tipo de padrão implícito a partir do banco de dados.

O produto principal de qualquer ferramenta de apoio à decisão é o conhecimento que ela pode fornecer. Existem inúmeras técnicas capazes de extrair conhecimento em banco de dados, mas em geral este conhecimento ainda é de grande volume, dificultando a tomada de decisões. Para viabilizar decisões eficientes, devem ser implementadas ferramentas de apoio à tomada de decisão capazes de extrair conhecimento novo e surpreendente a partir de banco de dados.

O termo "Descoberta de Conhecimento em Banco de Dados" (tradução de Knowledge Discovery in Databases - KDD) surgiu no primeiro workshop de KDD em 1989, para enfatizar que o produto final do processo de descoberta em banco de dados era o "conhecimento" (FAYYAD ET AL., 1996b).

KDD é uma área interdisciplinar específica que surgiu em resposta à necessidade de novas abordagens e soluções para viabilizar a análise de grandes bancos de dados. Particularmente, KDD tem obtido sucesso na área de marketing, onde a análise de banco de dados de clientes revela padrões de comportamento e preferências que facilitam a definição de estratégias de vendas. (BERRY apud FAYYAD et al., 1996a).

Segundo Fayyad et al (1996b), KDD "é o processo não trivial de identificação, a partir de dados, de padrões que sejam válidos, novos, potencialmente úteis e compreensíveis". Na definição de Fayyad, KDD é descrito como um processo geral de descoberta de conhecimento composto por várias etapas, incluindo: preparação dos dados, busca de padrões, avaliação do conhecimento e refinamentos. O termo "não trivial" significa que envolve algum mecanismo de busca ou inferência, e não qualquer processamento de dados direto de uma quantidade pré-definida.

Nessa definição, um conjunto de dados representa fatos enquanto que os padrões podem ser interpretados como uma expressão em alguma linguagem capaz de descrever um subconjunto de dados ou um modelo aplicável a este subconjunto. Os padrões descobertos devem ser válidos

diante de novos dados com algum grau de certeza. Estes padrões podem ser considerados conhecimento dependendo de sua natureza.

Os padrões devem ser novos, compreensíveis e úteis, ou seja, deverão trazer algum benefício novo que possa ser compreendido rapidamente pelo usuário para tomada de decisão.

Para descobrir conhecimento que seja relevante, é importante estabelecer metas bem definidas. Segundo Fayyad et al. (1996b), no processo de descoberta de conhecimento as metas são definidas em função dos objetivos na utilização do sistema, podendo ser de dois tipos básicos: verificação ou descoberta.

Quando a meta é do tipo verificação, o sistema está limitado a verificar hipóteses definidas pelo usuário, enquanto que na descoberta o sistema encontra novos padrões de forma autônoma. A meta do tipo descoberta pode ser subdividida em: previsão e descrição, conforme a Figura 3.

Figura 3. Tipos de metas no processo de KDD (ROMÃO, 2002)

A descrição procura encontrar padrões, interpretáveis pelos usuários, que descrevam os dados. A previsão parte de diversas variáveis para prever outras variáveis ou valores desconhecidos (FAYYAD ET AL., 1996a).

Na previsão, o sistema irá encontrar padrões com o propósito de estimar o comportamento futuro de algumas entidades, enquanto que na descrição o sistema deverá encontrar padrões com o propósito de apresentá-los ao usuário em uma forma compreensível pelo homem. As fronteiras entre previsão e descrição não são bem definidas, mas em KDD a descrição tende a ser mais importante do que a previsão (FAYYAD ET AL., 1996b).

As metas de previsão e descrição são alcançadas através de alguma das seguintes tarefas de MD: classificação, regressão, agrupamento, sumarização, modelagem de dependência e identificação de mudanças e desvios, sendo a tarefa de classificação a mais empregada.

Na modelagem preditiva para classificação ou regressão podem ser utilizadas, dentre inúmeras outras formas de representação do conhecimento, árvores de decisão e regras.

No processo de KDD, apesar da Mineração de Dados ser a etapa principal, o processo de descoberta de conhecimento em banco de dados não se resume a minerar os dados. Exige-se a construção de mais dois estágios: pré-processamento e pós-processamento, conforme ilustra a Figura 4.

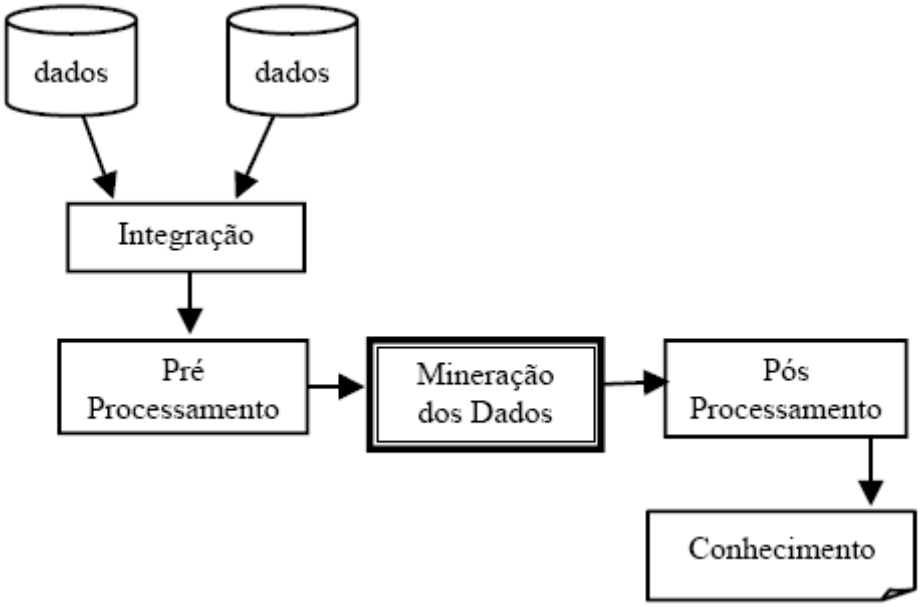


Figura 4. Etapas para descoberta de conhecimento (ROMÃO, 2002)

Na etapa de **pré-processamento** dos dados devem ser definidos os objetivos da análise. A definição dos objetivos necessita de um conhecimento prévio das possibilidades que a Mineração de Dados oferece, além de um profundo conhecimento das necessidades do negócio. Em seguida deve-se buscar conhecimento sobre as fontes dos dados, conhecendo sua estrutura e verificando como estes dados podem ser utilizados na mineração.

A partir da seleção dos dados, estes deverão ser organizados e armazenados em uma nova base de dados para análise, que pode ser mantida por um SGBD ou ser apenas um único arquivo texto. Durante a carga dos dados na nova base de dados, estes podem sofrer algum tratamento prévio para evitar resultados inesperados na mineração dos dados. É necessário tratar os dados quando existem distorções, como valores discrepantes gerados devido a erro na entrada de dados. Existem formas de amenizar os efeitos de problemas como falta de valores ou dados inconsistentes.

Ainda faz parte do pré-processamento a realização de uma análise prévia dos dados, através de alguns métodos estatísticos, para tentar identificar atributos mais relevantes ou dependência que possam facilitar ou dificultar a etapa de Mineração de Dados. Após a análise, pode ser necessário realizar a transformação dos valores de alguns atributos para melhorar os resultados obtidos com a Mineração de Dados.

Na etapa de **pós-processamento**, normalmente é possível a realização de previsões sobre fatos futuros. Para isso é necessário que alguém com conhecimento do negócio possa realizar a sua interpretação. Após a sua interpretação, o próximo passo é o uso no processo decisório. A principal meta dessa etapa é melhorar a compreensão do conhecimento obtido, em forma de relatórios demonstrativos, com a documentação e explicação das informações relevantes descobertas.

No pós-processamento pode-se chegar à conclusão de que o modelo obtido não atende às expectativas, ou seja, ao objetivo definido inicialmente. Neste caso, é necessário analisar todo o processo de KDD e identificar subseqüentes ao passo refeito também devem ser refeitos para que um novo modelo seja obtido e também avaliado. Existe ainda a possibilidade do usuário intervir em qualquer momento do processo e retornar a um passo anterior quando for detectado algum problema durante o seu desenvolvimento.

O processo de KDD é iterativo, uma vez que pode ser executado várias vezes até a obtenção do resultado desejado, e iterativo por permitir a interferência do usuário a qualquer momento e o retorno a passos anteriores.

Fayyad classifica o processo geral de KDD nas seguintes etapas:

- desenvolver um entendimento do domínio da aplicação, identificar o tipo de conhecimento que interessa, e identificar a meta do processo de KDD a partir do ponto de vista do usuário;
- realizar pré-processamento incluindo operações básicas, tais como: seleção de atributos relevantes, remoção de ruído, tratamento da ausência de valores de atributos e conversão de dados categóricos ou contínuos;
- reduzir os dados em função do objetivo da tarefa;
- escolher a tarefa de MD baseado no objetivo do processo de KDD;
- escolher o algoritmo de MD apropriado;
- realizar a mineração dos dados propriamente dita;
- interpretar os padrões descobertos, podendo retornar para um dos passos anteriores;
- consolidar o conhecimento descoberto, incluindo a conferência e a solução de possíveis conflitos com conhecimentos anteriores.

3.3 MINERAÇÃO DE DADOS

Mineração de Dados (MD) se refere ao meio pelo qual padrões são extraídos e enumerados a partir dos dados, enquanto que KDD envolve a avaliação e interpretação dos padrões para decidir o que é conhecimento e o que não é, incluindo a escolha do esquema de codificação, pré-processamento, amostragem e projeções realizadas antes da etapa de MD, bem como o pós–processamento naturalmente realizado depois da etapa de MD (FAYYAD ET AL, 1996a).

Técnicas de MD utilizam dados históricos para aprendizagem objetivando realizar alguma tarefa específica. Esta tarefa tem como meta responder alguma pergunta particular de interesse do usuário. Portanto, é necessário informar qual problema se deseja resolver.

Para encontrar respostas, ou extrair conhecimento relevante, existem diversas técnicas de MD disponíveis na literatura (CHEN ET AL., 1996; CHEUNG ET AL., 1996). As principais podem ser agrupadas em:

- Indução e/ou Extração de Regras;
- Redes Neurais;
- Algoritmos Evolucionários;
- Técnicas estatísticas (classificadores e redes Bayesianas, etc.); e
- Conjuntos Difusos.

A escolha da técnica depende, muitas vezes, do tipo de tarefa de KDD a ser realizada. A seguir algumas tarefas de KDD encontram-se comentadas.

- **Descoberta de Associação**: Abrange a busca por itens que freqüentemente ocorram de forma simultânea em transações do banco de dados. Um exemplo clássico e didático da aplicação desta tarefa é na área de marketing: durante um processo de descoberta de associações em sua vasta base de dados, uma grande rede de mercados norte-americana descobriu que um número razoável de compradores de fralda também comprava cerveja na véspera de finais de semana com jogos transmitidos pela televisão. Com uma análise mais detalhada sobre os dados, pode-se perceber que tais compradores eram, na realidade, homens que, ao comprarem fraldas para seus filhos, compravam também cerveja para consumo enquanto cuidavam das crianças e assistiam aos jogos na televisão durante o final de semana. Este exemplo ilustra a associação entre fraldas e cervejas. Esta empresa utilizou o novo conhecimento para aproximar as gôndolas de fraldas e cervejas na rede de mercados, incrementando assim a venda conjunta dos dois produtos. Algoritmos tais como o Apriori, GSP, DHP, entre outros, são exemplos de ferramentas que implementam a tarefa de descoberta de associações (GOLDSCHMIDT; PASSOS, 2005).

- **Classificação**: Consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de rótulos categóricos pré-definidos, denominados classes. Uma vez descoberta, tal função pode ser aplicada a novos registros de forma a prever a classe em que tais registros se enquadram. Como exemplo da tarefa de classificação, considere uma

financeira que possua um histórico com os dados de seus clientes e o comportamento destes clientes em relação ao pagamento de empréstimos contraídos previamente. Considere dois tipos de clientes: cliente que pagaram em dia e clientes inadimplentes. São as classes do problema. Uma aplicação da tarefa de classificação consiste em descobrir uma função que mapeie corretamente os clientes, a partir de seus dados, em uma destas classes. Tal função, uma vez descoberta, pode ser utilizada para prever o comportamento de novos clientes que desejem contrair empréstimos junto à financeira. Esta função pode ser incorporada a um sistema de apoio à decisão que auxilie na filtragem e concessão de empréstimos somente a clientes classificados como bons pagadores. Redes Neurais, Algoritmos Genéticos, Lógica Indutiva são exemplos de tecnologias que podem ser aplicadas na tarefa de classificação (MICHIE et al. 1994).

- **Regressão**: Compreende a busca por uma função que mapeie os registros de um banco de dados em valores reais. Esta tarefa é similar à tarefa de classificação, sendo restrita apenas a atributos numéricos. Como exemplo de aplicações de regressão, pode-se citar: predição da soma da biomassa presente em uma floresta; estimativa da probabilidade de um paciente sobreviver, dado o resultado de um conjunto de diagnósticos de exames; predição do risco de determinados investimentos, definição do limite do cartão de crédito para cada cliente em um banco; dentre outros. Estatística, Redes Neurais, dentre outras áreas, oferecem ferramentas para implementação da tarefa de regressão (MICHIE et al. 1994).

- **Clusterização ou Agregação**: Utilizada para separar os registros de uma base de dados em subconjuntos ou clusters, de tal forma que os elementos de um cluster compartilhem de propriedades comuns que os distingam de elementos em outros clusters. O objetivo nesta tarefa é maximizar similaridade intra-cluster e minimizar similaridade inter-cluster. Diferente da tarefa de classificação, que tem rótulos prédefinidos, a clusterização precisa automaticamente identificar os grupos de dados aos quais o usuário deverá atribuir rótulos (FAYYAD et al., 1996a). Por exemplo: uma empresa do ramo de telecomunicações pode realizar um processo de clusterização de sua base de clientes de forma obter grupos de clientes que compartilhem o mesmo perfil de compra de serviços. Na implementação

desta tarefa podem ser utilizados algoritmos tais como: K-Means, K-Modes, K-Prototypes, K-Medoids, Kohonen, dentre outros.

- **Sumarização**: Esta tarefa, muito comum em KDD, consiste em procurar identificar e indicar características comuns entre conjuntos de dados (GOLDSCHMIDT; PASSOS, 2005). Como exemplo considere um banco de dados com informações sobre clientes que assinam um determinado tipo de revista semanal. A tarefa de sumarização deve buscar por características que sejam comuns a boa parte dos clientes. Por exemplo: são assinantes da revista X, homens na faixa etária de 25 a 45 anos, com nível superior e que trabalham na área de finanças. Tal informação poderia ser utilizada pela equipe de marketing da revista para direcionar a oferta para novos assinantes. É muito comum aplicar a tarefa de sumarização a cada um dos agrupamentos obtidos pela tarefa de clusterização. Lógica Indutiva e Algoritmos Genéticos são alguns exemplos de tecnologias que podem ser aplicadas na implementação da tarefa de sumarização.

- **Detecção de Desvios**: Esta tarefa consiste em procurar identificar registros do banco de dados cujas características não atendam aos padrões considerados normais no contexto (GOLDSCHMIDT, 2003). Tais registros são denominados “outliers”. Como exemplo considere um banco de dados com informações sobre compras de clientes no cartão de crédito. A tarefa de detecção de desvios deve buscar por compras cujas características divirjam do perfil normal de compra do dono do cartão. A Estatística fornece recursos para a implementação desta tarefa.

- **Descoberta de Seqüências**: É uma extensão da tarefa de descoberta de associações onde são buscados itens freqüentes considerando-se várias transações ocorridas ao longo de um período. Consideremos o exemplo das compras no supermercado. Se o banco de dados possui a identificação do cliente associada a cada compra, a tarefa de descoberta de associação pode ser ampliada de forma a considerar a ordem em os produtos são comprados ao longo do tempo.

Seja qual for a tarefa a ser realizada, a aplicação cega de métodos de MD (chamada na literatura de estatística de "dragagem de dados") pode se tornar uma atividade perigosa e conduzir facilmente para a descoberta de padrões sem sentido (FAYYAD et al., 1996b).

Para a escolha da técnica mais adequada é estratégico saber alguma coisa a respeito do domínio da aplicação de MD: quais são os atributos importantes, quais os relacionamentos possíveis, o que é uma função útil para o usuário, que padrões já são conhecidos e assim por diante.

"Não há um método de Mineração de Dados 'universal' e a escolha de um algoritmo particular para uma aplicação particular é de certa forma uma arte" (FAYYAD et al., p. 86, 1996b).

Segundo Fayyad et al (1996b), os algoritmos de MD diferem primariamente nos critérios utilizados para avaliar o modelo e/ou no método de busca utilizado. Ele adverte que não há critérios estabelecidos para se decidir quais métodos devem ser usados em dada circunstância e que muitas abordagens são aproximações heurísticas para evitar o alto custo de processamento que seria necessário para se encontrar soluções ótimas.

Fayyad (1996b) identifica três componentes primários em algoritmos de MD:

- Representação do modelo: é a linguagem utilizada para descrever os padrões a serem descobertos;
- Critério de avaliação do modelo: afirmação quantitativa (ou função de aptidão) da qualidade que um padrão específico possui (um modelo e seus parâmetros) em alcançar as metas do processo de KDD. Modelos preditivos muitas vezes são julgados pela exatidão de previsão medida utilizando algum conjunto de dados de teste. Modelos descritivos podem ser avaliados pela novidade, utilidade e facilidade de compreensão do modelo obtido, além da exatidão;
- Método de busca: é constituído por dois componentes (busca de parâmetros e busca do modelo). Após a escolha da representação e do critério de avaliação do modelo, o problema de MD fica reduzido à tarefa de otimização (encontrar os parâmetros/modelos que satisfaçam o critério de avaliação).

Na busca, o algoritmo deve procurar os parâmetros que otimizem o critério de avaliação do modelo. A busca do modelo ocorre em um processo iterativo externo ao método de busca dos parâmetros.

3.4 MINERAÇÃO DE TEXTO

A área de Mineração de Textos ou Text Mining, também conhecida como Descoberta de Conhecimento em Textos (Knowledge Discovery in Text - KDT), surgiu com a finalidade de tratar os dados e as informações não-estruturadas considerando o alto nível de complexidade envolvida neste tipo de representação de informação (MONTEIRO; GOMES; OLIVEIRA, 2006).

Segundo o Text Mining Research Group, "Mineração de textos é a procura por padrões em um texto em linguagem natural e pode ser definido como o processo de análise do texto para extrair informação dele para um propósito em particular" (apud MONTEIRO; GOMES; OLIVEIRA, 2006).

A Mineração de Textos apresenta-se como uma ferramenta capaz de sumarizar um conjunto de documentos em agrupamentos, apresentando-os sob forma de gráficos indicativos das relações semânticas dos termos que os compõem. Assim, o usuário obtém uma idéia mais clara do assunto de que trata a coleção de páginas, sem precisar lê-las uma a uma (ARAÚJO JÚNIOR, 2007).

Para Pires (2008), no pré-processamento a coleção é carregada, processada e transformada numa representação numérica dos documentos, essa estrutura é chamada de Bag Of Words (BOW). No processamento é aplicado algum método de mineração sobre a BOW. O pós-processamento é a etapa que depende do objetivo do trabalho, nele podem ser feitos as análises dos resultados obtidos, visualizações, gráficos, etc.

A Mineração de Texto utiliza a BOW para representar um conjunto de documentos, com seus termos e a freqüência dos mesmos dentro de um documento, através de uma matriz de termo X

documento, onde as colunas representam os termos da coleção e as linhas são os documentos e os valores são as freqüências dos termos em cada documento.

3.4.1. Pré-Processamento

Para gerar uma BOW são necessárias quatro etapas: leitura e conversão, extração e limpeza dos termos, contagem de termos e cálculo de freqüência (WEISS et al apud PIRES, 2008).

- Leitura

Nessa etapa é definida uma coleção de documentos e cada documento pertencente a essa coleção terá seu conteúdo carregado na memória e seguirá pelas etapas seguintes.

- Extração e Limpeza dos termos

Cada documento da coleção vai ter o seu conteúdo dividido em termos, ou seja, cada palavra significante presente no documento. É composto por 3 sub-etapas.

  - Marcação

A marcação é utilizada para decompor o documento em cada termo que o compõe. Os delimitadores utilizados para marcação geralmente são: o espaço em branco entre os termos, quebras de linhas, tabulações, e alguns caracteres especiais.

  - Limpeza

Depois de fazer a marcação cada termo obtido passa pela etapa de limpeza. Primeiro são removidos as stopwords, depois é verificada a existência do sinônimo do mesmo no dicionário e por último é realizado o stemming do termo. Stopwords é uma lista de termos não representativos para um documento, geralmente essa lista é composta por: preposições, artigos, advérbios, números, pronomes e pontuação.

  - Stemming

Stemming é o método para redução de um termo ao seu radical, removendo as desinências, afixos, e vogais temáticas. Com sua utilização, os termos derivados de um mesmo radical serão contabilizados como um único termo.

- Contagem dos termos

Depois de extrair os termos representativos de cada documento, será calculado o número de ocorrências de cada termo num documento. Depois de concluída a contagem é criada uma lista com duas colunas: termo e quantidade de ocorrência.

- Cálculo da Freqüência

Após concluída a etapa de contagem de termos para cada documento da coleção, será calculada a freqüência dos termos. A medida escolhida para calcular a freqüência dos termos é a *tf-idf*. O *tf-idf* define a importância do termo dentro da coleção de documentos.

O *tf-idf* atribui um peso ao termo t_1 para cada documento da BOW. O peso é o número de ocorrências do termo no documento (*tf*), modificada por uma escala de importância do termo (*idf*), chamada de freqüência inversa do documento.

O tf do termo t_i no documento é:

$$tf(t_i) = \frac{n_i}{\sum_{j=1}^{k} n_j}$$

Onde n_i é o número de ocorrências do termo no documento e o denominador é o somatório de todos os números de ocorrências de todos os termos de um documento. O *idf* do termo t_i é:

$$idf(t_i) = \log \frac{N}{df(t_i)}$$

Onde N é o número total de documentos do conjunto e $df(t_i)$ é o número de documentos onde o termo t_i aparece, ou seja, $n_i \neq 0$. Então:

$$tfidf(t_i) = tf(t_i).idf(t_i)$$

Para obter uma freqüência com o *tf-idf* alta, ou seja, o termo ser representativo para o documento é necessário que o termo tenha um número alto de ocorrência no documento e um número baixo de ocorrência dentro da coleção.

A medida *tf-idf* é derivada da estatística e se baseia na freqüência de termos. No método das palavras-chave, as palavras mais freqüentes de um texto são consideradas representativas. Há, no entanto, o cuidado de não se considerar as palavras de domínio fechado, como artigos ou pronomes, que não carregam significado. Assim como essas palavras são muito freqüentes sem, no entanto, serem relevantes ou expressarem informações topicais, há palavras que aparecem muito constantemente em diversos textos, não sendo, portanto, úteis para expressar a individualidade do texto. A medida parte do princípio de que uma palavra será representativa em um texto se ocorrer diversas vezes no texto em questão e for pouco freqüente em outros textos (MARTINS et al, 2002).

Após ser calculada a freqüência dos termos de cada documento, o processo de transformação da coleção em dados numéricos estará concluído.

Depois da BOW ter sido gerada, são aplicados os métodos de processamento de mineração. classificação, análise de agrupamento, recuperação e predição são alguns dos processos de mineração. Depois de concluído o processamento, dependendo do objetivo, são aplicados no resultado obtido, métricas, ferramentas de análise e visualização, geração de gráficos entre outros, para extrair conhecimento.

3.4.2. Análise de Agrupamento

Análise de agrupamento é uma classificação não supervisionada de registros em grupos (JAIN; MURTY; FLYNN apud PIRES, 2008). O agrupamento de registros em grupos é feito baseado na similaridade entre os registros, assim os registros agrupados em um grupo são mais similares entre eles do que com algum registro pertencente a outro grupo.

A análise de agrupamento de dados é realizada por diversos métodos e cada um com uma diferente abordagem. Alguns deles são:

- Métodos hierárquicos produzem grupos aninhados;
- Métodos por particionamento produzem grupos isolados, como o K-means descrito a seguir;
- Métodos incrementais é possível criar um novo grupo quando um novo registro é apresentado durante o processo de agrupamento e não atende à taxa de similaridade exigida.

3.4.2.1. O algoritmo K-means

O algoritmo K-Means (também chamado de K-Médias) classifica as informações de acordo com os próprios dados. Esta classificação é baseada em análise e comparações entre os valores numéricos dos dados. Desta maneira, o algoritmo automaticamente vai fornecer uma classificação automática sem a necessidade de nenhuma supervisão humana, ou seja, sem nenhuma pré-classificação existente. Por causa desta característica, o K-Means é considerado como um algoritmo de Mineração de Dados não supervisionado (PICHILIANI, 2006).

Para entender como o algoritmo funciona, imagina-se uma tabela com linhas e colunas que contêm os dados a serem classificados. Nesta tabela, cada coluna é chamada de dimensão e cada linha contém informações para cada dimensão, que também são chamadas de ocorrências ou pontos. Geralmente, trabalha-se com dados contínuos neste algoritmo, mas nada impede que dados discretos sejam utilizados, deste que eles sejam mapeados para valores numéricos correspondentes.

O algoritmo vai analisar todos os dados desta tabela e criar classificações. Isto é, o algoritmo vai indicar uma classe (cluster) e vai dizer quais linhas pertencem a esta classe. O usuário deve fornecer ao algoritmo a quantidade de classes que ele deseja. Este número de classes que deve ser

passado para o algoritmo é chamado de “k” e é daí que vem a primeira letra do algoritmo: K-Means.

Para gerar as classes e classificar as ocorrências, o algoritmo faz uma comparação entre cada valor de cada linha por meio da distância. Geralmente utiliza-se a distância euclidiana para calcular o quão ‘longe’ uma ocorrência está da outra. A maneira de calcular esta distância vai depender da quantidade de atributos da tabela fornecida. Após o cálculo das distâncias o algoritmo calcula centróides para cada uma das classes. Conforme o algoritmo vai iterando, o valor de cada centróide é refinado pela média dos valores de cada atributo de cada ocorrência que pertence a este centróide. Com isso, o algoritmo gera k centróides e coloca as ocorrências da tabela de acordo com sua distância dos centróides.

Para simplificar a explicação de como o algoritmo funciona, Pichiliani (2006) apresenta o algoritmo K-Means em cinco passos:

PASSO 01: Fornecer valores para os centróides.
Neste passo os k centróides devem receber valores iniciais. No início do algoritmo geralmente escolhe-se os k primeiros pontos da tabela. Também é importante colocar todos os pontos em um centróide qualquer para que o algoritmo possa iniciar seu processamento.

PASSO 02: Gerar uma matriz de distância entre cada ponto e os centróides.
Neste passo, a distância entre cada ponto e os centróides é calculada. A parte mais ‘pesada’ de cálculos ocorre neste passo, pois se temos N pontos e k centróides teremos que calcular N x k distâncias neste passo.

PASSO 03: Colocar cada ponto nas classes de acordo com a sua distância do centróide da classe.
Aqui, os pontos são classificados de acordo com sua distância dos centróides de cada classe. A classificação funciona assim: o centróide que está mais perto deste ponto vai ‘incorporá-lo’, ou seja, o ponto vai pertencer à classe representada pelo centróide que está mais perto do ponto. É importante dizer que o algoritmo termina se nenhum ponto ‘mudar’ de classe, ou seja, se nenhum ponto for ‘incorporado’ a uma classe diferente da que ele estava antes deste passo.

PASSO 04: Calcular os novos centróides para cada classe.
Neste momento, os valores das coordenadas dos centróides são refinados. Para cada classe que possui mais de um ponto o novo valor dos centróides é calculado fazendo-se a média de cada atributo de todos os pontos que pertencem a esta classe.

PASSO 05: Repetir até a convergência.
O algoritmo volta para o PASSO 02 repetindo iterativamente o refinamento do cálculo das coordenadas dos centróides.

Desta maneira teremos uma classificação que coloca cada ponto em apenas uma classe. Assim, o algoritmo faz uma classificação *hard* (*hard clustering*) uma vez que cada ponto só pode ser classificado em uma classe. Outros algoritmos trabalham com o conceito de classificação *soft* onde existe uma métrica que diz o quão 'dentro' de cada classe o ponto está.

3.5 FERRAMENTAS DE MINERAÇÃO DE DADOS

Os primeiros softwares para Mineração de Dados começaram a ser desenvolvidos em meados da década de 90, ainda em ambiente acadêmico. Hoje em dia já existem algumas dezenas de ferramentas comerciais para Mineração de Dados, desenvolvidas por empresas como SAS (Enterprise Miner), IBM (Intelligent Miner) e SPSS (Clementine).

A maior parte dos sistemas para Mineração de Dados já demonstrou sua capacidade de servir como importante ferramenta de apoio no processo de tomada de decisões nas empresas. No entanto, é curioso perceber que a popularidade dos softwares de Mineração de Dados é relativamente baixa se comparada com a popularidade das ferramentas para data warehousing e OLAP, por exemplo. Artigos recentes apontam três motivos principais para explicar esta situação (GONÇALVES):

- Em função de seu alto potencial em relação ao Retorno Sobre Investimento (ROI), software de Mineração de Dados geralmente tem um custo elevado.
- Muitos softwares não conseguem realizar a Mineração de Dados diretamente sobre as tabelas de um SGBD. Em muitos casos é necessário exportar os dados para um repositório auxiliar.
- O terceiro motivo - freqüentemente apontado como o que mais contribui para a impopularidade da Mineração de Dados - está no fato de que as ferramentas de Mineração de Dados são muito difíceis de utilizar. Grande parte dos softwares exige que o usuário tenha algum conhecimento a respeito do funcionamento dos algoritmos de mineração implementados na ferramenta; outros softwares requerem usuários que possuam grande conhecimento de Estatística para que sejam manipulados.

Segue abaixo uma tabela comparativa de algumas ferramentas de Mineração de Dados.

Tabela 1. Comparação de ferramentas de Mineração de Dados (REZENDE apud BARROSO; FERREIRA NETO, 2006)

NOME	**TÉCNICAS DISPONÍIVEIS**	**FABRICANTE / SITE**	**TIPO DE APLICATIVO**
PolyAnalyst	classificação, regressão, regras associativas, clustering, sumarização, e modelagem de dependência	Megaputer Intelligence www.megaputer.com	pacote
Magnum Opus	regras associativas	Rule Quest www.rulequest.com	específico
XpertRule Miner	classificação, regressão, regras associativas e clustering	Attar Software Ltd. www.attar.com	pacote
DataMite	regras associativas	Dr Philip Vasey através do LPA PROLOG	específico
Microsoft Data Analyser 2002	classificação e clustering	Microsoft Corp. www.Microsoft.com	pacote
Oracle 9i Data Mining	classificação, regressão, associativas	Oracle Corp. www.oracle.com	pacote
Darwin	classificação, regressão e clustering	Oracle Corp. www.oracle.com	pacote

MineSet	classificação, regressão, regras associativas e clustering	Silicon Graphics Inc. www.sgi.com	pacote
WEKA	classificação, regressão, regras associativas e clustering	University of Waikato www.waikato.ac.nz	pacote
Inteligent Miner	regras associativas, padrões seqüenciais, classificação, clustering, sumarização e modelagem de dependência	IBM Corp. www.ibm.com	pacote
MLC++	classificação, regressão e Clustering	Silicon Graphics Inc. www.sgi.com/tech/mlc	biblioteca
See5	Classificação	Rule Quest www.rulequest.com	específico
Cubist	Regressão	Rule Quest www.rulequest.com	específico
Clementine	classificação, regras associativas, clustering e padrões seqüenciais	SPSS Inc. www.spss.com	pacote
Data-Miner Software Kit	classificação e regressão	Data-Miner PTY LTD www.data-miner.com	específico

3.5.1. Rapid Miner

O desenvolvimento da maior parte dos conceitos do RapidMiner começou em 2001 na Universidade de Inteligência Artificial de Dortmund. Vários membros das unidades começaram a implementar e realizar esses conceitos lançando a primeira versão do RapidMiner em 2002. Desde 2004, a versão aberta do RapidMiner é hospedada pela SourceForge. Desde então, um grande número de sugestões e expressões de desenvolvedores externos são embutidas no RapidMiner. Hoje, todas as versões de softwares livres (open-source) e softwares proprietários do RapidMiner são mantidos pela Rapid-I.

3.6 BASE DE DADOS DE ARTIGOS DE PERIÓDICOS CIENTÍFICOS (ABCDM)

A base ABCDM foi criada em 2001 com o propósito de identificar a literatura científica, publicada em periódicos, capaz de atender às necessidades acadêmicas de pesquisadores, professores, estudantes e público em geral com interesse nas áreas de Arquivologia, Biblioteconomia e Ciência da Informação e Documentação. Ela é uma ferramenta que facilita a identificação e conseqüentemente o acesso a todos os artigos de periódicos das revistas publicadas no Brasil e em Portugal das áreas de informação, citadas acima.

Foi iniciada como exercício final da disciplina "Planejamento e Elaboração de Bases de Dados" do Departamento de Ciência da Informação e Documentação da Universidade de Brasília. O projeto teve como atividades iniciais a identificação dos títulos de periódicos das áreas citadas e o estabelecimento de uma prioridade de tratamento desses títulos, bem como a definição de um formato de dados baseado no formato MARC 21. Foi escolhido o "CDS/ISIS for Windows" como sistema a ser utilizado.

Desde a sua criação até os dias atuais, a ABCDM vem sendo atualizada e alimentada por meio de projetos de monografia (MATOS, 2003; OLIVEIRA, 2003; SILVA, 2005; VIEIRA, 2005), Projetos de Atividade Complementar – PAC (SOUZA, 2005; SOUZA, 2006b; BALBINO, 2006) e projetos de pesquisa concluídos ou em andamento no CID. A execução desses projetos resultou em modificações no formato da base, atualizações no formato dos registros, inclusão de novos títulos, melhoramentos no aplicativo e no controle de qualidade dos dados, além da inclusão dos artigos da área de Museologia, quando a base passou a se chamar ABCDM.

No Anexo A encontra-se a descrição do formato de entrada de dados da base ABCDM.

A base ABCDM contém 4.786 registros (dados de 22/10/2008) de referências bibliográficas de artigos de 31 títulos de periódicos científicos publicados no Brasil e em Portugal, das áreas de informação.

Os seguintes os títulos de periódicos (e suas respectivas siglas no âmbito da ABCDM) estão cobertos pela base, entre títulos específicos e de áreas correlatas:

- Acervo: Revista do Arquivo Nacional (ARAN);
- Anais do Arquivo Público do Pará (AAPP);*
- Arquivística.net (ANET);
- Arquivo & Administração (AA);
- Arquivo e História (AH);*
- Biblos: Revista do Departamento de Biblioteconomia e História (BDBH);
- Cadernos de Biblioteconomia e Arquivística (CBA) (Portugal);
- Cadernos de Biblioteconomia, Arquivística e Documentação (CBAD) (Portugal);
- Cadernos de Biblioteconomia (CB);
- Cadernos Museológicos (CAMU);*
- Ciência da Informação (CI);
- Ciências em Museus (CIMU);*
- DatagramaZero (DGZ);
- Em Questão: Revista da Faculdade de Biblioteconomia e Documentação da UFRGS (EQ);
- Encontros Bibli (EB);
- Estudos Históricos (EH);
- Informação & Informação (II);
- Informação & Sociedade: estudos (ISE);
- Informare: Cadernos do Programa de Pós-graduação em Ciência da Informação (ICPCI);
- Perspectivas em Ciência da Informação (PCI);
- Revista ACB: Biblioteconomia em Santa Catarina (RACB);
- Revista Brasileira de Biblioteconomia e Documentação (RBBD);
- Revista Brasileira de Museus e Museologia (MUSAS);
- Revista da Escola de Biblioteconomia da UFMG (REBU);
- Revista de Biblioteconomia & Comunicação (RBC);
- Revista de Biblioteconomia de Brasília (RBB);

- Revista de Museologia (RM);*
- Revista Digital de Biblioteconomia e Ciência da Informação (RDBCI);
- Revista do Patrimônio Histórico e Artístico Nacional (RPHAN);*
- Revista Museu (RM);
- Transinformação (TRI).

(*) Títulos em fase de inclusão na base de dados.

Os registros são em sua maior parte de artigos científicos, embora existam registros de outra natureza como palestras, trabalhos de congressos, entre outros, que também são publicados nos periódicos científicos. Todos os artigos publicados foram registrados na base de dados, porém os outros tipos de trabalhos não foram todos incluídos. Esta inconsistência ocorreu devido à mudança na política de seleção dos trabalhos a serem incluídos na base. No princípio o objetivo da base era apenas suprir as necessidades de informação dos estudantes e dos pesquisadores da área. Depois o objetivo principal da base passou para ser fonte de informações para pesquisas sobre a produção de artigos científicos.

Desde 2006 são realizados estudos bibliométricos nessa base de dados focados nos periódicos brasileiros e são feitas análises quantitativas relacionadas com a produção de artigos e de ocorrências de artigos em autoria múltipla (AM), ou seja, artigos de dois ou mais autores.

Os primeiros resultados foram divulgados por Souza (2006a), cobrindo o período de 1972 a 2005. A autora destaca que o crescimento da autoria múltipla não é regular, acentuando-se a partir de 1996, saindo do patamar de cerca de 20% dos artigos nos primeiros 24 anos (1972-1995), para cerca de 30% dos artigos nos cinco anos seguintes (1996-2000) e passando para o patamar de cerca de 40% dos artigos nos cinco últimos anos (2001-2005) do estudo.

Estudos posteriores, incluindo os artigos publicados em 2006, foram divulgados por Vilan Filho e Souza (2007) e conclui que a média anual de produção brasileira foi de 175 artigos (2000-2006); a produtividade média anual é de 16 artigos/ano/periódico (2000-2006); o percentual de artigos em co-autoria em 2006 (49,16%) está próximo do percentual de artigos em autoria única

(50,84%); e 85% da co-autoria é de artigos com dois ou três autores (1972-2006). Conclui levantando algumas questões sobre possíveis causas desses resultados.

Vilan Filho e Souza (2007) apresentam um gráfico com a evolução da produção de artigos em valores absolutos (Gráfico 1).

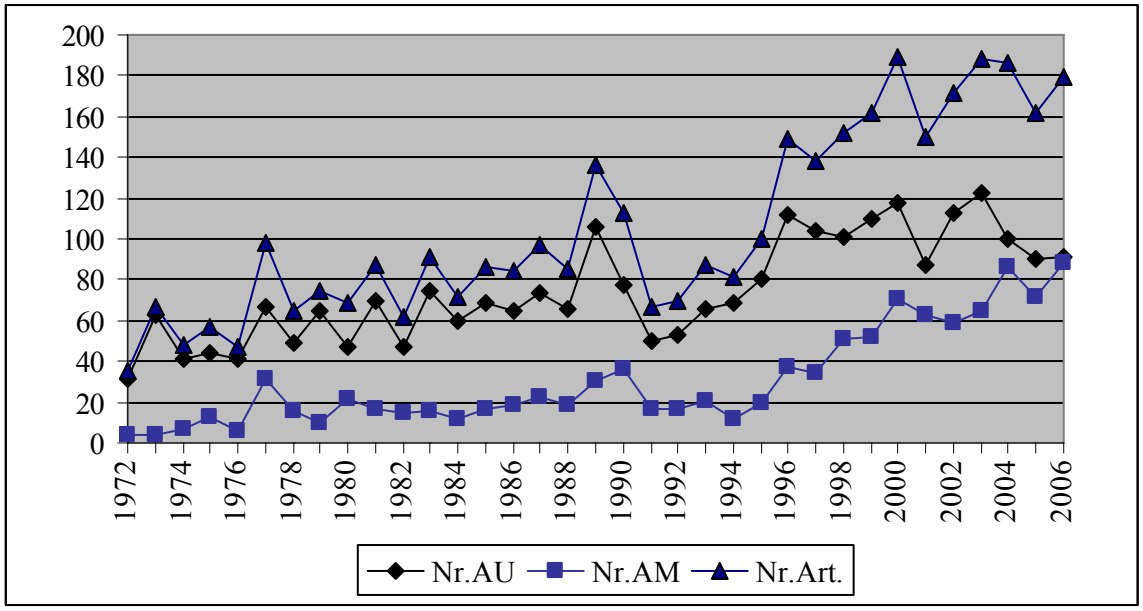


Gráfico 1 - Número de Artigos de Periódicos Científicos das Áreas de Informação no Brasil por Tipo de Autoria (1972-2006)

Observa-se no Gráfico 1 a linha de cima que representa a evolução do número total de artigos publicados. A linha do meio representa os artigos em autoria única (AU) e a última linha, os artigos em autoria múltipla (AM).

> "A partir de 2000, a produção se estabiliza no patamar médio de 175 artigos/ano. Pode-se notar ainda que a evolução da autoria única acompanha a evolução da produção total de artigos até 1996, quando o número de artigos em autoria única se estabiliza na faixa entre 80 e 120 artigos/ano e começa a se distanciar da produção total de artigos enquanto aumenta o número de artigos em autoria múltipla. Quanto à produção de artigos em autoria múltipla, permanece quase 20 anos (1977-1995) próximo da faixa de 20 artigos por ano até iniciar subida em 1996 e chegar ao patamar médio próximo de 80 artigos/ano a partir de 2004. O número de artigos em autoria única (91) e autoria múltipla (88) chegam a valores absolutos bem próximos em 2006." (VILAN FILHO, SOUZA, 2007).

Em todos os estudos realizados, os dados foram extraídos do WinISIS para arquivos do tipo texto (.txt), por meio do comando de impressão, e foram posteriormente inseridos no MS-Excel onde foram produzidas as tabelas e gráficos para análise.

Este trabalho pretende dar continuidade aos estudos bibliométricos realizados na base de dados ABCDM usando uma ferramenta de Mineração de Dados com o objetivo de identificar os principais assuntos abordados em cada década, de 1971 a 2007.

4 OBJETIVO

Aplicar técnicas de Mineração de Dados na base de dados ABCDM para identificar os principais assuntos abordados nos títulos dos artigos de cada década (1971 a 2007).

5 METODOLOGIA

Abaixo segue a descrição das etapas executadas para atingir os objetivos citados. Detalhes relacionados com a utilização da ferramenta Rapid Miner deverão ser consultados no Anexo C deste trabalho, onde se encontra a tradução (produto deste trabalho) do tutorial disponível na própria ferramenta.

5.1 ESCOLHA DA FERRAMENTA

Após análise da Tabela 1 de comparação das ferramentas de Mineração de Dados, foi escolhida a ferramenta Rapid Miner (evolução do Weka) para a aplicação das técnicas de mineração de texto deste trabalho. O Rapid Miner, além de ser gratuito e de código aberto, é bastante intuitivo.

O download do Rapid Miner pode ser feito através do site http://rapid-i.com/. É necessário também fazer o download do plugin de texto, pelo link "Download Rapid Miner Plugins", e copiá-lo na pasta de trabalho que a ferramenta cria ao ser instalada, chamada "rm_workspace".

5.2 EXTRAÇÃO DOS DADOS DA FERRAMENTA CDS/ISIS

Os dados foram exportados no dia 16 de setembro de 2008. Foram selecionados os atributos: MFN (número de identificação do artigo), ano de publicação do artigo (campo 265), título do artigo (campo 240) e subtítulo do artigo (campo 241). Para exportar apenas os conteúdos pertinentes para esse estudo foi criado o seguinte formato de saída:

MFN '; ' v265 '; ' v240 ': ' v241

Foi necessário também separar os artigos publicados nas revistas portuguesas, pois este trabalho pretende avaliar apenas a produção científica brasileira. Para isso foram utilizadas as seguintes expressões de busca na tela "Expert Search":

```
#1  ? v440: 'biblioteconomia,'
#2  ? p(v240)
#3  #2 ^ #1
```

A expressão #1 selecionou os artigos publicados nos periódicos "Cadernos de Biblioteconomia, Arquivística" e "Cadernos de Biblioteconomia, Arquivística e Documentação", que são de Portugal. A expressão #2 selecionou os artigos que tinham o campo 240 (título do artigo) presente, ou seja, todos os artigos da base. A expressão #3 excluiu o grupo de artigos recuperados na expressão #1 do grupo de artigos totais. Dessa forma obteve-se o grupo de estudo.

Ao mostrar o resultado da pesquisa #3, através do menu "Opções" e "Print Current Records", a amostra foi exportada para um arquivo".txt", no formato de saída definido anteriormente.

5.3 PREPARAÇÃO DOS DADOS

Os dados do arquivo de texto gerado na etapa anterior foram copiados e colados em uma planilha do MS Excel. Na opção "Texto para colunas" do menu "Dados", as informações foram distribuídas em colunas, onde o caractere ";" marcou a divisão. Assim a tabela ficou com três colunas: MFN, ano, título com subtítulo.

As colunas foram ordenadas pelo ano de publicação e os dados foram separados em quatro arquivos diferentes, cada um com os dados de uma década (Arquivo1 (1971-1979), Arquivo2 (1980-1989), Arquivo3 (1990-1999), Arquivo4 (2000-2007)). Os poucos artigos publicados em 2008 foram excluídos do estudo.

Com os dados divididos e os arquivos identificados pela década de publicação, as colunas que indicavam o ano de publicação dos artigos foram excluídas, pois estas não seriam necessárias na a mineração.

5.4 ESCOLHA DOS OPERADORES E DEFINIÇÃO DOS PARÂMETROS NA FERRAMENTA RAPID MINER

Com base nas etapas de pré-processamento mencionadas por Pires (2008) foram escolhidos os seguintes operadores (visualizados no modo “Expert”):

- ExcelExampleSource

Leitura de dados de planilhas do MS Excel. Nesse operador foram selecionados os arquivos com os dados, porém um de cada vez.

- StringTextInput

Geração de vetores. Foi utilizada a definição padrão dos parâmetros, inclusive o TF/IDF como método de criação de vetores.

- StringTokenizer

Decomposição dos documentos em termos. Não há parâmetros a serem definidos nesse operador.

- TokenLenghFilter

Seleção das palavras com um valor mínimo e máximo de caracteres. Foi determinado o mínimo de 4 caracteres.

- ToLowerCaseConvert

Conversão dos caracteres para caixa baixa. Não há parâmetros a serem definidos nesse operador.

- StopwordFilterFile

Remoção de termos não representativos para o documento. Foi utilizada a definição padrão dos parâmetros, sendo que o arquivo selecionado foi criado em formato “.txt” com as palavras desnecessárias (como artigos e preposições) listadas.

- K-Means

Agrupamento dos registros. Foi utilizada a definição padrão dos parâmetros, sendo que foi definida a quantidade de 8 grupos (K), pois a ANCIB (Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação) classifica a área de Ciência da Informação em 8 grupos de trabalho: GT 1: Estudos Históricos e Epistemológicos da Informação; GT 2: Organização e Representação do Conhecimento; GT 3: Mediação, Circulação e Uso da Informação; GT 4: Gestão da Informação e do Conhecimento nas Organizações; GT 5: Política e Economia da Informação; GT 6: Informação, Educação e Trabalho; GT 7: Produção e Comunicação da Informação em CT&I; GT 8: Informação e Tecnologia.

- ExcelExampleSetWriter

Exporta o resultado dos dados para uma planilha no MS Excel. Nos atributos basta selecionar o arquivo de saída do resultado.

- ResultWriter

Exporta o resultado dos dados e dos agrupamentos para um arquivo ".txt". Nos atributos basta selecionar o arquivo de saída do resultado.

A Figura 5 mostra os operadores, à esquerda, definidos na ferramenta.

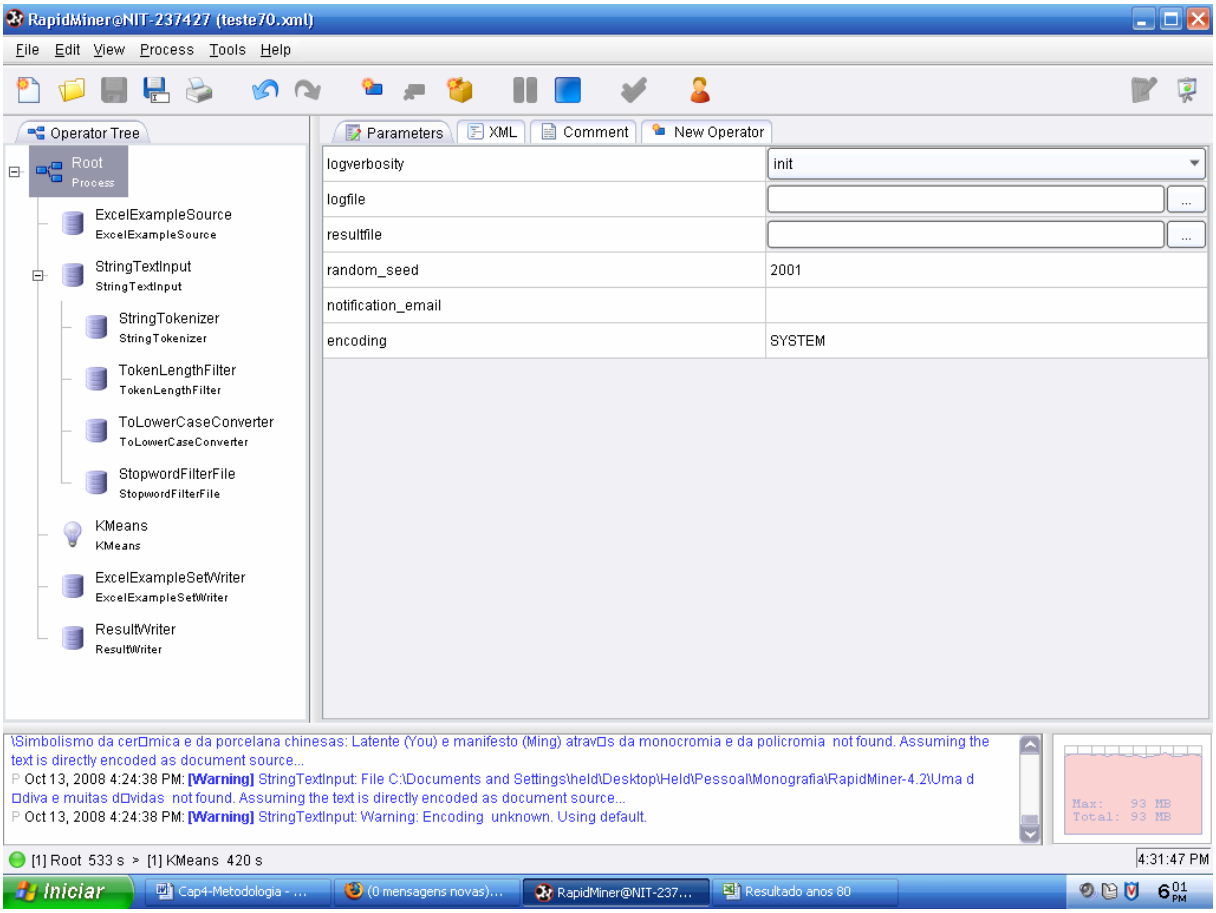


Figura 5. Operadores definidos na ferramenta Rapid Miner

Após a execução do experimento, na tela de resultados, a ferramenta mostra os resultados dos índices TF/IDF de cada termo em cada documento (Figura 6), que também é exportado para uma planilha no MS Excel, e mostra também o agrupamento dos registros (Figura 7), que também é exportado para o arquivo ".txt". Na opção de visualização "Folder View" (Figura 8) foram selecionados os 5 primeiros registros de cada cluster.

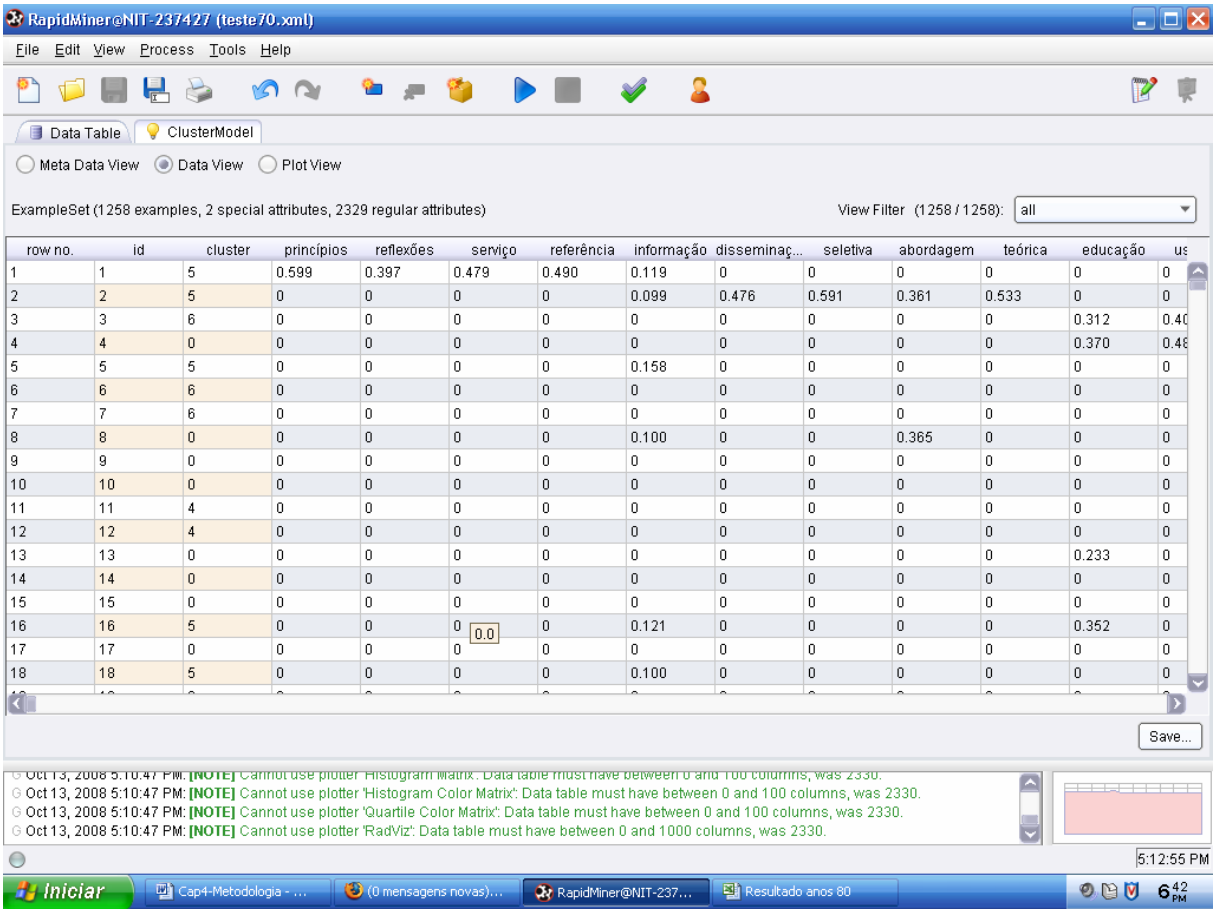


Figura 6. Resultado do experimento com os registros da década de 90: índice TF/IDF

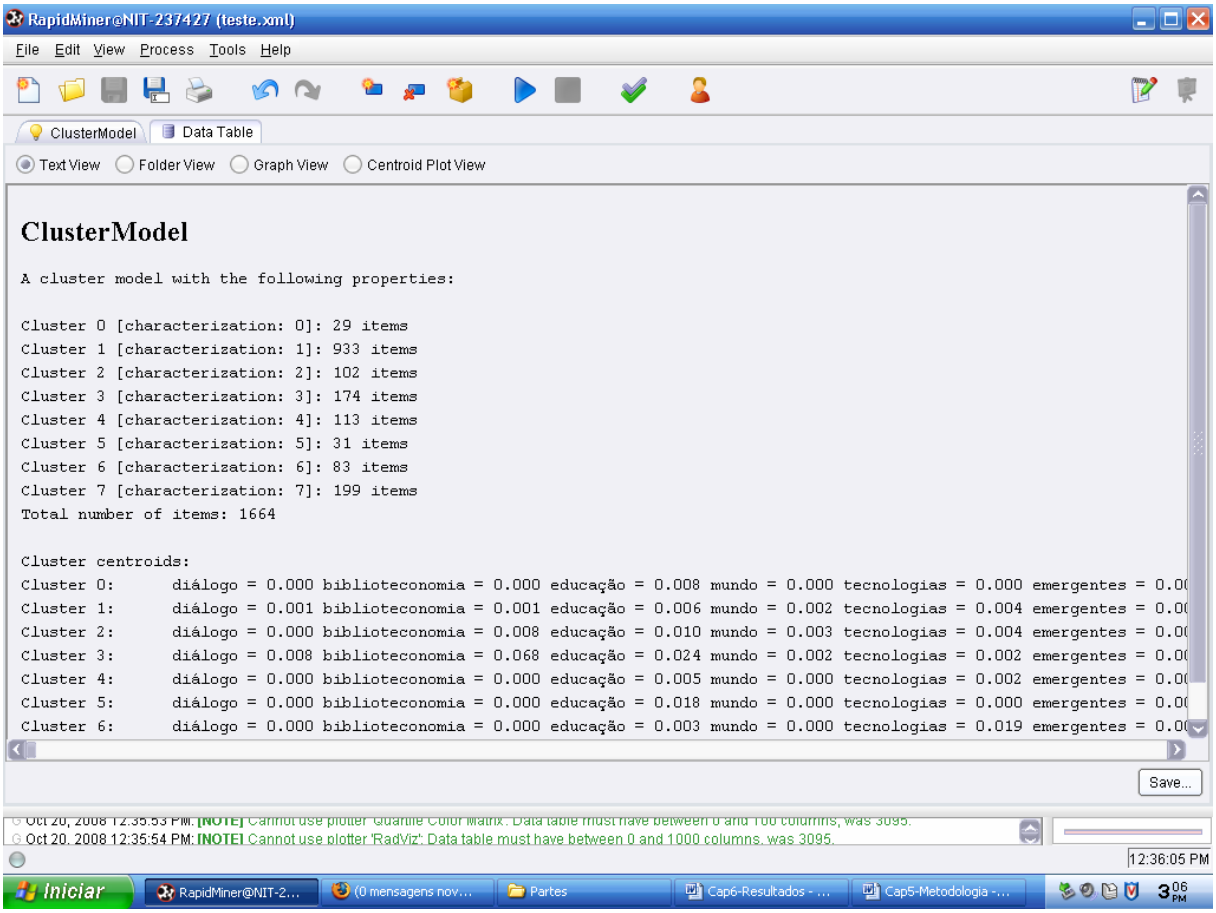


Figura 7. Resultado do experimento com os registros dos anos 2000: agrupamento dos registros

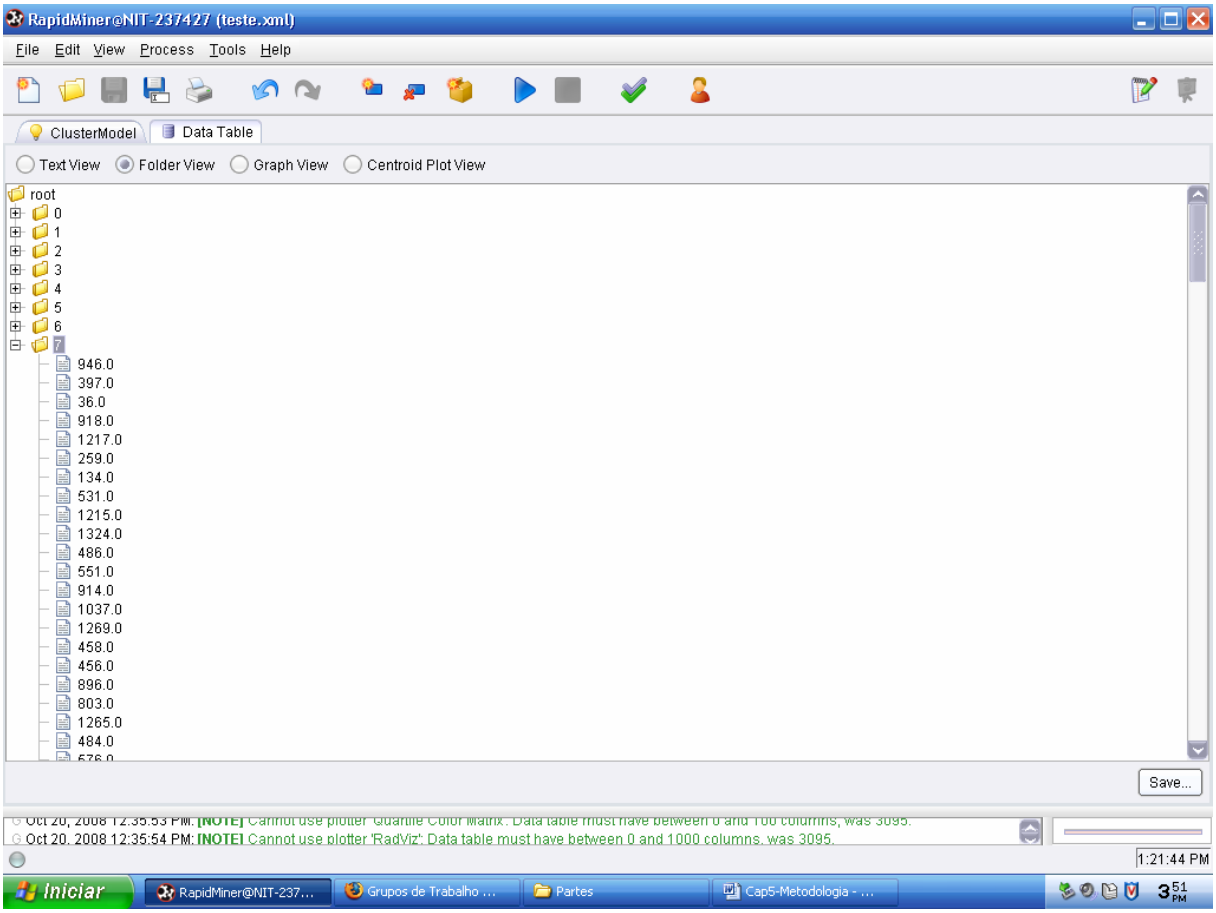


Figura 8. Resultado do experimento com os registros dos anos 2000: registros agrupados em cada cluster

5.5 TRATAMENTO DOS DADOS

Após a execução do experimento obteve-se 8 arquivos: 4 com os resultados dos agrupamentos (um arquivo referente a cada década) e 4 com os resultados dos índices TF/IDF dos termos (um arquivo referente a cada década).

Nos arquivos com os resultados dos índices TF/IDF, as colunas foram somadas para se obter o peso das marcas em relação à coleção de documentos. Depois os termos foram ordenados decrescentemente e foram obtidos os termos mais relevantes de cada coleção.

Para os agrupamentos, foi utilizado apenas o arquivo referente à coleção dos anos 2000. Nesse arquivo foram selecionados os 10 centróides mais relevantes de cada cluster com seus índices TF/IDF e a quantidade de registros de cada cluster. Na opção de visualização “Folder View” foram selecionados os 5 primeiros registros de cada cluster.

6 RESULTADOS

6.1 TERMOS MAIS RELEVANTES

Tabela 2. Índices TF/IDF dos 15 termos mais relevantes de cada década

DÉCADA							
70		80		90		2000	
TERMOS	**TF/IDF**	**TERMOS**	**TF/IDF**	**TERMOS**	**TF/IDF**	**TERMOS**	**TF/IDF**
biblioteca(s)	30,3609	biblioteca(s)	40,4659	informação(ões)	46,3521	informação	59,0041
informação(ões)	18,6440	informação(ões)	33,8240	biblioteca(s)	36,9688	biblioteca(s)	39,2129
pública(s)	11,8287	biblioteconomia	19,5343	biblioteconomia	20,8046	ciência	34,9069
brasileira(s)	9,8589	bibliotecário(s)	17,7470	brasil	19,8594	conhecimento	30,4812
biblioteconomia	8,9636	arquivo(s)	17,2708	ciência	18,5023	sociedade	24,1498
ciência(s)	8,3253	pesquisa	16,8996	bibliotecário(s)	15,4466	gestão	23,3840
nacional	8,0801	estudo(s)	16,3498	serviço(s)	14,9858	profissional(is)	22,7289
bibliografia(s)	7,7739	ciência(s)	14,9195	estudo(s)	14,4245	tecnologia(s)	21,1875
brasil	6,8850	brasileira	13,7330	profissional(is)	14,0407	estudo(s)	20,7921
documentação	6,8022	avaliação	13,6661	pesquisa(s)	13,0568	brasil	19,7284
bibliotecário	6,2041	brasil	13,5634	conhecimento	13,0555	análise	19,4812
pesquisa	6,1914	usuário(s)	12,3234	memória	12,7800	pesquisa	19,1501
sistemas	5,9677	serviço(s)	12,1712	desenvolvimento	12,1067	científica	16,9950
classificação(ões)	5,8067	análise	11,4095	educação	12,0771	rede(s)	15,4974
literatura	5,6497	ensino	11,2936	científica	11,1646	digital	14,8435

A Tabela 2 mostra os índices TF/IDF dos 15 termos mais relevantes de cada década. Esses índices também podem ser observados nos gráficos a seguir.

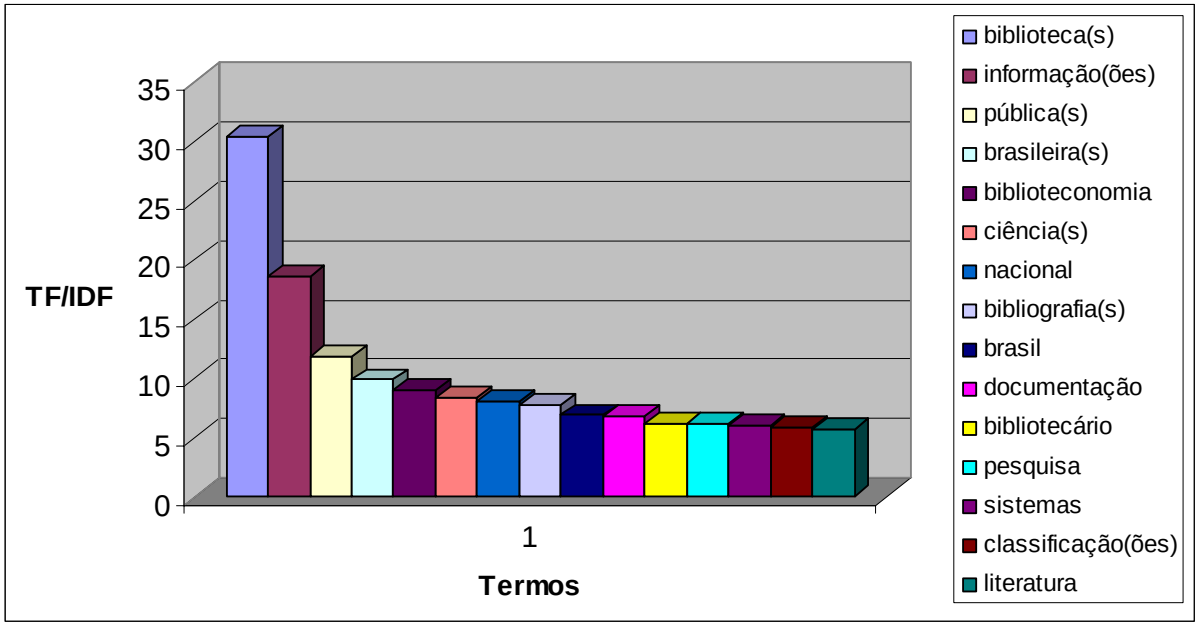


Gráfico 2. Termos mais relevantes nos títulos dos artigos publicados na década de 70

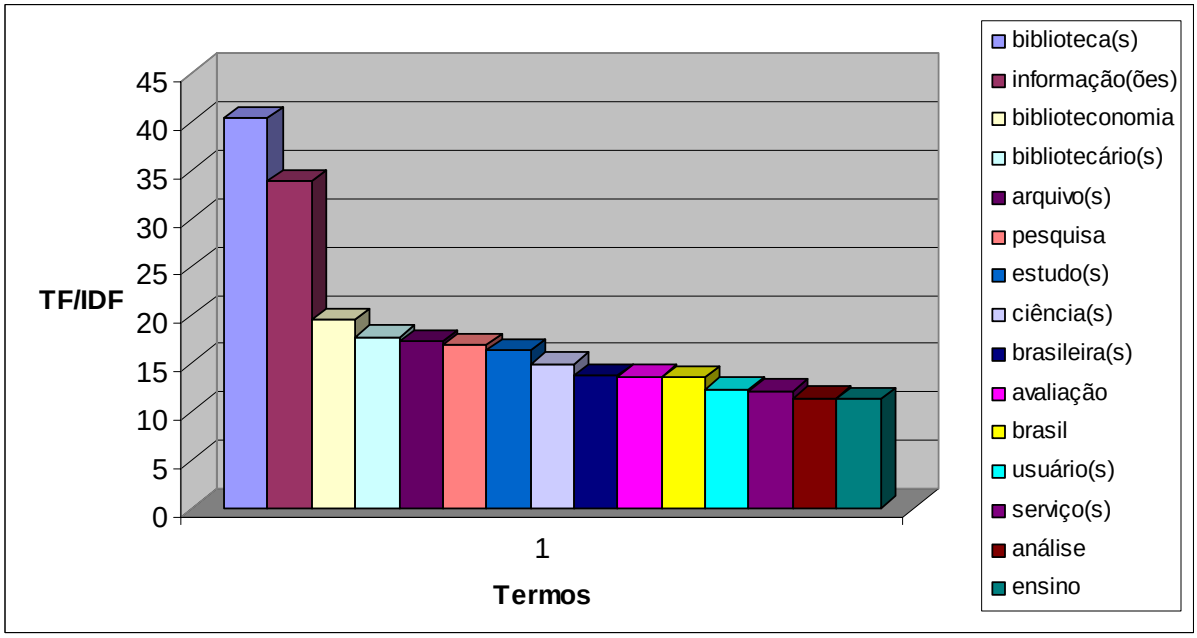


Gráfico 3. Termos mais relevantes nos títulos dos artigos publicados na década de 80

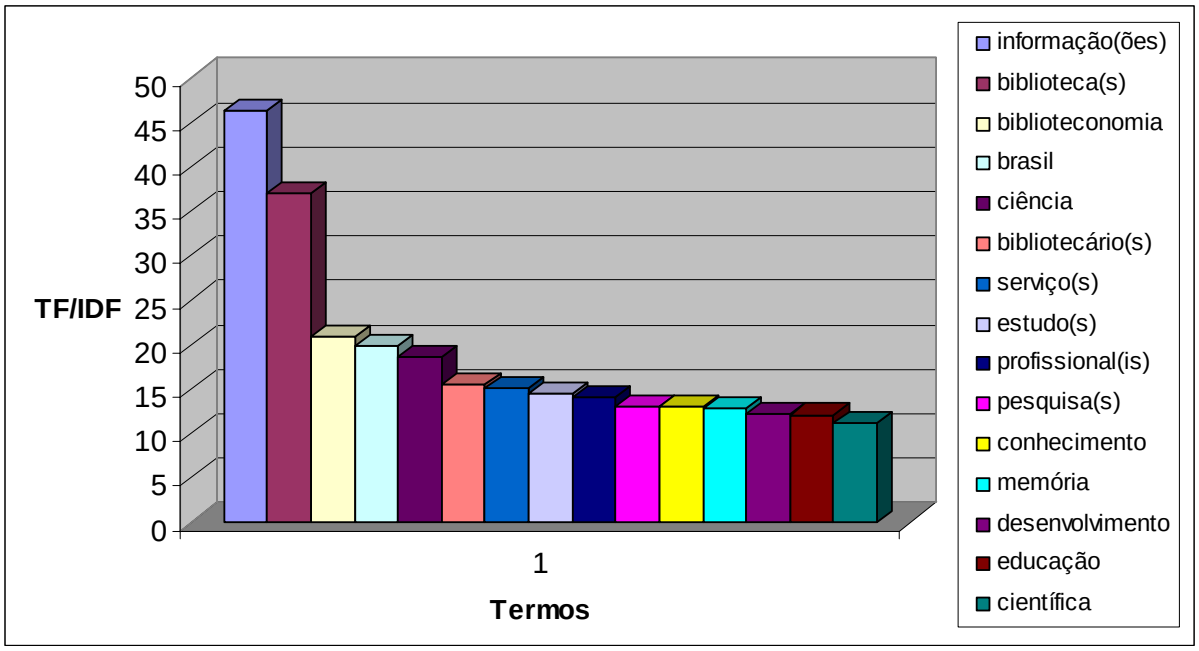


Gráfico 4. Termos mais relevantes nos títulos dos artigos publicados na década de 90

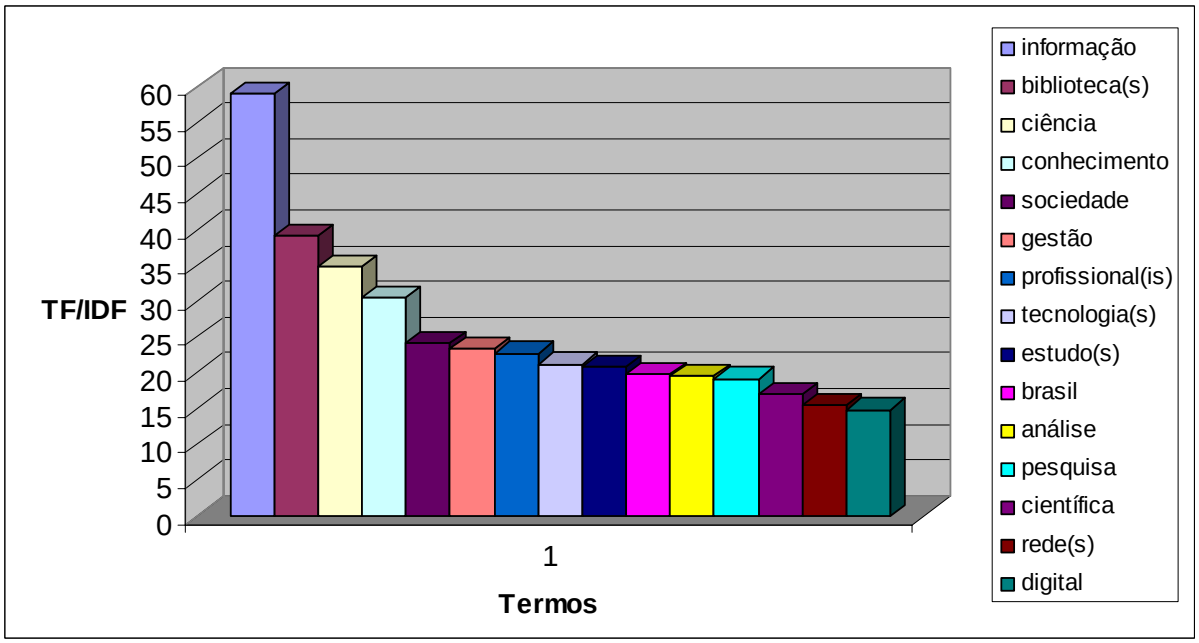


Gráfico 5. Termos mais relevantes nos títulos dos artigos publicados nos anos 2000

Conforme mencionado no item 2.4.1, que descreve o pré-processamento de uma mineração de textos, o índice TF/IDF define a representatividade do termo dentro da coleção de documentos.

Observa-se na Tabela 2 que o índice do termo “biblioteca(s)” permaneceu na mesma faixa (entre 30 e 40) enquanto que o termo “informação(ões)” cresceu, passando de 18 para 59, se tornando o termo mais importante da coleção de forma bastante expressiva. É interessante notar também que o termo “informação” não foi abordado no plural nos anos 2000.

O termo “ciência(s)” teve um crescimento gradual (de 8 a 34), passando a ocupar uma posição significativa entre os termos mais importantes. Também deixou de ser usando no plural desde a década de 90.

O termo “biblioteconomia” subiu de posição relativa na década de 80, manteve na década de 90, mas nos anos 2000 caiu, não constando na relação dos 15 mais representativos. O termo “bibliotecário(s)” ocorreu, no grupo dos 15, até a década de 90. Nessa década começa a aparecer o termo “profissional(is)”, que cresce nos anos 2000, quando o termo “bibliotecário(s)” decresce ao ponto de sair da lista dos 15.

Termos como “brasil”, “nacional” e “brasileira(s)” estiveram sempre presentes, em especial na década de 70.

Nos anos 2000 alguns termos tiveram destaque por aparecerem pela primeira vez entre os 15 mais representativos, como: “conhecimento”, “sociedade”, “gestão”, “tecnologia”, “rede(s)” e “digital”.

Alguns termos podem ser relacionados, como: biblioteca pública (na década de 70); estudo de usuários, serviços de informação (década de 80), serviços de informação, profissional da informação (década de 90), profissional da informação, tecnologias da informação, gestão do conhecimento, biblioteca digital, sociedade da informação e redes de biblioteca (nos anos 2000); além da certa relação de termos para a expressão Ciência da Informação.

O índice decrescente do termo “bibliotecário(s)” simultâneo ao crescimento do índice do termo “profissional(is)” pode significar uma substituição de um termo pelo outro nos títulos dos artigos. Essa substituição pode ser observada no cotidiano de arquivistas e bibliotecários, pois é mais comum o uso da expressão “profissionais da informação” para mencionar os atuantes dessas profissões.

O grande crescimento do termo “informação(ões)” é completamente compreensível, pois todos os registros são artigos de áreas de informação. Porém, o crescimento do termo “ciência”, que pode estar relacionado com a expressão “ciência da informação”, está relacionado com o estudo realizado por Vilan Filho e Oliveira (2008). Este estudo compara a produção de artigos de periódicos em cada área de informação. Pode-se observar no Gráfico 6 que a produção de artigos das revistas de Biblioteconomia & Ciência da Informação, e das revistas somente da área de Ciência da Informação, tiveram um crescimento significativo.

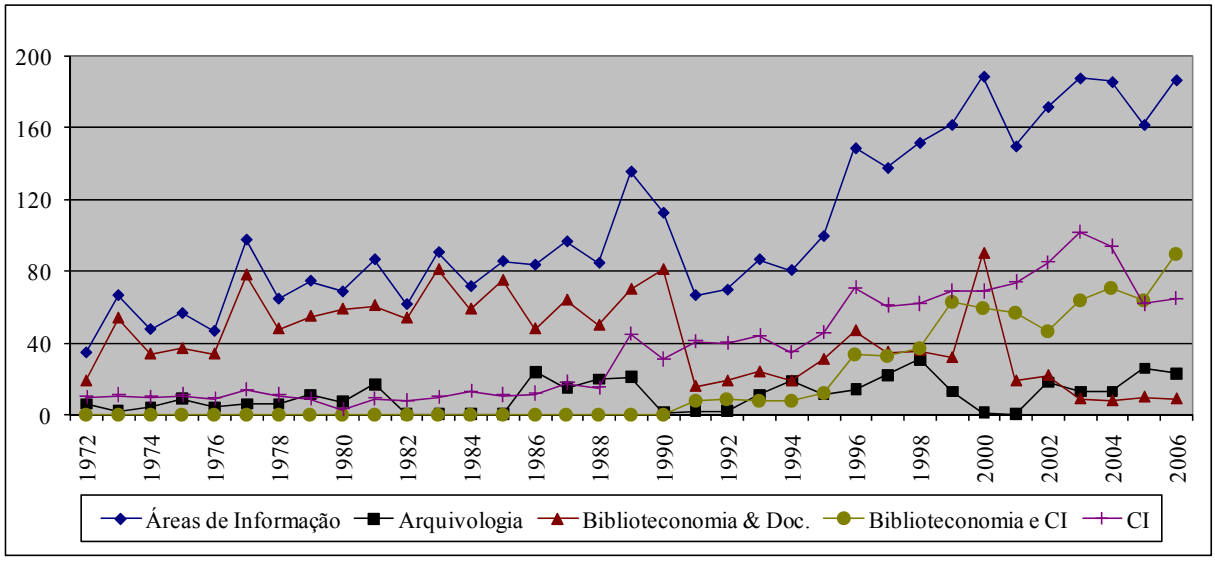


Gráfico 6. Produção de artigos científicos de periódicos das áreas de informação no Brasil (1972-2006) (VILAN FILHO; OLIVEIRA, 2008).

Outro fato relacionado com esses dados, é que no final da década de 90 ocorreram mudanças nos títulos de alguns periódicos, quando eles passaram a abordar expressamente não só artigos da área de Biblioteconomia, mas também da área de Ciência da Informação.

Portanto, a expressão “Ciência da Informação” de fato cresceu em importância nos anos 2000.

6.2 AGRUPAMENTO DOS REGISTROS

A coleção de 1664 artigos dos anos 2000 foi classificada em 8 grupos pelo algoritmo K-Means. O resultado do processo mostrou os 8 clusters com seus os centróides e seus respectivos índices TF/IDF. Na Tabela 3 pode-se observar a quantidade de itens agrupados pelo algoritmo em cada cluster, a linha do registro na tabela, o título dos 5 primeiros registros agrupados, e os 10 centróides mais relevantes do cluster com seus índices TF/IDF.

Tabela 3. Agrupamentos de registros realizados pelo algoritmo K-Means na década de 2000

CL	LN	TÍTULO	CENTRÓIDES
0 29 itens	1595	Informação e software livre no capitalismo contemporâneo	perspectiva - 0.226 contemporâneo - 0.101 histórica - 0.083 ciência - 0.057 síntese - 0.048 informação - 0.045 brasil - 0.040 moderno - 0.034 pesquisa - 0.031 atuais - 0.031
	615	A pesquisa histórica no ensino: saberes necessários à prática docente	
	1200	A formação do arquivista contemporâneo numa perspectiva histórica: impasses e desafios atuais	
	1384	Arqueologia Histórica nas Lavras do Abade: uma proposta de gestão do patrimônio	
	1134	O quarteto antropofágico: da redescoberta ao moderno e ao contemporâneo	
1 933 itens	410	Crime e castigo: as civilizadas práticas jurídicas de uma Idade Moderna	informação - 0.017 biblioteca - 0.016 museu - 0.013 museus - 0.012 memória - 0.011 científica - 0.010 análise - 0.010 digital - 0.009 avaliação - 0.009 informações - 0.009
	1645	Arte coletiva: um problema para arte-educadores?	
	709	O fio de Ariadne e a arquitetura da informação na WWW	
	1615	Uma abordagem sistêmica aplicada à arquivística	
	488	A aplicação de biblioterapia em crianças enfermas	

2 102 itens	625	Disseminação de informação para a cidadania no Brasil: uma análise da cobertura das matérias sobre indicadores sociais na mídia impressa	brasil - 0.167 história - 0.094 informação - 0.033 ciência - 0.028 pesquisa - 0.027 arte - 0.022 sociedade - 0.021 fontes - 0.021 grandense - 0.021 estudo - 0.020
	1612	A Bibliografia arquivística no Brasil: análise quantitativa e qualitativa	
	605	Correspondência e escrita da história na trajetória intelectual de Afonso Taunay	
	578	Comunidades científicas e infra-estrutura tecnológica no Brasil para uso de recursos eletrônicos de comunicação e informação na pesquisa	
	1614	Automação de arquivos no Brasil: os discursos e seus momentos	
3 174 itens	661	Terminologia da Ciência da Informação: abordagem da análise do discurso	ciência - 0.157 informação - 0.072 biblioteconomia - 0.068 pesquisa - 0.060 ensino - 0.037 curso - 0.027 análise - 0.025 educação - 0.024 redes - 0.022 tecnologia - 0.020
	289	A pesquisa científica na Ciência da Informação: análise da pesquisa financiada pelo CNPq	
	136	Cronologia da Escola de Biblioteconomia da UFMG - 1950/2000	
	489	Divulgação do curso de biblioteconomia da FURG nos municípios de Rio Grande, Santa Vitória do Palmar e São José do Norte	
	1312	A temática do desenvolvimento sustentável em grupos de pesquisa	
4 113 itens	511	Análise contrastiva: memória da construção de uma metodologia para investigar a tradução de conhecimento científico em conhecimento público	conhecimento - 0.244 gestão - 0.078 informação - 0.042 representação - 0.037 construção - 0.032 científico - 0.031 comunicação - 0.025 organização - 0.025 sociedade - 0.020 inovação - 0.019
	224	Transparência e gestão do conhecimento por meio de um banco de teses e dissertações: a experiência do PPGEP/UFSC	
	936	Profissionais da informação e o mapeamento do conhecimento nas organizações: o caso da kpmg Brasil	
	1484	Gestão do conhecimento ou gestão de organizações da era do conhecimento?: um ensaio teórico-prático a partir de intervenções na realidade brasileira	
	1079	A representação metafórica nos caminhos do conhecimento em tempos de comunicação globalizada	

5 31 itens	1022	Pesquisa em inteligência competitiva organizacional: utilizando a análise de conteúdo para a coleta e análise de dados - Parte II	inteligência - 0.353 competitiva - 0.265 organizacional - 0.115 processo - 0.071 organizações - 0.056 conhecimento - 0.040 informação - 0.038 gestão - 0.037 sociedade - 0.034 métodos - 0.032
	514	Inteligência competitiva em organizações: dado, informação e conhecimento	
	37	Sociedade da Informação e inteligência em unidades de informação	
	1026	Gestão do conhecimento como parte do processo de inteligência competitiva organizacional	
	274	Entre a Sociedade da Informação e a inteligência coletiva: educação e (in)formação para a ação emancipatória	
6 83 itens	457	Estudo de usuários em bibliotecas públicas e universitárias: em foco as dissertações defendidas no CMCI/UFPB	bibliotecas - 0.224 universitárias - 0.110 serviços - 0.049 digitais - 0.034 públicas - 0.032 brasileiras - 0.029 tecnologias - 0.019 internet - 0.019 periódicos - 0.019 virtuais - 0.019
	114	Impacto da automação sobre os funcionários das bibliotecas da Universidade Federal de Pernambuco	
	1257	Padronização da coleta de dados nas bibliotecas do SIBi/USP	
	1206	Avaliação de websites de bibliotecas universitárias da região Sul	
	81	Disponibilização do catálogo do acervo das bibliotecas da UNICAMP na web, utilizando o Altavista Search Intranet	
7 199 itens	946	Identidades, valores e mudanças: o poder da identidade profissional. Os bibliotecários subsistem na era da informação?	informação - 0.096 sociedade - 0.090 profissional - 0.073 bibliotecário - 0.054 gestão - 0.034 cidadania - 0.030 formação - 0.030 unidades - 0.029 atuação - 0.023 mercado - 0.022
	397	A formação profissional no século XXI: desafios e dilemas	
	36	Um estudo do poder na Sociedade da Informação	
	918	Novas mídias, cidadania e exclusão digital no contexto da sociedade da informação	
	1217	A mediação do profissional da informação nas florestas da sociedade da informação	

CL=Cluster; LN=Linha de origem do registro

Na Tabela 3 pode-se identificar os principais assuntos pelos quais foram feitos os agrupamentos. Nesta classificação é possível visualizar a ocorrência de assuntos mais específicos se compararmos à identificação realizada no item anterior.

Os principais assuntos de cada cluster podem ser identificados, com exceção do cluster 0, onde as palavras mais relevantes não são utilizadas como assunto principal, e do cluster 1, onde os índices dos principais termos são baixos e muito próximos. Os principais assuntos identificados foram:

- Cluster 2 – Aspectos Históricos
- Cluster 3 – Ciência da Informação e Biblioteconomia
- Cluster 4 - Gestão do Conhecimento
- Cluster 5 - Inteligência Competitiva
- Cluster 6 – Biblioteca Pública
- Cluster 7 - Sociedade da Informação

7 CONCLUSÃO

Este trabalho propôs a aplicação de técnicas de Mineração de Texto na base de dados ABCDM para identificar os principais assuntos abordados nos artigos de periódicos científicos brasileiros por década.

O objetivo foi alcançado, visto que foram identificados os 15 termos mais representativos de cada década (1971-2007), com uma análise mais específica na década de 2000, através da classificação feita pelo algoritmo K-Means. Entre os termos dos anos 2000 identificados na primeira etapa, alguns apareceram pela primeira vez entre os 15 mais representativos, em especial os termos relacionados com tecnologia da informação. O K-Means permitiu visualizar melhor estas ocorrências. Estudos posteriores podem analisar mais especificamente cada década através do agrupamento dos registros.

O K-Means se mostrou relativamente impreciso no agrupamento dos clusters 0 e 1, o primeiro por apresentar centróides de pouco significado e o segundo por apresentar centróides de índices muito baixos e muito próximos, ou seja, muitos centróides diferentes e relevantes. Estudos posteriores poderão testar o agrupamento com o K maior e comparar a precisão dos resultados obtidos.

Os resultados obtidos com a Mineração de Texto na base de dados ABCDM foram válidos, já que confirmaram algumas tendências já percebidas pelos pesquisadores das áreas de informação, como o crescimento da área de Ciência da Informação e o uso crescente de tecnologias da informação. Porém a análise dos resultados foi realizada de maneira superficial. Análises mais aprofundadas poderão convergir para outros resultados também interessantes, além dos que já foram identificados.

A experiência com a aplicação de técnicas de Mineração de Dados em dados bibliográficos foi muito interessante devido ao potencial da ferramenta. Esses estudos podem contribuir de forma significativa para a área de Comunicação Científica e Bibliometria, revelando comportamentos da produção científica brasileira que dificilmente seriam obtidos de outras formas. Além da base

de dados ABCDM, outras bases de dados bibliográficas podem ser utilizadas para descoberta de padrões, como o Repositório Institucional da UnB, a Plataforma Lattes e a Biblioteca Digital de Teses e Dissertações (BDTD).

O curso de Gestão de Tecnologias da Informação teve uma importância muito grande em nossa formação. Todas as disciplinas de uma maneira geral contribuíram muito para a realização desta pesquisa, em especial as disciplinas de "Cenários de Tecnologias da Informação" que abordou a gestão das tecnologias, o contexto organizacional e os conceitos de tomada de decisão; a disciplina de "Plataformas de Sistemas de Informação" que deu uma visão geral sobre as novas tecnologias; a disciplina de "Projetos de Sistemas de Informação" que auxiliou a elaboração do projeto da monografia; a disciplina de "Administração de Banco de Dados" e a disciplina de "Sistemas de Apoio a Decisão e Análise Informacional" que abordou conceitos que dão suporte ao processo de tomada de decisão, como Mineração de Dados, KDD e Data Warehouse.

O curso de uma maneira geral também contribuiu para o crescimento e amadurecimento profissional, abrindo muitas portas na área acadêmica e no mercado de trabalho.

8 REFERÊNCIAS BIBLIOGRÁFICAS

ARAÚJO JÚNIOR, Rogério Henrique de. **Precisão no processo de busca e recuperação da informação**. Brasília: Thesaurus, 2007.

BALBINO, Laysse Noleto. **Projeto de Atividade Complementar**: atualização da base ABCID (2006/1). Brasília, UnB/CID, 2006.

BARROSO, Bruno da Costa; FERREIRA NETO, Pedro Nolasco. **Descoberta de conhecimento na base de dados de uma locadora de filmes**. Belém, Universidade Federal do Pará, Monografia de Graduação de Bacharelado em Ciência da Computação, 2006.

BOENTE, Alfredo N. P.; GOLDSCHMIDT, Ronaldo R.; ESTRELA, Vânia V. **Uma metodologia para apoio ä realização do processo de descoberta de conhecimento em bases de dados**. In: Workshop de Computação Científica da UENF, 2. Disponível em: <http://www.boente.eti.br/ publica/artigocompleto0.pdf> Acesso em 15 de outubro de 2008.

BOENTE, Alfredo N. P; OLIVEIRA, Fabiano Saldanha Gomes de; ROSA, José Luiz dos Anjos. **Utilização de Ferramentas de KDD para Integração de Aprendizagem e Tecnologia em Busca da Gestão Estratégica do Conhecimento na Empresa**. Disponível em: <http://www.boente.eti.br/publica/artigocompleto1.pdf> Acesso em 15 de outubro de 2008.

CARVALHO, Débora Ribeiro et al. **Ferramenta de Pré e Pós-processamento para Data Mining**. Disponível em: <http://www.inf.furb.br/seminco/2003/artigos/97-vf.pdf> Acesso em 15 de outubro de 2008.

CHAPULA, C. Macias. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. **Ciência da Informação**, Brasília, v.27, n.2, p. 134-140, maio/ago. 1998.

CHEN, M.S.; HAN, J.; YU, P. S. **Data mining**: an overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, v. 8, n. 6, p.886-883, 1996.

CHEUNG, D. W.; NG, V. T.; FU, A. W. **Efficient mining of association rules in distributed databases**. IEEE Transactions on Knowledge and Data Engineering, v.8, n. 6, p. 911-922, 1996.

FAPESP. **Análise da produção científica a partir de indicadores bibliométricos**. 2005. Disponível em: <http://www.fapesp.br/indicadores2004/volume1/cap05_vol1.pdf> Acesso em: 13 de agosto de 2008.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G. & SMYTH, P. ***Advances in knowledge discovery & data mining*.** Chapter 1: From data mining to knowledge discovery: an overview. AAAI/MIT, 1996a.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G. & SMYTH, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. ***Second International Conference on KD & DM***. Portland, Oregon, 1996b.

GOLDSCHMIDT, R. **Assistência Inteligente à Orientação do Processo de Descoberta de Conhecimento em Bases de Dados**. Rio de Janeiro, Tese de Doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, 2003.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: Um Guia Prático**. Rio de Janeiro: Elsevier, 2005.

GONÇALVES, Eduardo Corrêa. **Data Mining**: novos recursos nos sistemas de banco de dados. Disponível em: <http://www.devmedia.com.br/articles/viewcomp.asp?comp=5892> Acesso em: 13 de agosto de 2008.

KING, Donald W.; TENOPIR, Carol. A publicação de revistas eletrônicas: economia da produção, distribuição e uso. **Ciência da Informação**, v.27, n.2, 1998.

LIEBTEIN, Lourdes Helene. **Data Mining**: teoria e prática. Disponível em: <http://www.inf. ufrgs.br/~clesio/cmp151/cmp15120021/artigo_lourdes.pdf> Acesso em 15 de outubro de 2008.

MARTINS, Camilla Brandel et al. **Introdução à sumarização automática**. 2002. Disponível em: <http://www.icmc.usp.br/~taspardo/RTDC00201-CMartinsEtAl.pdf> Acesso em: 09 de outubro de 2002.

MATOS, Aline Lima. **Aperfeiçoamento do catálogo bibliográfico automatizado de periódicos científicos nas áreas de Arquivologia, Biblioteconomia, Ciência da Informação e Documentação (ABCID)**. Brasília, UnB/CID, Monografia de Graduação do Bacharelado em Biblioteconomia, julho 2003.

MEADOWS, A. J. **A comunicação científica**. Brasília: Briquet de Lemos, 1999.

MENEGHINI, Rogério. Avaliação da produção científica e o Projeto SciELo. Brasília, **Ciência da Informação**, v. 27, n. 2, p. 219-220, maio/ago. 1998

MICHIE, D.; SPIEGELHALTER, D.; TAYLOR, C. **Machine Learning, Neural and Statistical Classifications**. Ellis Horwood, 1994.

MUELLER, Suzana P. Machado. O impacto das tecnologias de informação na geração do artigo científico: tópicos para estudo. **Ciência da Informação**, v. 23, n. 3, set./dez. 1994.

OLIVEIRA, Alessandra Marchiori. **Aplicação de algumas técnicas de Data Mining em bancos de dados utilizando o Weka**. Disponível em: <http://materdei.ceicom.com.br/arquivos/ Aplicação%20de%20Algumas%20Técnicas%20de%20Data....pdf> Acesso em 15 de outubro de 2008.

PICHILIANI, Mauro. Data Minig na prática: o algoritmo. K-Means. **iMasters**, 2006. Disponível em: < http://imasters.uol.com.br/artigo/4709/> Acesso em: 25 set. 2008.

PIRES, Marina Melo. **Agrupamento incremental e hierárquico de documentos**. Rio de Janeiro, Universidade Federal do Rio de Janeiro, Dissertação de Mestrado em Engenharia Civil, 2008.

ROMÃO, Wesley. **Descoberta de conhecimento relevante em banco de dados sobre ciência e tecnologia**. Florianópolis, Universidade Federal de Santa Catarina, Tese de Doutorado em Engenharia de Produção, fevereiro 2002.

SILVA, Sebastião Dimas Justo da. **Aperfeiçoamento da Base de Artigos de Periódicos Científicos das Áreas de Arquivologia, Biblioteconomia, Ciência da Informação e Documentação (ABCID)**: melhoria da qualidade e atualização dos nove títulos já catalogados. Brasília, UnB/CID, Monografia de Graduação do Bacharelado em Biblioteconomia, março 2005

SOUZA, Held Barbosa de. **O reflexo da colaboração científica nos periódicos**: uma análise da co-autoria em artigos das áreas de Arquivologia, Biblioteconomia, Ciência da Informação e Documentação publicada no Brasil. Brasília, UnB/CID, Monografia de Graduação de Bacharelado em Biblioteconomia, dezembro 2006a.

SOUZA, Held Barbosa de. **Projeto de Atividade Complementar**: atualização da base ABCID (2006/1). Brasília, UnB/CID, 2006b.

SOUZA, Held Barbosa de; SILVA, Alessandra Marinho da. **Projeto de Atividade Complementar**: atualização da base ABCID (2005/2). Brasília, UnB/CID, 2005.

SPINAK, E. Indicadores Cienciométricos. **Ciência da Informação**, Brasília, v. 27, n. 2, p. 121-148, maio/ago., 1998.

TENOPIR, Carol; KING, Donald W. A importância dos periódicos para o trabalho científico. **Revista de Biblioteconomia de Brasília**, v.25, n. 1, jan./jun. 2001.

VIEIRA, José Ronaldo. **Aperfeiçoamento da Base de Artigos de Periódicos Científicos das Áreas de Arquivologia, Biblioteconomia, Ciência da Informação e Documentação (ABCID)**: obtenção de um catálogo com todas as 20 revistas em língua portuguesa. Brasília, UnB/CID, Monografia de Graduação do Bacharelado em Biblioteconomia, março 2005.

VILAN FILHO, Jayme Leiro**. Manual de manutenção da base ABCDM em CDS/ISIS**. Brasília, 2008.

VILAN FILHO, Jayme Leiro; SOUZA, Held Barbosa de. Artigos de periódicos científicos das áreas de informação no brasil: evolução da produção e da autoria múltipla. In: Encontro Nacional de Pesquisa em Ciência da Informação, 8, 2007, Salvador. **Anais...** Salvador: ENANCIB, 2007.

VILAN FILHO, Jayme Leiro; SOUZA, Held Barbosa de; MUELLER, Suzana. Artigos de periódicos científicos das áreas de informação no Brasil: evolução da produção e da autoria múltipla. **Perspectivas em Ciência da Informação**, v. 13, p. 2-17, 2008.

VILAN FILHO, Jayme, OLIVEIRA, Eliane Braga de. A produção de artigos nos periódicos científicos brasileiros de Arquivologia (1972-2006). In: Congresso Brasileiro de Arquivologia, 15. **Anais...** Goiânia, 2008.

WORMELL, Irene. Informetria: explorando bases de dados como instrumentos de análise. Brasília, **Ciência da Informação**, v. 27, n. 2, mai./ago. 1998

ANEXO A

Relação de campos e forma de preenchimento da base de dados ABCDM (VILAN FILHO, 2008).

(8) Idioma do Artigo

(100) Autor Principal Pessoal

^a nome do autor do artigo

^b último sobrenome do autor do artigo

^c afiliação do autor

^d notas do autor

^e endereço eletrônico do autor

(110) Autor Principal Corporativo

^a nome da entidade principal autora do artigo seguido de sua sigla entre parênteses

^b nome da entidade subordinada autora do artigo seguido da sua sigla entre parêntese

^c local

^d notas de autor corporativo

^e endereço eletrônico do autor

(240) Título do Artigo

(241) Subtítulos do Artigo

(242) Data do Recebimento do Artigo

(243) Data de Aceitação do Artigo

(250) Título em Outro Idioma

(251) Subtítulos em Outro Idioma

(260) Local de Publicação

(261) Editora(s)

(262) Volume

(263) Número do Fascículo

(264) Período do Fascículo

(265) Ano do Fascículo

(267) Ano Final de Fascículo Cumulativo

(300) Paginação

(440) Título da Publicação

(441) Subtítulo da Publicação

(442) Seção da Publicação

(445) Título Abreviado da Publicação

(446) Sigla da Publicação

(447) ISSN

(448) e-ISSN

(500) Notas Gerais

(520) Resumo

(521) Abstract

(522) Resumen

(523) Resumé

(600) Palavras-Chave

(601) Keywords

(602) Palabras-Clave

(603) Mots-Clef

(690) Área do Conhecimento:

- ‘A’ - Arquivologia;
- ‘B’ - Biblioteconomia;
- ‘C’ - Ciência da Informação;
- ‘D’ - Documentação;
- ‘M’ - Museologia;
- ‘O’ - Outros. Deve ser usado para especificar especialmente as áreas correlatas: Administração, História, Ciência da Computação, Cultura, Artes, Educação, Sociologia, entre outros, quando o artigo tiver mais de uma área do conhecimento. Não deve ser usado sozinho em um registro;
- ‘X’ – Indefinido. Deve ser usado quando o catalogador não tiver certeza da área do conhecimento.

(700) Autor Secundário Pessoal

(710) Autor Secundário Corporativo

(850) Acesso Eletrônico

^a endereço completo para acesso às versões eletrônicas do artigo

^b tipo de arquivo

^c tamanho do arquivo

(990) MFN na ABCID

(999) Mensagem de Abertura

APÊNDICE A

Tutorial da ferramenta Rapid Miner.

Tutorial

RapidMiner

Tutorial do Usuário

SISTEMA RAPIDMINER

Tutorial RapidMiner, versão 4.2

Manual do Usuário
1ª Edição – setembro/2008

Traduzido e adaptado do tutorial on-line
por Daniela Leite Naglis e Held Barbosa de Souza

Sumário

Prefácio

Esta é a versão em Português, traduzida e adaptada, do Tutorial on-line do sistema RapidMiner, versão 4.2 por Daniela Leite Naglis e Held Barbosa de Souza. O tutorial está em processo de tradução e deverá passar por uma revisão futura. As partes ainda não traduzidas foram destacadas em itálico.

Sobre o RapidMiner

O RapidMiner é o primeiro software livre de solução universal para data mining, devido à combinação de tecnologias e alcance da funcionalidade. As aplicações do RapidMiner cobrem uma extensa cadeia de palavras reais das tarefas de data mining.

Mais de 400 operações de data mining podem ser usadas em combinações quase arbitrárias. A instalação é descrita por arquivos XML que podem facilmente ser criados com a interface gráfica. Esses arquivos XML são baseados em uma linguagem de script, tornando o RapidMiner uma ferramenta desenvolvida em um ambiente integrado para uma maquina de aprendizagem de data mining. O RapidMiner ultrapassou o conceito de primeiro *rapid prototyping* muito rápido para obter o resultado desejado. Alem disso, o RapidMiner pode ser usado como uma biblioteca de data mining Java.

ÍCONES
Procedimento
Ação do Sistema
Observação
Importante

Bem vindo ao tutorial Rapidminer

Bem vindo ao tutorial RapidMiner!
Este tutorial mostra conceitos básicos do RapidMiner que é um processo simples de ser executado. O usuário deve ter alguns conhecimentos e domínio da máquina de aprendizagem e mineração de dados.
Sempre que este tutorial referir-se ao "Tutorial RapidMiner", isso significa que a versão impressa, que está disponível em http://rapid-i.com

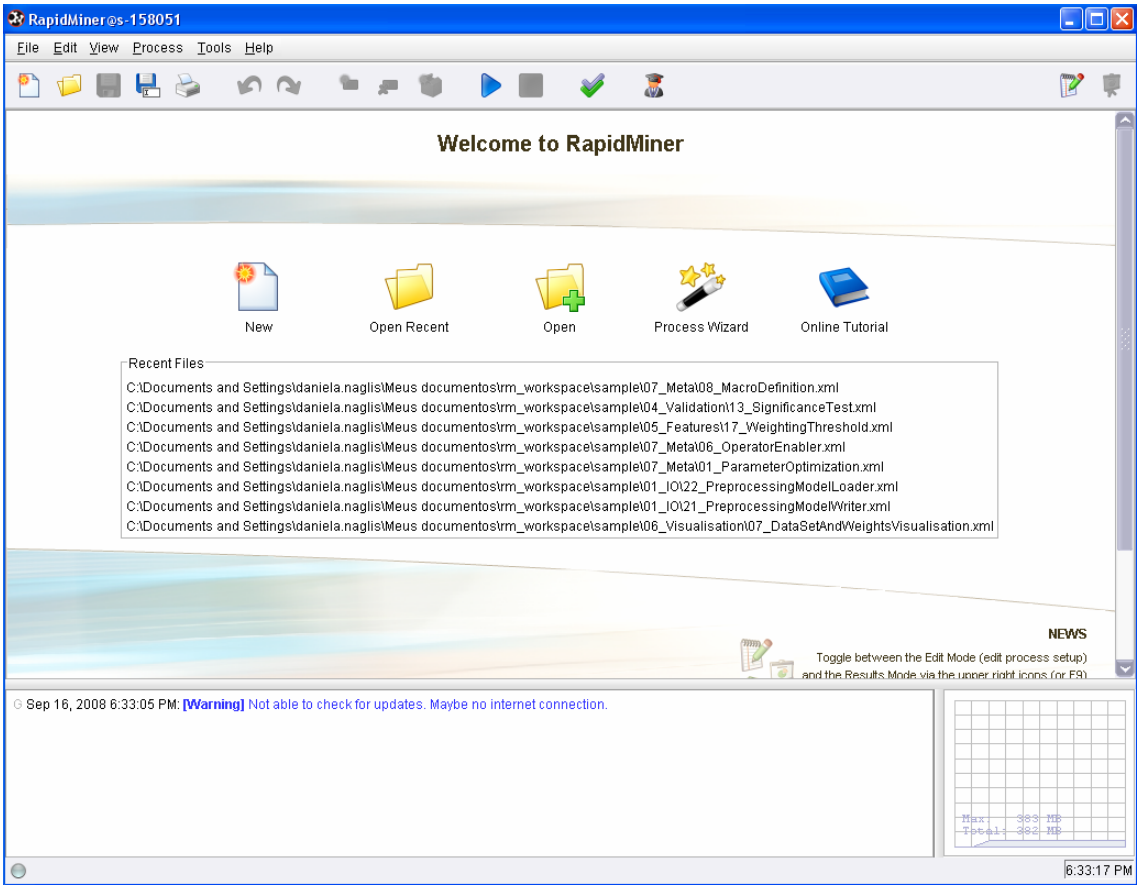
Você deve ler o primeiro capítulo do Tutorial RapidMiner para uma maior motivação, mas você também pode tentar aprender com o tutorial on-line sem ler a versão impressa. Por favor, leia o texto com atenção e tente realizar, pelo menos, os passos sugeridos. O tutorial on-line levará cerca de uma hora.

Por favor, note:
A maior parte do RapidMiner fornece informações adicionais caso você mantenha o ponteiro do mouse por alguns minutos na Ferramenta aparecerá a dica de textos (*Tooltip textos*). Desta forma todos os operadores e os parâmetros são descritos também.

ÍCONES
Procedimento
Ação do Sistema
Observação
Importante

- Para acessar o Tutorial Online do Rapidminer, você poderá clicar sobre o ícone 

que é exibido na tela inicial do sistema ou acessar o menu Help e o sub-menu RapidMiner_Tutorial, conforme a figura abaixo:


RapidMiner@s-158051
File Edit View Process Tools Help
Welcome to RapidMiner
New
Open Recent
Open
Process Wizard
Online Tutorial
Recent Files
C:\Documents and Settings\daniela.naglis\Meus documentos\rm_workspace\sample\07_Meta\08_MacroDefinition.xml
C:\Documents and Settings\daniela.naglis\Meus documentos\rm_workspace\sample\04_Validation\13_SignificanceTest.xml
C:\Documents and Settings\daniela.naglis\Meus documentos\rm_workspace\sample\05_Features\17_WeightingThreshold.xml
C:\Documents and Settings\daniela.naglis\Meus documentos\rm_workspace\sample\07_Meta\06_OperatorEnabler.xml
C:\Documents and Settings\daniela.naglis\Meus documentos\rm_workspace\sample\07_Meta\01_ParameterOptimization.xml
C:\Documents and Settings\daniela.naglis\Meus documentos\rm_workspace\sample\01_IO\22_PreprocessingModelLoader.xml
C:\Documents and Settings\daniela.naglis\Meus documentos\rm_workspace\sample\01_IO\21_PreprocessingModelWriter.xml
C:\Documents and Settings\daniela.naglis\Meus documentos\rm_workspace\sample\06_Visualisation\07_DataSetAndWeightsVisualisation.xml
NEWS
Toggle between the Edit Mode (edit process setup)
Sep 16, 2008 6:33:05 PM: [Warning] Not able to check for updates. Maybe no internet connection.
6:33:17 PM


Help
About RapidMiner...
RapidMiner Tutorial...
RapidMiner GUI Manual...
Need Support?

1 – Operador Root

Clique sobre o operador denominado "root" 

na exibição em árvore. Este é o operador Raiz de uma árvore muito simples. Do lado direito da tela parâmetros globais podem ser alterados.

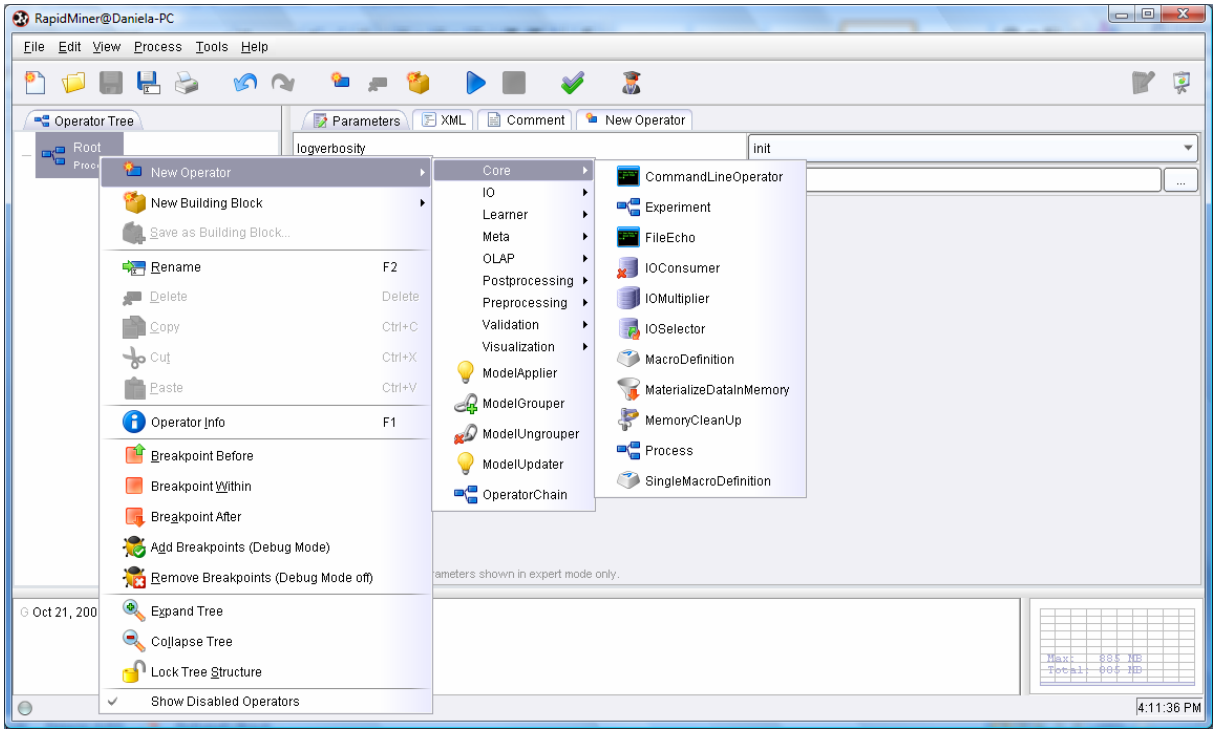
Duas perspectivas básicas ou modo de visualização estão disponíveis no RapidMiner: primeiro, modo de edição exibindo o operador em árvore, os parâmetros e todas as ferramentas necessárias para o projeto de experimentos. Segundo, a tela mostra resultados intermediários do atual experimento. Eles serão exibidos depois da experiência concluída.

Você pode mudar para o modo de edição pressionando o botão  localizado no canto direito da barra de ferramentas.

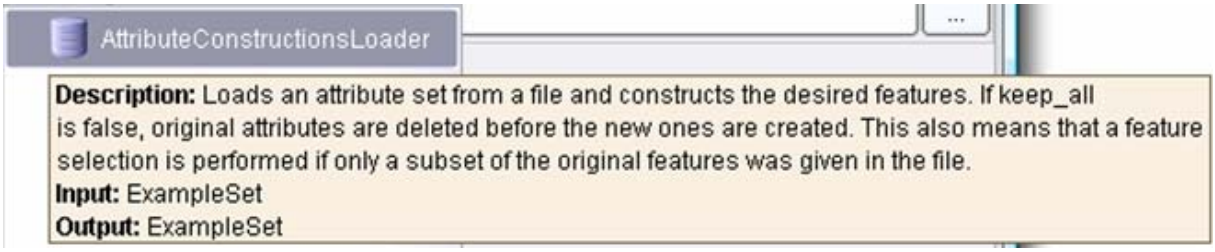
O Modo resultado é ativado ao pressionar esse botão  . A tela de resultado é ativada automaticamente no final desta experiência

Por favor, tente o seguinte:

• Mude de volta para visualização em árvore  . Clique com o botão direito do mouse, sobre o operador Raiz. Um menu de contexto permite algumas ações. Tente adicionar outro operador. Você tem que selecionar o item do menu "New Operator" e um operador de um dos subgrupos.



Breve informação sobre os operadores são exibidas pela ferramenta como dica de texto. Note que os novos operadores só podem ser adicionados ao operador em cadeia.

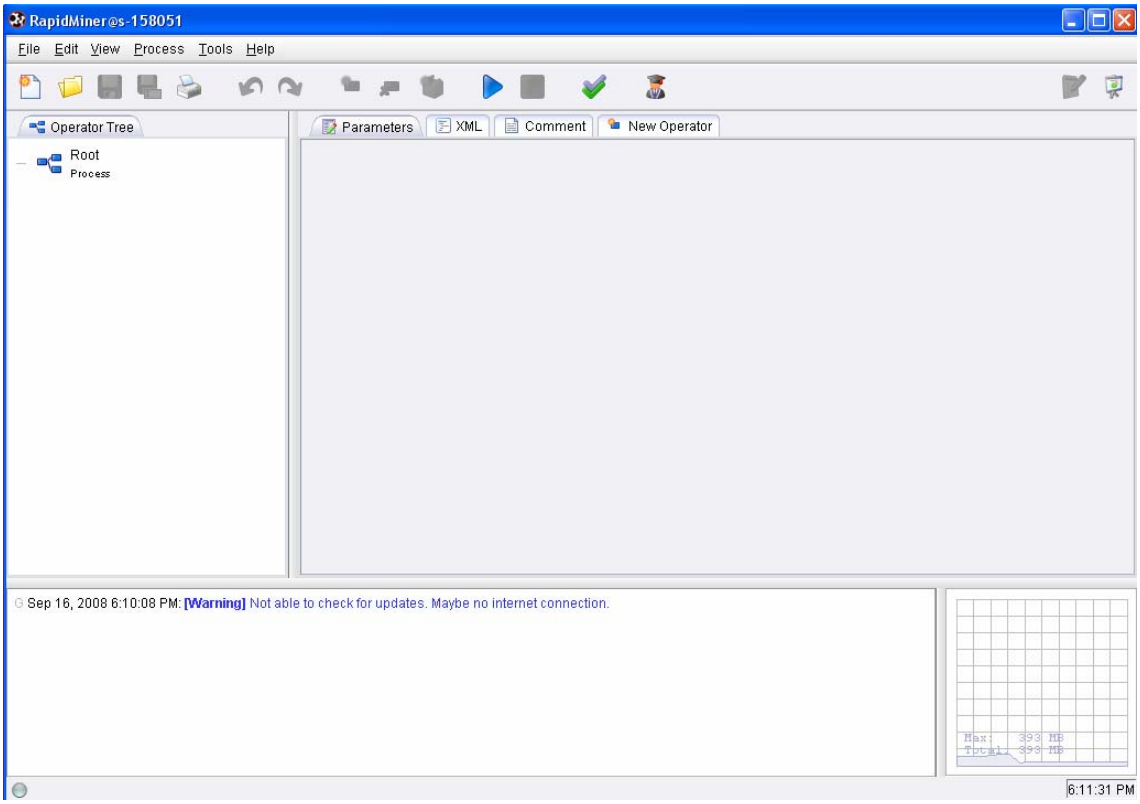


Selecione um operador de um dos subgrupos "*New operator*" do menu de contexto de um operador de cadeia selecionado.

• Após adicionar um ou vários operadores arbitrários clique no botão "*Validate Experiment*" na barra de ícones  (o que fica a direita com a marca de verificação).

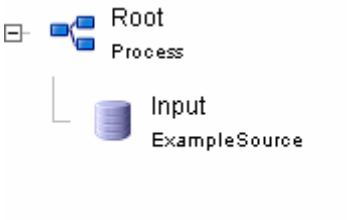
Talvez, algumas mensagens de erro serão exibidas na parte inferior da tela. Nas próximas etapas vamos ver como podem ser criadas experiências válidas.

Validate experiment setup
Quando acabar você pode avançar para a próxima experiência.

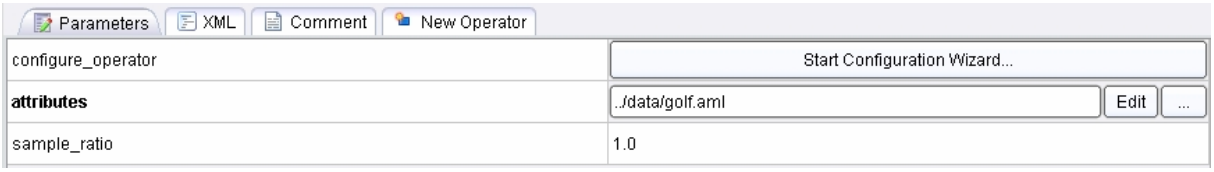


2 – Guias “Parameters”, “XML”, “Comment” e “New Operator”

Antes de podermos analisar os dados e tentar encontrar uma hipótese que explique a estrutura inerente, os dados são carregados a partir do arquivo. Por conseqüência, o primeiro operador *child* chamado "*Input*" é utilizado.

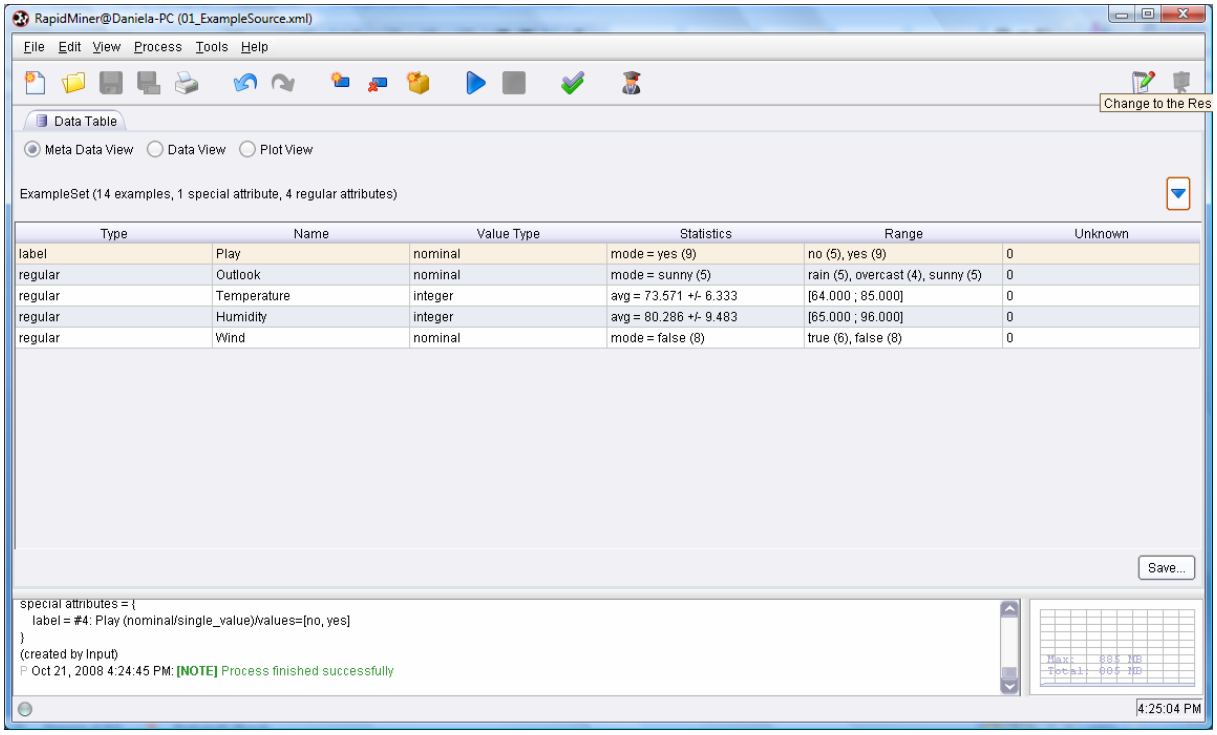


Se você selecionar o operador da entrada da propriedade da tela ao lado direito será exibido alguns parâmetros deste operador. Se um desses parâmetros for exibido com uma fonte diferente, isto significa que este parâmetro é obrigatório e deve ser definido para a utilização deste operador. Um experimento só pode ser iniciado se todos os parâmetros obrigatórios foram definidos.



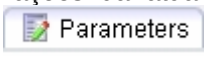
Experimente o seguinte:

• Pressione o botão "*Play*"  para iniciar a experiência. Após alguns momentos a experiência deve terminar e a interface de usuário mudará automaticamente para o modo "*Results*".

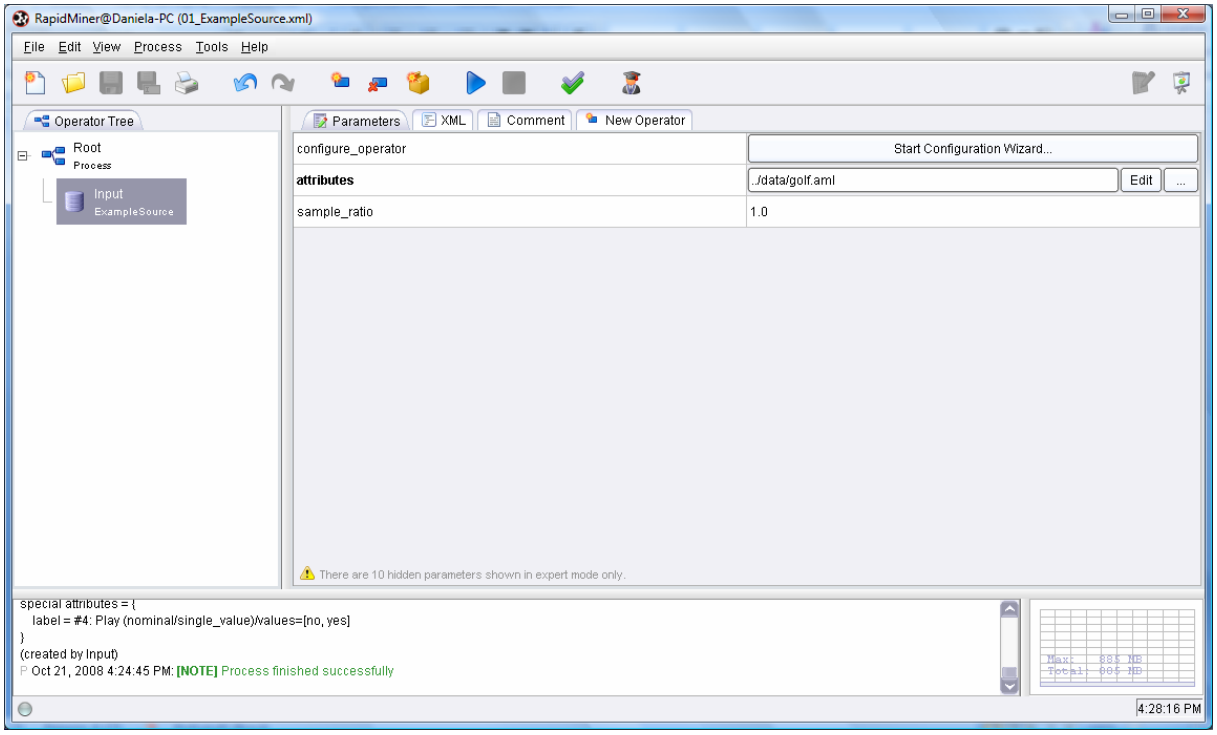


Os dados carregados (exemplo dado) são mostrados juntamente com algumas estatísticas básicas.

• Volte para visualizar a tela de edição  e veja as mudanças entre as guias "*Parameters*", "*XML*", "*Comment*", e "*New Operator*".

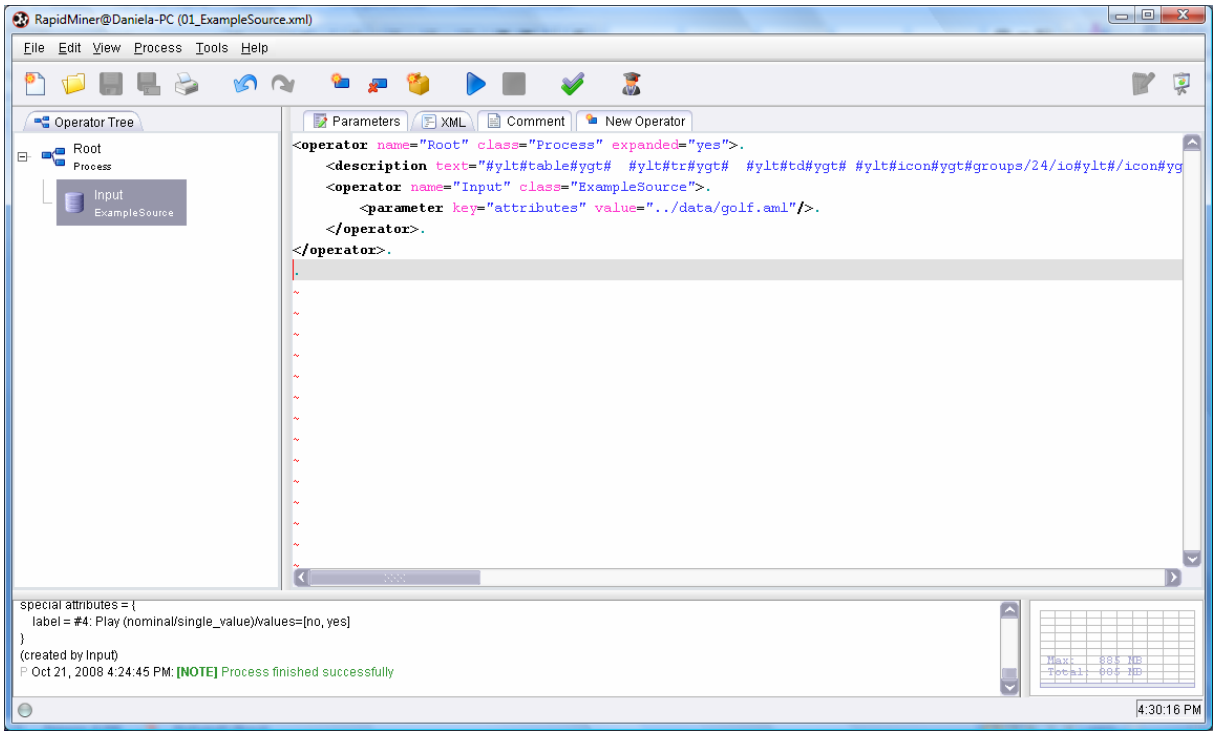
As primeiras três guias podem ser usadas para fazer as alterações da atual experiência ou do operador selecionado no momento. *Parameters* 

permitem a adaptação dos parâmetros do operador selecionado no momento.



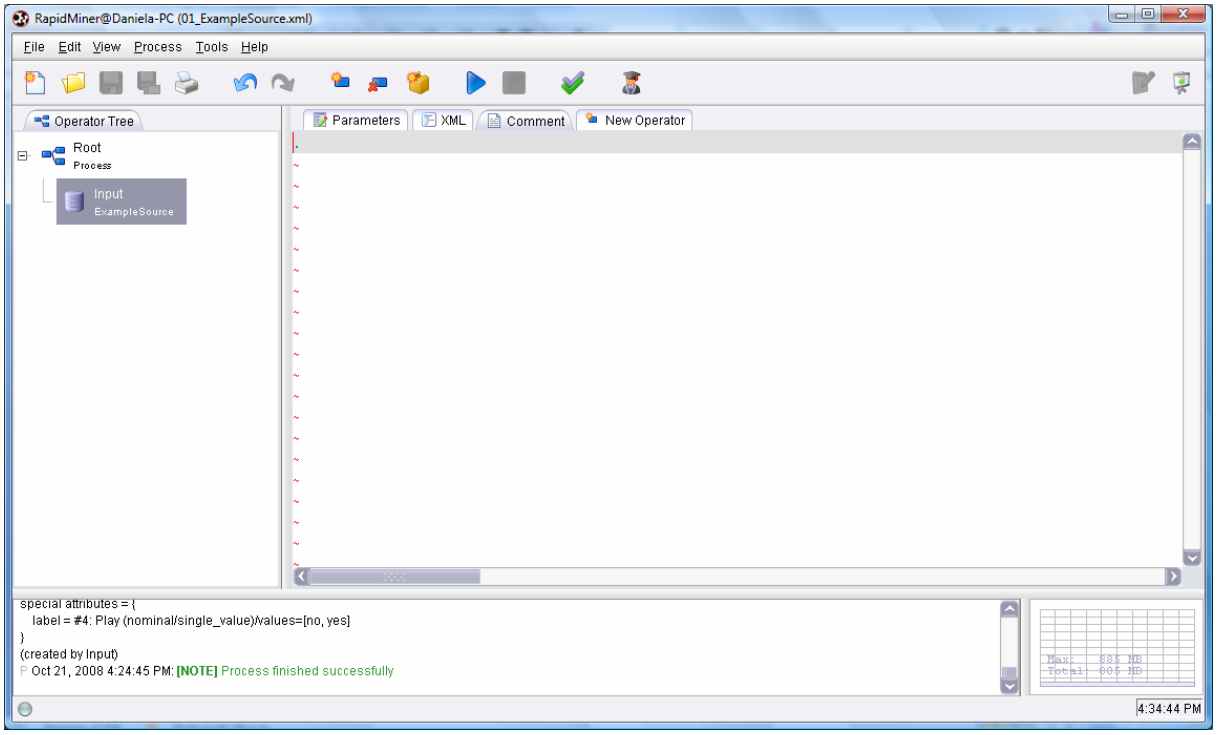
A visualização do *XML* 

pode ser usada para mudar rapidamente a configuração da experiência.



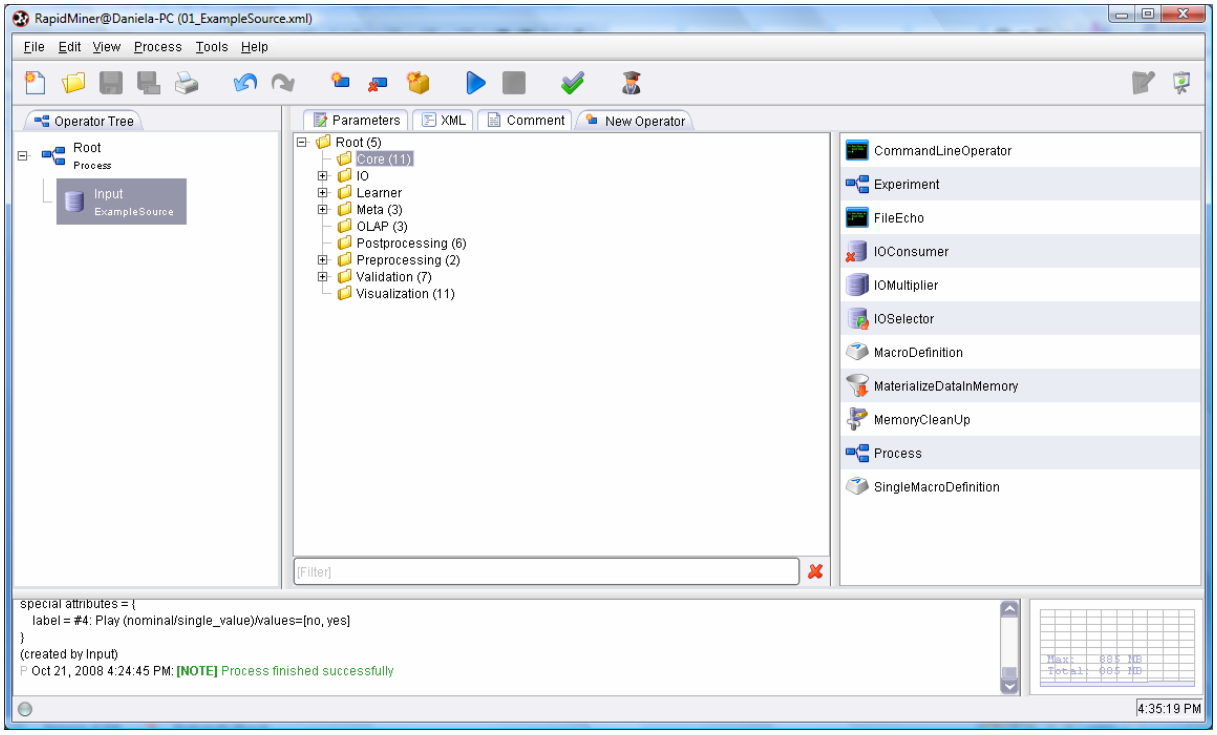
A guia *Comment* 

pode ser utilizada para descrever os comentários do usuário para o atual operador.



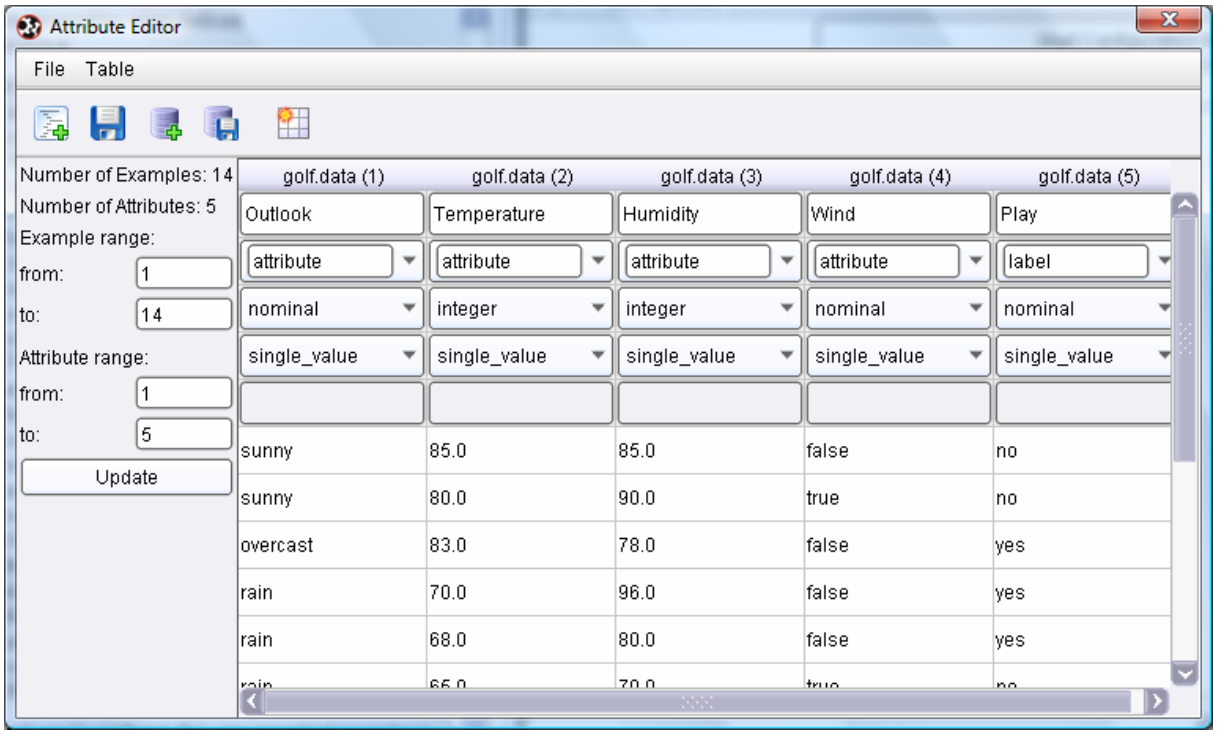
A guia *New Operator* 

fornece todos os operadores do RapidMiner agrupados em um repositório que permite visualizar novos operadores que podem ser arrastados com o mouse para árvore de operadores.

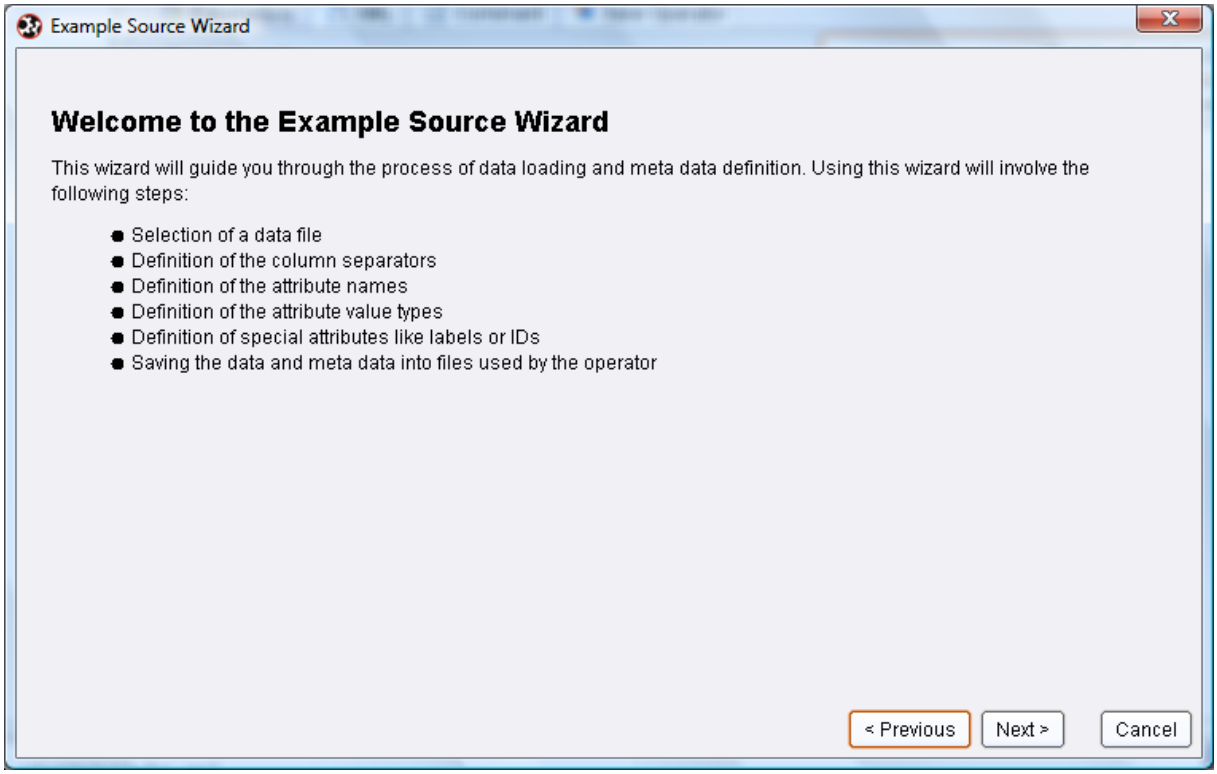


• Selecione o operador de entrada. O arquivo é determinado pelo parâmetro "*Atributes*" que é um atributo de descrição do arquivo que descreve os dados.

Clique no botão Edit para começar a alterar o atributo. Com o atributo editor você pode facilmente carregar seus dados e criar o atributo de descrição arquivos utilizados pelo RapidMiner.

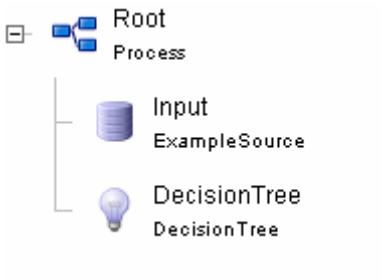


A maneira mais fácil de configurar esse importante operador será utilizar o assistente de configuração que será iniciado ao clicar no botão do topo da tabela parâmetro! Start Configuration Wizard...



3 – Inicio da Experiência com Operadores de aprendizagem e numéricos

Esta experiência começa carregando os dados. Após terminar de digitar os dados é realizada a etapa típica do operador de Aprendizagem. Aqui, uma execução de uma árvore de decisão é utilizada também para lidar com valores numéricos (semelhante ao bem conhecido algoritmo C4.5).



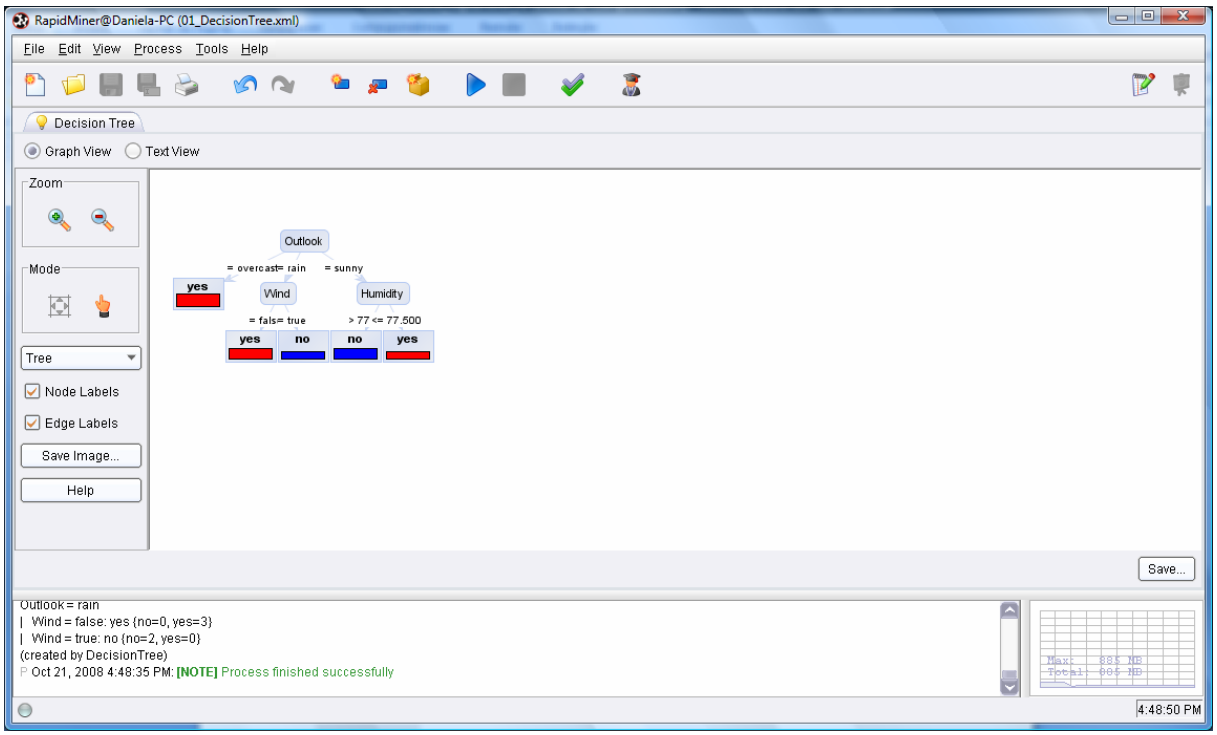
Cada operador poderá exigir algumas entradas e oferecer algumas saídas. Estes tipos de entrada e saída são transferidos entre os operadores. Neste exemplo o primeiro operador "input" não demanda entrada e fornece um exemplo que é definido como saída. Esse exemplo é utilizado pelo *Learner* que entrega o resultado final: “*operator chain*”.

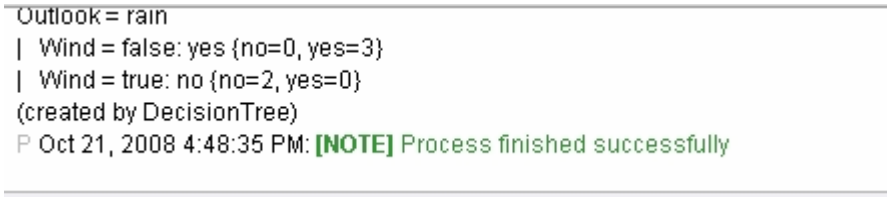
Uma vez que se trata de um fluxo linear de dados dessa experiência é chamado de "*Operator Chain*". Mais tarde veremos experiências mais sofisticadas, sob a forma de operador em árvore.

Experimente o seguinte:

• Pressione a tecla "Play"  , na barra o ícone que fica no topo da tela.

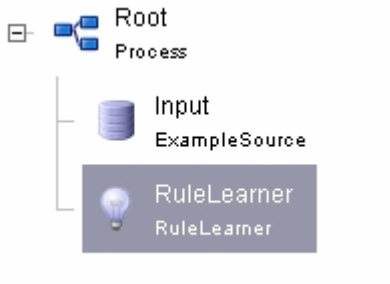
O experimento deverá iniciar e após um curto período de tempo, a mensagem na parte inferior da tela mostra que a experiência foi concluída com êxito. O quadro “*Results*” mostra as principais alterações realizadas na árvore de decisão de aprendizagem (chamada Modelo no RapidMiner).



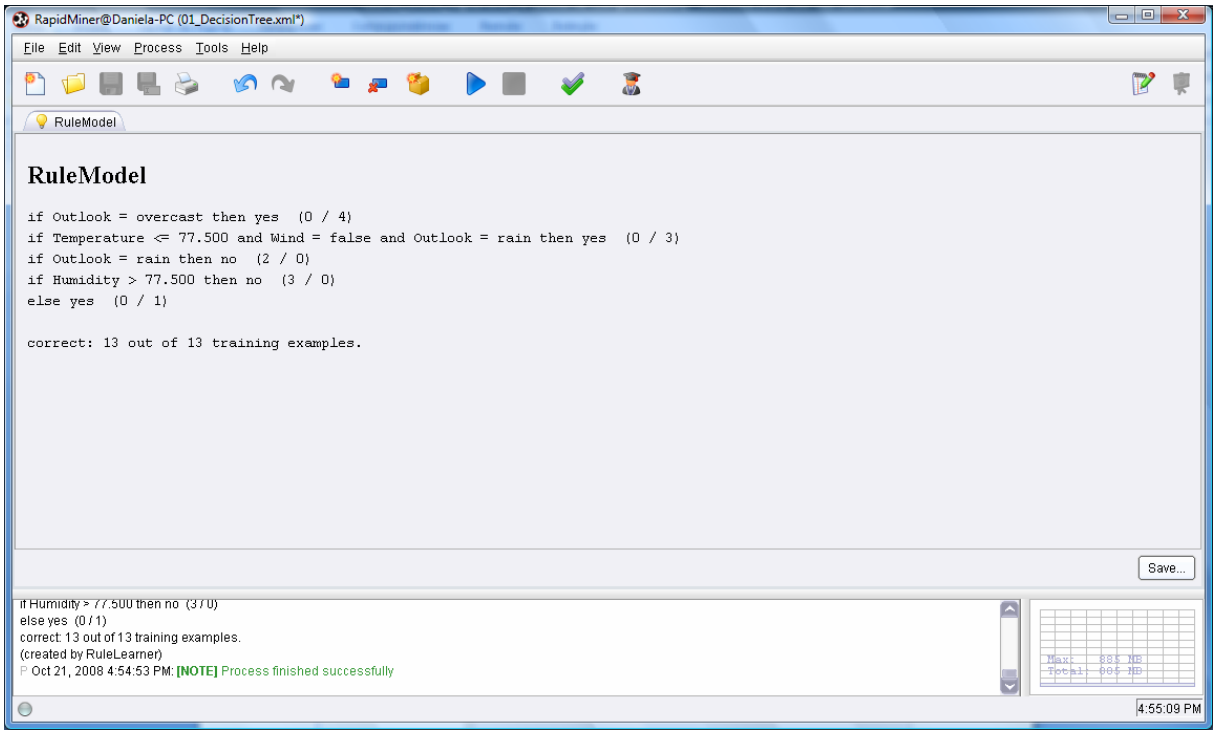


Pressione o botão "*Play*"  para iniciar a experiência

• Volte ao modo de edição (quer seja através do menu Exibir (View) entrada, ícone no canto superior direito  , ou através da tecla de atalho F9). Substitua o *Learner another learning scheme* para classificação de tarefas (clique com o botão direito na árvore de decisão de aprendizagem e substitua o operador). Você pode usar o *RuleLearner* por exemplo.



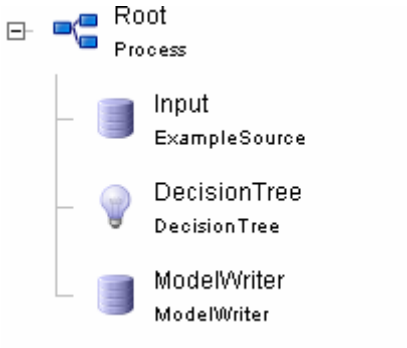
Após executar o experimento o novo modelo é apresentado.



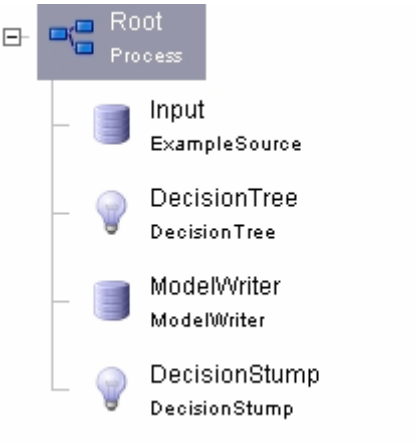
Substitua a árvore de decisão de aprendizagem pelo *RuleLearner*.

4 – Construção da árvore de decisão

Muitas vezes é necessário gravar o modelo em um arquivo, a fim de aplicá-lo mais tarde em novos dados invisíveis. Nesta experiência, uma árvore de decisão é construída a partir de um conjunto de dados, e escritos em um arquivo com o operador *ModelWriter*.

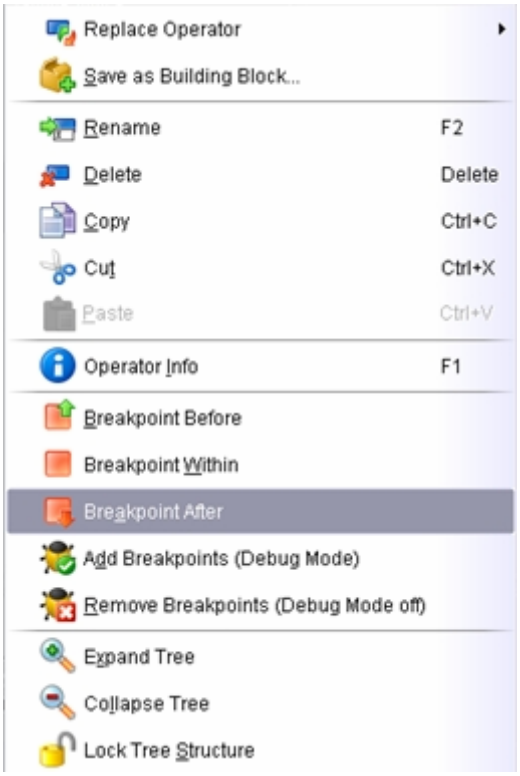


• Escolha o *Learner* na exibição em árvore e substitua-o por outro esquema de aprendizagem que podem lidar com atributos numéricos e nominais como "*Decision Stump*". Todos os *Learners* RapidMiner - incluindo todos os *Learners* da biblioteca de aprendizagem Weka - podem ser encontrados em subgrupos do "*Learner*".



Um *Learner* é utilizado a fim de construir um modelo de dados. Todos os *Learners* da Weka e muitos outros operadores Weka são totalmente integrados ao RapidMiner.

• Clique com o botão direito sobre a entrada de dados e selecione "*Breakpoint After*" (você também pode clicar duas vezes sobre o operador e ativar ou desativar o Breakpoint ). Comece a experiência.

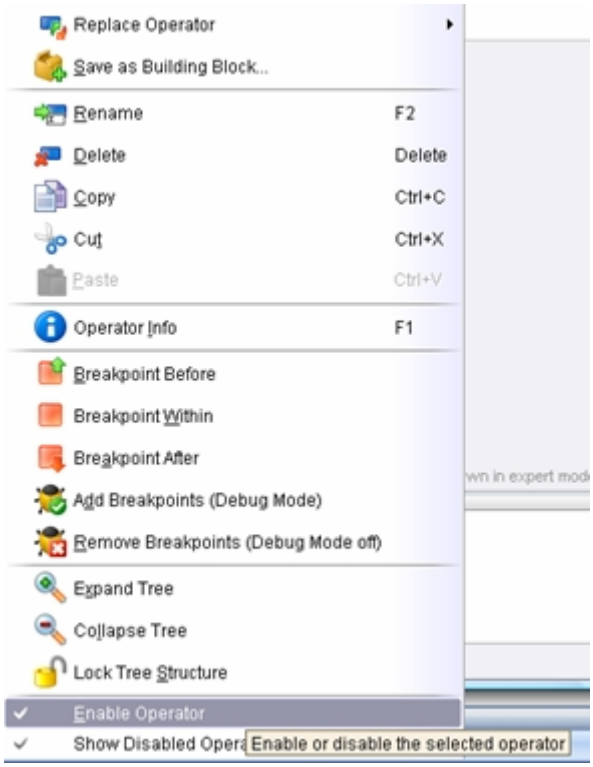


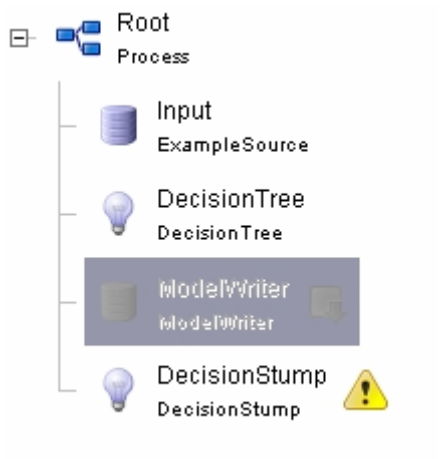
Após um curto período de tempo é exibida a mensagem "Breakpoint alcançado", e você pode mudar para a guia *Results* para verificar os resultados intermediários. Neste caso, o resultado intermediário é um conjunto de exemplos já conhecidos.

Imagens semelhantes a esta 

aparecem na árvore ao lado do operador como um *Breakpoint*

• Pressionando o ícone "*Resume*"  na barra de ícones (Pausa o símbolo que será usado clicando na tecla "*Play*") é permitido retomar sua experiência. Você também pode desativar os operadores, por exemplo, o operador *Learner*. Operadores desabilitados não executam qualquer ação.



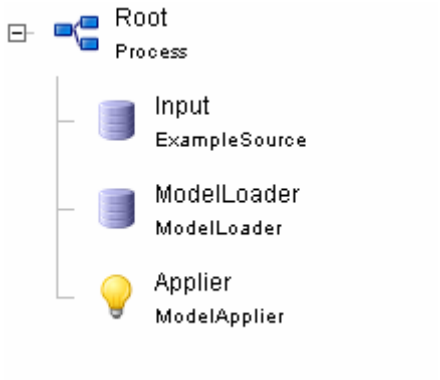


Pressione este ícone  para retomar o experimento após parar em um *breakpoint*

Note que muitos operadores como *Learners* assumem a sua entrada padrão. Esse comportamento pode ser mudado para muitos desses operadores usando um parâmetro *keep_* ***. Alternativamente, o operador *IOMultiplier* poderia ser utilizado antes do operador ser aplicado.

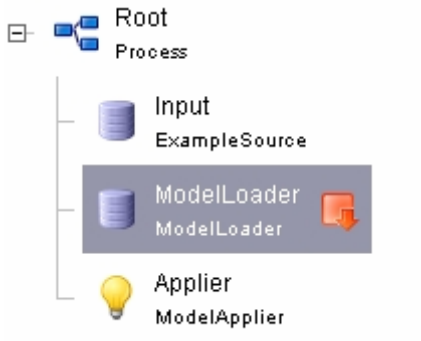
5 – Dados de Entrada, breakpoint e modos Expert e Iniciante

Esta experiência carrega os modelos da experiência anterior. Outro conjunto de dados também é carregado e o modelo é aplicado para os dados de entrada. Após realizar esta experiência a guia de resultados mostra um conjunto de exemplos com uma coluna de previsão.



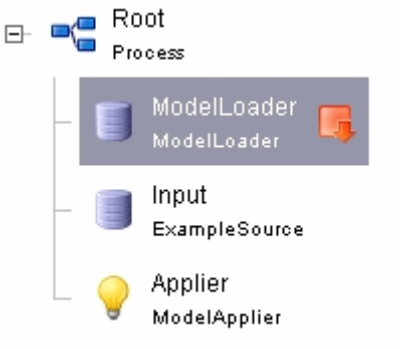
Experimente o seguinte:

• Clique com o botão direito sobre o modelo carregado na exibição em árvore e defina um *breakpoint* após este operador (duplo clique sobre o operador). Quando você muda a visualização do resultado depois de ser atingido o *breakpoint*, você pode ver tanto o exemplo dado na entrada quanto o modelo carregado.



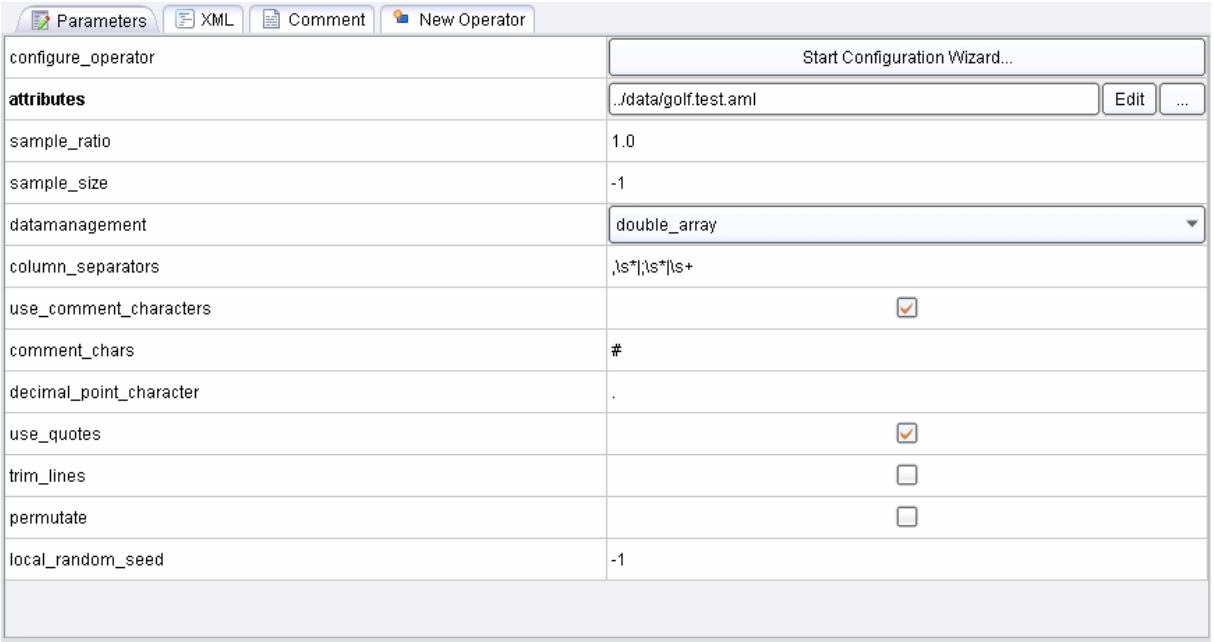
Concluindo o experimento, pressione o botão retomar .

• Remova o *breakpoint* (pelo menu de contexto ou clicando duas vezes novamente) e selecione o modelo carregado pelo operador de exibição em árvore. Agora você pode arrastar o operador e movê-lo para uma nova posição, soltando-o antes da do operador de entrada. Recomece a experiência.



O resultado é o mesmo, a seqüência de operadores não importa, o modelo aplicado, nem os tipos de entrada (conjunto de exemplos e um modelo), são entregues.

• Selecione o operador de Entrada na exibição em árvore. Pressione o ícone com o símbolo de usuário ou com o símbolo de uma pessoa na barra de ícones .



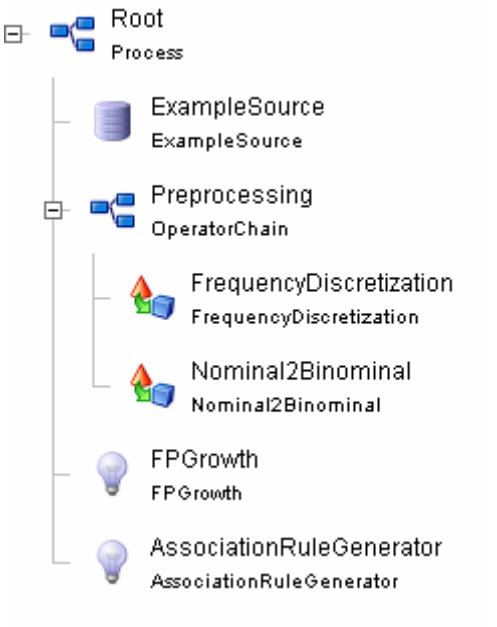
O número de parâmetros mudou. O RapidMiner fornece dois modos de usuário. No modo expert  todos os parâmetros dos operadores são mostrados. No modo iniciante  são exibidos apenas os parâmetros mais importantes.

Mude para o modo expert  *(todos os parâmetros são mostrados).*

Desative o modo expert  *(modo "iniciante", só os parâmetros mais importantes são mostrados).*

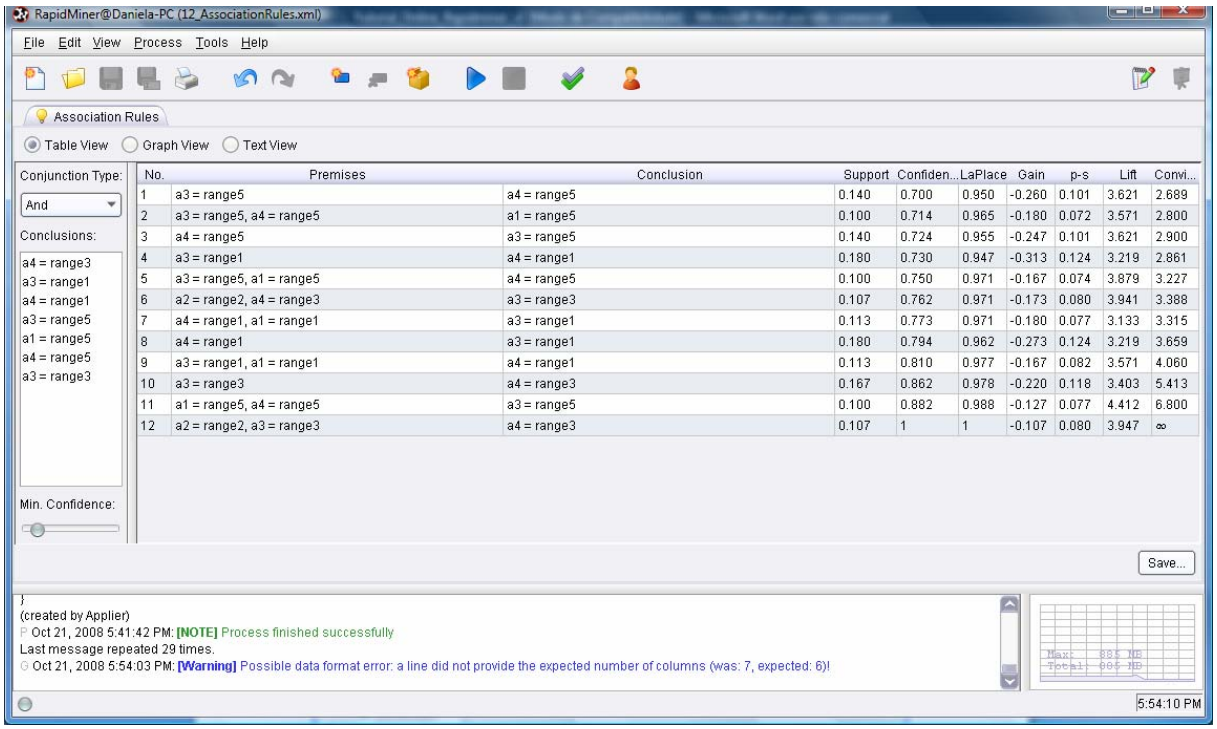
6 – Pré-Operadores: Operador de *Discretization* e Nominal

Este experimento utiliza dois importantes pré-operadores: em primeiro lugar a freqüência do operador de *discretization*, com *discretizes* numéricas de atributos, colocando os valores em *bins* de igual tamanho. Em segundo lugar, o filtro *operator nominal to binominal* criando para cada possível *nominal value* de um atributo *polynominal* uma nova entrada *binominal* (binário) característica que é verdadeiro se o exemplo tiver um valor *nominal* particular.



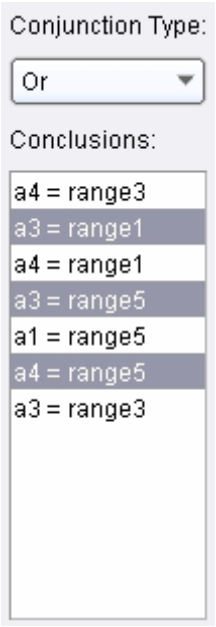
Estes pré-operadores são necessários visto que *learning schemes* não podem manipular atributos de certos tipos de valor. Por exemplo, um item eficiente freqüentemente operado por um conjunto FPGrowth utilizados neste processo de instalação só pode tratar recurso binário e não numéricos ou polinominal.

O próximo operador é um item freqüente do *set mining* FPGrowth. Este operador eficientemente calcula conjunto de valores de atributos muitas vezes ocorridos em conjunto. A partir destes freqüentes conjuntos de itens são chamadas as regras *confident* são calculadas com associação de regras geradas.



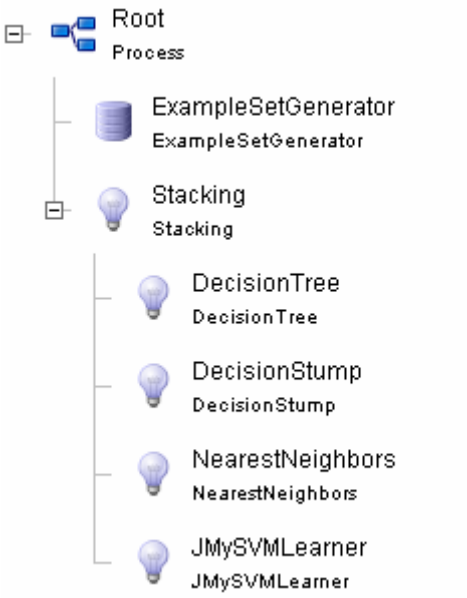
No.	Premises	Conclusion	Support	Confiden...	LaPlace	Gain	p-s	Lift	Convi...
1	a3 = range5	a4 = range5	0.140	0.700	0.950	-0.260	0.101	3.621	2.689
2	a3 = range5, a4 = range5	a1 = range5	0.100	0.714	0.965	-0.180	0.072	3.571	2.800
3	a4 = range5	a3 = range5	0.140	0.724	0.955	-0.247	0.101	3.621	2.900
4	a3 = range1	a4 = range1	0.180	0.730	0.947	-0.313	0.124	3.219	2.861
5	a3 = range5, a1 = range5	a4 = range5	0.100	0.750	0.971	-0.167	0.074	3.879	3.227
6	a2 = range2, a4 = range3	a3 = range3	0.107	0.762	0.971	-0.173	0.080	3.941	3.388
7	a4 = range1, a1 = range1	a3 = range1	0.113	0.773	0.971	-0.180	0.077	3.133	3.315
8	a4 = range1	a3 = range1	0.180	0.794	0.962	-0.273	0.124	3.219	3.659
9	a3 = range1, a1 = range1	a4 = range1	0.113	0.810	0.977	-0.167	0.082	3.571	4.060
10	a3 = range3	a4 = range3	0.167	0.862	0.978	-0.220	0.118	3.403	5.413
11	a1 = range5, a4 = range5	a3 = range5	0.100	0.882	0.988	-0.127	0.077	4.412	6.800
12	a2 = range2, a3 = range3	a4 = range3	0.107	1	1	-0.107	0.080	3.947	∞

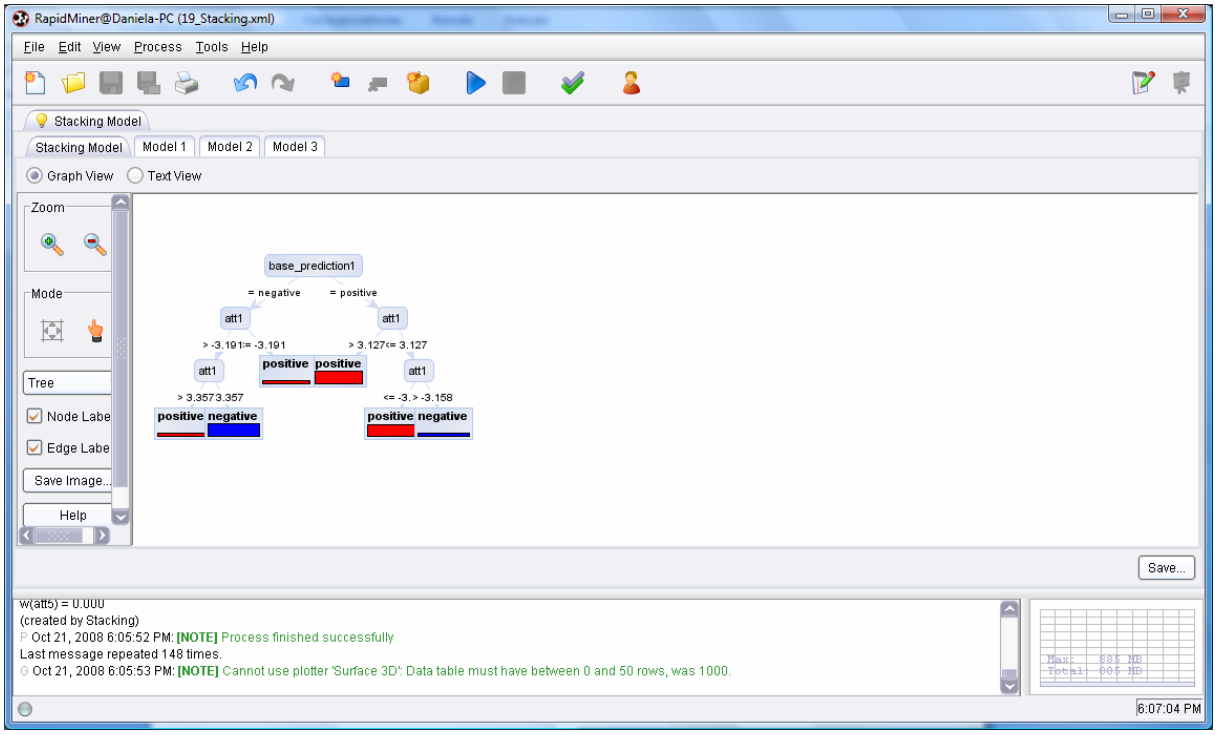
O resultado será exibido em um navegador onde a conclusão desejada pode ser selecionada em uma lista do lado esquerdo. Tal como para todas as outras telas disponíveis no RapidMiner você pode classificar as colunas clicando no cabeçalho da coluna. Pressionando CTRL durante esses cliques é permitida a seleção de até três colunas.



7 – Operadores de Meta Aprendizagem e Stacking

O RapidMiner suporta *MetaLearning* apoiando a inserção de uma base com vários *Learners* em operador de *MetaLearning*. Neste exemplo temos que gerar um conjunto de dados com o operador ExampleSetGenerator e aplicar uma versão melhorada do *Stacking* sobre este conjunto de dados. O operador de *Stacking* contém quatro operadores interiores, o primeiro é o Learner que deve aprender a *Stacking* o modelo de previsões dos outros quatro *child* operadores (base *Learners*).

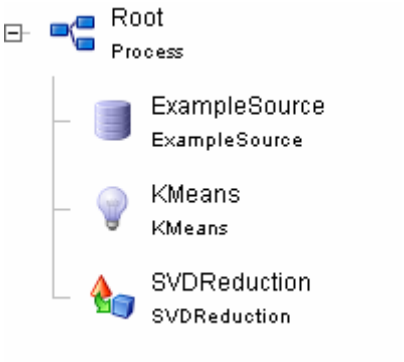




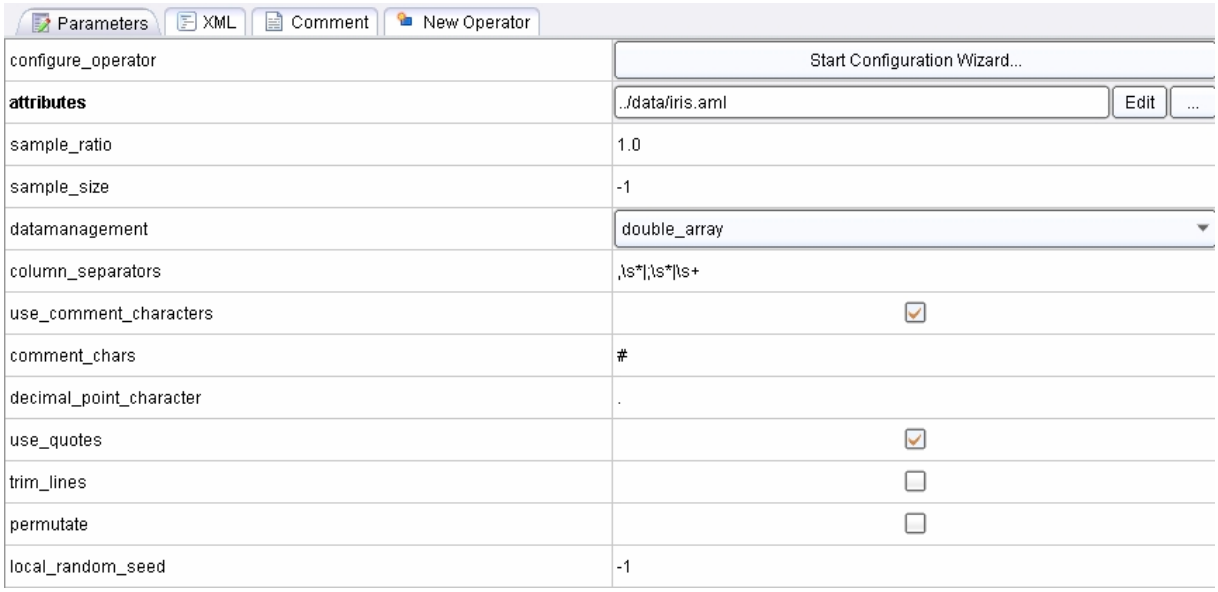
Outra meta learning schemes like Boosting or Bagging contem apenas um operador de aprendizagem interior. Em ambos os casos os parâmetros learning schemes são fixados diretamente para a base de operadores de aprendizagem. Não há necessidade de *to cope* com diferentes estilos de parâmetros para o interior e de operador de meta learning.

8 - Clustering

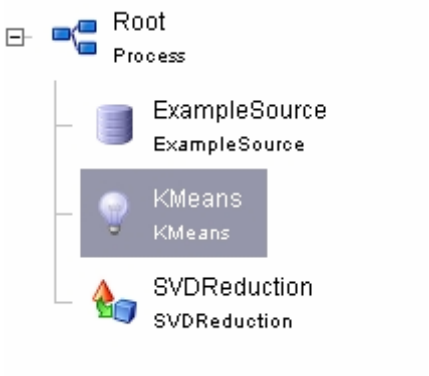
Em muitos casos, nenhum atributo *label* pode ser definido, e os dados devem ser agrupados automaticamente. Este procedimento é chamado de "*Clustering*". O RapidMiner suporta uma ampla gama de esquemas de *clustering*, que podem ser usados da mesma forma como qualquer outro *meta learning*. Isso inclui a combinação com todos os pré-operadores.



Nesta experiência, o conhecido conjunto de dados Iris é carregado (o texto é carregado, também, mas ela só é usada para visualização e comparação e não para a construção de "*clusters*" em si).



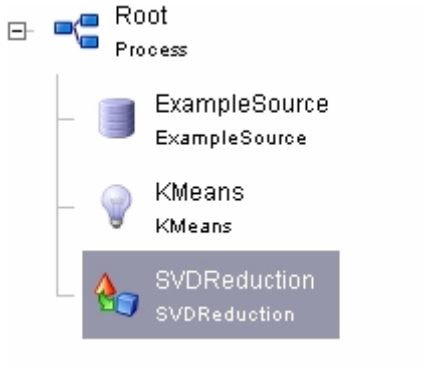
Um dos mais simples sistemas de *clustering*, chamado *KMeans* é então aplicado a este conjunto de dados.



Parameters | XML | Comment | New Operator

keep_example_set	☑
add_cluster_attribute	☑
add_characterization	☐
k	3
max_runs	10
max_optimization_steps	100
local_random_seed	-1

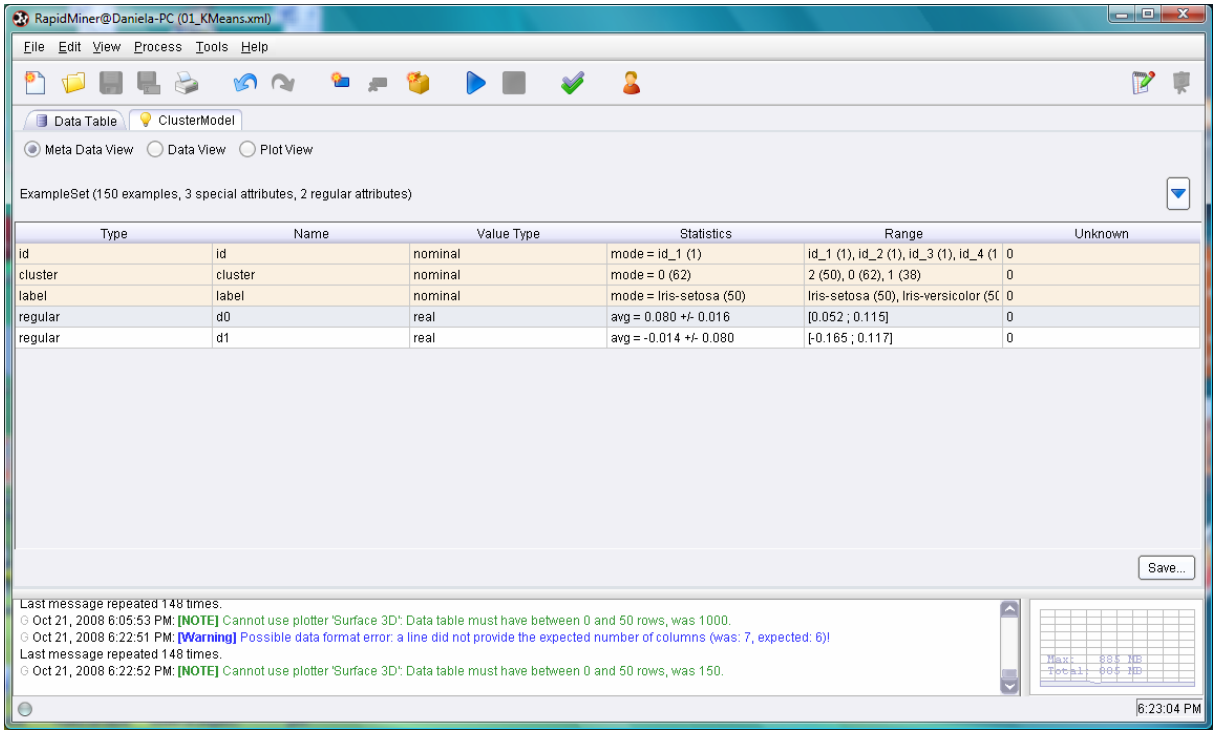
Posteriormente, a redução da dimensionalidade é realizada, a fim de melhorar a visualização do conjunto de dados em duas dimensões.

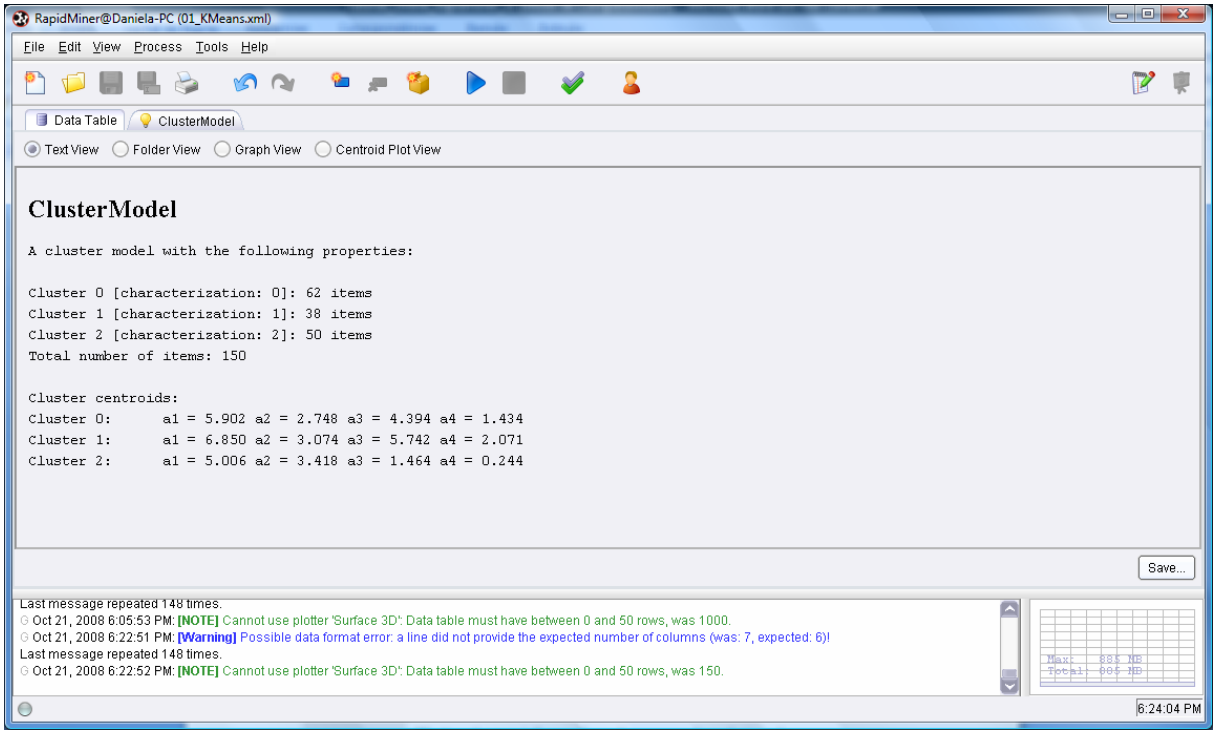


Parameters | XML | Comment | New Operator

keep_example_set	☐
return_preprocessing_model	☐
dimensions	2

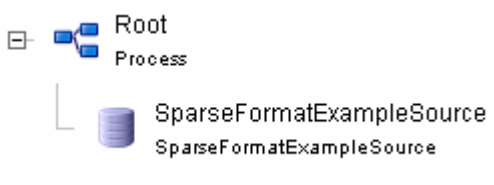
Basta executar a experiência e comparar o resultado do *clustering* do texto original (por exemplo, na tela o exemplo dado). Você também pode visualizar o próprio modelo de *cluster*.



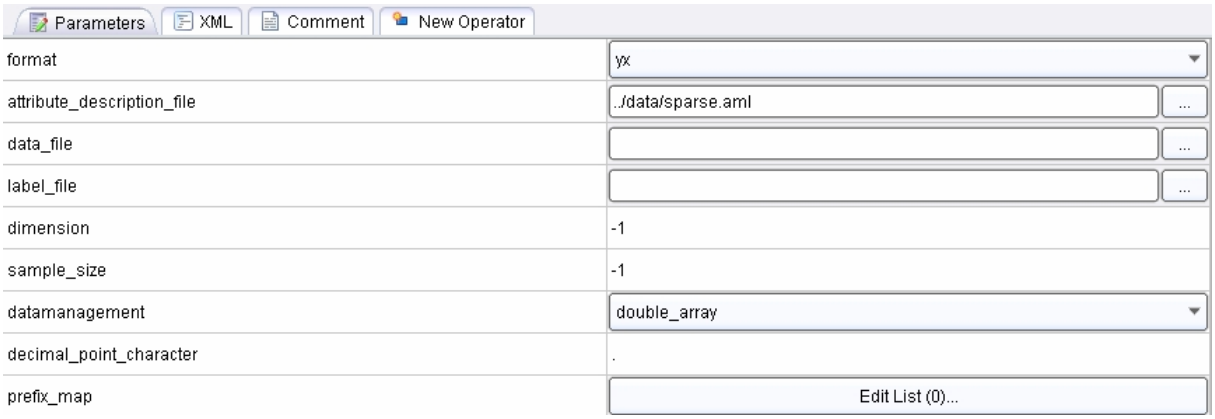


9 – Dados *Sparses*

Antes de começarmos com as experiências mais complexas, vamos mostrar algumas outras maneiras para carregar seus dados. Esta experiência carrega amplamente os dados utilizados para o formato conhecido a partir de dados *sparse* do *Support Vector Machines*. Este formato é especialmente útil para dados de texto ou outros dados em que muitos valores atribuídos são 0.

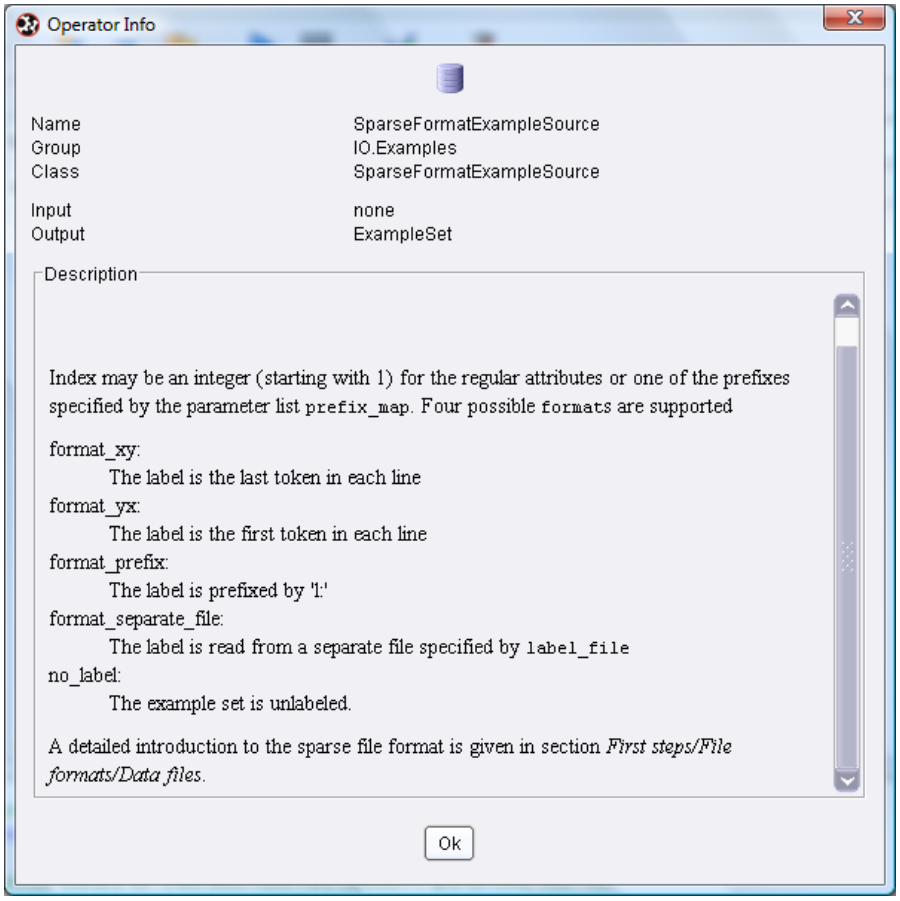


• Quando você muda do modo experts  para o modo iniciantes  é que você pode ver que o *SparseFormatExampleSource* usa muitos parâmetros. Consulte o Tutorial *RapidMiner* para mais explicações.



• O operador referencia o capítulo do Tutorial RapidMiner e oferece uma descrição detalhada de todos os operadores e os seus parâmetros. Uma breve descrição de um operador é emitida em um operador *info dialog* (no menu de contexto do operador em uma árvore ou pressionando F1).

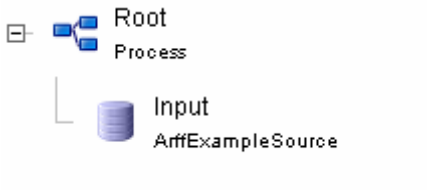
Mostra algumas informações sobre o operador selecionado



• Algumas informações sobre os parâmetros aparecem como caixa de texto quando você mantém o ponteiro do mouse em uma área de propriedade da tabela.

10 – Operadores Weka

O formato de arquivo ARFF do Weka é também usado muitas vezes para descrever os dados. Este experimento utiliza um operador *ArffExampleSource* para ler dados e descrições do atributo de arquivos em formato ARFF.

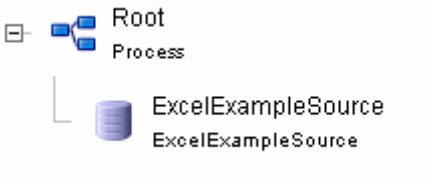


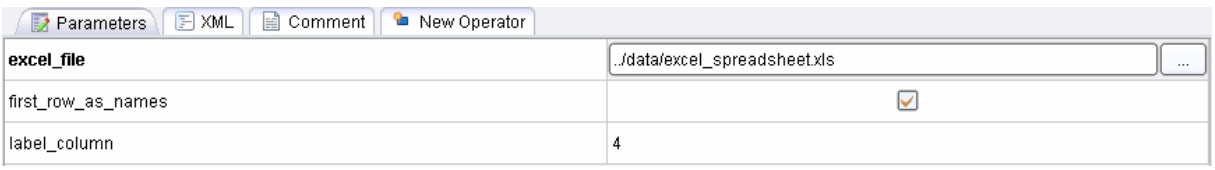
Existem várias maneiras de carregar seus dados no RapidMiner, por exemplo:

• O operador *ExampleSource* pode ser usado pela maior parte dos formatos de dados. Ele permite uma definição *arbitrary* de separar caracteres. Os valores padrões devem ser suficientes para a maioria das aplicações e de conjuntos de dados. Este operador também é usado para criar arquivos de descrição de atributos de seus dados (botão *Edit*). Este operador também fornece um assistente de configuração para carregar conjuntos de dados.
• O operador *ExcelExampleSource* pode ser usado para carregar os dados diretamente a partir do Excel.
• O *SparseFormatExampleSource* pode ler dados mais *sparse* conhecidos como formatos de SVMs. Tal como para o operador *ExampleSource*;
• O operador *ArffExampleSource* pode ser usado para os formatos de dados ARFF que é usado pelo Weka.
• Outros operadores *XXXExampleSource* existem para carregar dados que sejam em outros formatos especiais como CSV, C4.5, BibTeX, ou dBase.
• O *DatabaseExampleSource* pode ler seus dados diretamente a partir de uma tabela da base de dados. Para mais exemplos consulte o tutorial RapidMiner.
• O *ExampleSetGenerator* cria exemplos de conjuntos aleatórios para fins de teste baseados em uma função definida pelo usuário alvo.

11 – Operador *ExcelExampleSource*

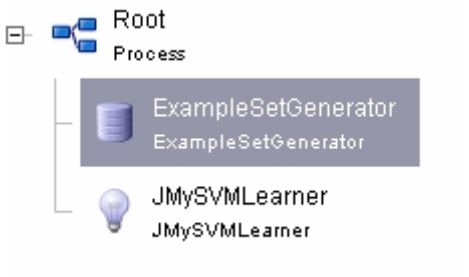
O operador ExcelExampleSource pode ser usado para carregar os dados diretamente a partir de planilhas de arquivos do Microsoft Excel.





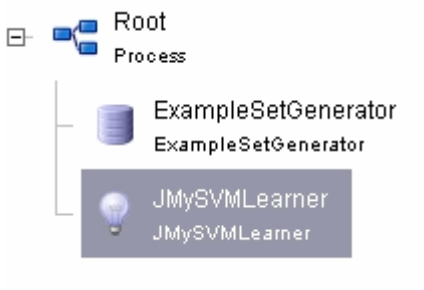
12 – Ferramentas de Apoio *Vector Machines* (SVM) e de outros modelos do *Kernel*

Esta experiência demonstra as possibilidades de visualização de ferramentas de Apoio *Vector Machines* (SVM) e de outros modelos do *kernel* baseados em margens largas.



Parameters | XML | Comment | New Operator

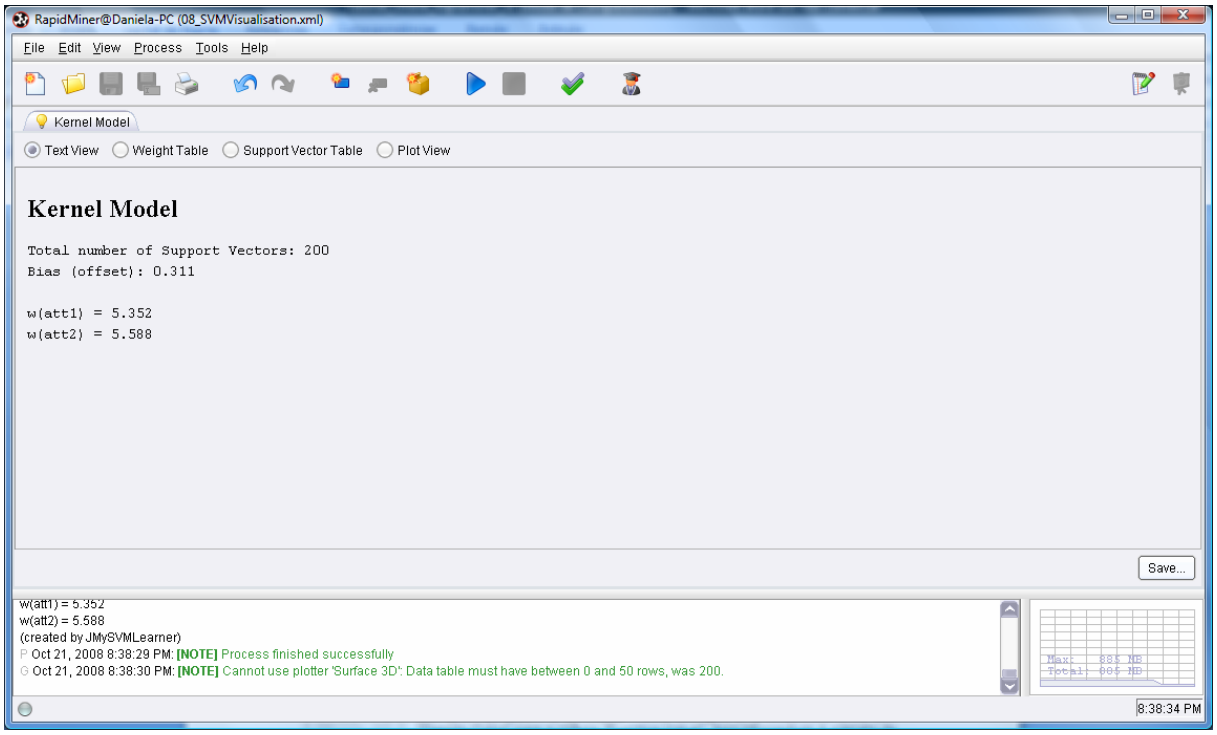
target_function	sum classification
number_examples	200
number_of_attributes	2



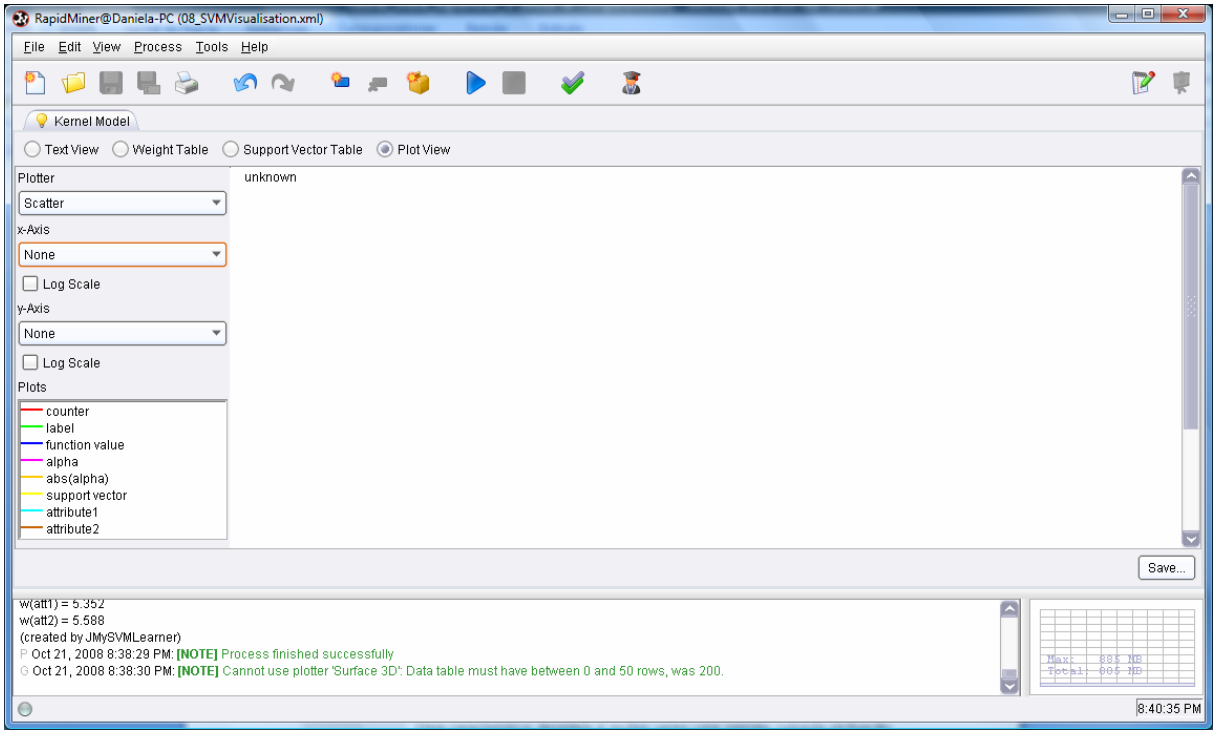
Parameters | XML | Comment | New Operator

kernel_type	dot
kernel_gamma	1.0
kernel_sigma1	1.0
kernel_sigma2	0.0
kernel_sigma3	2.0
kernel_shift	1.0
kernel_degree	2.0
kernel_a	1.0
kernel_b	0.0
C	10.0

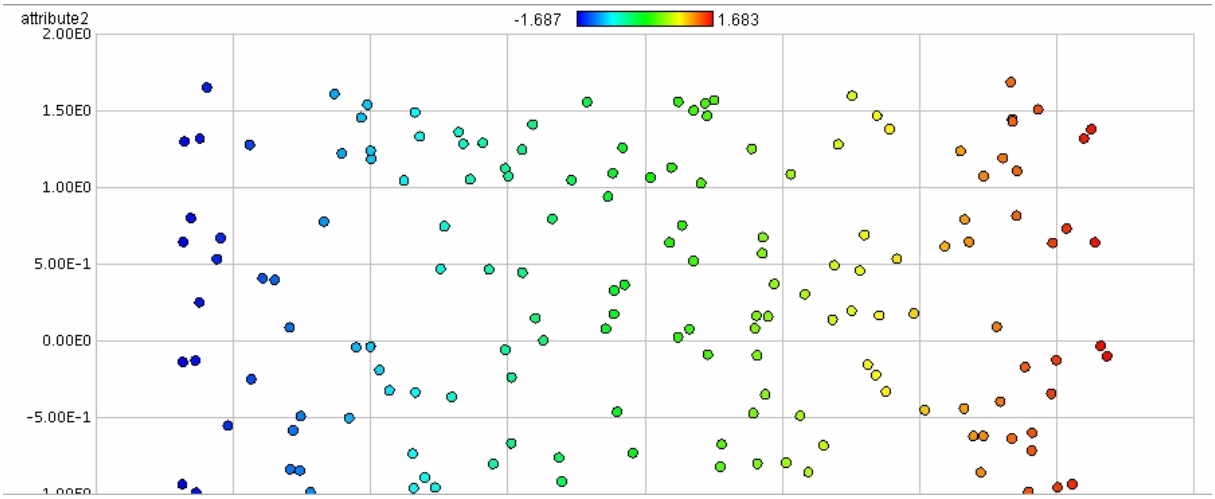
O resultado desta experiência será um modelo SVM para o qual você pode mudar para *plot view*. Várias dimensões são fornecidas para *plotting* objetivos, incluindo o treinamento de *set labels*, os valores alfa (Language multiplier), a informação se um exemplo de formação é um vetor apoio, os valores funcionais (previsões) para a formação de exemplos e, naturalmente, atribuição de valores para todos os exemplos de formação. Estes dados juntamente com o poderoso mecanismo do RapidMiner permitem traçar diferentes tipos de visualizações SVM. Basta experimentar algumas delas.

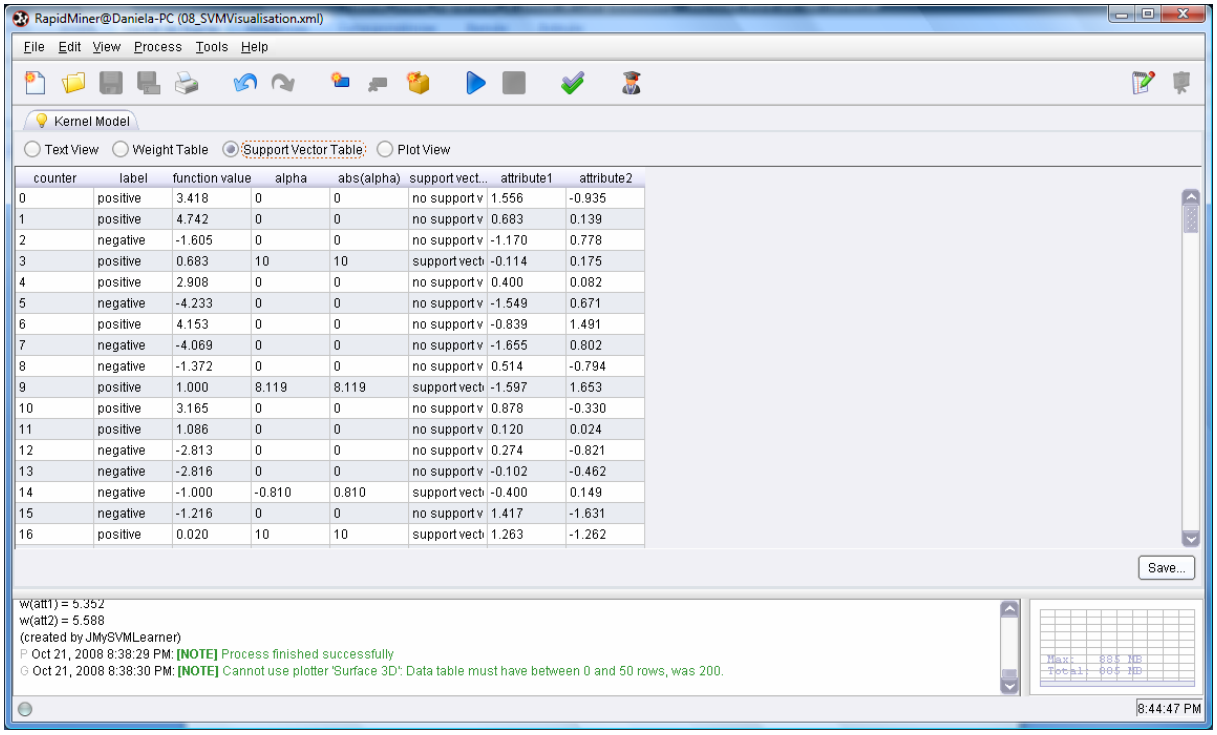


Sugerimos que você tente pelo menos *to plot* a "função valores" ao contrário da "alpha" com valores habituais *scatter plot*. O que lhe pode dar uma boa dica e a função *kernel*, usada para o seu conjunto de dados. O mesmo se aplica para *quartile plots* da função e os valores alfas coloridas pelo *label*.



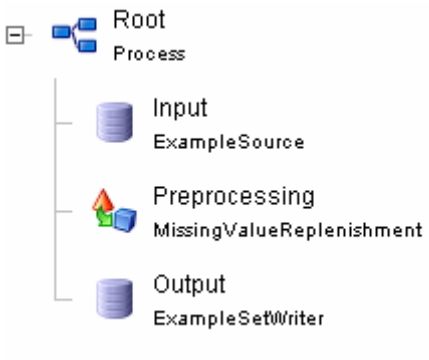
Uma característica desejada é muitas vezes uma parcela colorida da função valores. Você pode conseguir isto utilizando a parcela de modelos da SVM alterando o *plotter* de "*Density*", selecionando dois atributos para o e eixo x-y, por exemplo, "attributo1" e "attributo2", neste exemplo, e no estabelecimento dos "*Density Color*" para a coluna "*Function Value*". Isso irá conduzir à parcela de densidade desejada. Se você definir o "*Point Color*" para "*Support of vector*" ou "alpha", irá também obter explicações dos pontos que são vetores apoio.





13 – Operadores de Entrada, Pré-processamento e Saída

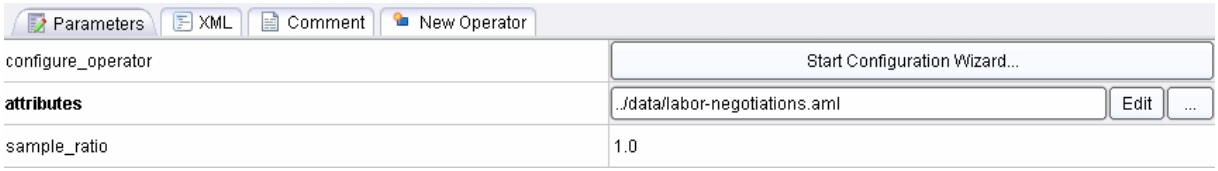
Normalmente muito tempo de mineração de dados é gasto para pré-processar os dados. O RapidMiner oferece vários operadores de leitura de dados de diversas fontes, e também os operadores que processam os dados de fácil *learning*.
Em muitas aplicações de dados faltam valores. Um dos operadores disponíveis realiza uma pré-substituição com a média / min / max do atributo. Outros operadores também podem lidar com valores infinitos.



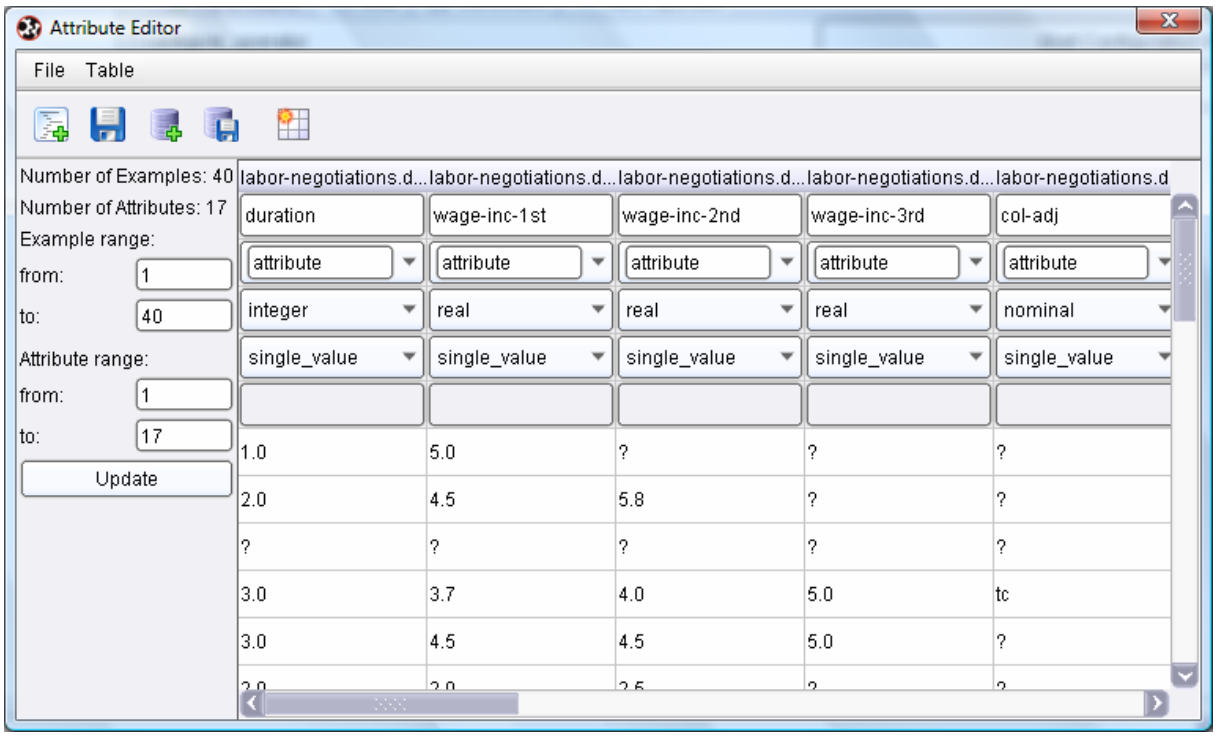
Experimente o seguinte:

• Selecione o operador de Entrada. A propriedade da tabela do quadro do lado direito mostra os parâmetros deste operador. Pressione o botão "Edit" 

do "atributo" parâmetro.



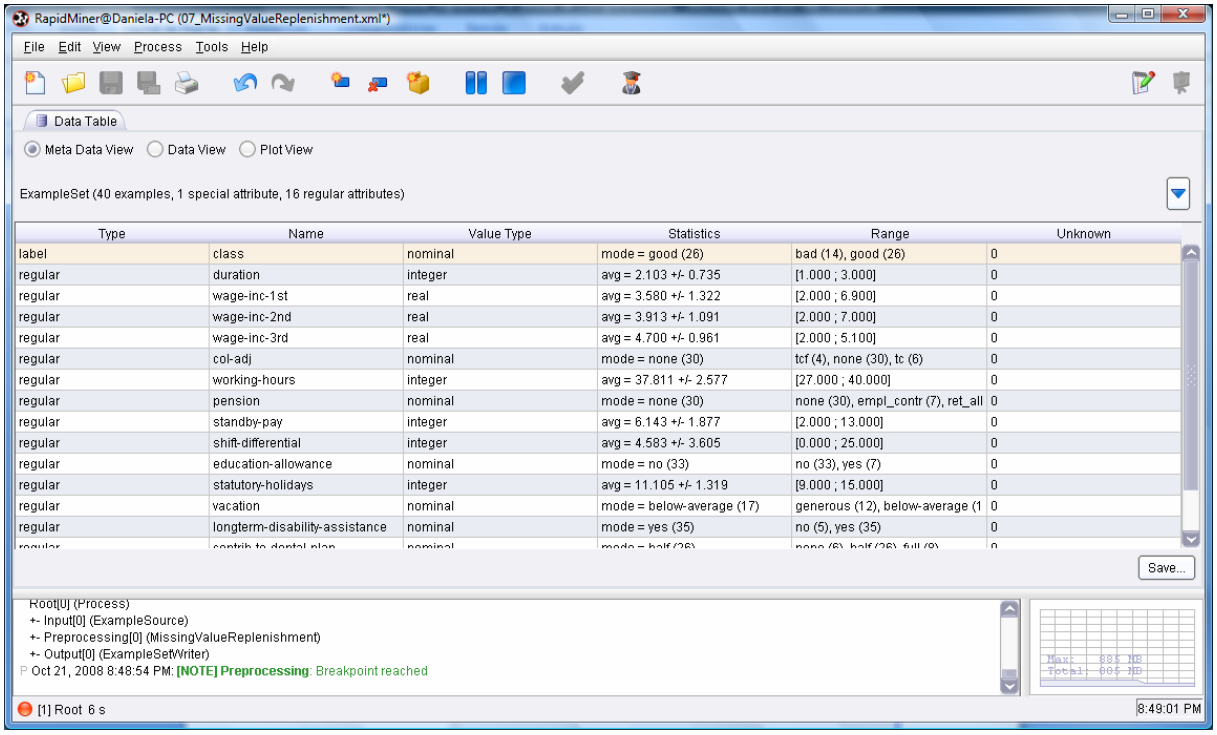
O atributo *editor* exibe uma amostra dos dados. Observe as marcas que representam dados desconhecidos.



Feche o editor de atributos.

O editor de atributo também pode ser usado para criar arquivos de descrição de atributos (.AML) para os conjuntos de dados.

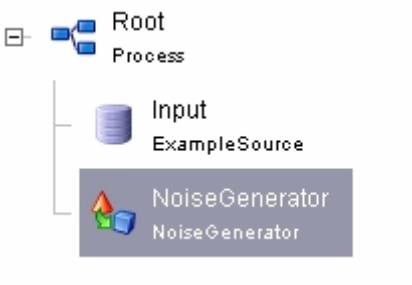
• Utilize um *breakpoint* após o operador de Entrada e execute a experiência. Compare os dados antes e depois do pré-processamento.



• O operador de saída exibe os dados de volta em um arquivo. Você pode ver esse arquivo com um editor de texto *arbitrary*. Consulte o Tutorial RapidMiner para obter mais informações sobre como usar o *ExampleSetWriter*.

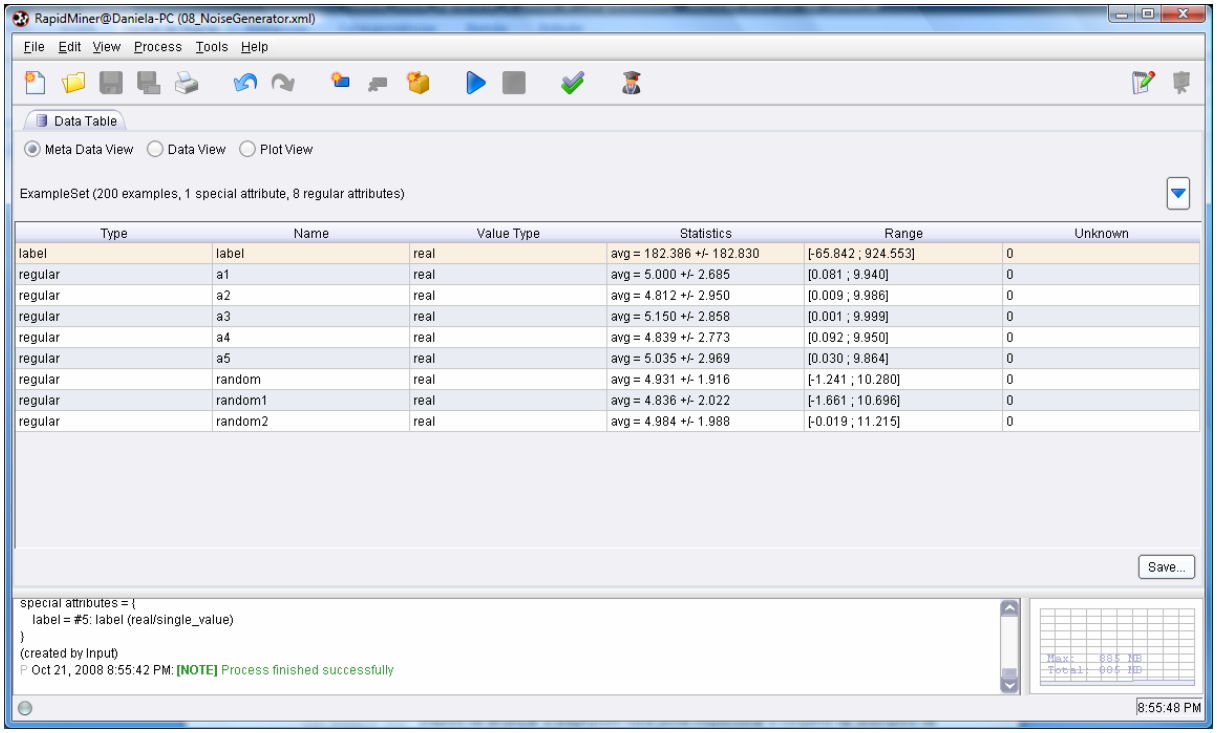
14 – *NoiseOperator*

O *NoiseOperator* pode ser usado para adicionar controladores de *noise or noisy feature* para o seu conjunto de dados. Isso é especialmente útil, a fim de avaliar o desempenho de um pré-processamento ou a *robustness* de um *Learner* específico.



Parameters | XML | Comment | New Operator

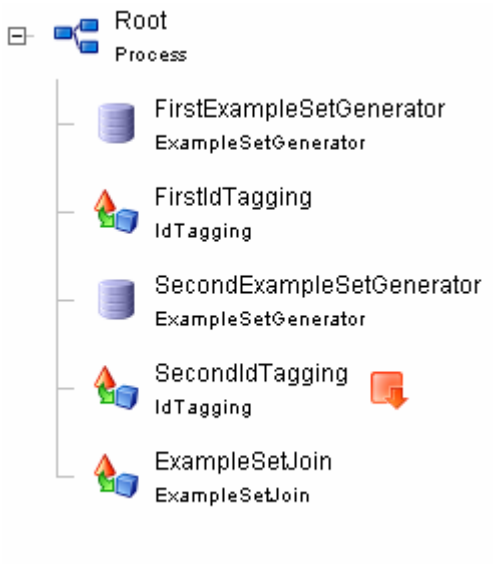
random_attributes	3
label_noise	0.05

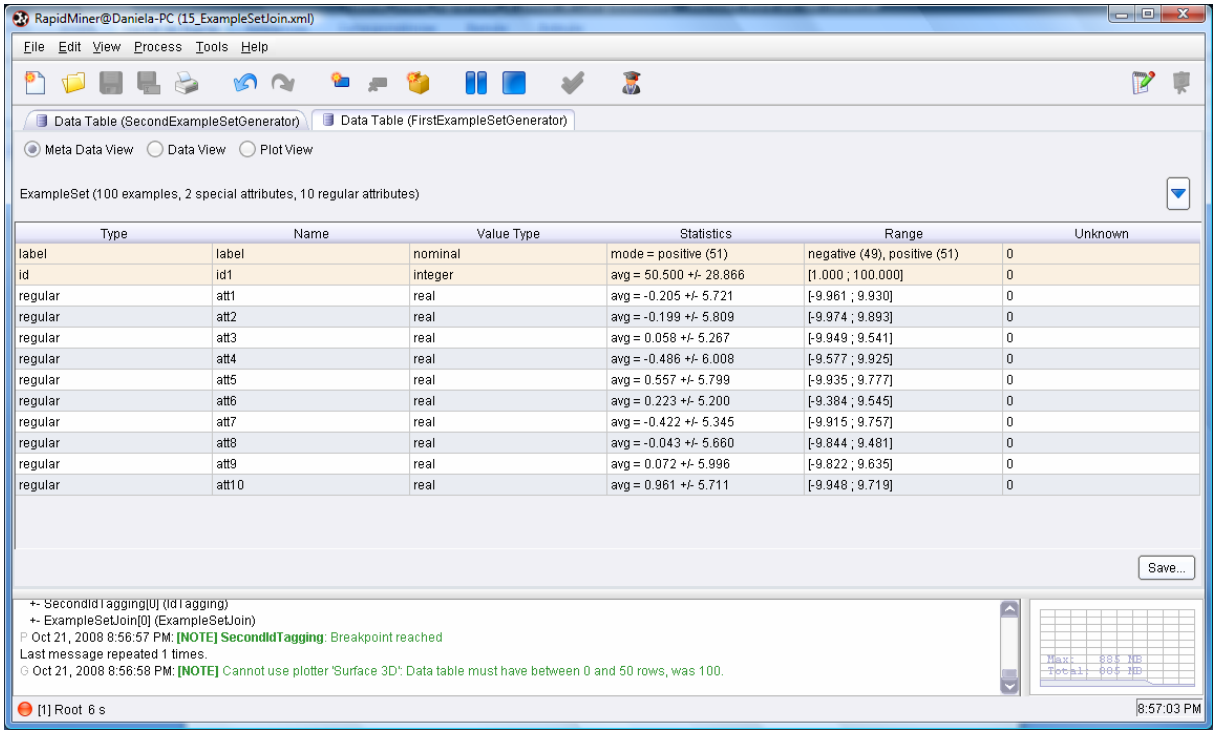


O RapidMiner também oferece muitos outros operadores, incluindo um pré-filtro do TFIDF, ofuscando movimentação de valores em série e muito mais.

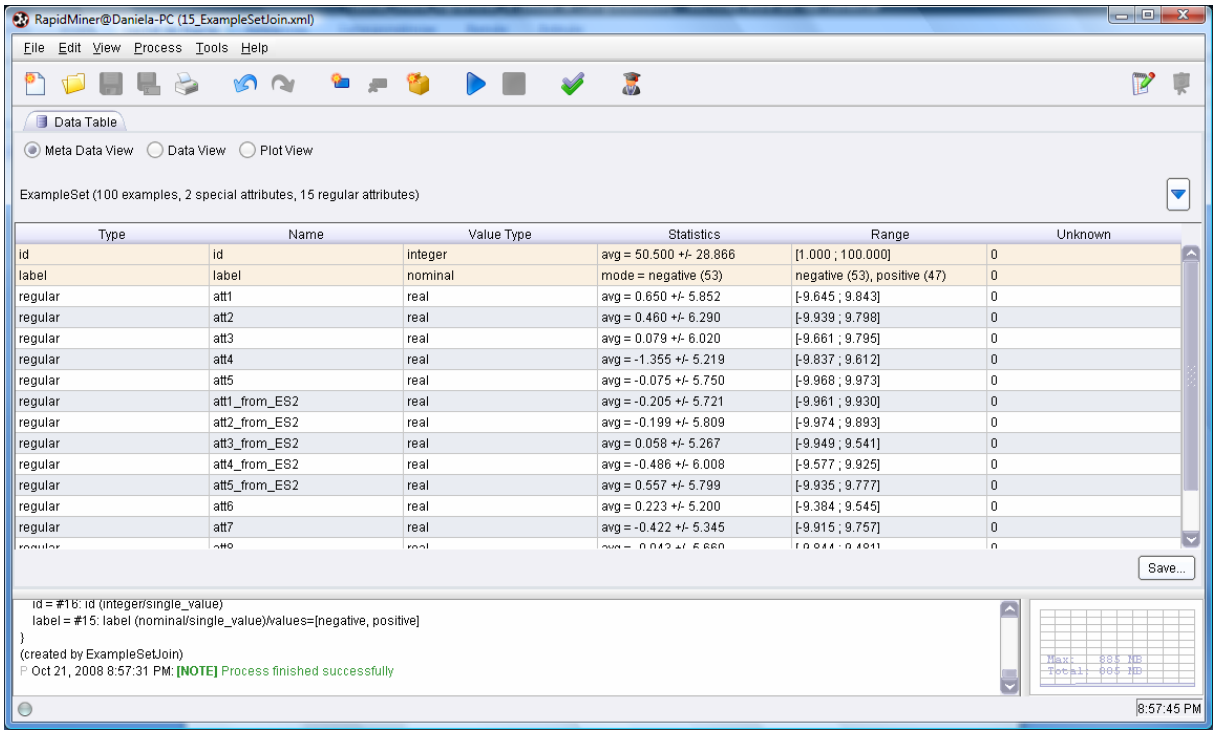
15 – Operador *ExampleSetJoin*

O operador *ExampleSetJoin* neste experimento de consumo adere dois conjuntos de exemplos. Note que atributos com nomes iguais serão renomeados durante o processo. O conjunto de exemplos deve fornecer um ID do atributo, a fim de determinar exemplos correspondentes.



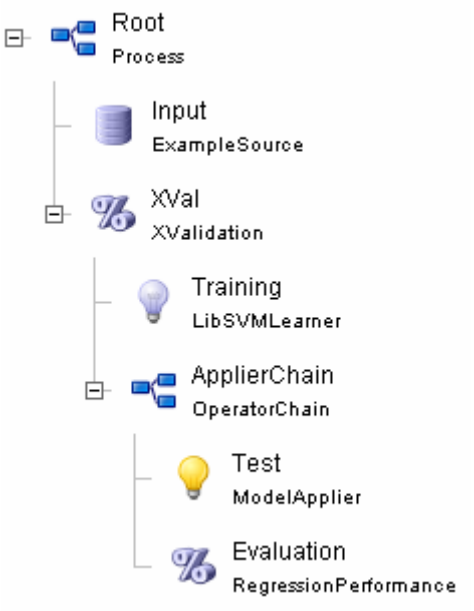


Depois de alcançar o breakpoint  você pode inspecionar o conjunto de exemplos de entrada. Depois retome a experiência  e o exemplo dado será o resultado.



16 – Validação Cruzada do RapidMiner

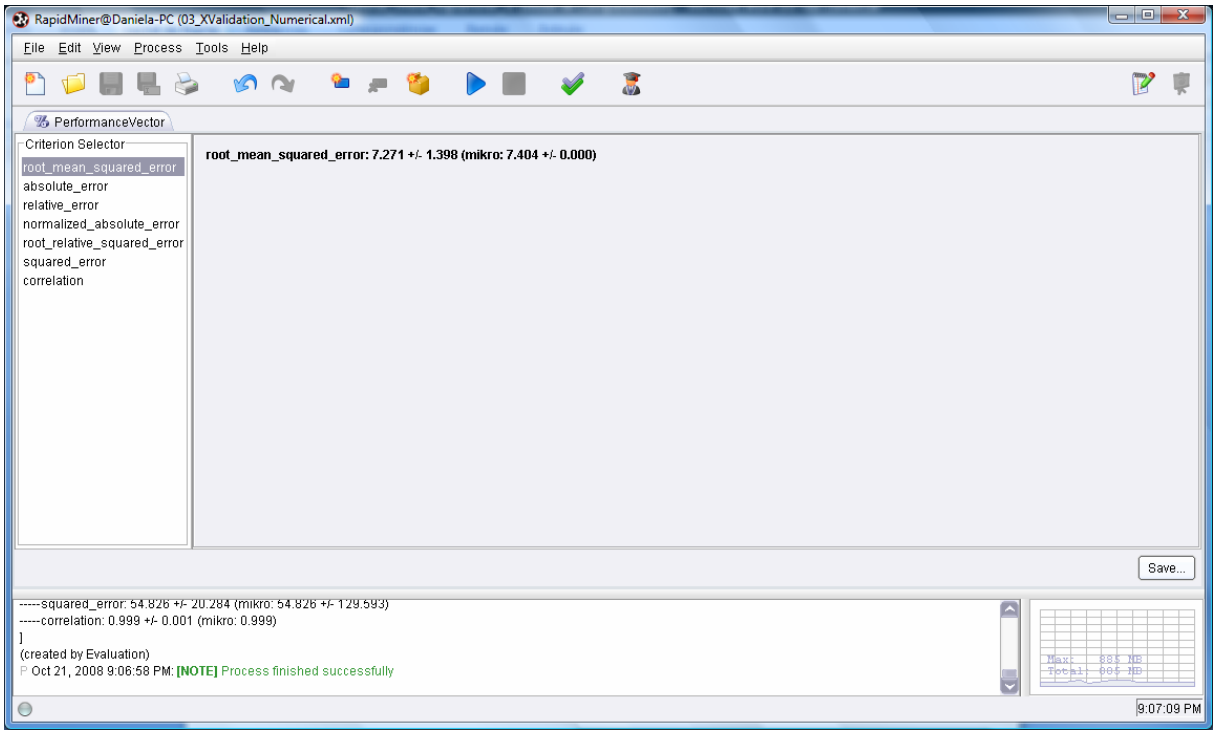
Em muitos casos, é interesse não aprender o modelo, mas a precisão do modelo. Uma possível solução para estimar a previsibilidade do modelo aprendido é aplicá-lo aos dados marcados como teste e calcular a previsão do número de erros (ou outros critérios desempenho). Uma vez que os dados estejam marcados o que é raro, outras abordagens para estimar o desempenho de performance de *learning* são muitas vezes utilizados. Esta experiência demonstra "validação cruzada" no RapidMiner.



Depois a validação de dados é dividida em conjuntos *labelled* de formação e de teste. Os modelos são *learned on training* e aplicados em dados de teste. Os erros são calculados e é feita a média da previsão para todos os subgrupos. Este bloco pode ser usado como operador de vários *wrappers* como características geradas / seleção de operadores.

Este é o primeiro exemplo de uma experiência mais complexa. Os operadores constroem uma árvore estruturada. Agora é o suficiente para aceitar que a validação cruzada de operadores demande um exemplo dado como entrada e forneça um vetor de desempenho valores como saída. Além disso, gere a divisão em treinamento e exemplos de teste. Este modelo e exemplos de teste fazem o teste na entrada do aplicador em cadeia que proporciona o desempenho para estes conjuntos de testes. Os resultados de todos os conjuntos de testes possíveis são recolhidos pelo operador de validação cruzada. Por último, a média é construída é entregue como resultado.

Uma das coisas mais difíceis para o iniciante no RapidMiner é muitas vezes se tem a idéia do fluxo de dados. A solução é surpreendentemente simples: fluxo dos dados lembra uma *depth-first-search* da pesquisa com estrutura em árvore. Por exemplo, após a transformação do treinamento em conjunto com o primeiro *child* do cruzamento da validação do modelo aprendido, é entregue ao segundo *child* (o aplicador cadeia). Este fluxo de dados básicos é sempre o mesmo para todas as experiências utilizando esse fluxo se tornará muito conveniente para o utilizador experiente.

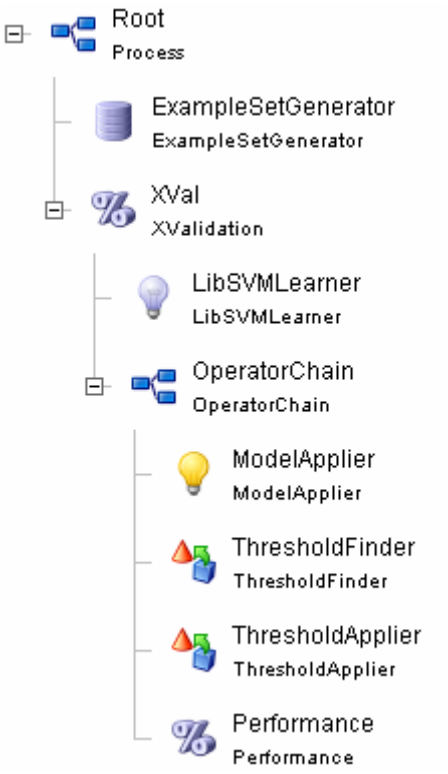


Experimente o seguinte:

• Comece a experiência. O resultado é uma estimativa do desempenho do sistema de aprendizagem sobre os dados de entrada.
• Selecione o operador de avaliação e selecione outro dos critérios de desempenho. O principal critério é utilizado para comparar o desempenho, por exemplo, em um *wrapper*.
• Substitua a validação por cruzamento "xval" por outros sistemas de avaliação e de executar o experimento com eles. Alternativamente você pode verificar o modo como outros *Learners* utilizam esses dados e substituir o operador de Formação.

17 – Operador de Confiança

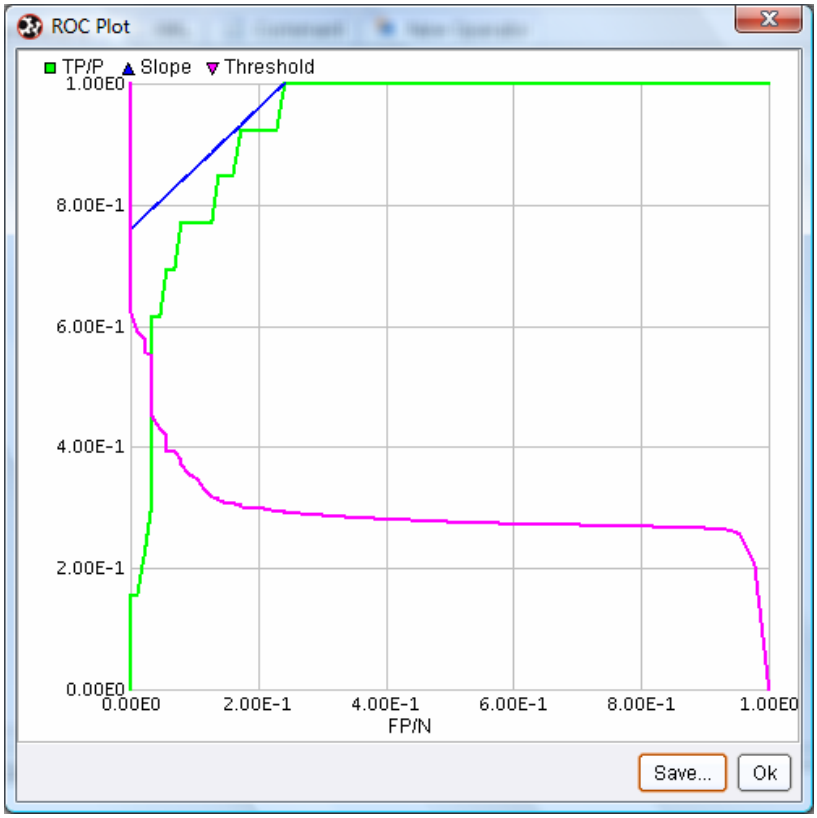
Nós usamos valores de confiança entregues pelo *Learner* neste experimento. Todos os *Learners* RapidMiner confiam na entrega destes valores além dos valores previstos. Eles podem ser lidos como uma espécie de garantia de que o *Learner* realizou uma previsão de fato verdadeira. Assim é chamado de confiança.



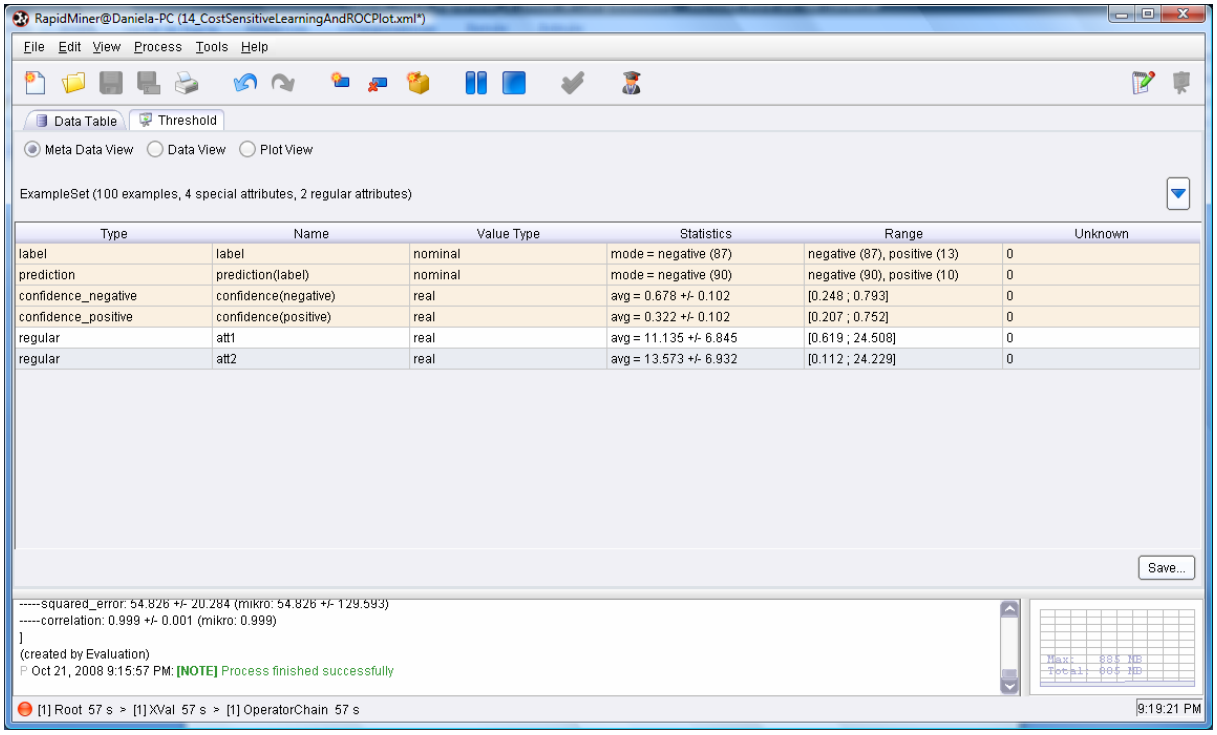
Em muitos cenários de classificação binária um erro para uma previsão errada não causa os mesmos custos para ambas as classes. Um esquema de aprendizagem deve-se levar em conta estes custos assimétricos. Ao usar a previsão de confiança podemos transformar toda a classificação dos *Learners* em custos sensíveis. Por isso, ajustamos a confiança limiar para fazer algumas previsões (geralmente 0,5).

Um *ThresholdFinder* pode ser usado para determinar o melhor limite no que diz respeito a pesos de classes. Os seguintes mapas *ThresholdApplier* fazem previsões de confiança claras e classificações no que diz respeito à determinação de valores limiares.

O *ThresholdFinder* também pode produzir uma curva ROC para vários limiares. Esta é uma boa visualização para o desempenho de um sistema de aprendizagem.



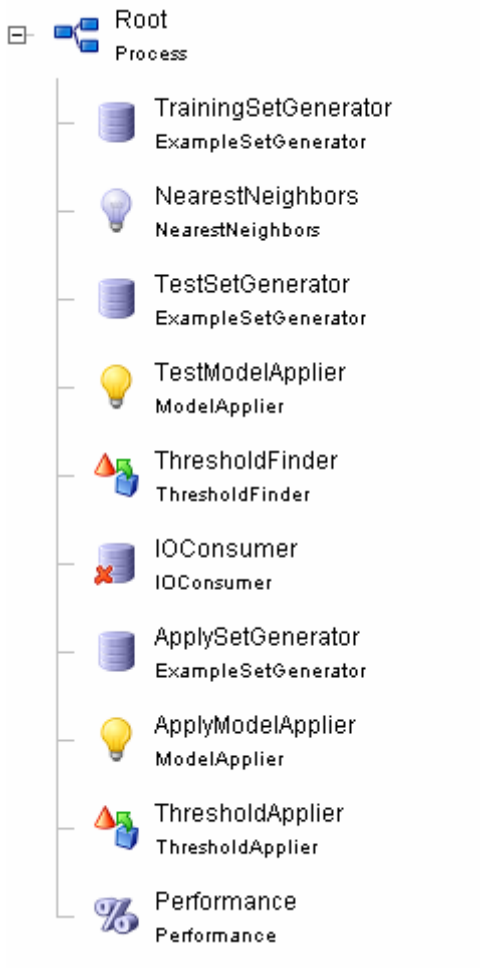
O experimento pára cada vez que a curva ROC é gerada até que você pressione o botão OK (5 vezes). O parâmetro "*show_ROC_plot*" determina se a parcela ROC deve ser exibida em todos.



Mais informações sobre a validação de operadores utilizados neste experimento podem ser encontradas no diretório correspondente, naturalmente, em referência ao operador no Tutorial RapidMiner.

18 – Classificação *Soft e Crisp*

Esta experiência demonstra como um limiar pode ser obtido a partir de um classificador *soft* e aplicado a um conjunto de testes independentes.



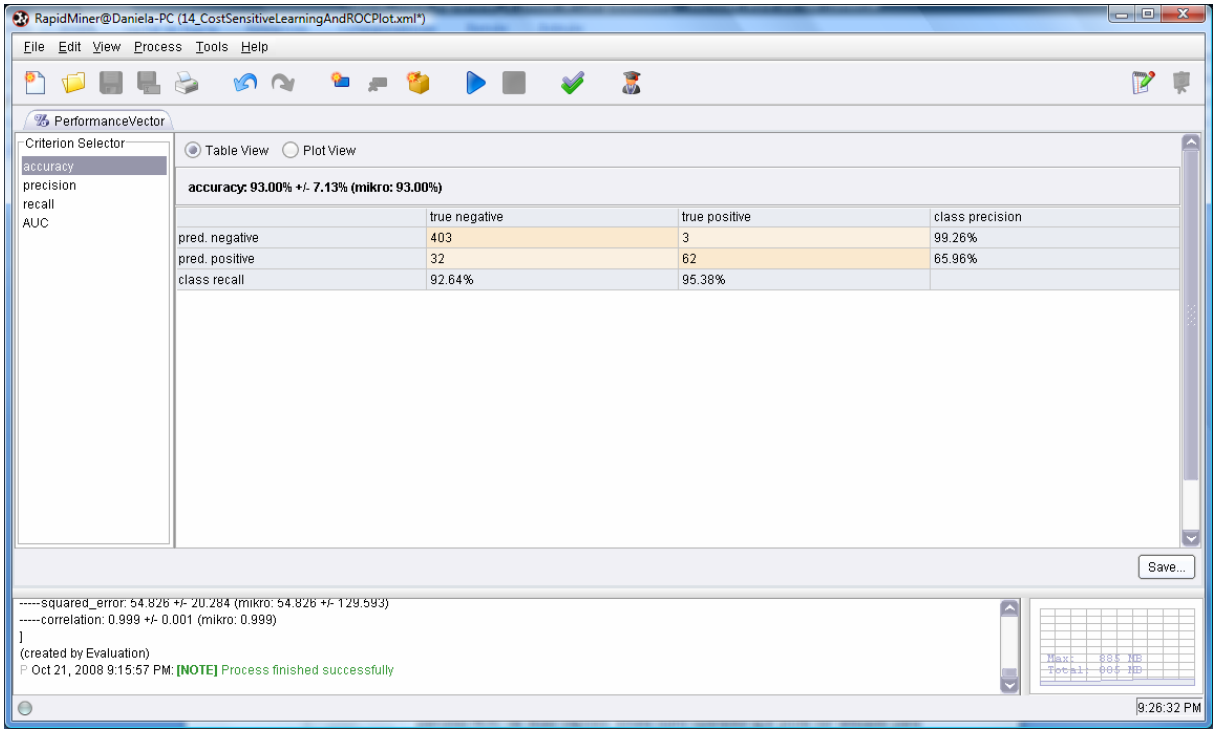
1. O *Learner* utilizado neste experimento faz previsões *softs*, em vez de classificações *crisps*. As previsões e confiança entregues por todos os *Learners* do RapidMiner que são capazes de lidar com a *labels* nominais (classificação) que serão usados como previsões *softs*.

2. O *ThresholdFinder* é utilizado para determinar o melhor limite no que diz respeito à classes de pesos. Neste caso, uma classificação errada da primeira classe (negativo) irá causar custo cinco vezes maior do que o outro erro.

3. Por favor, note que um *ModelApplier* deve ser realizado com um conjunto de testes antes de encontrar um limite. Uma vez que este modelo deve ser aplicado novamente mais tarde, o modelo aplicado mantém o modelo de entrada.

4. O *IOConsumer* garante que a previsão é feita em conjunto com os dados corretos.

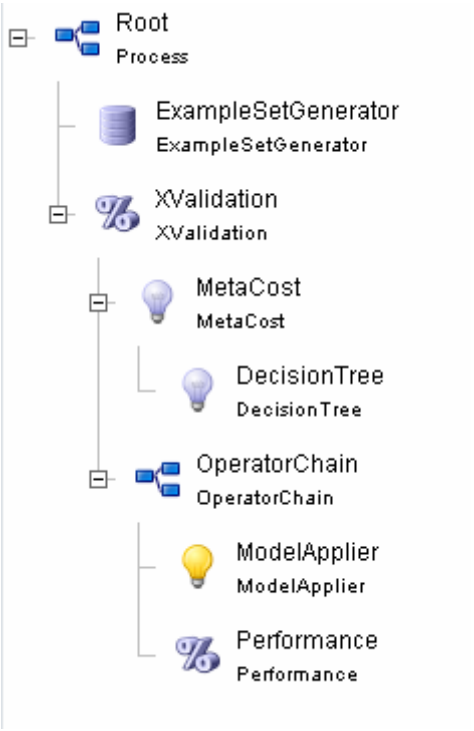
5. Os últimos passos aplicados no modelo são os dados sobre o limiar fixado à mão.

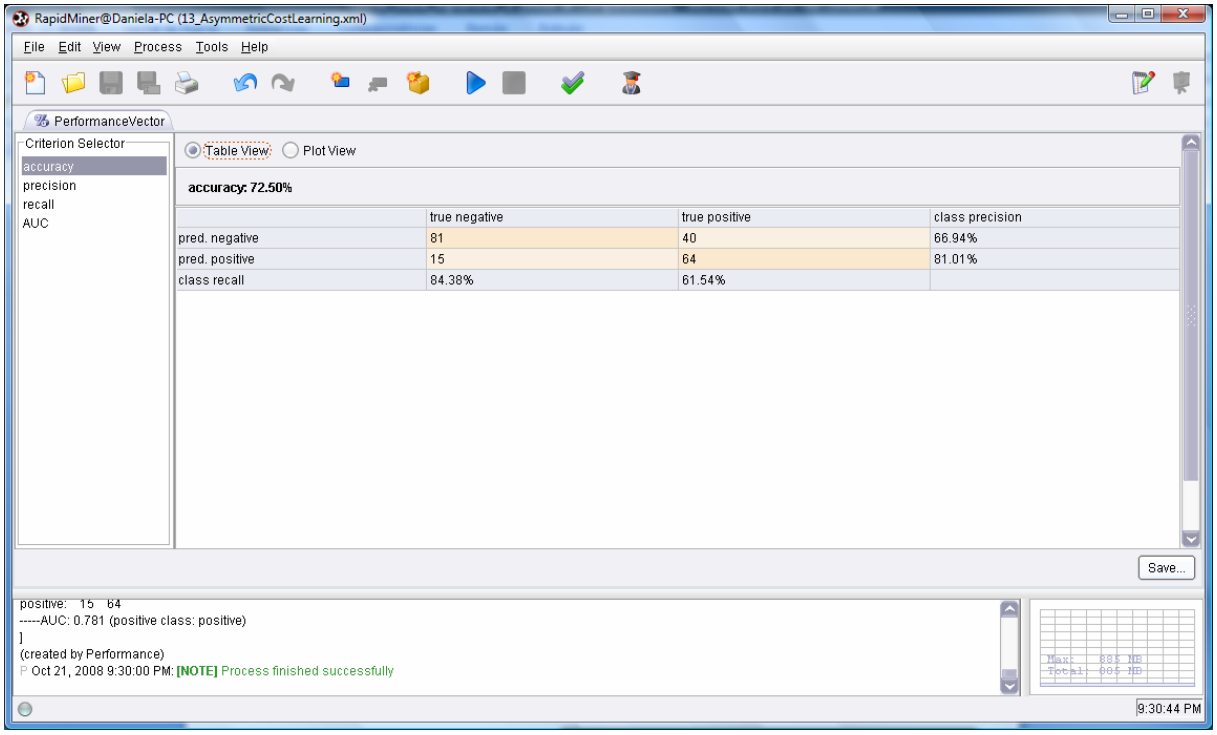


19 – Custo de Aprendizagem

Esta experiência é outro exemplo de custo sensível de aprendizagem, ou seja, para um caso em que diferentes previsões de erro poderiam causar diferentes custos. Ao lado do pré-operador *ThresholdFinder*, que também é capaz de entregar parcelas ROC de duas classes, existe outro operador que pode ser utilizado para custos sensíveis de aprendizagem.

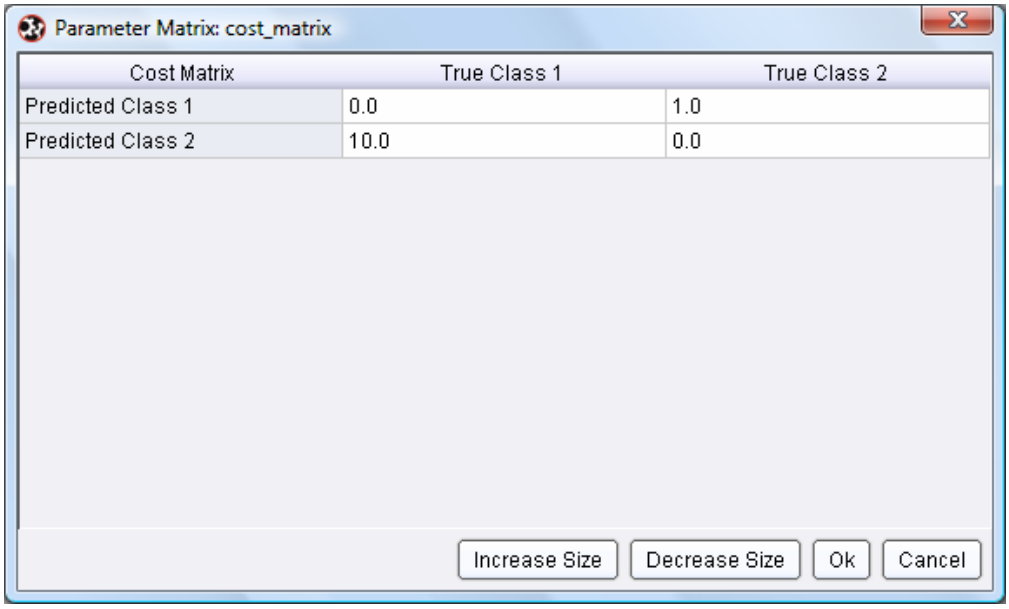
Este operador faz parte do aprendizado - Meta grupo e é chamado *MetaCost*. É utilizado como qualquer outro processo de *meta learning* e deve conter o operador de aprendizagem interior, neste caso, a árvore de decisão do *Learner* é utilizada.





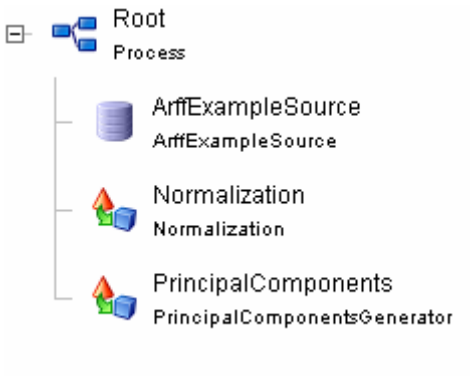
O custo matrix usado para aprendizagem de custo pode ser definido através do editor de matrix (basta pressionar o botão para o parâmetro do operador *cost_matrix* do MetaCost).

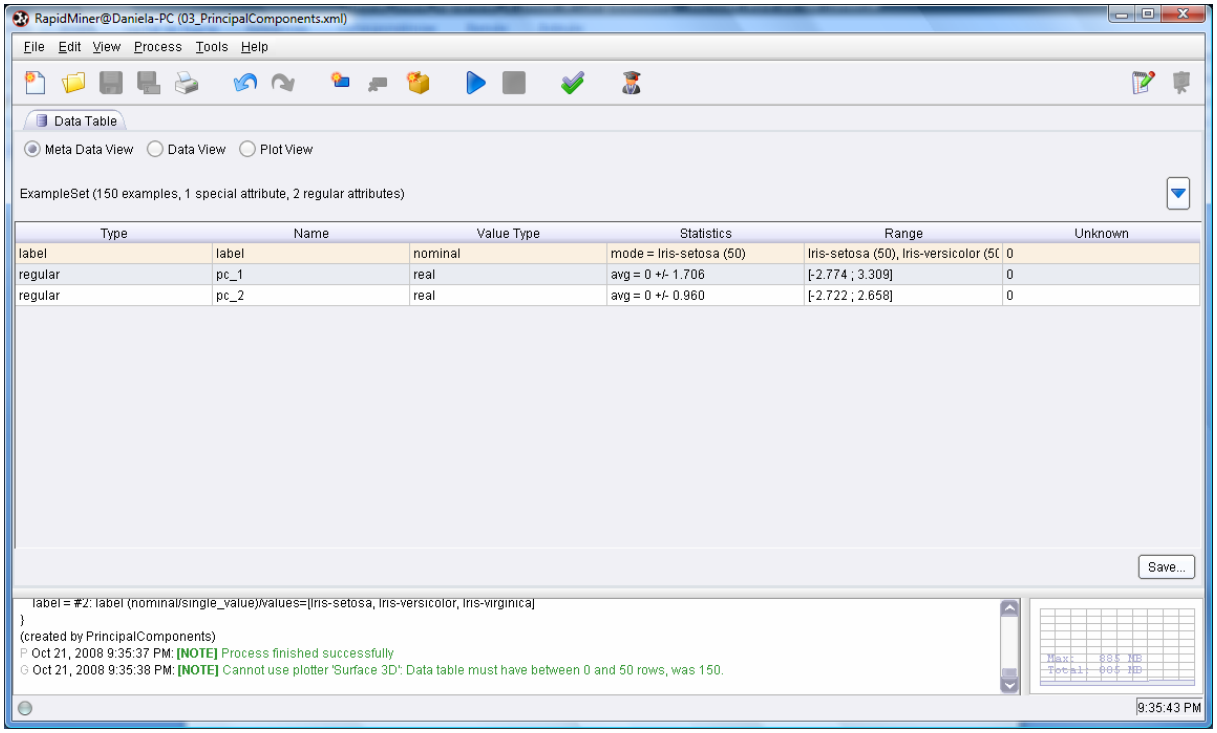
O formato básico para o parâmetro de custo-matriz é [k11 ... K1M; k21 ... k2m; ... ; Kn1 ... kNm], por exemplo 2x2 custo de uma matriz de um problema de classificação binário [0 1; 10 0]. Este exemplo significa que os custos para o erro de previsão da primeira classe como a segunda são dez vezes mais elevados do que o outro tipo de erro


Parameter Matrix: cost_matrix
Cost Matrix
True Class 1
True Class 2
Predicted Class 1
0.0
1.0
Predicted Class 2
10.0
0.0
Increase Size
Decrease Size
Ok
Cancel

20 – Ponto de Vista do Conjunto de Dados Iris

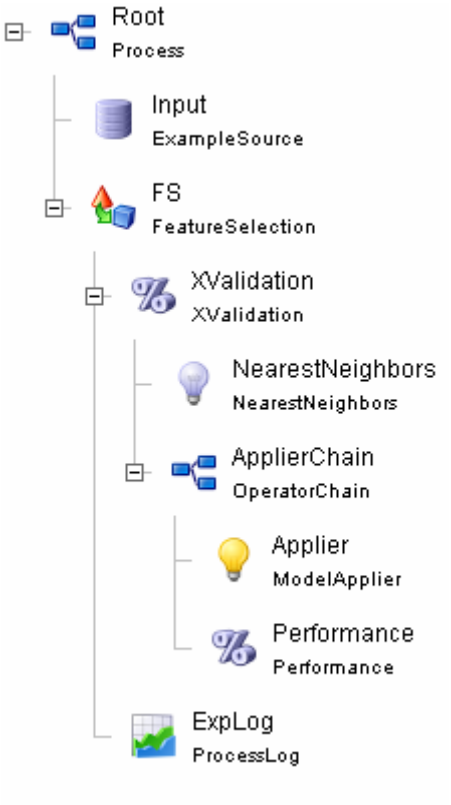
O cálculo dos componentes principais é muitas vezes usado como um recurso do passo de pré-transformação. Isto pode reduzir a dimensionalidade do conjunto de dados na mão, enquanto que a maior variância dos dados é preservada. Realize o experimento e verifique o ponto de vista do conjunto de dados Iris carregados e transformados por essa experiência.





21 – Kernel

As transformações do atributo *space* podem facilitar a aprendizagem de uma forma, que aquele simples sistema de aprendizagem possa ser capaz de aprender funções complexas. Essa é a idéia básica do *trick* kernel. Mas, mesmo sem a aprendizagem do kernel ser baseada na transformação dos esquemas de espaço do recurso ele pode ser necessário para alcançar bons resultados da aprendizagem.

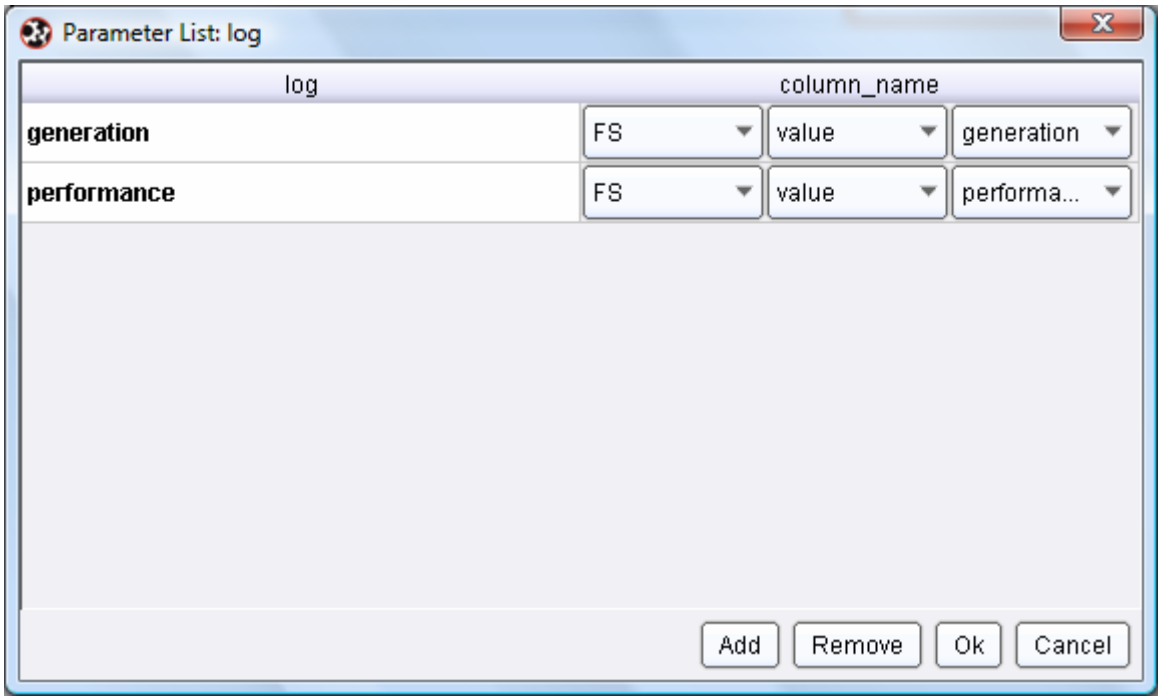


O RapidMiner oferece várias característica de seleção, construção, e de métodos de extração. Este experimento de seleção (o melhor conhecido como em frente de seleção) usa um cruzamento de validação de desempenho de estimativa de performance. Isto serve como alicerce de avaliação de todos os conjuntos de características dos candidatos. Uma vez que o desempenho de um determinado esquema de aprendizagem é levado em conta, nos referimos a este tipo de experiências como "*wrapper approaches*".

Além disso, o experimento com operador de log exibe parcelas intermediárias de resultados. Você pode inspecionar-los on-line na guia de Resultados. Para mais detalhes da consulta de experiências ou para mais detalhes consulte o tutorial RapidMiner.

Experimente o seguinte:

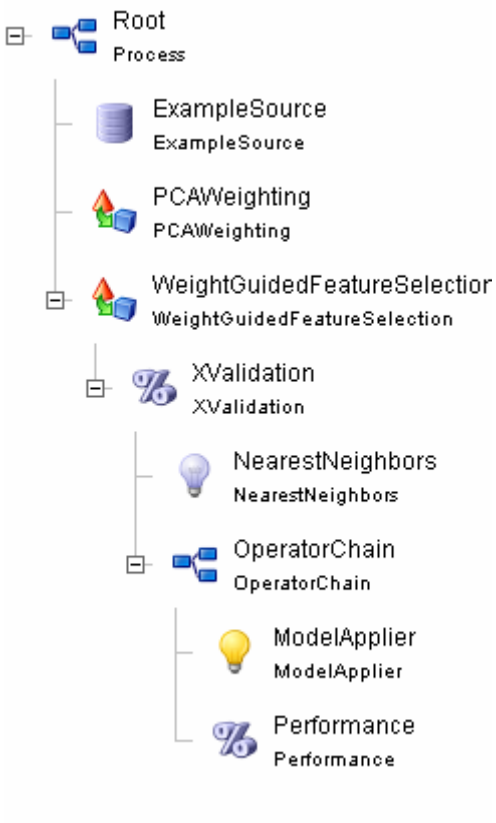
• Comece a experiência mudando para a visualização do "*Result*". A parcela não pode estar selecionada. *Plot* a "performance" contra "geração" de recursos do operador de seleção.
• Selecione o recurso de operador na exibição em árvore. Altere o diretório de busca em frente (*forward selection*) para trás (*backward elimination*). Reinicie o experimento. Todos os recursos serão selecionados.
• Selecione o recurso do operador de seleção. Clique com o botão direito do mouse para abrir o menu de contexto e mude o operador especificando outra característica do esquema de seleção (por exemplo, um algoritmo genérico).
• De uma olhada na lista de experimentos do operador de log. Cada vez que é aplicado o coletor dos dados especificados. Consulte o tutorial RapidMiner para mais explicações. Após alterar a abordagem de seleção do recurso do operador do algoritmo genérico, você tem que especificar os valores corretos.



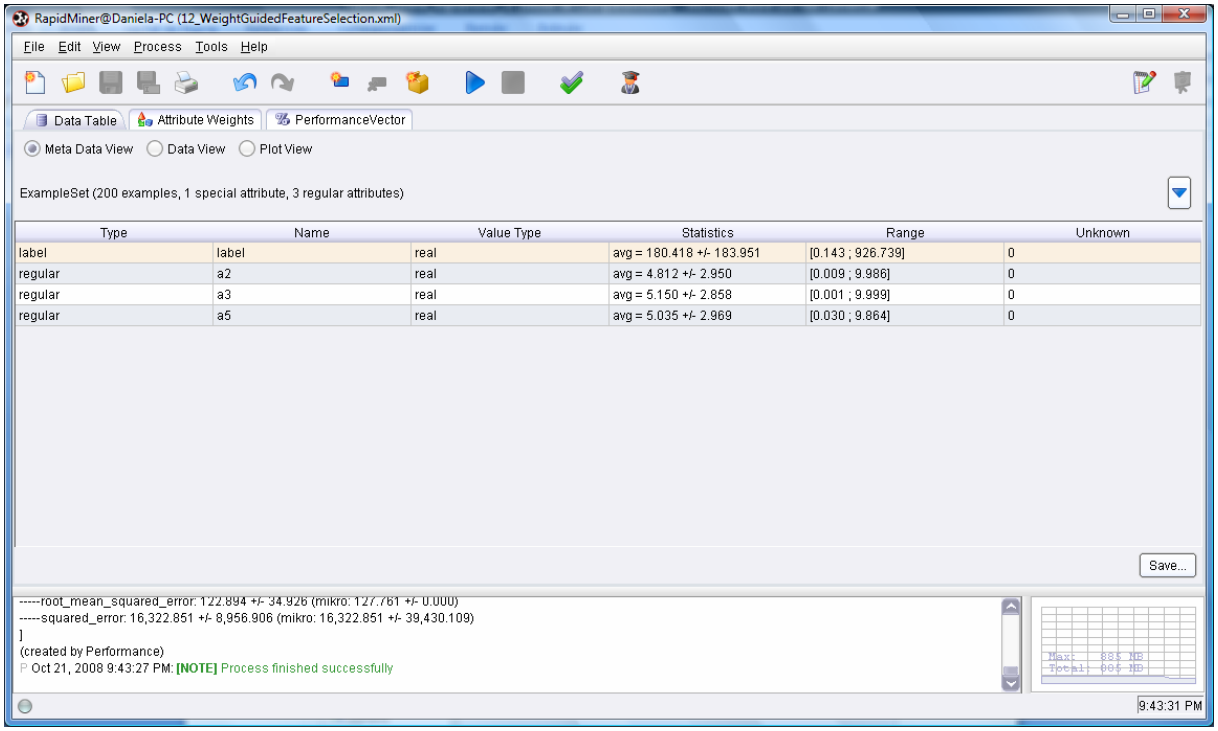
Use o experimento do operador de log para valores de log on-line

22 – Operador *WeightGuidedFeatureSelection*

O operador *WeightGuidedFeatureSelection* utiliza dados de entrada *AttributeWeights* para determinar a ordem de atributo adicionando.

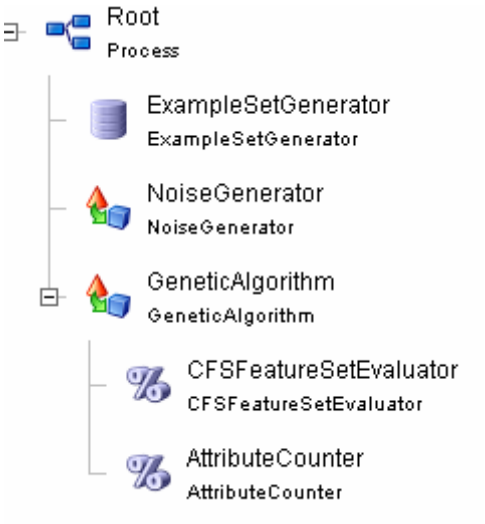


Nesta experiência, nós usaremos um cruzamento de 10 validações de um sistema de aprendizagem como de avaliação (*o operador interior*) e combinam com um atributo de filtragem abordagem de *wrapper*.

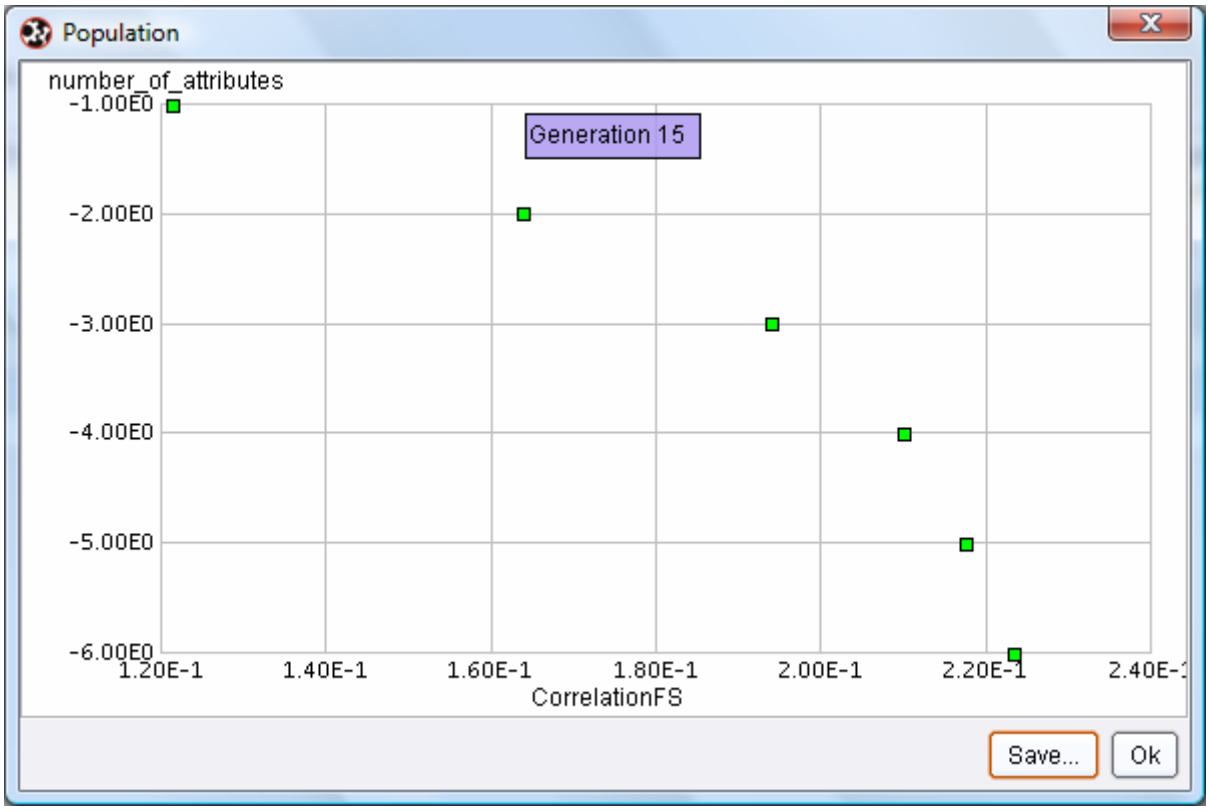


23 - Pareto

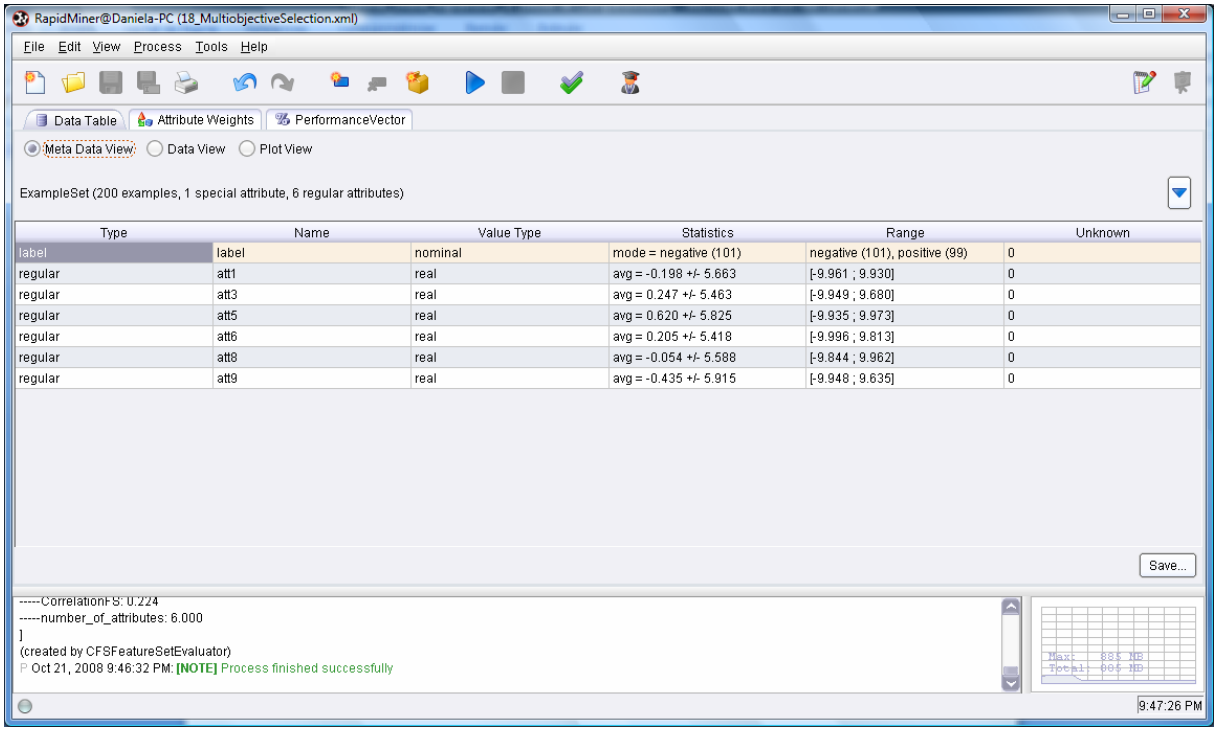
Essa é outra característica genérica muito simples do esquema de abordagem de seleção. Devido a outro esquema de seleção, o operador de recurso de seleção não só tenta maximizar o desempenho emitido pelo avaliador do conjunto de recursos, mas também tenta minimizar o número de recursos. O resultado é uma frente de *Pareto plotados* durante a otimização.



Assim que a otimização tiver terminado, o usuário poderá clicar duas vezes sobre o *Pareto-ótimo* e ver as soluções que são representadas pelo conjunto de característica.



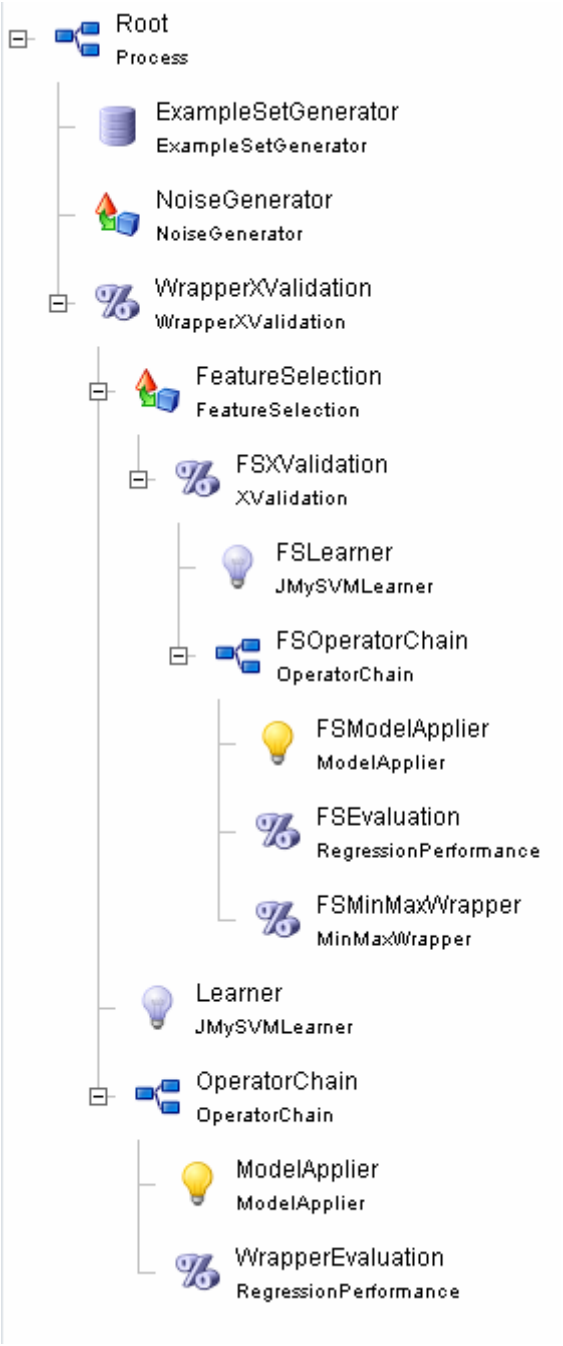
A frente de Pareto não só da uma visão sobre o número total de recursos necessários, mas também para o *trade-off* entre o número de recursos e o desempenho e em um *ranking* das características


RapidMiner@Daniela-PC (18_MultiobjectiveSelection.xml)
File Edit View Process Tools Help
Data Table
Attribute Weights
PerformanceVector
Meta Data View
Data View
Plot View
ExampleSet (200 examples, 1 special attribute, 6 regular attributes)
Type Name Value Type Statistics Range Unknown
label label nominal mode = negative (101) negative (101), positive (99) 0
regular att1 real avg = -0.198 +/- 5.663 [-9.961 ; 9.930] 0
regular att3 real avg = 0.247 +/- 5.463 [-9.949 ; 9.680] 0
regular att5 real avg = 0.620 +/- 5.825 [-9.935 ; 9.973] 0
regular att6 real avg = 0.205 +/- 5.418 [-9.996 ; 9.813] 0
regular att8 real avg = -0.054 +/- 5.588 [-9.844 ; 9.962] 0
regular att9 real avg = -0.435 +/- 5.915 [-9.948 ; 9.635] 0
Save...
-----number_of_attributes: 6.000
]
(created by CFSFeatureSetEvaluator)
Oct 21, 2008 9:46:32 PM: [NOTE] Process finished successfully
9:47:26 PM

24 – Operadores de Pré-processamento

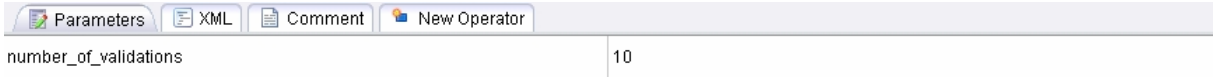
Assim como a aprendizagem, pode ocorrer durante o pré-processamento, para estimar o desempenho da generalização de um método de pré-processamento em que o RapidMiner suporta vários operadores de validação de pré-etapas. A idéia básica é a mesma para todos os outros operadores de validação com uma ligeira diferença: o primeiro operador deve produzir um conjunto de transformação no conjunto de exemplos, a segunda deve produzir um modelo de transformação deste conjunto de dados e o terceiro operador deve produzir um vetor de performance desse modelo em *hold-out* de conjunto de teste transformado, da mesma forma.

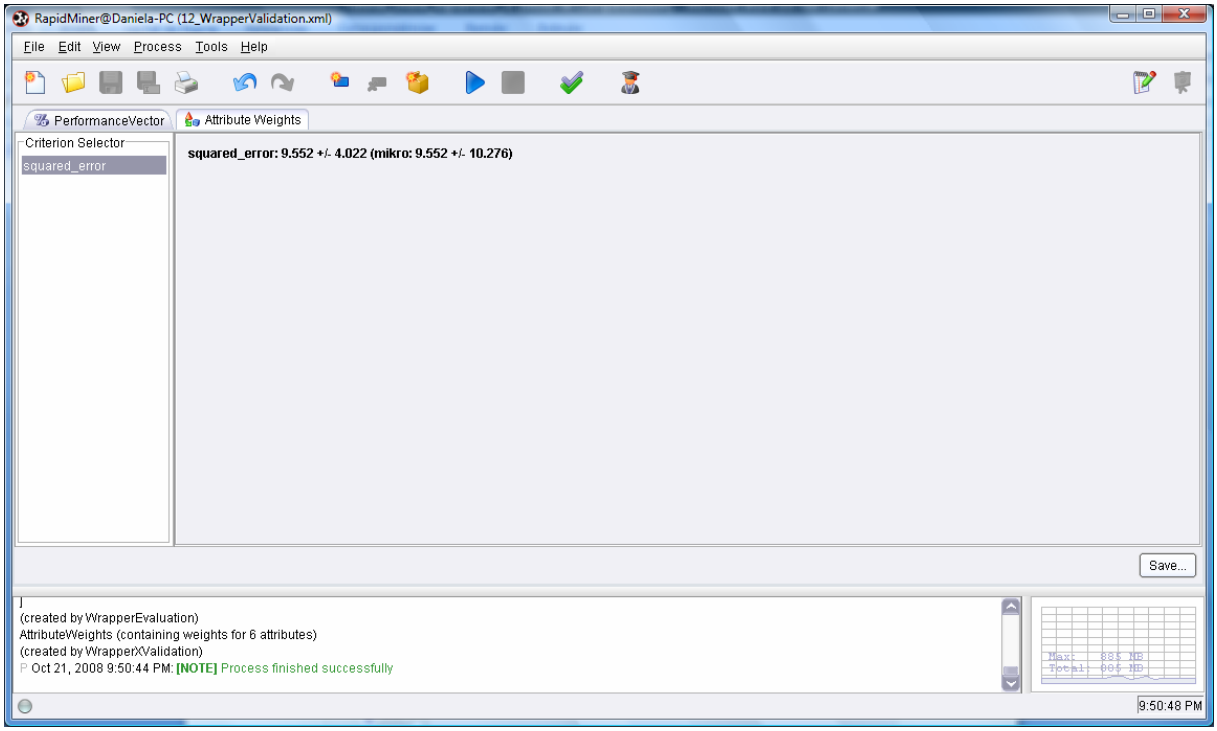
Esta é uma experiência que demonstra a mais complexa capacidade do RapidMiner de construir experiências a partir de *building blocks* já conhecidos. Especialmente nesta experiência existe uma variante de uma validação de cruzamento de operadores que são utilizados para calcular o desempenho de uma característica da transformação do espaço, ou seja, o simples recurso seleção frente a este caso.



A completa funcionalidade da seleção de *building blocks* é agora o primeiro operador de entrada de um *WrapperXValidation* como validação cruzada normal que usa um subconjunto de transformação de funcionalidades do espaço e de aprendizagem baseado em determinada característica definida. Uma segunda cadeia de aplicação é utilizada para calcular um conjunto de testes que não foram utilizados para a aprendizagem no recurso de seleção. Estima-se que o desempenho é um atributo de peso retornando como resultado um vetor.

Observe o *MinMaxWrapper* após a avaliação do desempenho de entrada. Este operador encapsula os dados do critérios ao desempenho de tal forma que já não apenas valores médios, mas também valores mínimos sejam calculados durante a validação cruzada.

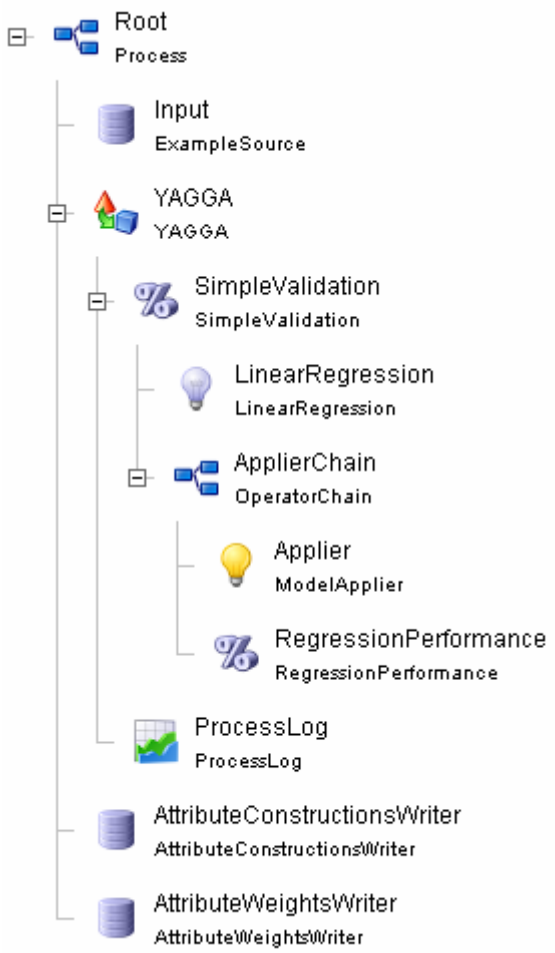




Arbitrariamente poderão ser realizadas combinações lineares do mínimo e da média normal que alcança melhor a generalização de capacidades. Basta alterar a parâmetro de ponderação para 0,0 ou desativar o operador no menu de contexto ou excluí-lo a partir da experiência e ver o efeito. O desempenho diminui rapidamente quando só o desempenho médio é utilizado como critério de seleção.

25 – *YAGGA*

Às vezes a seleção dos recursos por si só é insuficiente. Nestes casos, outras transformações do recurso de espaço devem ser realizadas. A geração de novos atributos a partir do dado do atributo estende a funcionalidade espaço. Talvez uma hipótese possa ser facilmente encontrada no espaço de recurso estendido.



YAGGA (Geração de mais um algoritmo genérico) é uma característica híbrida de seleção / *wrapper* de geração. A estimativa de desempenho é feita com um cruzamento interior de *building block*. Claro outras formas de desempenho de estimativa são também possíveis. A probabilidade de geração de recurso depende da probabilidade do recurso remoção. Isso garante que a duração média de recurso fixo mantenha-se até a mais curta ou a mais longa característica de conjuntos que revelar-se melhor.

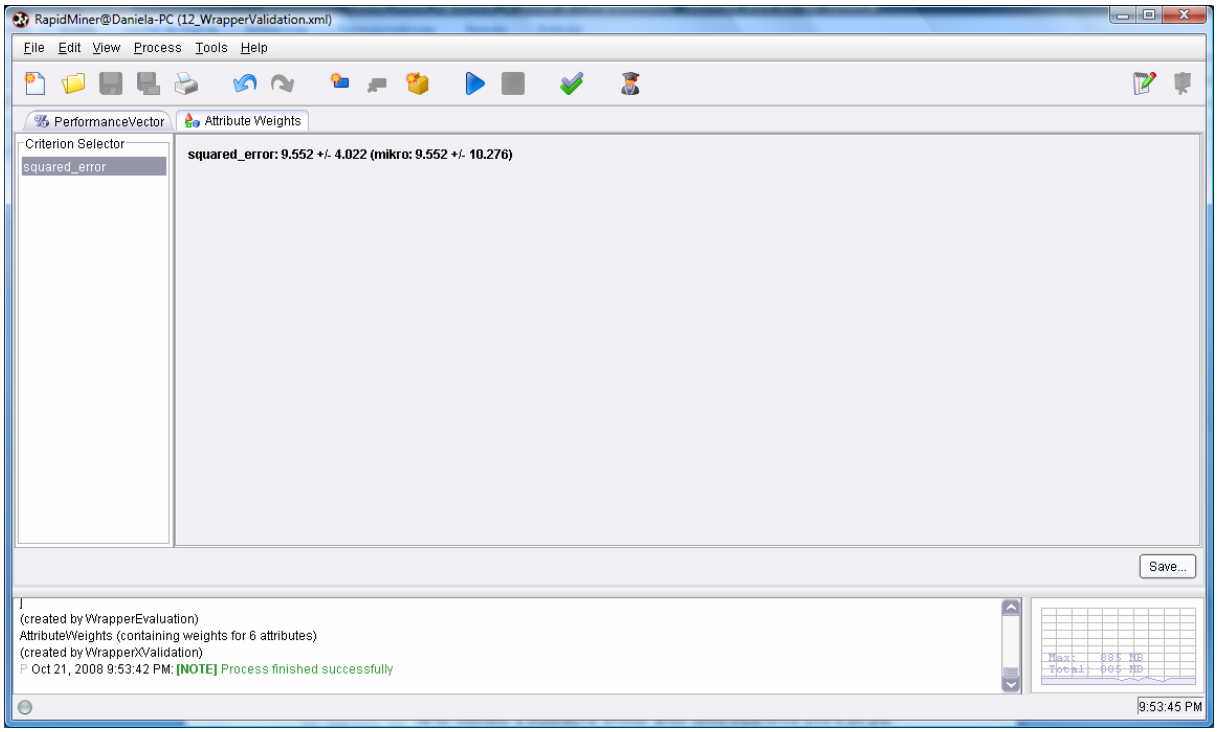
Quando YAGGA termina a transformação, foram construídas novas funcionalidades. Em muitos casos, esta característica do conjunto ótimo deve ser usada em outros dados também. Por isso o melhor conjunto de atributos é escrito em um arquivo. No próximo exemplo, vamos ver como estes arquivos podem ser utilizados para transformar novos dados para a representação de aprendizagem ótima.

Experimente o seguinte:

• Comece a experiência. A transformação de um conjunto de exemplos de dados de entrada, a estimativa do desempenho, pesos e um vetor são entregue como resultado. Todos os operadores como característica YAGGA tem um parâmetro "*apply_best_weights*" (apenas em modo expert). Como é o resultado das mudanças quando se usa este parâmetro?
• Tente adicionar um experimento de operador de log. A YAGGA não só permite um operador interior, você tem que acrescentar um simples operador da cadeia (a partir do "núcleo" do grupo) para YAGGA. Clique com o botão direito do mouse sobre o operador da validação cruzada e experimente cortar e colar para acrescentar o armazenamento da validação para a cadeia. Adicione um experimento de operador de log na cadeia. Adicione os valores que gostaria de utilizar como parâmetro da lista do experimento do operador de log. Consulte o tutorial RapidMiner para mais explicações.
Um simples operador de cadeia para fundir vários operadores.

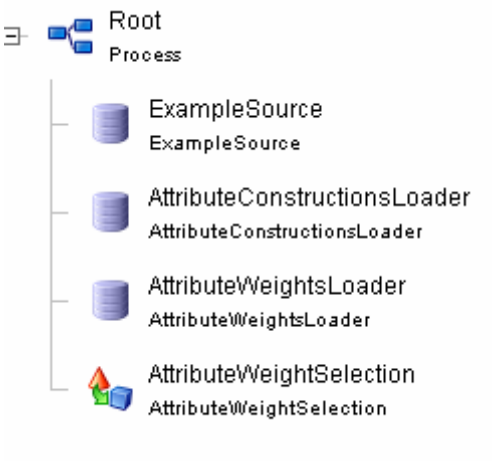
Retire o operador, de um operador árvore.

Cole o operador cortado anteriormente para o operador de cadeia selecionado.

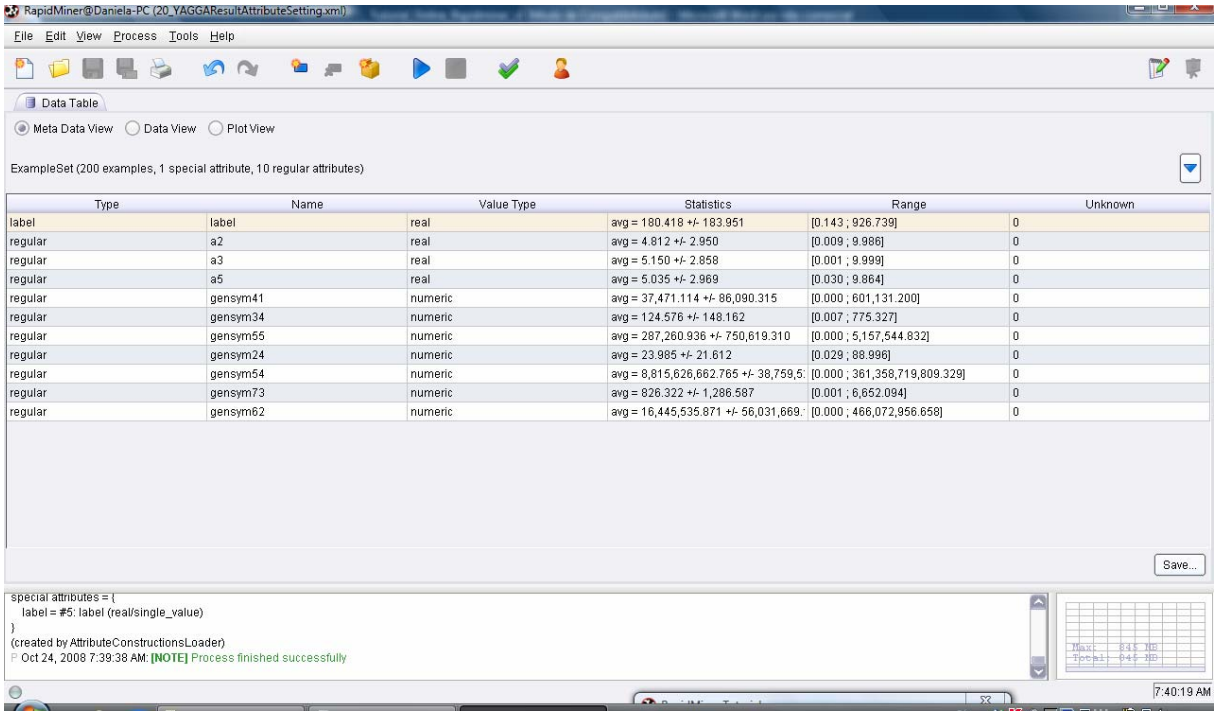


26 – Conjunto de Atributo Ideal

Na experiência anterior foi um conjunto do atributo ótimo pesquisado (Certifique-se de ter realizado a experiência anterior antes desta experiência pois é um pré-requisito).



Este conjunto ideal de atributo é carregado e aplicado a outra entrada de dados. Isto é necessário para aplicar um modelo que foi aprendido a partir de dados com a mesma representação de entrada.

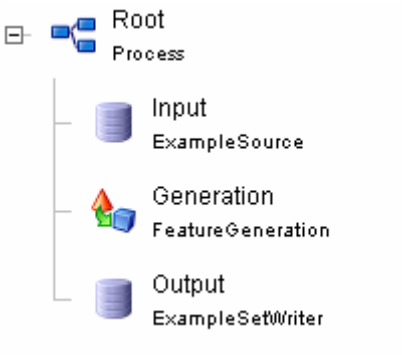


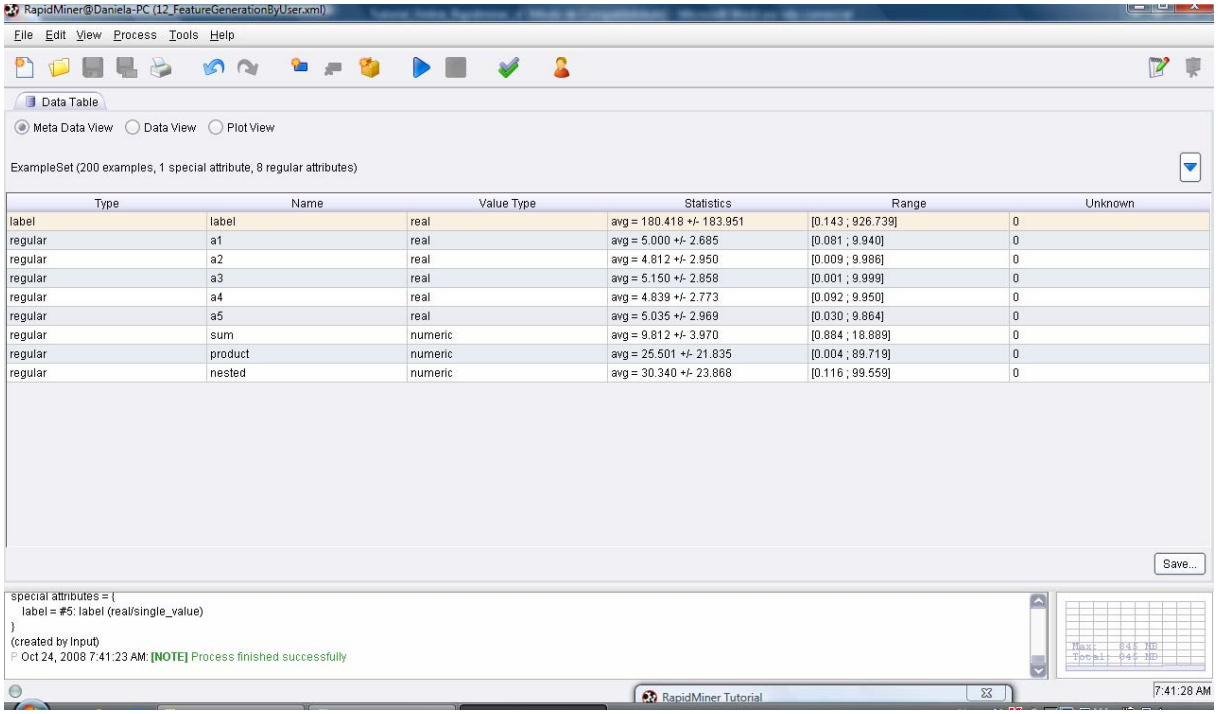
Experimente o seguinte:

• Comece a experiência. Depois de alguns instantes o exemplo de entrada de dados usa o recurso de representação ótima que foi encontrado na experiência anterior.

27 – Operadores de geração

Esta experiência carrega dados de arquivos numéricos e gera alguns atributos com a característica de operadores de geração. Este operador também pode gerar atributos de acordo com um atributo arquivo que foi salvo no arquivo de antemão. Desta forma, não é só possível usar o atributo criado automaticamente, mas também atributos definidos pelo usuário. Por isso o parâmetro lista "funções" da geração de operador deve ser editado.



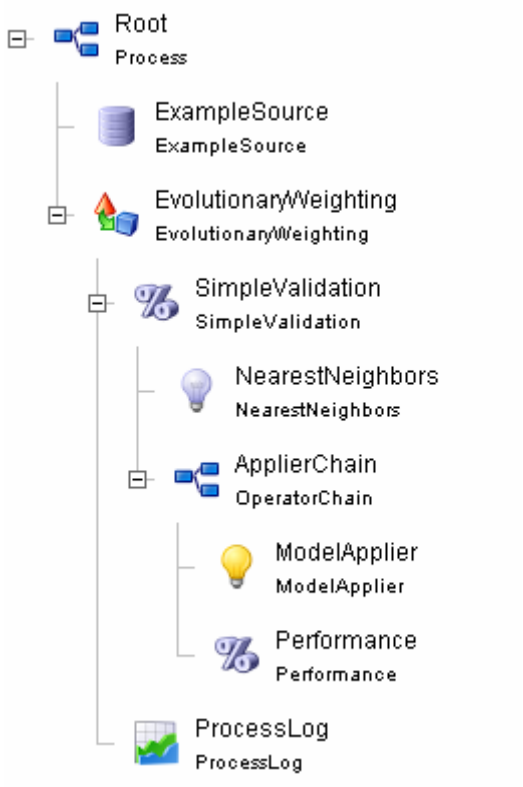


Experimente o seguinte:

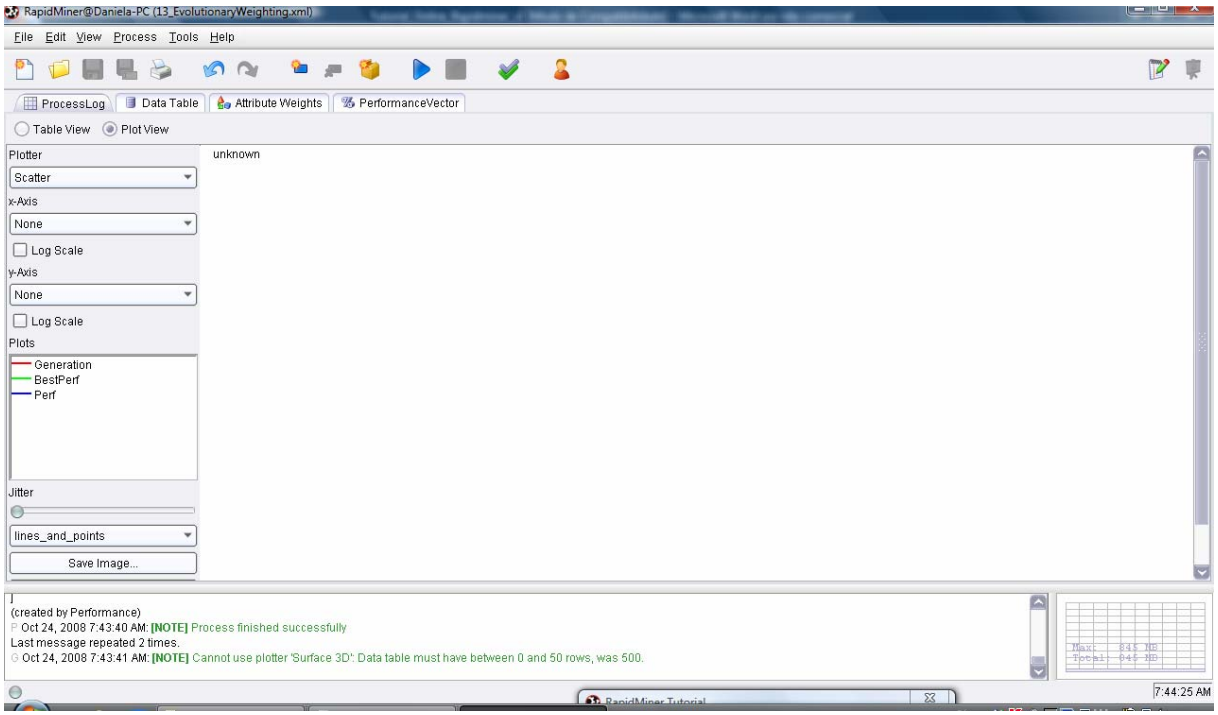
• Comece a experiência. Use o *breakpoint* para verificar a geração do passo. O parâmetro "*keep_all*" define se todos os atributos devem ser utilizados para o exemplo dado como resultado ou apenas o recém-gerado.
• Edite o parâmetro lista "*functions*" e acrescente algumas outras funções. As funções são escritas em ordem de prefixadas e a maioria das funções matemáticas definidas pelo Java podem ser usadas. Valores constantes são definidos por "*const* [valor]()". Não esquecer o vazio entre colchetes.

28 – *Validação da cadeia interna*

Esta é outra amostra de experiência mais complexa. Ela usa uma validação interior da cadeia (neste caso uma simples validação ao invés da validação cruzada) para estimar o desempenho de um *Learner* no que diz respeito aos pesos dos atributos. Estes são adaptados com uma ponderação de abordagem evolutiva.



Como você pode ver, a estrutura geral do experimento é muito semelhante ao recurso a seleção e geração de experimentos. Em todos os casos, uma validação da cadeia interna é usada como alicerce para a estimativa de desempenho. O operador ("*EvolutionaryWeighting*", neste caso) realiza algumas operações em conjunto com as características que são avaliadas pelo operador (simples validação).



Experimente o seguinte:

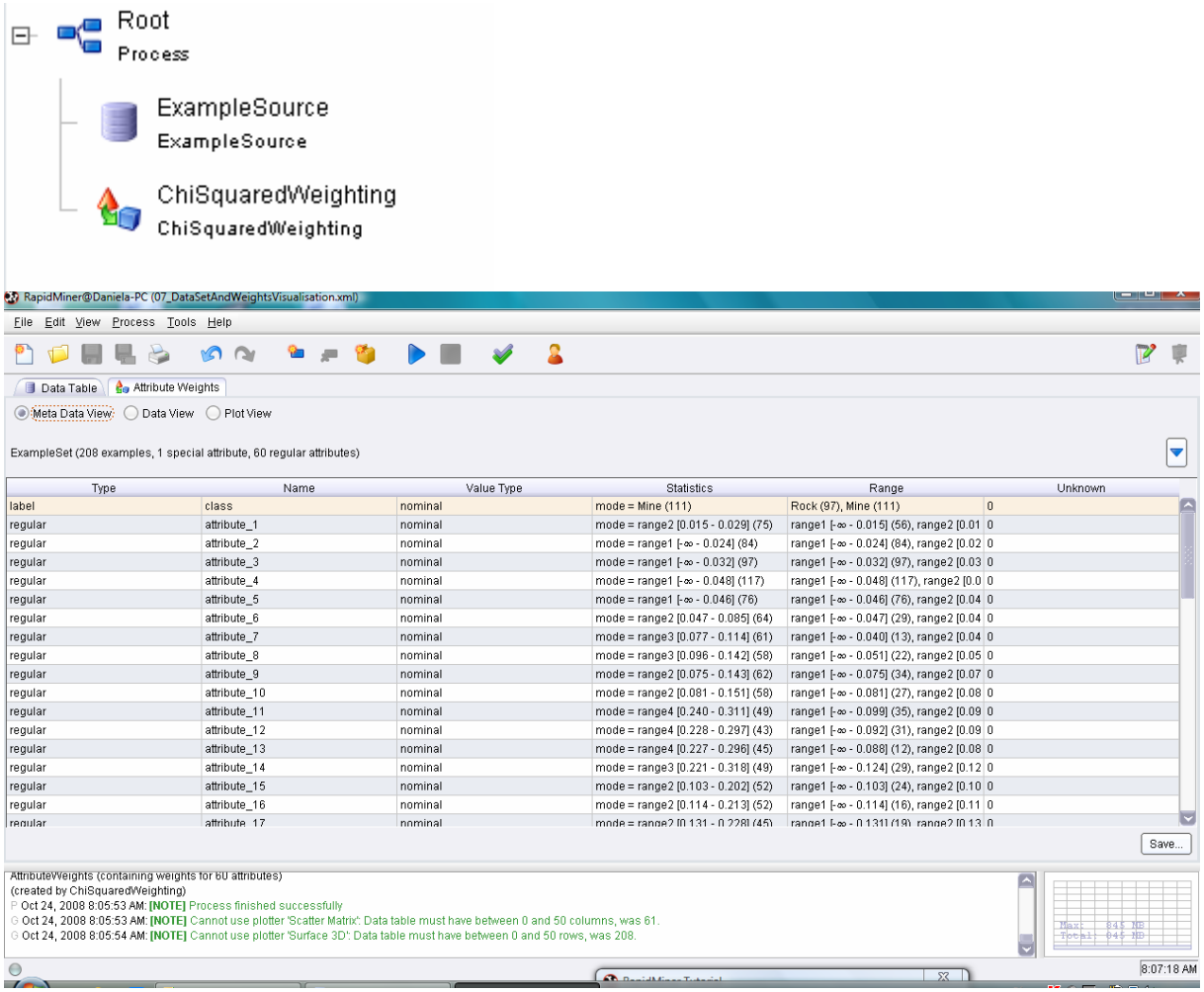
• Comece a experiência. Mude para "*Results*" e veja linha *plotter*. Pressione o ícone parar na barra de ícones para parar a experiência. O operador atual irá parar o seu funcionamento em segundo plano e pode durar algum tempo até que a experiência seja completamente interrompida.

Embora você possa alterar a atual experiência e reiniciá-la, ela irá mais devagar até a antiga experiência estar totalmente parada.

Pressione o ícone "*stop*" para interromper a experiência

29 – Combinação de Resultados

Nesta experiência, carregaremos um conjunto de dados e aplicaremos uma ponderação dos esquemas de recurso disponíveis no RapidMiner sobre este conjunto de dados.

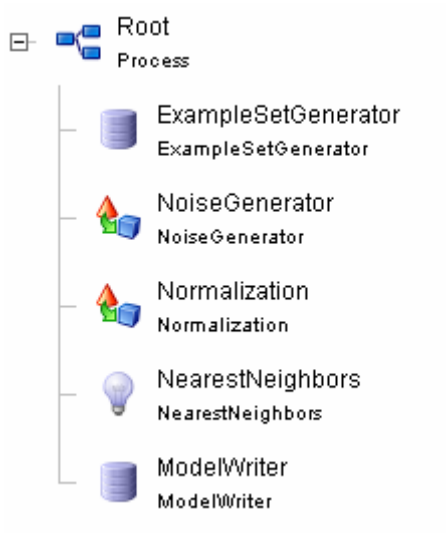


Assim que a experiência tiver terminada, basta mudar a *plot view* do exemplo do conjunto e conferir o high-dimensional disponível como *plot* paralelo, vistoria parcela, *RadViz* ou *GridViz* parcela, *histogram matrix*, *quartile matrix* e as variantes desta *plots* colorida. Você vai ver que algumas das colunas são marcadas por uma cor amarelada, por exemplo, por um retângulo em torno ou diretamente na parcela. Estas marcas amarelas indicam que o peso dos atributos e os correspondentes, a cor é mais intensa se o peso correspondente é superior.

Esta experiência demonstra a capacidade do RapidMiner de apresentar vários resultados através da combinação deles. É claro que você ainda pode ter que olhar a tabela ou os pesos de diferentes pontos de vista do *plot* do atributo de pesos.

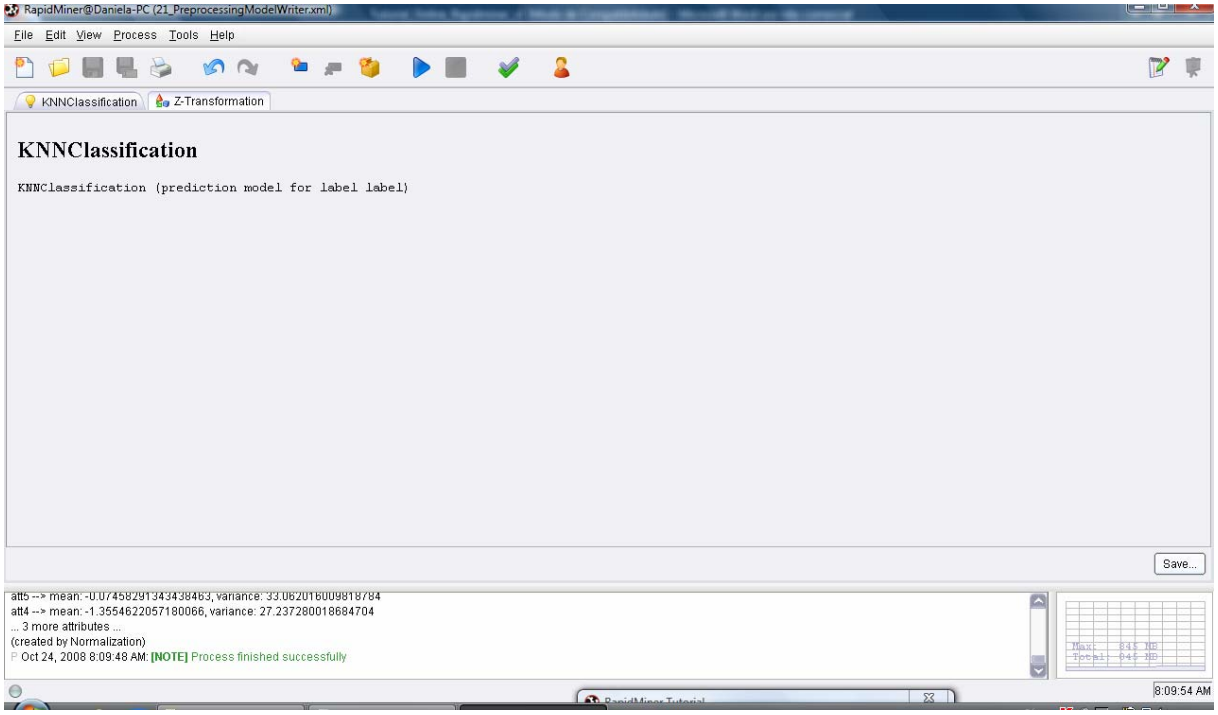
30 – Operador de Normalização

Esta experiência gera um conjunto de dados e executa uma normalização (ou seja, a média é de 0 e 1 depois do desvio-padrão).



Parameters | XML | Comment | New Operator

return_preprocessing_model	☑
create_view	☐
z_transform	☑
min	0.0
max	1.0



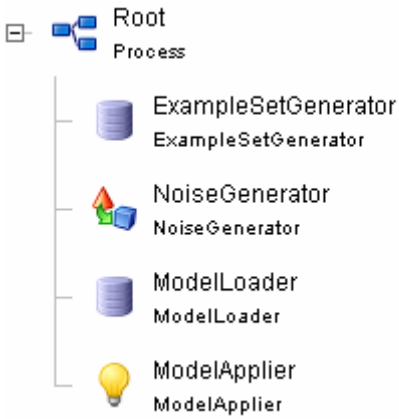
Observe que alguns pré-operadores como o operador de Normalização também são capazes de produzir um modelo, ou seja, um modelo prévio. O parâmetro "*return_preprocessing_model*" do operador de Normalização deve ser verificado de forma a criar um modelo desse tipo.

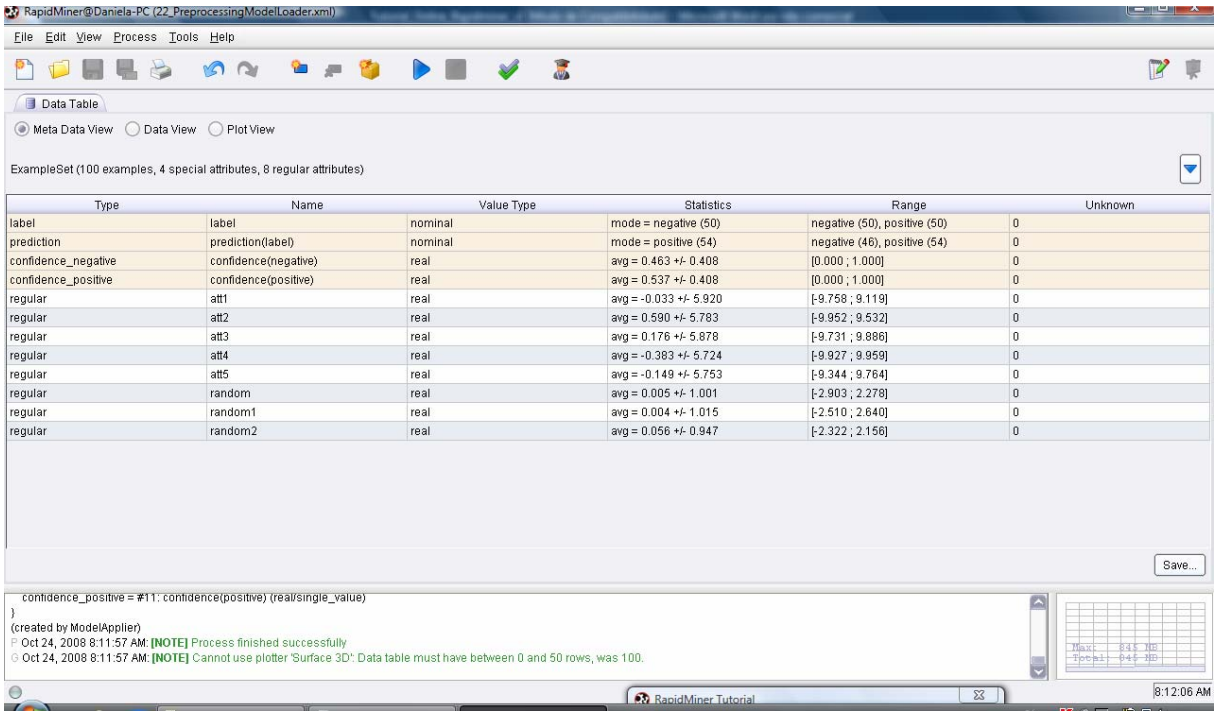
Este modelo pode ser utilizado, a fim de aplicar a mesma transformação nos dados de testes que não tenham sido vistos antes. O modelo prévio é automaticamente pré-combinado com uma previsão modelo, neste experimento com vizinhos mais próximos ao modelo entregue pela IBk. A combinação do modelo pode ser salva em um arquivo e, mais tarde, recarregada e aplicada a novos conjuntos de dados. Combinando modelos, assim, garantindo que os mesmos passos prévios com as mesmas configurações também sejam aplicados aos novos conjuntos de dados.

Observe que, assim que a experiência tiver terminado você pode ver o modelo na guia resultado combinado e utilizar o modelo seletor com um clique do botão esquerdo do mouse para escolher o modelo que deve ser exibido.

31 – Combinação de modelos diferentes de arquivos

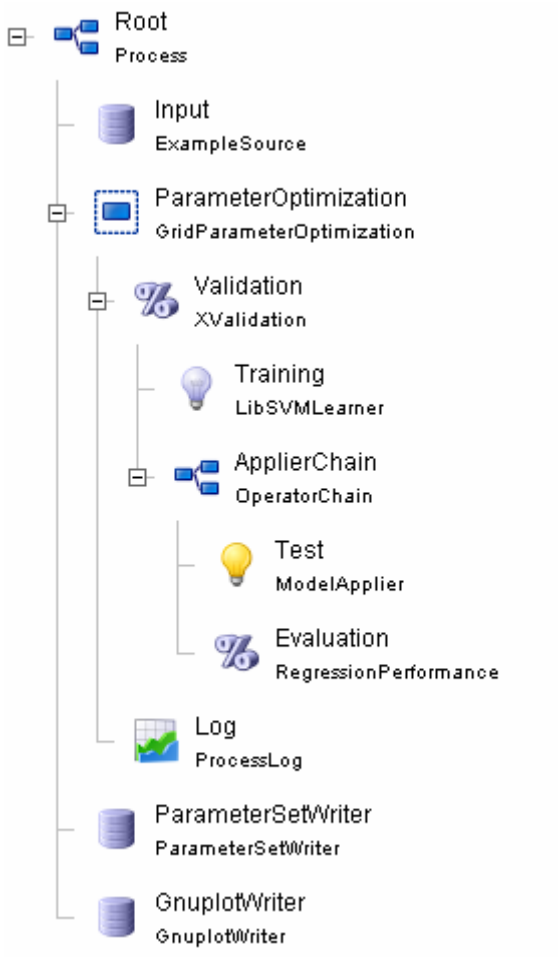
Este experimento faz uso da experiência anterior e carrega o texto e combinando modelo semelhante de arquivo após a geração, mas não iguais, do conjunto de dados (provocado por diferentes sementes aleatórias na experiência de operadores em árvore). A combinação do modelo é então aplicada ao conjunto de dados exibido.





32 – Operador de Otimização

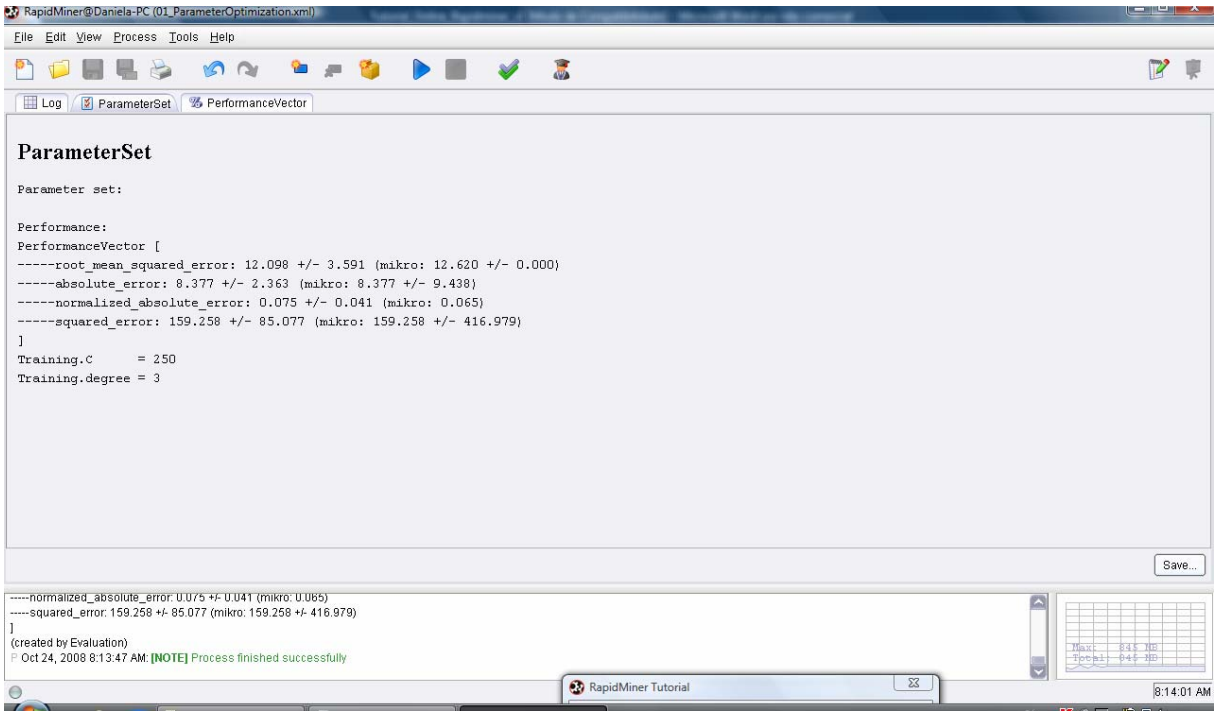
Muitas vezes os diferentes operadores têm muitos parâmetros e não ficam claros quais valores do parâmetro são melhores para a aprendizagem da tarefa pela frente. O parâmetro do operador de Otimização ajuda a encontrar um parâmetro ótimo estabelecido para os operadores usados.



A validação cruzada inserida das estimativas do desempenho de cada um dos parâmetros estabelecidos. Nesta experiência dois parâmetros da SVM estão sintonizados. O resultado pode ser *plotted* em 3D (utilizando o *gnuplot*) ou em modo cores.

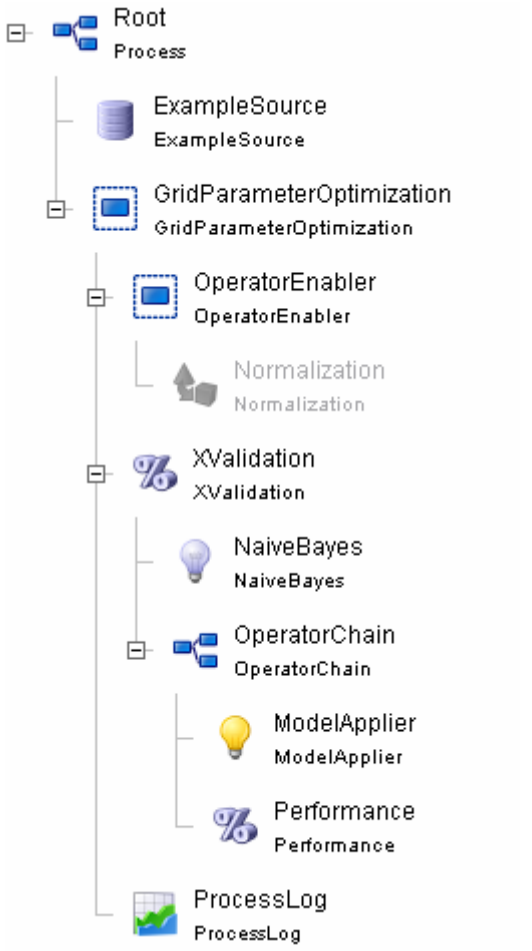
Experimente o seguinte:

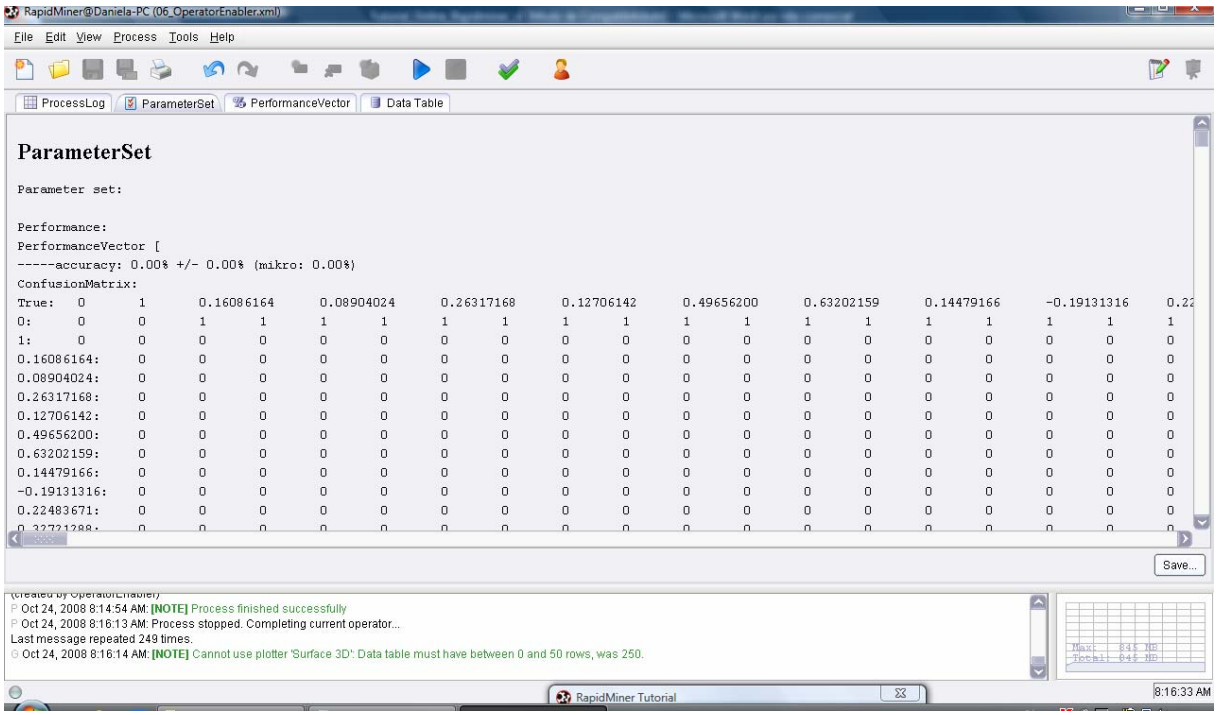
• Comece a experiência. O resultado é o melhor parâmetro estabelecido em que o desempenho foi conseguido com este parâmetro estabelecido.
• Edite o parâmetro da lista do operador *ParameterOptimization* e encontre outro parâmetro estabelecido.

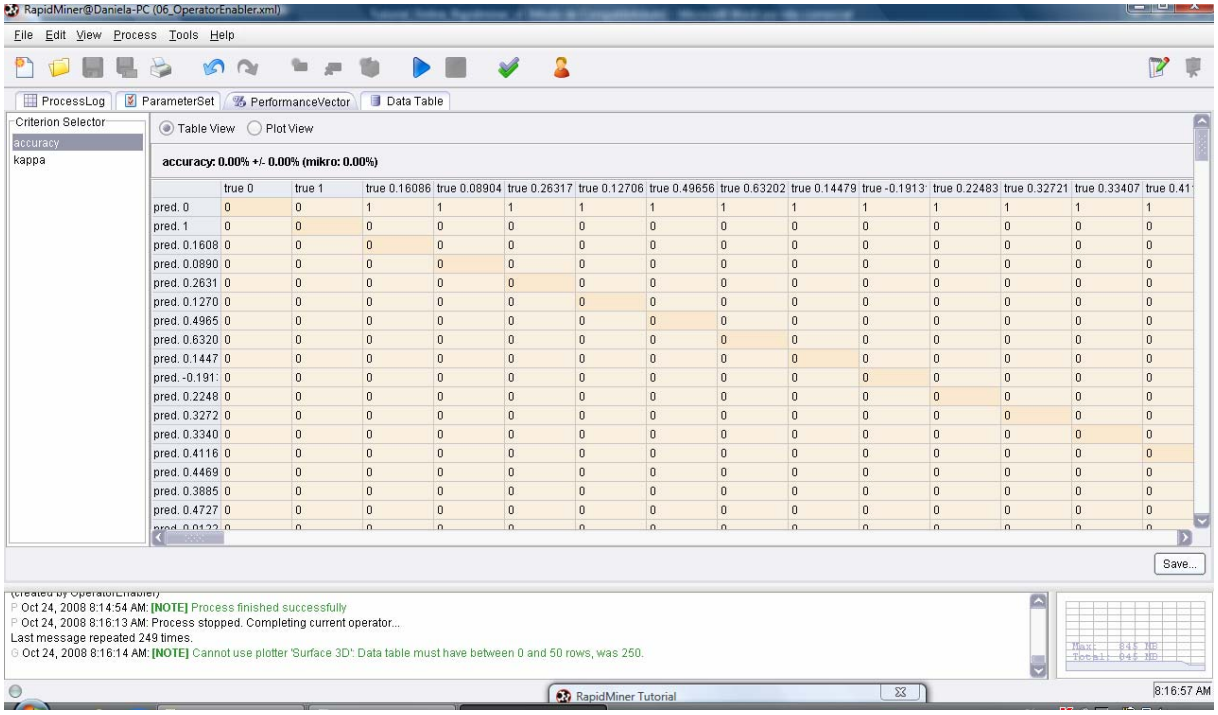


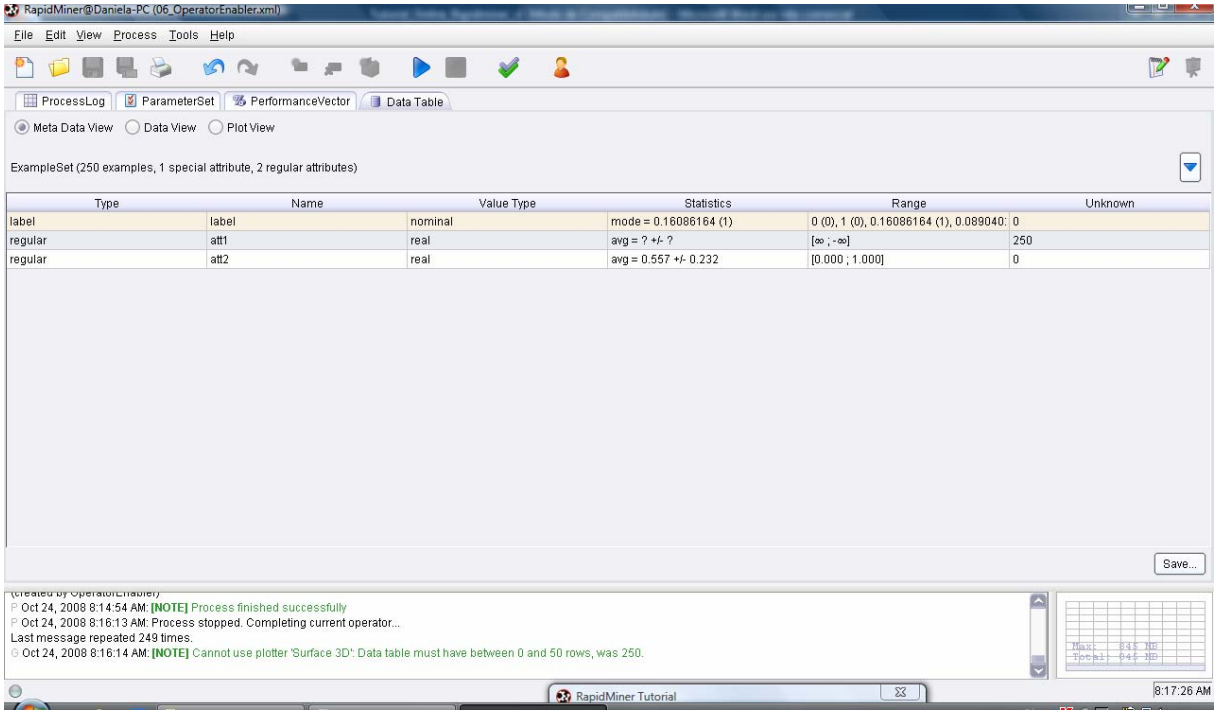
33 – Operador Enabler

Esta meta experiência mostra automaticamente outra possibilidade para otimizar a experiência layout. O operador "*OperatorEnabler*" pode ser usado para ativar ou desativar um dos seus filhos.


Root
Process
ExampleSource
ExampleSource
GridParameterOptimization
GridParameterOptimization
OperatorEnabler
OperatorEnabler
Normalization
Normalization
XValidation
XValidation
NaiveBayes
NaiveBayes
OperatorChain
OperatorChain
ModelApplier
ModelApplier
Performance
Performance
ProcessLog
ProcessLog


RapidMiner@Daniela-PC (06_OperatorEnabler.xml)
File Edit View Process Tools Help
ProcessLog
ParameterSet
PerformanceVector
Data Table
ParameterSet
Parameter set:
Performance:
PerformanceVector [
-----accuracy: 0.00% +/- 0.00% (mikro: 0.00%)
ConfusionMatrix:
Save...
Oct 24, 2008 8:14:54 AM: [NOTE] Process finished successfully
Oct 24, 2008 8:16:13 AM: Process stopped. Completing current operator...
Last message repeated 249 times.
Oct 24, 2008 8:16:14 AM: [NOTE] Cannot use plotter 'Surface 3D': Data table must have between 0 and 50 rows, was 250.
RapidMiner Tutorial
8:16:33 AM

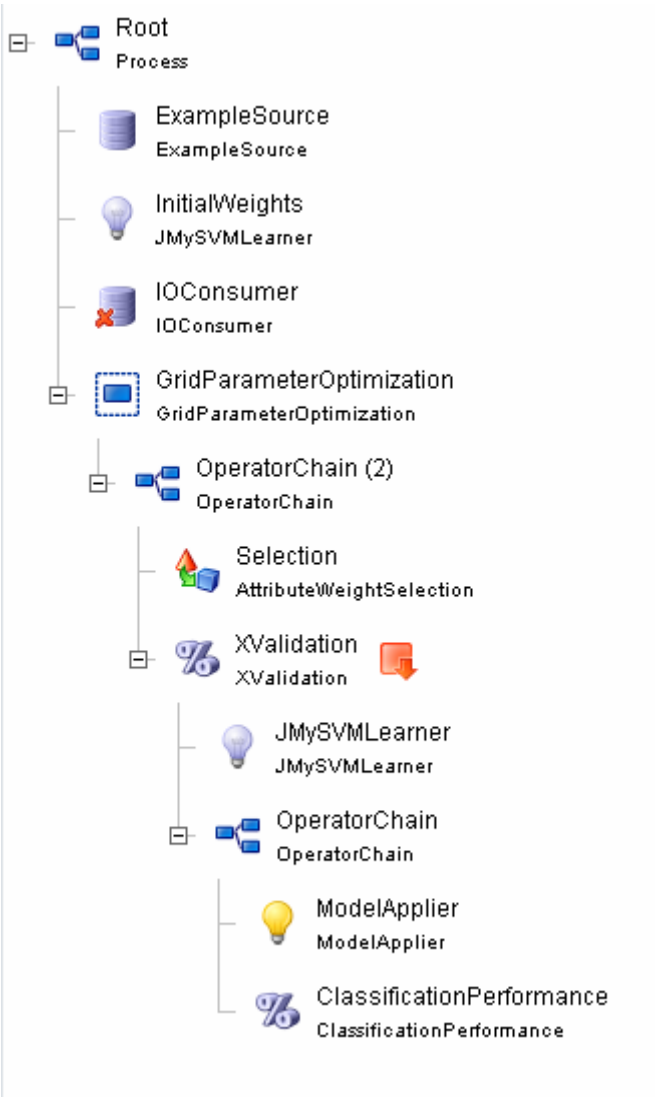


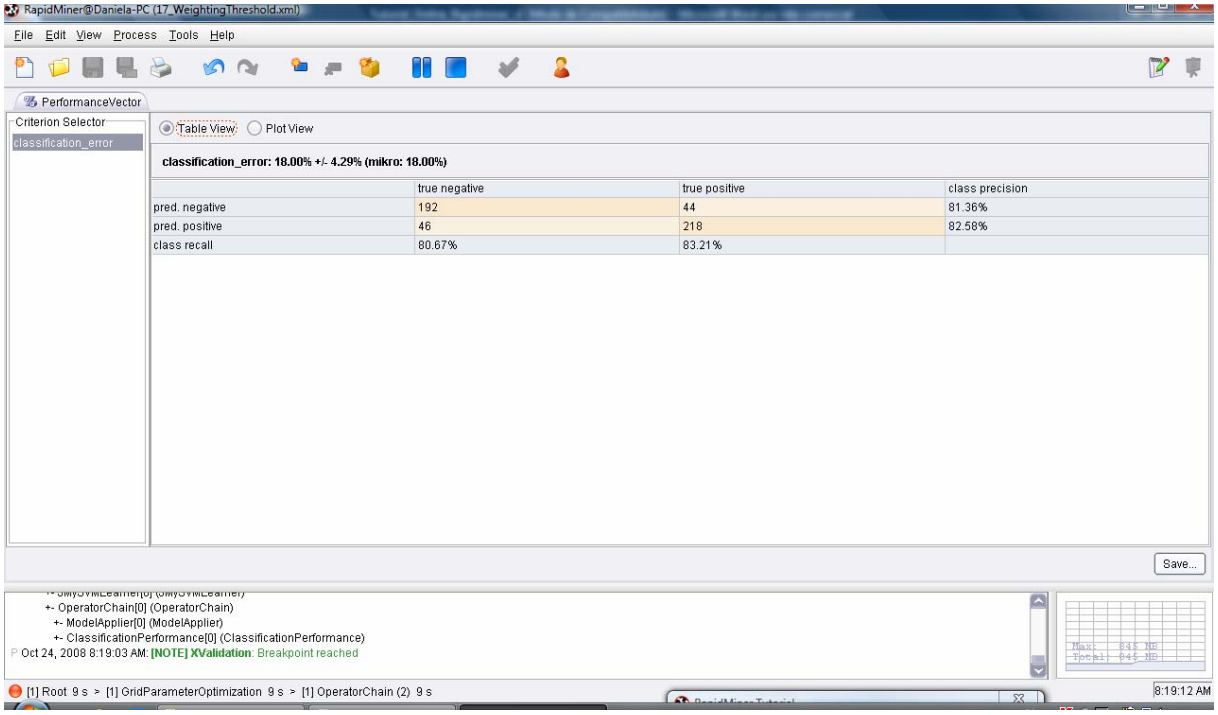


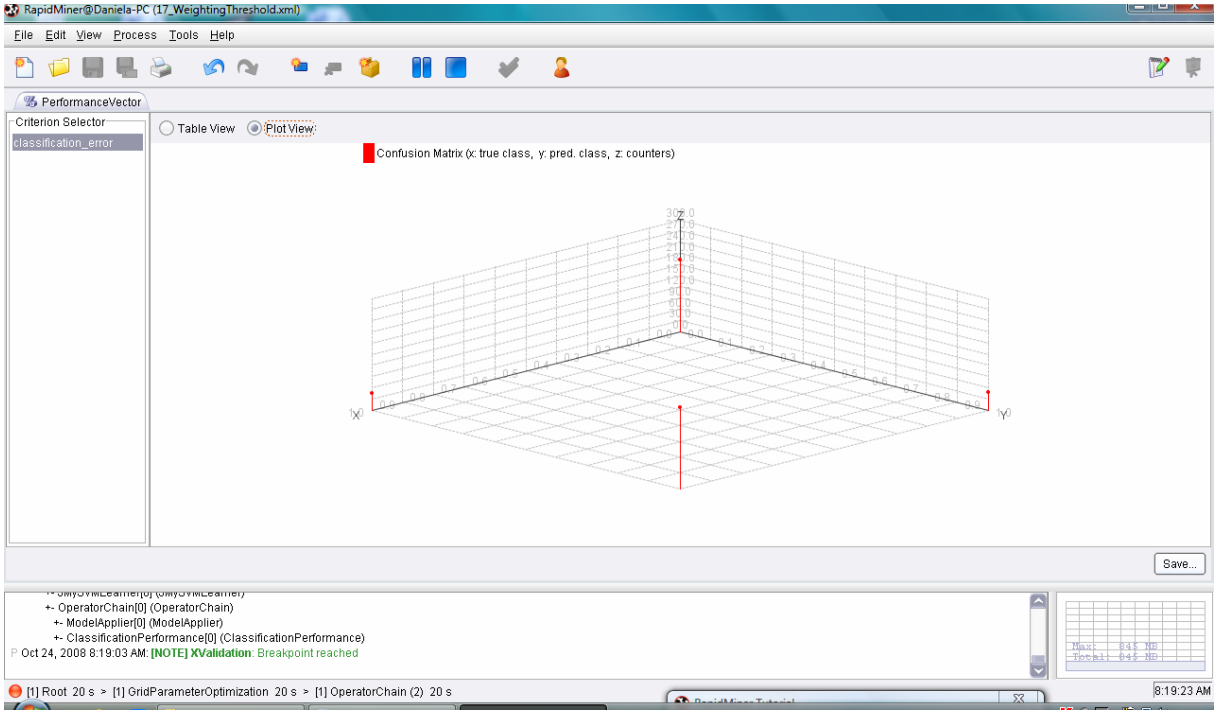
Juntamente com um dos operadores de otimização deste parâmetro pode ser usado para verificar que os operadores devem ser utilizados para melhores resultados. Isto é especialmente útil, a fim de determinar quais os operadores prévios devem ser usados para determinar um conjunto de dados – *Learner combination.*

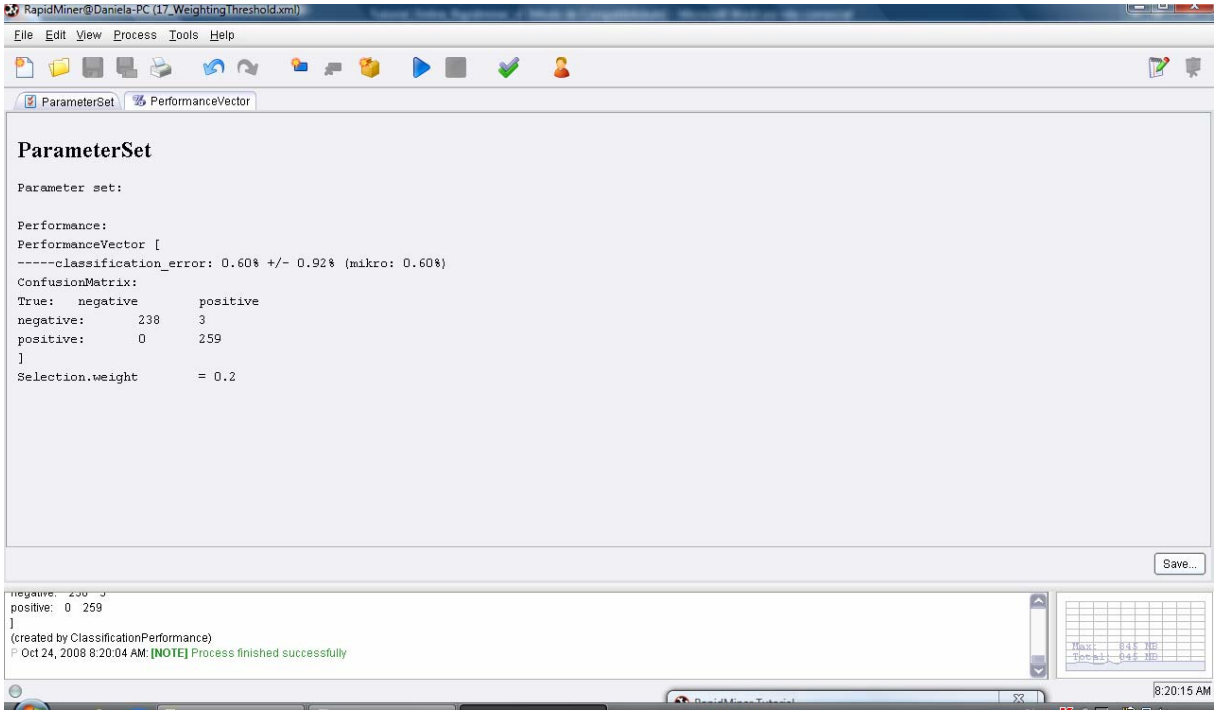
34 – Experimento de Otimização de Operadores

Essa experiência tenta encontrar a melhor seleção limiar para os pesos fornecidos por um SVM *Learner*. Os pesos e o exemplo é encaminhado a um parâmetro de otimização. O parâmetro "peso" da Seleção operador é otimizado com uma pesquisa de grid. O desempenho deste limiar é avaliado com a validação cruzada de *building block*. Consulte a meta amostra de experimentos para mais detalhes sobre o parâmetro de otimização de operadores.




RapidMiner@Daniela-PC (17_WeightingThreshold.xml)
File Edit View Process Tools Help
PerformanceVector
Criterion Selector
classification_error
Table View
Plot View
classification_error: 18.00% +/- 4.29% (mikro: 18.00%)
true negative
true positive
class precision
pred. negative
192
44
81.36%
pred. positive
46
218
82.58%
class recall
80.67%
83.21%
Save...
+- OperatorChain[0] (OperatorChain)
+- ModelApplier[0] (ModelApplier)
+- ClassificationPerformance[0] (ClassificationPerformance)
Oct 24, 2008 8:19:03 AM: [NOTE] XValidation: Breakpoint reached
[1] Root 9 s > [1] GridParameterOptimization 9 s > [1] OperatorChain (2) 9 s
8:19:12 AM

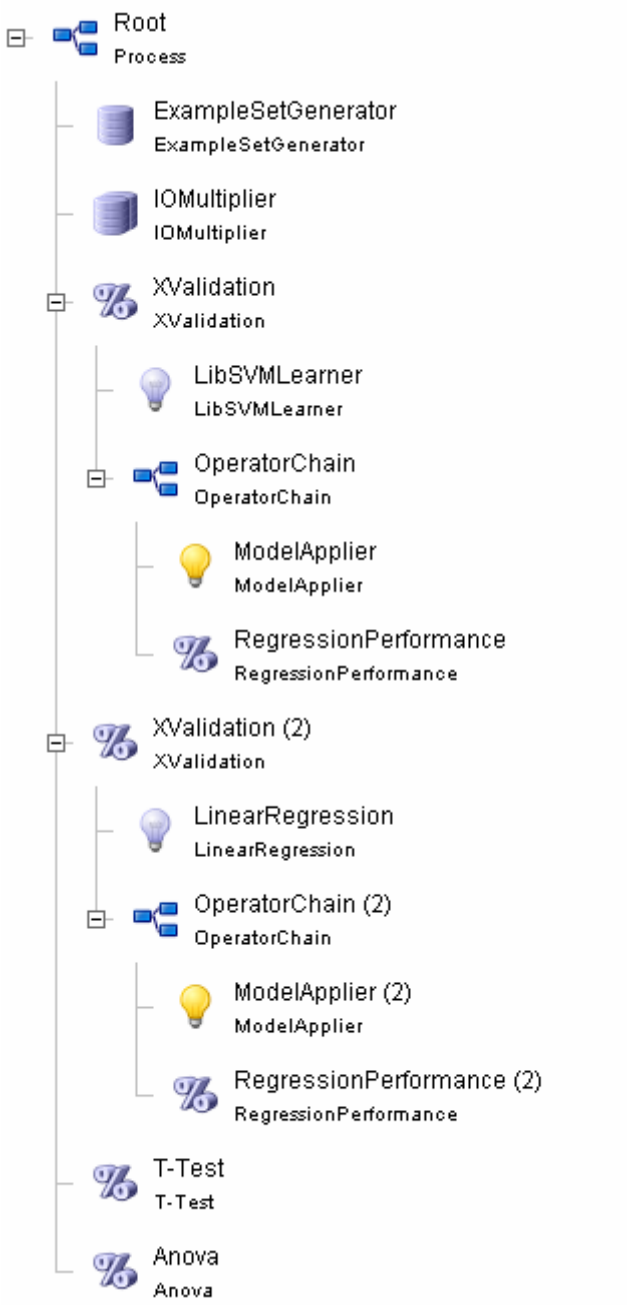

RapidMiner@Daniela-PC (17_WeightingThreshold.xml)
File Edit View Process Tools Help
PerformanceVector
Criterion Selector
classification_error
Table View
Plot View
Confusion Matrix (x: true class, y: pred. class, z: counters)
Save...
+- OperatorChain[0] (OperatorChain)
+- ModelApplier[0] (ModelApplier)
+- ClassificationPerformance[0] (ClassificationPerformance)
Oct 24, 2008 8:19:03 AM: [NOTE] XValidation: Breakpoint reached
[1] Root 20 s > [1] GridParameterOptimization 20 s > [1] OperatorChain (2) 20 s
8:19:23 AM

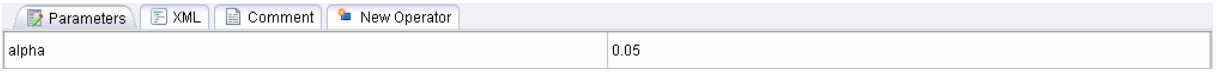


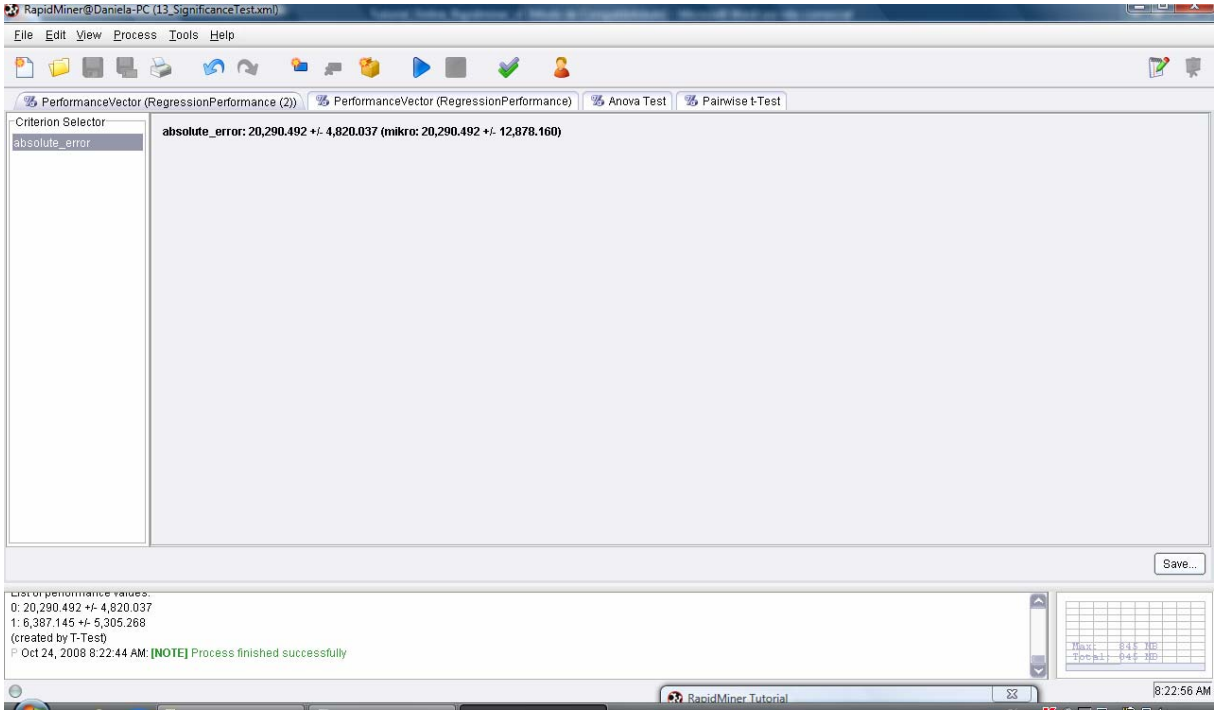
35 – Operadores de validação de desempenho de um *Learner*

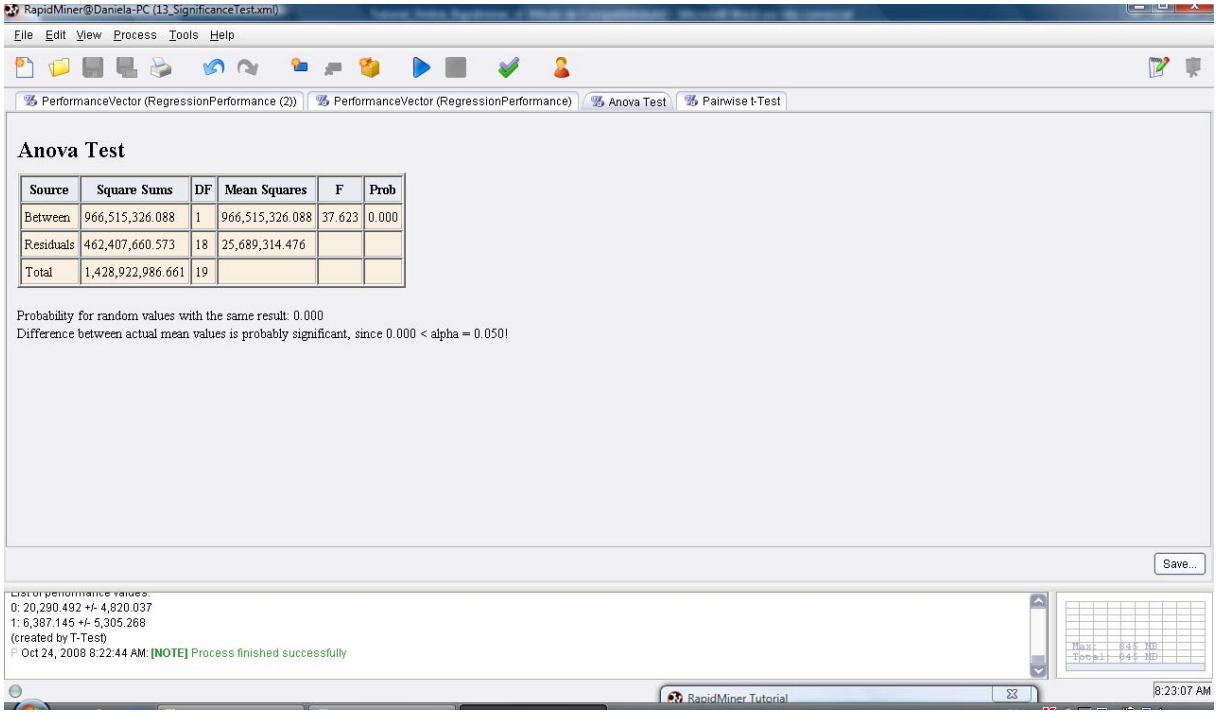
Muitos operadores RapidMiner podem ser usados para estimar o desempenho de um *Learner*, um passo prévio, ou uma característica espaço em um ou vários conjuntos de dados. O resultado da validação destes operadores é um vetor de desempenho que coleta os valores utilizando um conjunto de critérios de desempenho. Para cada critério, o valor médio e desvio-padrão são dados no diz respeito ao parâmetro de otimização operadores.

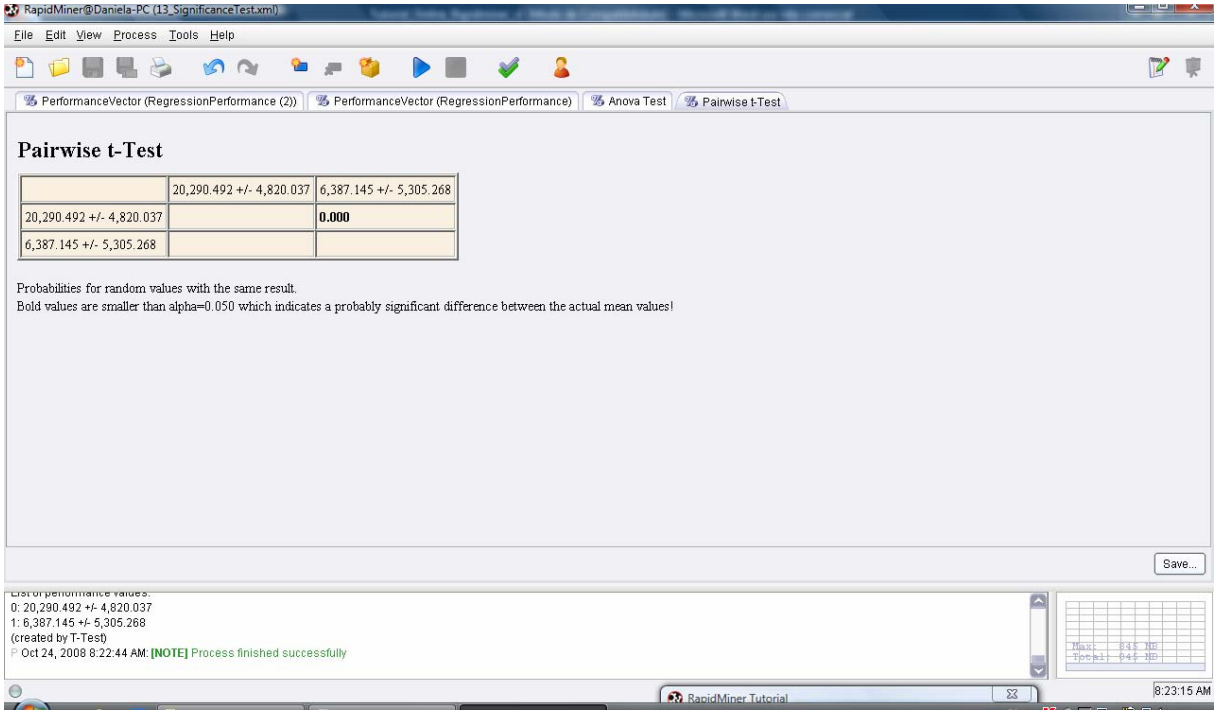
A questão é saber como esses vetores podem ter seus desempenhos comparados? Testes de estatística de significância como *ANOVA* ou *pairwise t-testes* podem ser usados para calcular a probabilidade de que os valores médios reais sejam diferentes.


Root
Process
ExampleSetGenerator
ExampleSetGenerator
IOMultiplier
IOMultiplier
XValidation
XValidation
LibSVMLearner
LibSVMLearner
OperatorChain
OperatorChain
ModelApplier
ModelApplier
RegressionPerformance
RegressionPerformance
XValidation (2)
XValidation
LinearRegression
LinearRegression
OperatorChain (2)
OperatorChain
ModelApplier (2)
ModelApplier
RegressionPerformance (2)
RegressionPerformance
T-Test
T-Test
Anova
Anova


Parameters
XML
Comment
New Operator
alpha
0.05


RapidMiner@Daniela-PC (13_SignificanceTest.xml)
File Edit View Process Tools Help
PerformanceVector (RegressionPerformance (2))
PerformanceVector (RegressionPerformance)
Anova Test
Pairwise t-Test
Criterion Selector
absolute_error
absolute_error: 20,290.492 +/- 4,820.037 (mikro: 20,290.492 +/- 12,878.160)
Save...
0: 20,290.492 +/- 4,820.037
1: 6,387.145 +/- 5,305.268
(created by T-Test)
Oct 24, 2008 8:22:44 AM: [NOTE] Process finished successfully
RapidMiner Tutorial
8:22:56 AM


RapidMiner@Daniela-PC (13_SignificanceTest.xml)
File Edit View Process Tools Help
PerformanceVector (RegressionPerformance (2))
PerformanceVector (RegressionPerformance)
Anova Test
Pairwise t-Test
Anova Test
Source | Square Sums | DF | Mean Squares | F | Prob
Between | 966,515,326.088 | 1 | 966,515,326.088 | 37.623 | 0.000
Residuals | 462,407,660.573 | 18 | 25,689,314.476 | |
Total | 1,428,922,986.661 | 19 | | |
Probability for random values with the same result: 0.000
Difference between actual mean values is probably significant, since 0.000 < alpha = 0.050!
Save...
0: 20,290.492 +/- 4,820.037
1: 6,387.145 +/- 5,305.268
(created by T-Test)
Oct 24, 2008 8:22:44 AM: [NOTE] Process finished successfully
RapidMiner Tutorial
8:23:07 AM



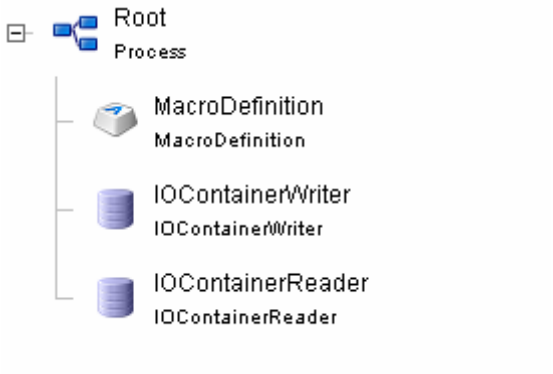
Presumindo que você tenha atingido o desempenho de vários vetores e queira compará-los. Nesta experiência, utilizaremos o mesmo conjunto de dados para ambas as validações cruzadas (daí o *IOMultiplier*) e faremos uma estimativa de desempenho de um sistema linear e de uma aprendizagem baseada *RBF SVM*.

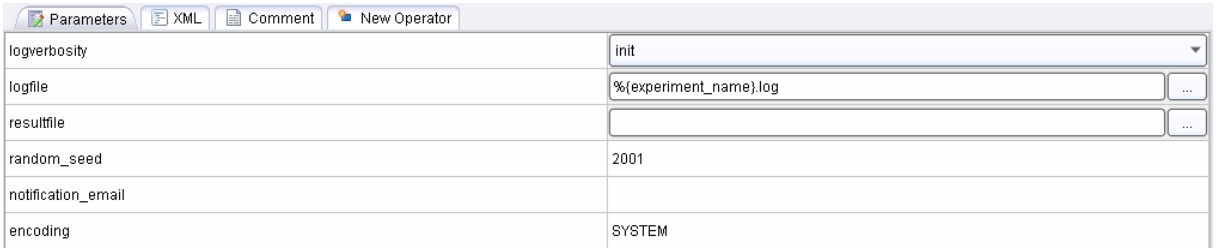
Executar a experiência e comparar os resultados: as probabilidades de uma diferença significativa são iguais, uma vez que só foram criados dois vetores de desempenho. Neste caso, SVM é provavelmente melhor adaptado para o conjunto de dados a mão desde que os valores médios reais sejam provavelmente diferentes.

Por favor, note que os vetores de desempenho como todos os outros objetos que possam ser transmitidos pelo operador RapidMiner podem ser escritos e carregados de um arquivo.

36 – Criação de arquivo de Log a partir da experiência predefinida na macro

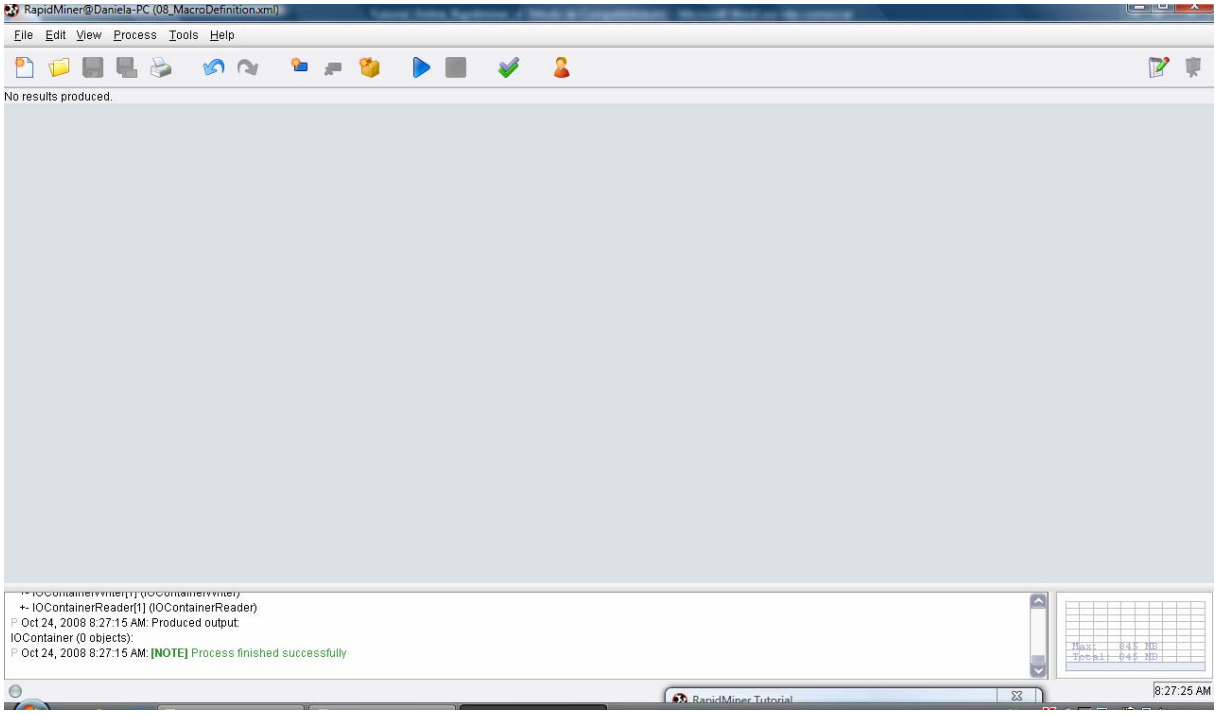
Nesta experiência, várias macros são utilizadas. O operador utiliza a experiência raiz predefinida na macro (% {experiment_name}, a fim de criar um arquivo de log com o mesmo nome como arquivo base de experiência arquivo. O operador MacroDefinition é então usado para definir uma macro para um arquivo de entrada e saída e que é usado em vários lugares do experimento (neste exemplo de apenas duas vezes). A macro é definida no parâmetro da lista "macros" deste operador.





Os usuários podem definir macros arbitrariamente para além das macros predefinidas. Todas as macros são fechadas com % (e) durante o uso.

Para além destas macros de alto nível, existem várias extensões predefinidas como parâmetro igual a a% (substituído pelo número de vezes que o operador era chamado) ou% t (o tempo atual). Por favor, consulte o tutorial escrito para obter mais informações sobre este tópico.



Parabéns!

Você terminou o tutorial RapidMiner online. Você deve ser capaz de executar muitos dos possíveis processos. Agora, como você sabe os mais importantes *building block* de possíveis processos de definições data mining. Evidentemente esses *block* podem ser arbitrariamente encaixados no RapidMiner desde que os seus tipos de entrada e de saída se encaixem. Para uma referência de todos os operadores, por favor, consulte o Tutorial RapidMiner. Verifique também os outros processos de configurações que podem ser encontrados na amostra do diretório RapidMiner.

Acrescentamos muitos passos pré conhecidos de aprendizagem e operadores para RapidMiner. A maior parte dos formatos de dados também podem ser tratados. Se você precisar adaptar o RapidMiner você deve ler o capítulo do tutorial RapidMiner que descreve a criação de operadores, bem como o mecanismo de extensão. O RapidMiner pode facilmente ser estendido. Divirta-se!