UNIVERSIDADE DE BRASÍLIA
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
CURSO DE MESTRADO EM INFORMÁTICA

Mineração de Dados em Base de Germoplasma

GILBERTO DE OLIVEIRA HIRAGI

Dissertação submetida à avaliação
como requisito parcial para a obtenção do grau de
Mestre em Informática

Prof. Dr. Marcelo Ladeira
Orientador

Brasília, Distrito Federal

Março de 2008

CIP - Catalogação na Publicação

Hiragi, Gilberto de Oliveira

Mineração de Dados em Base de Germoplasma / Gilberto de Oliveira Hiragi. - Brasília: CIC da UnB, 2008.

108p.: il.

Dissertação (mestrado) – Universidade de Brasília. Programa de Mestrado em Informática, Brasília, BR – DF, 2008. Orientador: Ladeira, Marcelo.

1. Base de germoplasma. 2. Mineração de dados. 3. SIBRARGEN. 4. Metodologia de mineração de dados 5. CRISP/DM. 6. HaDog. I. Ladeira, Marcelo.

UNIVERSIDADE DE BRASÍLIA
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
CURSO DE MESTRADO EM INFORMÁTICA

Mineração de Dados em Base de Germoplasma

GILBERTO DE OLIVEIRA HIRAGI

Dissertação submetida à avaliação
como requisito parcial para a obtenção do grau de
Mestre em Informática

Prof. Dr. Marcelo Ladeira (Orientador)
CIC/UnB- Universidade de Brasília

Prof. Dr. Paulo Martins Engel
II/UFRS – Universidade Federal do Rio Grande do Sul

Prof. Dr. Marcos Mota do Carmo Costa
Embrapa – Empresa Brasileira de Pesquisa Agropecuária

Alba Cristina Magalhães Alves de Melo
Coordenadora do Mestrado de Informática

Brasília, Distrito Federal

Brasília, 25 de fevereiro de 2008

Resumo

Os bancos de germoplasma do SIBRARGEN (Sistema Brasileiro de Informações em Recursos Genéticos) funcionam como um grande catálogo das espécies vegetais e de seus acessos (tipos característicos dentro de um grupo ou variabilidades dentro da espécie), contendo mais de 100 mil acessos catalogados. Esses bancos incluem a identificação do acesso (passaporte), descrição dos aspectos genótipos (caracterização) e descrição dos aspectos fenótipos (avaliação) e permitem aos pesquisadores dessa área realizarem consultas SQL mas recuperando apenas os dados armazenados, resultantes da resolução das expressões booleanas utilizadas como critérios de busca. Essas consultas não facilitam a descoberta de novos conhecimentos ou a construção de modelos de previsão ou descrição.

Essa pesquisa propõe uma metodologia de mineração de dados, derivada do modelo de referência CRISP/DM, que auxilie a exploração dessas bases de dados por pesquisadores não vinculados à área de informática (por exemplo, biólogos ou agrônomos) visando facilitar a realização de tarefas previstas nas seguintes fases do CRISP/DM: entendimento do negócio, compreensão dos dados, preparação de dados, modelagem, avaliação dos modelos gerados e colocação em uso. Para materializar a metodologia proposta e automatizar a sua utilização por parte de não-informatas, foi implementada a ferramenta HaDog (Hiragi Approach for Data Mining of Germoplasm). HaDog foi implementada utilizando a linguagem Java, banco de dados Oracle® versão 10g release 2 e é acessível através de uma interface Web, disponível aos pesquisadores credenciados para acesso ao SIBRARGEN. A metodologia de mineração de germoplasma proposta foi avaliada de forma experimental através de dois estudos de casos conduzidos com o apoio de pesquisadores da Embrapa Recursos Genéticos e Biotecnologia: determinação de acessos representativos de uma espécie ou grupo de espécies e proposição de coletas direcionadas, ambos problemas típicos de interesse do curador (pesquisador responsável pelo banco de germoplasma de uma espécie). Essa avaliação experimental mostrou que é possível introduzir os especialistas na área na utilização de técnicas de mineração de dados na base de germoplasma sem requerem que eles se envolvam em atividades de programação. Os resultados experimentais obtidos até o momento demonstram que o HaDog pode se constituir em um importante facilitador para a mineração das bases do SIBRARGEN, visando, principalmente, a descoberta de novos conhecimentos pelos especialistas.

PALAVRAS-CHAVES: Base de germoplasma, mineração de dados, SIBRARGEN, metodologia de mineração, CRISP/DM, HaDog.

Abstract

The banks of germplasm of the SIBRARGEN (Brazilian Information System in Genetic Resources) function as a great catalogue of the vegetal species and of its accesses (characteristic types inside of a group or variabilities inside of the species), contend more than 100 thousand catalogued accesses. These banks include the identification of the access (passport), description of the genotypes aspects (characterization) and phenotype description (evaluation) and allow researchers of this area to carry through SQL queries but recouping only the stored data, resultant of the resolution of the used boolean expressions as criteria search. These queries don't facilitate to the discovery of new knowledge or the construction of forecast models or description. This research considers a data mining methodology, derived from the model of reference CRISP/DM, that assists the exploration of these databases for researchers tied with the computer science area (for example, biologists or agronomists) aiming to facilitate the accomplishment of tasks foreseen in the following phases of the CRISP/DM: business understanding, data understanding, data preparation, modeling, evaluation of the generated models and deployment. To materialize the methodology proposal and to automatize its use by people who aren't of the computer science area, the HaDog tool was implemented (Hiragi Approach of Data Mining of Germplasm). HaDog was implemented using the Java language, database Oracle® version 10g release 2 and is accessible through a Web interface, available to the credential researchers for access to the SIBRARGEN. The methodology of mining of germplasm proposal was evaluated of experimental form through two studies of cases lead with the support of researchers of the Embrapa (Genetic Resources and Biotechnology: determination of representative accesses of a species or group of species and proposal of directed collections, both typical problems of interest of the custodian (responsible researcher for the Bank of germplasm of a species). This experimental evaluation showed that it is possible to introduce the specialists in the area in the use of techniques of mining of data in the base of germplasm without require that they become involved themselves in activities of programming. The experimental results obtained so far show that HaDog can be a major facilitator for the mining of foundations of SIBRARGEN, targeting mainly, the discovery of new knowledge by specialists.

Key words: base of germplasm, data mining, SIBRARGEN, mining methodology, CRISP/DM, HaDog.

Navegar é preciso, minerar não é preciso!
Marcelo Ladeira

Agradecimentos

Agradeço aos meus familiares pelo incentivo e apoio constante em toda trajetória de minha vida. Devo a vocês toda a minha gratidão.

Ao meu orientador, Marcelo Ladeira, que me deu a honra e oportunidade de execução desta pesquisa, assim como soube dosar as cobranças e elogios ao longo desta caminhada que teve altos e baixos.

A César de Oliveira Hiragi pela idéias de *Layout* da aplicação implementada e também pelas contribuições no *framework tau*.

A Emerson Lopes Machado pela boa vontade de continuar as reuniões durante o merecido descanso do meu orientador.

A todos aqueles que buscam formação acadêmica apesar das dificuldades do caminho. Que seja dificuldade financeira, dificuldade de tempo, dificuldade de espaço e localização. A todos que o caminho é mais difícil, que tenham êxito.
A nós.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1 Introdução

Este capítulo apresenta a especificação geral do problema abordado, sua relevância e as áreas de pesquisas relacionadas. A Seção 1.1 introduz o problema a ser abordado. A Seção 1.2 revela o principal objetivo desta pesquisa. A Seção 1.3 expõe as áreas de pesquisa relacionadas e a Seção 1.4 apresenta a importância do tema e a contribuição científica esperada.

1.1 Definição do Problema

Hoje existe uma base de dados operacional em funcionamento para a área de recursos genéticos sediada na Embrapa Recursos Genéticos e Biotecnologia. Este repositório central armazena dados sobre passaporte (identificação), caracterização (descrição dos aspectos genéticos), avaliação (descrição dos aspectos fenótipos), conservação e intercâmbio de acessos (tipos característicos dentro de um grupo ou variabilidades dentro da espécie). Esta base é alimentada por equipes especializadas da Embrapa e parceiros pré-definidos. Após o registro das informações o sistema atual fornece os dados necessários para o controle do germoplasma, seu histórico, sua localização, suas características, sua disponibilidade e outras formas de controle operacional do material genético vegetal. É possível fazer consultas e emitir relatórios pré-determinados através de interfaces amigáveis para ferramentas de software que utilizam instruções SQL de consulta para recuperar os dados armazenados. Estas consultas e relatórios são limitados a filtros baseados em expressões booleanas. Apesar do grande conjunto de tarefas que são executadas pelo sistema, o potencial de exploração dos dados pode ser incrementado, pois hoje se limita a estratégias diretas tanto de armazenamento, quanto de extração dos dados.

Um outro aspecto importante e que tem impacto sobre a mineração de dados é a forma de armazenamento das informações neste sistema. Por ser um repositório de informações sobre espécies diferentes, os atributos que descrevem cada espécie são específicos. A forma encontrada para deixar o sistema genérico para a parte de caracterização e avaliação, funcionando parametrizado para cada espécie, foi constituir um módulo que permite a alimentação do sistema com metadados, os quais descrevem os atributos e suas peculiaridades para cada espécie. De posse destes metadados o sistema poderá receber os dados de caracterização e avaliação propriamente ditos sobre os acessos da espécie. Em termos de mineração de dados deve-se considerar este aspecto na fase de preparação dos dados, sendo necessário levar em consideração os metadados no momento de extrair os dados. O termo linearização é utilizado para designar o processo de extrair dados das tabelas do SIBRARGEN levando em consideração, se necessário, os metadados.

É grande a gama de problemas cujas soluções podem ser potencialmente obtidas a partir dos dados nas bases de germoplasma. Uma primeira aproximação para solução destes problemas pode vir dos dados armazenados no SIBRARGEN. Como qualquer aproximação contribui para melhorar o entendimento de um problema em bases de germoplasma, a mineração de dados, ao fornecer esta aproximação, pode ser bastante útil para o especialista, contribuindo para reduzir custos e tempo de resposta para as pesquisas nesta área.

Os especialistas envolvidos com este sistema de informações estão distribuídos geograficamente, desta forma é necessário fornecer um ferramental que atenda a este requisito. Escolhemos a Web como meio de distribuição do modulo de mineração de dados do SIBRARGEN.

1.2 Objetivo

O objetivo principal deste trabalho é aplicar uma metodologia que contemple a mineração de dados nas suas principais fases, utilizada em um primeiro momento sobre bases de germoplasma no que tange aos dados de passaporte, caracterização e avaliação, com o objetivo de facilitar aos especialistas desta área, leigos em informática (por exemplo, biólogos e agrônomos), aplicarem técnicas de mineração aos dados armazenados no SIBRARGEN.

Devemos levar em conta nesta metodologia não apenas o caráter da explicação minuciosa, detalhista e rigorosa que envolve o trabalho de minerar dados, mas complementarmente prover um ferramental Web que seja intuitivo e de fácil uso por leigos em informática.

1.3 Áreas de Pesquisas Relacionadas

A mineração de dados constitui uma área de pesquisa ampla que utiliza técnicas de aprendizagem de máquina e estatística aplicadas à banco de dados para a construção de modelos e aplicações sobre bases de dados. Ela também pode ser vista como um dos principais passos do processo de extração de conhecimento em bases de dados (KDD - *Knowledge Discovery in Databases*) [FAYYAD, 1997]. Existem diversas metodologias para mineração de dados tais como a metodologia de KKD proposta em [FAYYAD, 1997], mas nessa pesquisa será utilizada apenas a denominada CRISP/DM [SPSS, 1999] por ser a mais difundida ao nível mundial.

O CRISP/DM é uma metodologia de mineração de dados baseada em um modelo hierárquico de processos. Este modelo consiste em um conjunto de tarefas descritas em quatro níveis de abstração: do geral para o específico temos a fase, a tarefa genérica, a tarefa especializada e o processo [CHAPMAN et. al, 2000]. Essa metodologia pode ser aplicada a qualquer domínio do conhecimento no qual se deseje minerar. No entanto, nessa pesquisa, somente será considerado o domínio de base de germoplasmas.

A Embrapa Recursos Genéticos e Biotecnologia inclue entre os seus métodos de conservação de recursos genéticos o manejo de um Banco de Germoplasma de Sementes ou Coleção de Base de Sementes (Colbase) que, no âmbito nacional, faz parte da Curadoria de Germoplasma da Embrapa, juntamente com os Bancos Ativos (BAG) [FAIAD et al., 1998].

As atividades de caracterização e avaliação do germoplasma conservado são essenciais para o seu uso nos programas de melhoramento. Tais ações começam com a correta identificação e classificação na espécie, registro do acesso, caracterização biológica por meio de descritores, e avaliação preliminar agronômica e zootécnica. Uma subseqüente avaliação mais profunda é feita através de comparações com outros cultivares através de

parâmetros conhecidos. A caracterização biológica envolve estudos nos aspectos reprodutivos e a discriminação da variabilidade está aumentando com as novas técnicas de marcadores moleculares. Nesse contexto, a biologia, a agronomia e a zootecnia podem ser consideradas áreas correlatas ao tema central desta pesquisa.

O principal objetivo do SIBRARGEN é armazenar e tornar acessíveis informações para tomada de decisões sobre os recursos genéticos vegetais disponíveis no Brasil para a pesquisa agropecuária. [HIRAGI, et. al, 2001]. No contexto da tomada de decisão, como últimas áreas de pesquisas relacionadas, cita-se a estatística, a pesquisa operacional e a análise de decisão que fornecem o ferramental teórico para a exploração de dados tais como análise de variância, regressão, teste de hipóteses, a teoria de decisão propriamente dita, etc. No caso dessa pesquisa, é utilizado o suporte teórico fornecido por essas áreas do conhecimento na construção e na colocação em uso de modelos obtidos por mineração de dados.

A base de dados em estudo está hospedada em um SGBD Oracle®, utilizamos a forma relacional deste SGBD para armazenamento dos dados. Iremos utilizar como ferramenta de desenvolvimento do protótipo o NetBeans juntamente com a linguagem de programação Java integradas com servidor Web Tomcat. De forma complementar e residual, a área de banco de dados pode ser considerada como relacionada a essa pesquisa, sendo, no entanto, utilizada como ferramenta e não explorada do ponto de vista teórico.

1.4 Importância da Pesquisa

Os recursos genéticos hoje são fonte de pesquisas incessantes e seu valor é incalculável, com grande potencial para pesquisa, ainda não sabemos onde podemos chegar com a combinatória entre os genes disponíveis [LOPES, 2006].

Muitas áreas do conhecimento humano têm se beneficiado da informática. É comum a existência de bases de dados de milhares de registros formadas através de sistemas de informações com propósito operacional.

Na área de recursos genéticos no Brasil de hoje temos a informática sendo aplicada nos seus diversos estágios. Desde o armazenamento de dados em planilhas eletrônicas, passando por sistemas de informações, até pesquisas que envolvem técnicas apuradas de informática. Ainda não temos um sistema de informação estabelecido, mas sim uma tentativa com relativo sucesso de organizar os dados de recursos genéticos vegetais.

A comunidade cientifica que trabalha com estes dados tem alimentado o SIBRARGEN. Este tem servido como base operacional para diversas tarefas burocráticas e de organização da informação. Dizemos que este sistema de informação está no primeiro estágio, quando os objetivos principais são coletar, validar, organizar e disponibilizar dados. Uma forma de consolidar este sistema e suprir a necessidade de uma comunidade acadêmica atuante é partir para um segundo estágio, onde os dados serão usados para o propósito da descoberta de novos conhecimentos.

A comunidade cientifica composta de mestres e doutores que interage com o SIBRARGEN tem potencial e desejo de explorar os dados armazenados. As consultas SQL atualmente disponíveis apresentam os dados armazenados como filtragens e

sumarizações. Porém estes dados têm o potencial de serem minerados, assim encontrando relações diferentes. Estes achados estariam ao alcance da comunidade cientifica da Embrapa e parceiros, podendo entrar diretamente no fluxo da pesquisa, dinamizando e até diminuindo o custo do processo.

Um ferramental para mineração de dados, de fácil utilização e baseado em uma metodologia de mineração de dados aceita mundialmente, que seja integrado com o sistema SIBRARGEN irá possibilitar que o pesquisador em recursos genéticos tenha acesso à área de mineração, explorando os dados, criando seus modelos, avaliando-os e inclusive os colocando em uso, ampliando assim o leque de técnicas que possa utilizar em suas pesquisas.

1.5 Organização desta Dissertação

Esta dissertação está organizada da seguinte forma:

No Capítulo 2 é apresentada a organização atual do sistema. Suas peculiaridades e perspectivas futuras.

No Capítulo 3 é apresentada a metodologia de mineração de dados em bases de germoplasma alinhada ao modelo de referência CRISP/DM. Detalhamos as fases que serão abarcadas nesta pesquisa.

No Capítulo 4 é apresentada a ferramenta HaDog que materializa as seguintes etapas de mineração de dados no domínio de germoplasma: compreensão dos dados, preparação de dados, modelagem, avaliação dos modelos gerados e colocação em uso. Esta ferramenta está integrada com o sistema atual e é coerente com a metodologia apresentada no capítulo anterior.

No Capítulo 5 é apresentado um estudo de caso que envolve encontrar acessos representativos de um determinado grupo (espécie). Aqui desejamos explorar a tarefa de agrupamento. Os estudos de caso validam tanto a metodologia apresentada, quanto o protótipo proposto nos capítulos anteriores.

No Capítulo 6 é apresentado outro estudo de caso onde desejamos direcionar coletas de acessos, ou seja, a partir de dados de latitude e longitude indicar uma provável região para processar novas coletas com bases em características de acessos desejadas. Aqui desejamos explorar uma tarefa básica de mineração, a sumarização.

No Capítulo 7 temos conclusões e trabalhos futuros.

Capítulo 2 Sistema Atual e Perspectivas

Neste capítulo é apresentado o sistema atual. São abordadas características peculiares e soluções propostas para integração com a metodologia de mineração de dados que será proposta. Este capítulo está organizado da seguinte maneira: na Seção 2.1 é apresentado o SIBRARGEN e seus diversos módulos com os dados que são abarcados por cada um. Na Seção 2.2 temos a explicação de como são organizados os dados de caracterização e avaliação, com o uso de metadados. Na Seção 2.3 é apresentada uma solução de linearização dos dados por SQL. Na Seção 2.4 são abordadas as perspectivas em relação à integração do sistema atual com a proposta de mineração de dados.

2.1 O SIBRARGEN

O SIBRARGEN (Sistema Brasileiro de Recursos Genéticos) dispõe do repositório central de informações sobre recursos genéticos. Este sistema de informações foi desenvolvido sobre plataforma Oracle, utilizando SGBD Oracle® e Developer/2000®. Sediado na unidade Embrapa Recursos Genéticos e Biotecnologia em Brasília, este sistema é utilizado por outras unidades da Embrapa e por parceiros distribuídos geograficamente no Brasil. Além do sistema de alimentação de dados, o SIBRARGEN apresenta uma serie de consultas SQL disponíveis em páginas Web desenvolvidas sobre plataforma Java.

Conforme informado em www.cenargen.embrapa.br/recgen/sibrargen/objetivos.html, o objetivo principal do SIBRARGEN é armazenar e tornar acessíveis informações sobre os recursos genéticos vegetais, animais e microbianos disponíveis no Brasil para a pesquisa agropecuária.

Dentre os demais objetivos destacam-se:

- automatizar o fluxo de informação sobre os recursos genéticos;
- estabelecer uma gerência efetiva e eficiente das informações sobre os recursos genéticos;
- fornecer informações para o processo de tomada de decisão nas ações sobre recursos genéticos;
- centralizar o acesso às informações sobre os recursos genéticos disponíveis para pesquisa;
- facilitar a padronização de descritores;
- disponibilizar de forma instantânea, pela Internet, as informações de interesse mais geral da comunidade científica sobre os recursos genéticos;
- contribuir para intensificar o intercâmbio de informação e uso do germoplasma na agropecuária e
- fortalecer a Rede de Conservação de Recursos Genéticos do SNPA.

Notamos ainda que atualmente é incipiente o item relativo ao auxílio ao processo de tomada de decisão. Um dos resultados esperados nessa pesquisa é melhorar o processo que subsidie este item.

As consultas hoje realizadas no sistema utilizam instruções SQL, sendo, portanto, baseadas em expressões booleanas. Estas consultas atendem aos objetivos mais operacionais, envolvendo gestão da informação, mas não ao objetivo relacionado ao apoio à tomada de decisão. Também são disponibilizadas algumas formas de sumarização, porém é possível incrementar a extração de informações, assim como a descoberta de novas relações através do uso de técnicas de mineração de dados. Facilitar a tarefa de descoberta de conhecimento a partir dos dados do SIBRARGEN, constitui um dos objetivos desta pesquisa.

2.1.1 Módulos do SIBRARGEN

O SIBRARGEN trabalha atualmente com dados de espécies vegetais. Um objeto neste domínio é chamado de acesso e representa um indivíduo para o sistema. Uma espécie pode ter muitas variedades chegando a milhares de acessos por espécie. Os dados são organizados e classificados segundo os módulos descritos abaixo. Para cada módulo são indicados quais são os atributos comuns.

- *Passaporte*: são dados que individualizam o acesso, sua origem e forma básica. É o conjunto de atributos que formam a identidade do acesso dentro do sistema. É o mínimo de informação requerida para o acesso ser catalogado no sistema. Atributos comuns: código do acesso, taxonomia, origem e denominações.

- *Caracterização*: são dados que descrevem o genótipo do acesso, ou seja, descrevem as características perenes. Atributos comuns: dependem da espécie que será analisada, por exemplo, para a espécie *Manihot esculenta Crantz* (mandioca) temos tipo de superfície da película da raiz, cor da película da raiz, hábito de ramificação, cor do broto terminal, entre outras.

- *Avaliação*: são dados que descrevem o fenótipo do acesso, ou seja, descrevem características que variam conforme o ambiente no qual o acesso esta contido. Atributos comuns: dependem da espécie que será analisada, por exemplo, para a espécie *Manihot esculenta Crantz* temos produção de raiz em kg/ha, vigor inicial, peso da parte aérea da planta, comprimento médio da raiz, número de estacas comerciais por planta, entre outras.

- *Conservação*: são dados sobre o armazenamento do acesso para fins de conservação a médio e longo prazo. Atributos comuns: localização, quantidade armazenada, período de armazenamento, forma de armazenamento, entre outros.

- *Intercâmbio*: são dados sobre a importação e exportação de acessos. Atributos comuns: instituição receptora, instituição fornecedora, quantidade, entre outros.

- *Coleta*: são dados sobre a obtenção de acessos através de expedições. Atributos comuns: dados geográficos, quantidade, tipo de material coletado, entre outros.

Estes dados têm no centro o **acesso** que participará como ator principal em cada um destes módulos de informações. Como o acesso é identificado no passaporte, é necessário primeiro alimentar o módulo de passaporte para depois utilizar os outros módulos que sempre se ligarão ao passaporte, estendendo-o.

Em um primeiro momento iremos tratar dos módulos de passaporte, caracterização e avaliação, porém a proposta apresentada pode ser estendida sem prejuízo de semântica para os outros módulos do sistema.

A seguir acompanhe um esquema que mostra a interligação entre os módulos do SIBRARGEN:

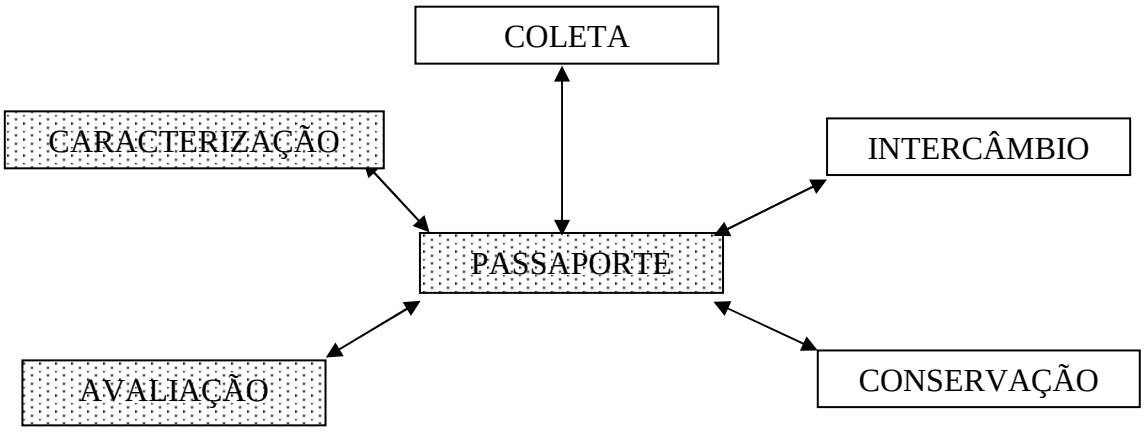


Figura 2.1: Módulos do SIBRARGEN

Os dados do SIBRARGEN estão armazenados de forma relacional em uma instância do SGBD Oracle®. Alguns destes dados podem ser extraídos seguindo a linha de linearização dos dados, desmembrando as relações de chave primária e chave estrangeira do modelo tradicional, o que é possível nos dados de passaporte, conservação, intercâmbio e coleta.

Os dados de caracterização e avaliação seguem uma outra linha que será detalhada na próxima seção.

2.2 Metadados de Caracterização e Avaliação

Os dados dos módulos de passaporte, conservação, intercâmbio e coleta são fixos, independentemente da espécie em questão. Estes dados, por possuírem uma estrutura mais rígida, tiveram um tratamento tradicional de modelagem e implementação. A extração de dados para mineração ocorre nestes módulos na seguinte seqüência:

1. escolha dos atributos a serem extraídos
2. leitura das relações de chaves do modelo relacional
3. junção das tabelas envolvidas com explosão dos estados dos atributos
4. geração de uma nova tabela

Já nos módulos de caracterização e avaliação temos uma peculiaridade, os atributos não são fixos: estes variam de espécie para espécie. A solução implantada no SIBRARGEN

foi à criação de metadados que descrevem os atributos para cada espécie. Primeiro há uma alimentação dos metadados onde são informados, individualmente para cada espécie, quais são os atributos, seus tipos e seus estados. Em um momento posterior são alimentados os dados propriamente ditos, sempre aderentes a uma espécie cujos metadados já foram alimentados. Acompanhamos no esquema a seguir:

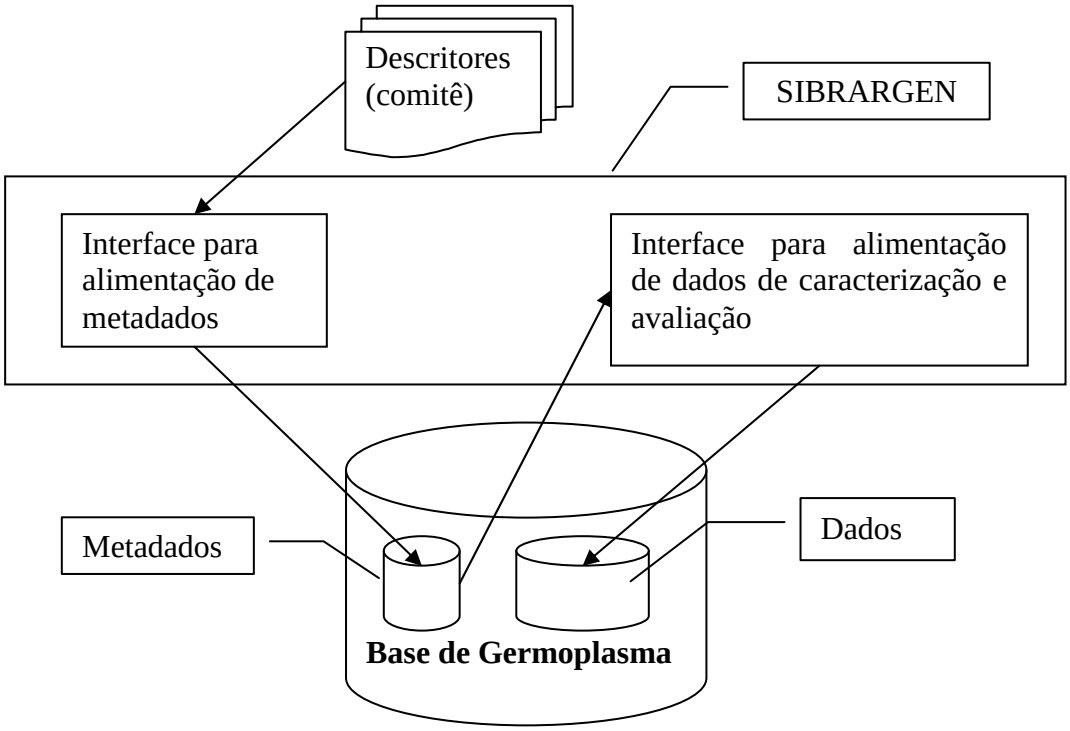


Figura 2.2: Alimentação de Caracterização e Avaliação

A extração de dados para mineração nestes módulos deve levar em conta o passo da leitura e interpretação dos metadados. Por exemplo, para consultar os dados de caracterização de um acesso é necessário:

1. descobrir qual a espécie do qual o acesso faz parte,
2. ler os dados sobre os atributos de caracterização da espécie obtida no passo anterior (ler metadados),
3. ler os dados propriamente ditos de caracterização do acesso,
4. interpretar os dados conforme os metadados, e
5. apresentá-los

Portanto a extração de dados de caracterização e avaliação, além de levar em conta os passos tradicionais (escolha de atributos até geração de nova tabela) já descritos no início dessa seção, deve se preocupar com a sistemática de leitura dupla (metadados e dados), sua interpretação para então fazer as junções necessárias entre as tabelas. Só então partimos para explosão dos estados dos atributos, resultando na linearização.

No Apêndice A encontramos a estrutura dos dados de passaporte, avaliação e caracterização, com a definição dos seus tipos e tamanhos, como são encontrados no SIBRARGEN.

Na próxima seção iremos explorar o processo proposto de linearização por SQL para as tabelas tradicionais, ou seja, aquelas que representam entidades do modelo relacional. Também mostraremos a visão dos metadados acoplados aos módulos de caracterização e avaliação.

2.3 Processo de Linearização por SQL

A base do SIBRARGEN é construída sobre um modelo relacional clássico onde as instâncias são ligadas através de chaves. Como esta base esta pautada na álgebra relacional propomos o uso de instruções SQL para derivar a linearização. Esta abordagem tem alguns benefícios:

- é mais rápida, pois os dados estão disponíveis diretamente no *kernel* de processamento do SGBD.

- é mais segura, pois não estamos retirando dados do SGBD e sim trabalhando todas as etapas dentro do próprio, sendo este o único repositório de informações.

- a linguagem SQL é poderosa o suficiente para tratar condições booleanas e efetuar os cálculos necessários para gerar dados a partir dos metadados.

Para ilustrar o processo de linearização que foi automatizado através de instruções SQL iremos acompanhar o exemplo a seguir.

Neste exemplo, um acesso tem um conjunto de dados de identificação (código único, taxonomia e outros) e pode ter várias formas de conservação. A estrutura simplificada deste subconjunto é apresentada na Tabela 2.1.

Tabela 2.1: Subconjunto de Atributos de Acesso

TABELAS/ATRIBUTOS	OBRIGATÓRIO	TIPO(TAMANHO)
TABELA-ACESSOS		
ACESSOID	NOT NULL	NUMBER(8)
TAXNO	NOT NULL	NUMBER(8)
TABELA-CONSERVACOES		
ACESSOID	NOT NULL	NUMBER(8)
DFORM	NOT NULL	VARCHAR(2)
TABELA-DFORM		
DFORM	NOT NULL	VARCHAR(2)
NDFORM	NOT NULL	VARCHAR(15)
TABELA-TAX		
TAXNO	NOT NULL	NUMBER(8)
TAXON	NOT NULL	VARCHAR2(150)

Desejamos linearizar dados para mineração, ou seja, pretendemos extrair um conjunto de dados com os seguintes atributos:

ACESSOID, Taxonomia (tax.taxon), Forma de Conservação (dform.ndform)

Neste caso cada um dos atributos vem de uma tabela diferente. Ainda devemos levar em consideração a tabela associativa “conservacoes” que servirá de ponte para chegarmos no atributo (dform.ndform).

No algoritmo genérico tivemos que tratar os casos “1-N” e “N-N” levando em conta as chaves primarias e estrangeiras. Consideramos ainda que o modelo relacional estava correto.

Os passos para alcançarmos a linearização são:

1. selecionar os atributos que devem ser extraídos;

2. selecionar as tabelas que possuem os atributos extraídos;

3. descobrir a tabela base, ou seja, aquela que não recebe relacionamentos. Neste exemplo é a tabela “acessos”;

4. relacionar a tabela base com as tabelas associadas via chave estrangeira (processo de descodificação). No exemplo, a tabela base é “acessos” a qual é relacionada à tabela associada “tax”, através do campo “taxno”;

5. relacionar a tabela base com as tabelas associativas pela chave primaria. No exemplo, a tabela base “acessos” com a tabela associativa “conservações”, através do campo “acessoid”;

6. relacionar as tabelas associativas pela chave estrangeira (processo de descodificação). No exemplo, a tabela associativa “dform”, através do campo “dform”, presente nessa tabela e na tabela “conservações”.

Ainda devemos lembrar que no caso de dados de caracterização e avaliação foi necessário fazer a transposição de linhas para colunas, transformando os metadados em estrutura de dados, para então posteriormente seguir no mesmo algoritmo de linearização.

2.4 Integração com Mineração de Dados

O público alvo do sistema são biólogos e agrônomos, normalmente com formação acadêmica na área de pesquisa. A maioria com titulação de mestrado ou doutorado. Este público não é especialista em informática, sendo que este aspecto deve ser levado em conta.

Para o sucesso desta pesquisa é necessário que a integração venha com um conjunto de ferramentas de fácil uso, integrada e similar à forma de trabalho do sistema atual. Para alcançar estes objetivos temos que:

- propor uma metodologia de mineração de dados clara e adequada para uso no sistema SIBRARGEN.
- fornecer ferramentas disponíveis na Internet para materializar a metodologia proposta.
- dispor de um sistema de autenticação de usuários que garanta o acesso seguro e individualizado aos bancos de germoplasma.

- construir ferramentas amigáveis que sejam de fácil utilização por pessoal não especializado em informática e mineração de dados.

A metodologia de mineração de dados proposta será baseada no CRISP/DM e será discutida em detalhes no Capítulo 3.

A ferramenta que materializa a metodologia proposta é um modulo batizado de HaDog (*Hiragi approach for Data mining of Germoplasm*) e será abordado em detalhes no Capítulo 4.

Capítulo 3 Metodologia de Mineração de Dados

Este capítulo apresenta a metodologia de mineração de dados proposta para bases de germoplasma. Esta metodologia está baseada no modelo de referência CRISP/DM. Na Seção 3.1 discutimos o modelo de referência CRISP/DM, contextualizando a metodologia que desejamos implantar. Na Seção 3.2 delimitamos as fases do modelo CRISP/DM que serão tratadas nesta pesquisa. Na Seção 3.3 é apresentada a metodologia proposta, considerando as peculiaridades da área de bancos de germoplasma. Por fim na Seção 3.4 apresentamos um fluxograma com o *workflow* das atividades previstas na metodologia proposta.

3.1 Modelo de Referência CRISP/DM

Segundo Carvalho (2005), mineração de dados é definida como o uso de técnicas automáticas de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano.

Apesar das técnicas utilizadas em mineração de dados serem antigas, estas só recentemente passaram a ser usadas na exploração de dados, devido às questões tecnológicas e de competição empresarial, assim como a existência de ferramentas comerciais.

Houve inegável evolução da atividade de mineração ao longo do tempo, porém esta continua sendo artesanal, no sentindo de que tanto o tratamento dos dados quanto a escolha do que minerar fica muito mais a critério do ser humano. Além de que, na conclusão do trabalho computacional, o resultado gerado deve ser analisado por seres humanos para atestar sua viabilidade e tomar as decisões com base no que foi descoberto.

Uma visão clara e bem definida de que passos seguir a cada fase do processo de mineração de dados é assunto de destaque. A nível mundial, há um esforço de definição de metodologia visando alcançar um nível maior de sistematização tornando o trabalho de mineração menos empírico.

Atualmente a metodologia de mineração de dados mais difundida é a baseada no modelo de referência chamado CRISP/DM (*Cross Industry Process Model for Data Mining*) que define de modo hierárquico as atividades comuns em um processo de mineração de dados [SPSS, 1999]. Este modelo pode ser aplicado a casos específicos, pois sua descrição fala sobre tarefas genéricas que podem ser ajustadas a problemas distintos.

Podemos dizer que o CRISP/DM representa a experiência de casos de sucesso na área de mineração. Proposto por um conjunto de grandes empresas, este modelo incorpora as atividades de mineração com quatro níveis de abstração, desde o modelo mais genérico até instâncias de processos específicos.

As fases do modelo CRISP/DM são as seguintes:

- entendimento do negócio
- compreensão dos dados
- preparação dos dados
- modelagem
- avaliação
- colocação em uso

O processamento que ocorre em uma fase gera produtos ou subsídios que alimentam a fase posterior. Uma possível revisão de fases anteriores faz com que o fluxo seja retroalimentado.

Passamos a descrever cada uma das fases propostas no modelo de referência CRISP/DM.

O entedimento do negócio compreende o entendimento dos objetivos, as metas que se deseja alcançar e também os requerimentos do projeto, sempre na perspectiva do domínio sendo tratado.

A compreensão dos dados é a experimentação com a massa de dados que será minerada, visando maior familiaridade com os dados e descobertas preliminares. Compreende a coleta do conjunto inicial de dados e a familiarização, desde a inspeção até a verificação da qualidade dos dados. É o primeiro contato com os dados que serão minerados.

A preparação dos dados envolve tarefas de informática que auxiliam na construção da base de dados final a partir dos dados brutos iniciais. O resultado desta fase será o conjunto de dados que servirá de subsídio para mineração dos dados. Aqui ocorre a seleção de atributos, o tratamento de valores faltantes, erros nos dados, integração de fontes de dados, formatações, divisão dos dados em, pelo menos, um conjunto de treinamento e um conjunto de avaliação, entre outras. O conjunto de treinamento é utilizado para construir modelos, os quais são avaliados com relação à performance que apresentam no conjunto de avaliação.

A modelagem é a parte que envolve processos de inteligência artificial e estatística de forma mais significativa. Inicialmente devemos escolher a tarefa de mineração de dados a ser usada, sempre com base no domínio de conhecimento e tipos de dados. Então de posse da tarefa (por exemplo, classificar, estimar, descrever ou visualizar) iremos selecionar a ferramenta de inteligência artificial ou estatística que implemente a técnica escolhida. Em muitos casos, neste momento, o problema será resolvido por ferramental de inteligência artificial tais como: redes neurais artificiais, algoritmos genéticos, árvores de decisão, regras de associação, entre outras. É importante para o usuário leigo que o uso deste ferramental seja transparente durante o processo de mineração de dados.

A avaliação é o momento de mensurar a qualidade da mineração de dados realizada, a partir da análise de performance dos modelos obtidos. Também são verificados se os objetivos do negócio foram alcançados. Normalmente, com base nos resultados obtidos na avaliação o processo de mineração é revisado podendo ser retomadas fases anteriores do CRISP/DM.

A colocação em uso é o momento em que a mineração de dados pode ser revertida em ganhos, tanto no processo decisório, como em ações futuras. Em geral consiste na colocação em produção do modelo que apresentou a melhor performance do ponto de vista dos objetivos do negócio. A colocação em uso pode ser vista como utilizar resultados obtidos pela aplicação (a um novo conjunto de dados) do modelo selecionado para apoiar uma tomada de decisão por parte do decisor que o utiliza. Como a performance do modelo pode degradar se as características dos dados mudarem, é necessário estabelecer um plano de acompanhamento da performance e manutenção do modelo, visando correções incrementais no mesmo ou construção de um novo modelo através de uma nova mineração de dados com os dados atuais.

Fazendo um paralelo com uma pesquisa, o CRISP/DM compilou nas suas fases, o desenvolvimento que ocorre em uma pesquisa ou estudo. É necessário contextualizarmos o objeto de interesse (compreensão do domínio), entender este objeto de interesse buscando outras fontes de conhecimento ou uma exploração mais detalhada (compreensão dos dados), ajustar o conjunto de informações obtidas, compilando e adequando as necessidades previstas (preparação dos dados). Desenvolver a solução, modelando cenários, ensaiando com várias alternativas (modelagem). Testar a viabilidade das soluções, escolher entre as alternativas, simplificar o processo (avaliação) e por fim executar, tornar disponível a melhor solução ou a solução encontrada (colocação em uso).

Até por representar bem a forma estruturada de conduzir uma pesquisa, adequada ao processo de mineração de dados, foi escolhido o modelo de referência CRISP/DM para formar um paralelo com a mineração de dados em bases de germoplasma, resultando na metodologia de mineração proposta nesta pesquisa.

3.2 Delimitação da Metodologia Proposta

O modelo de referência CRISP/DM atua sobre todo o processo de mineração de dados. Nesta seção iremos delimitar o escopo da nossa metodologia. Iremos abarcar principalmente as cinco últimas fases quando estivermos falando de apoio por *software*: compreensão dos dados, preparação dos dados, modelagem, avaliação e colocação em uso.

No contexto dessa pesquisa a ferramenta não aborda a fase de entendimento do domínio, visto que o sistema de informação já é estabelecido. Os futuros usuários da metodologia são pesquisadores especializados que já utilizam o sistema SIBRARGEN, já ambientados com o assunto, suas implicações e os dados armazenados.

Apesar do escopo menor em termos de ferramenta a fase de entendimento do negócio deve ser vista como prioritária. É nesta fase que o planejamento nasce, que encontramos os principais objetivos. É nesta fase que determinamos as principais estratégias com base no modo como enxergarmos o domínio tratado.

A compreensão dos dados também não será tratada por completo na ferramenta, visto que estes usuários constituem comitês especializados em produtos agrícolas e estes comitês servem para definir que atributos que irão descrever uma determinada espécie ou produto. Sendo assim os dados disponíveis para mineração são aqueles definidos e já bem conhecidos pelos usuários. Assim apenas as tarefas de coleta do conjunto inicial de

dados e a verificação da qualidade dos mesmos são consideradas nesta fase de compreensão dos dados.

As demais fases que serão abarcadas também exigem uma interação maior com a informática. A idéia é diminuir a complexidade de execução de todas as fases, tornando possível que os usuários atuem na mineração de dados mesmo não sendo especialistas em informática. Para tanto, a metodologia proposta será materializada através da ferramenta HaDog. Assim, as possíveis instâncias de processos em cada uma das fases tratadas serão pautadas por opções do HaDog. Desejamos testar a viabilidade da pesquisa, que poderá ser expandida no futuro. Esta viabilidade abre caminho para que a comunidade científica, foco desta pesquisa, tenha mais uma ferramenta disponível para exploração dos dados.

3.3 Metodologia Proposta

A metodologia proposta irá partir de conhecimentos adquiridos na fase de entendimento do negócio, já a ferramenta estará ligada às atividades práticas da fase de compreensão dos dados. O entendimento conceitual dos dados já faz parte do conhecimento comum dos usuários envolvidos e como comentado na Seção 3.2 não fará parte da ferramenta proposta.

Durante a fase de entendimento do negócio não estaremos trabalhando na linha de tecnologia, mas sim, na parte de conhecimento do domínio. Os relatórios gerados nesta fase constituem a documentação do processo, seus pontos críticos e o claro entendimento dos objetivos.

Uma das contribuições do modelo de referência CRISP/DM é justamente elencar o entendimento do negócio como ponto de partida e então direcionar as outras fases conforme o conhecimento estabelecido.

Na fase de compreensão dos dados estaremos interessados em atuar sobre demandas de obtenção, visualização e prospecção de dados. Neste momento uma compreensão ainda mais clara e concreta dos dados é obtida, pois pela visualização e inspeção é possível fazer um planejamento quanto às necessidades de correções e ainda mensurar com maior clareza o que se tem disponível para mineração.

A fase de preparação dos dados já é totalmente abarcada pela metodologia proposta. Nesta etapa estaremos cuidando dos processos iniciais para adequação dos dados. Esta adequação visa formatar, selecionar, esculpir os dados para que possam ser analisados pelos algoritmos de mineração de dados.

O produto da fase de preparação dos dados é passado à fase de modelagem. Nesta fase temos a experimentação. É necessário que o objetivo da mineração esteja claro, para que o modelo seja criado sob o crivo das ferramentas e técnicas adequadas à solução do problema. Um modelo une um conjunto de dados iniciais e um conjunto de parâmetros para serem processados por um determinado algoritmo. Os modelos fornecem uma idéia inicial e servem principalmente para explicar um contexto ou serem aplicados em um novo conjunto de dados.

Na fase de avaliação estamos considerando os modelos gerados anteriormente. Como são possíveis as construções de vários modelos sobre um mesmo conjunto de dados com

um mesmo fim. Torna-se necessário ter uma forma de mensurar a performance de cada modelo para então optar pela aplicação de um em detrimento de outros. A avaliação é mais objetiva em algumas tarefas de mineração que em outras, visto que em muitos casos a escolha do modelo a ser aplicado passa por uma avaliação subjetiva. Possíveis indicadores de performance do modelo servirão de parâmetro para escolha de qual modelo usar.

Na fase de colocação em uso iremos aplicar o modelo que foi selecionado na fase de avaliação. Esta aplicação deve ocorrer sobre um novo conjunto de dados. Por exemplo, nesta fase iremos munir o modelo gerado com novos dados para então processar uma classificação dos novos elementos. Facilidades de aplicação, extração de regras e visualizações compõe instâncias no nível mais concreto desta fase.

A seguir iremos detalhar cada uma destas fases, exemplificando os níveis mais concretos de cada fase, segundo a hierarquia: tarefas genéricas, tarefas especializadas e instâncias de processo.

3.3.1 Entendimento do Negócio

Nesta fase abordarmos o negócio sob o ponto de vista dos objetivos da mineração de dados. Estaremos preocupados em mapear os objetivos a serem alcançados, os recursos disponíveis para obtenção dos resultados, os riscos envolvidos para então gerar um planejamento da mineração. As principais tarefas genéricas segundo o modelo de referência CRISP/DM são:

- determinar os objetivos de negócio
- descrever o cenário atual
- determinar os objetivos de mineração
- produzir um plano de projeto de mineração

Estas tarefas não serão abordadas na ferramenta, porém são de importância para o sucesso da mineração de dados. Muito do esforço de entendimento do negócio já foi produzido com a introdução do sistema SIBRARGEN, que permitiu a discurssão sobre documentação de recursos genéticos, principalmente na parte vegetal.

É importante salientar que sem o entendimento do negócio e seus objetivos o trabalho em mineração pode ser em vão, visto que pode não atender as verdadeiras demandas do negócio. Desta forma apesar desta fase envolver um nível menor de tecnologia da informação é extremamente necessário ter estes requisitos para dar prosseguimento no projeto de mineração.

Cada uma das tarefas e atividades desta fase compõe um conjunto de exercícios que darão ao analista de mineração o conhecimento necessário para dirigir o projeto de mineração na direção esperada pelo cliente. O balizamento do projeto de mineração se dá com base nos produtos desta fase.

A Figura 3.1 ilustra a cronologia das tarefas ligadas à fase de entendimento do negócio.

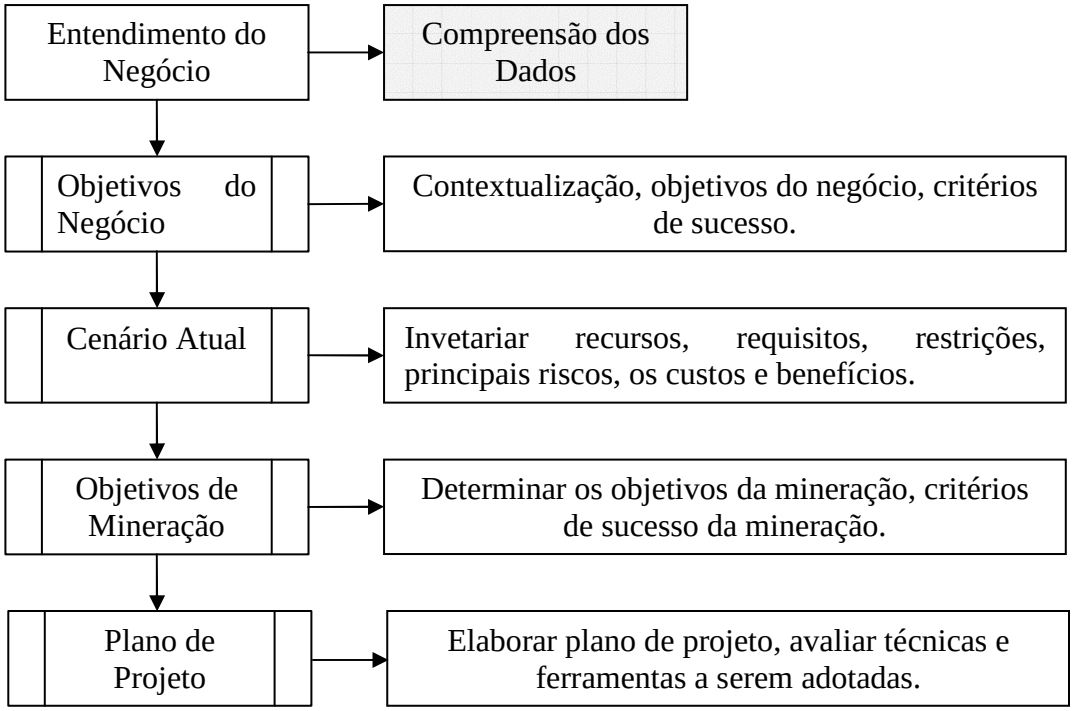


Figura 3.1: Tarefas Genéricas da Fase de Entendimento do Negócio

A seguir iremos detalhar cada uma das tarefas genéricas da fase de entendimento do negócio. Também serão listados alguns exemplos de tarefas específicas. As instâncias de processo referem-se à execução de uma tarefa específica e estas estão ligadas à prática da metodologia proposta.

Objetivos do Negócio

Em qualquer empreendimento é necessário que se tenha foco. Não é possível acompanhar, avaliar ou mesmo saber se esta sendo efetivo caso não se tenha claro o que se pretende.

Nesta tarefa desejamos obter informações sobre quais são os objetivos principais, assim como as restrições do negócio. Visto que as soluções em mineração de dados devem ser balizadas pelos objetivos de negócio e não ferir restrições impostas.

Conhecer o organograma da empresa, as pessoas chave em cada setor, a função de cada setor e suas interações são primordiais para o reconhecimento inicial do que deve ser feito.

Esta tarefa é necessária para que o trabalho de mineração seja feito dentro da perspectiva do negócio, atendendo a demanda verdadeira, pois para o sucesso do negócio é importante que além do sucesso no processo de mineração este esteja alinhado a meta estabelecida.

Nesta tarefa também é importante determinar critérios para avaliar o sucesso da mineração. É importante ter claramente os indicadores de sucesso, assim como os patamares a serem alcançados.

Cenário do Negócio

Nesta tarefa desejamos estar a par do que esta acontecendo em termos do negócio a ser tratado, note que o mais importante é o negócio, a mineração deve estar alinhada com o negócio.

Em um empreendimento real existem processos em produção, restrições de mudanças, contigência de recursos, entre outros. O levantamento destes pontos proporcionará uma visão geral ao minerador do que poderá ser efetivamente implantado.

Como primeiro ponto é importante elencar os recursos existentes: pessoas, *software*, *hardware*, o que estará efetivamente disponível para uso na mineração de dados. Em seguida é necessário listar as necessidades do projeto, quais são as premissas para que o projeto possa ser executado, neste ponto já é possível fazer uma analise entre os recursos existentes e requerimentos básicos do projeto. É importante também levantar as restrições de negócio, visto que algumas soluções podem ser impossíveis de execução.

É importante trabalhar com os riscos, atacando os riscos, prevendo a ocorrência deles é possível arquitetar planos de contigência. Durante um projeto é comum aparecer fatos incomuns e se estes já foram pensados a resposta é mais rápida e menos danosa ao projeto.

Um outro ponto importante é conhecer os termos chave, a nomenclatura utilizada no negócio e na mineração de dados. Esta tarefa tem como resultado a criação de um glossário com os principais termos envolvidos no projeto de mineração, tanto em termos de negócio, quanto de mineração de dados.

Por fim, é possível fazer uma analise de custo e benéficio, ou seja, um relatório elencando os recursos que serão utilizados e o retorno que o projeto dará. Os custos devem estar coerentes com o levatamento do cenário ou balizados nos recursos que serão investidos, já os benefícios devem estar alinhados aos objetivos de negócio.

Objetivos de Mineração

Com a determinação dos objetivos de negócio e a visão clara do cenário atual possamos a nos concentrar em como a mineração de dados poderá ajudar o negócio. Os objetivos da mineração são sempre auxiliares aos objetivos de negócio.

Nesta tarefa queremos transformar os objetivos de negócio em objetivos de mineração. O que é possível fazer diante dos recursos disponíveis para alcançar as metas estabelecidas. Aqui o objetivo é bem mais detalhado e já enfoca a massa de dados persistida nos bancos de dados.

Nesta tarefa também devemos determinar os indicadores que nos darão a noção se as metas foram cumpridas ou não, se o projeto de mineração conseguiu atender as demandas do negócio. Na medida do possível desejamos que os critérios de sucesso sejam objetivos, caso não sejam, é necessário documentar as opiniões subjetivas dos patrocinadores do projeto.

Plano de Projeto

Nesta tarefa iremos produzir um plano de trabalho com o conjunto de passos cronológicos necessários para alcançarmos os objetivos de mineração.

Note que esta tarefa é possível de ser executada após a conclusão das anteriores nesta fase. Com base nos produtos das tarefas anteriores é que iremos construir o plano do projeto.

No plano de projeto devemos ter os estágios de maturação do projeto, especificando o que deve ser feito, a duração de cada uma, os pré-requisitos e até mesmo uma previsão do que pode ser feito em paralelo. O plano de projeto segue a recomendação CRISP/DM de fases e tarefas e deve ser especificada com este formato.

Outro ponto importante é determinar uma metodologia de avaliação do andamento do projeto, fixando pontos de acompanhamento e limites de tempo. Os riscos também devem ser abordados e os possíveis imprevistos devem ser acomodados na agenda do projeto.

Neste plano também devem ser abordadas as técnicas de mineração e possíveis ferramentas que serão utilizadas. Apesar de parecer estranho tratar do assunto bem cedo no processo, isto é necessário, pois as técnicas e ferramentas irão influenciar diretamente as outras fases.

Uma determinada técnica ou ferramenta irá demandar ações específicas (atividades) em cada uma das fases posteriores, portanto a escolha previa possibilitará planejar corretamente a execução das fases de compreensão dos dados, preparação dos dados, modelagem, avaliação e colocação em uso. Por exemplo, na preparação de dados devemos trabalhar os dados para que possa ser utilizado na ferramenta escolhida. O formato, o tipo de dados e até a qualidade dos dados envolvidos irão influenciar, vetar ou se adequar à técnica e ferramenta escolhidos.

3.3.2 Compreensão dos Dados

Esta fase compreende a obtenção do conjunto de dados a ser trabalhado, sua inspeção e mensurações diversas para atestar a qualidade dos dados obtidos. As tarefas genéricas segundo o modelo de referência CRISP/DM nesta fase são:

- coleta dos dados
- descrição dos dados
- exploração ou inspeção dos dados
- mensuração da qualidade dos dados

Destas quatro tarefas genéricas a tarefa de descrição dos dados é desnecessária no contexto da ferramenta, visto que os usuários já conhecem os dados. As outras tarefas genéricas serão abordadas. A Figura 3.2 ilustra as tarefas genéricas ligadas à fase de compreensão dos dados. Note que as tarefas seguem uma linha seqüencial, sendo executadas uma após a outra, pois o resultado de uma tarefa é o subsidio para execução da próxima.

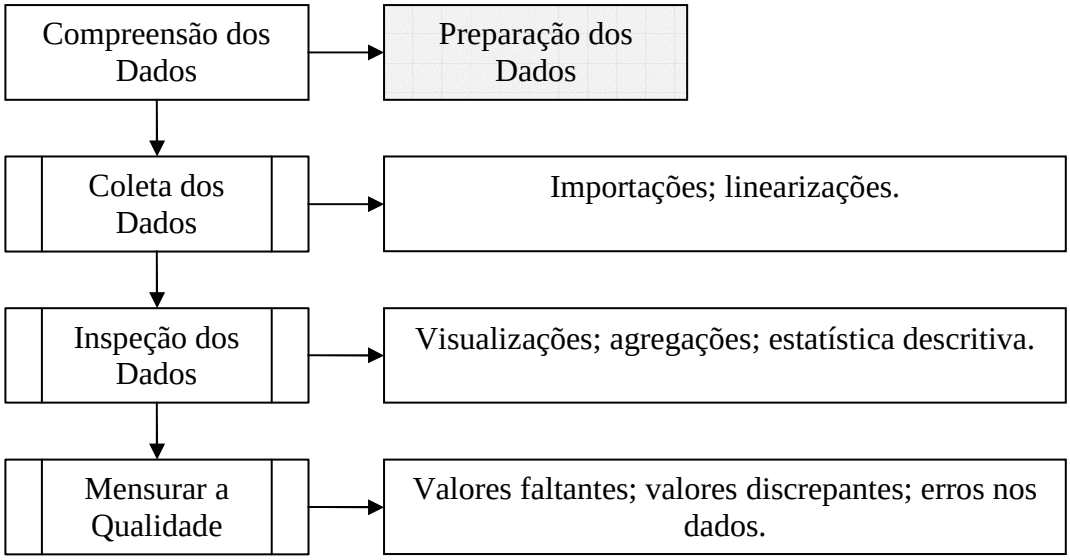


Figura 3.2: Tarefas Genéricas da Fase de Compreensão dos Dados

A seguir iremos detalhar cada uma das tarefas genéricas da fase de compreensão dos dados. Também serão listados alguns exemplos de tarefas específicas. As instâncias de processo referem-se à execução de uma tarefa específica e estas estão ligadas às aplicações em si do projeto de mineração.

Coleta dos Dados

Esta tarefa envolve obter os dados a serem trabalhados no processo de mineração. A obtenção destes dados é o passo inicial para poder extrair as informações buscadas. Muitas vezes estes dados estão disponíveis em várias fontes diferentes. Mesmo dentro de uma base de dados informatizada as informações normalmente estão distribuídos em tabelas distintas.

Alguns exemplos de tarefas específicas dentro desta tarefa genérica são: linearização de bases de dados relacionais e importação de dados. As instâncias de processo serão materializadas no HaDog.

Inspeção dos Dados

Esta tarefa envolve aumentar o conhecimento sobre os dados. Não no sentido conceitual dos dados, mas em relação ao conjunto de dados que foi extraído. O conhecimento da estrutura dos dados e o significado de cada atributo estão inclusos na tarefa "descrição dos dados" que julgamos desnecessária no contexto desta pesquisa. Já conhecer os dados em si, aqueles que foram extraídos para mineração na tarefa anterior é parte importante para aumentar o nível de entendimento dos dados, até para a tomada de decisões mais corretas no projeto de mineração.

Alguns exemplos de tarefas específicas são: visualização dos dados, geração de agregações, analise estatística descritiva e outras tarefas que visam dar uma idéia mais clara dos dados extraídos na etapa anterior.

Mensuração da Qualidade dos Dados

Nesta tarefa iremos analisar os dados para medir casos indesejáveis, que podem influenciar negativamente na mineração de dados. A quantificação dos valores faltantes e os tipos de erros nos dados indicarão se é necessário voltar à etapa anterior de aquisição de dados.

Alguns exemplos de tarefas específicas são: quantificação de valores faltantes, existência de valores discrepantes (*outliers*), se os dados são suficientes para o objetivo de mineração, etc. Todas estas tarefas têm o objetivo de quantificar as falhas para que se possa ter a noção exata de que passos são necessários na fase de preparação dos dados.

3.3.3 Preparação dos Dados

Esta fase reúne tarefas que irão construir o conjunto final de dados que será utilizado na fase de modelagem. Não é necessária a execução de todas as tarefas ligadas à fase de preparação dos dados. A decisão de qual tarefa será ou não executada vem da fase anterior, pois o estudo preliminar dos dados extraídos gera as demandas de transformação que são agrupadas nesta fase. As tarefas possíveis são diversas e vão depender do tipo de necessidade observado. A preparação dos dados deve estar alinhada ao objetivo de mineração; normalmente envolve tarefas de seleção, transformação, formatação e eliminação de dados.

Nessa fase, as tarefas genéricas, segundo o modelo de referência CRISP/DM, são:

- seleção de dados
- limpeza de dados
- construção de dados
- integração de dados
- formatação de dados
- construção de conjuntos de treinamento e teste

Cada uma das tarefas genéricas pode ser mapeada para tarefas específicas, utilizando as mais diversas técnicas nas instâncias de processo. Não é necessário que seja seguida uma ordem pré-determinada de execução de tarefas da fase de preparação de dados. Do conjunto de tarefas propostas somente algumas serão utilizadas em cada projeto de mineração específico. A Figura 3.3 mostra as tarefas genéricas ligadas à fase de preparação de dados.

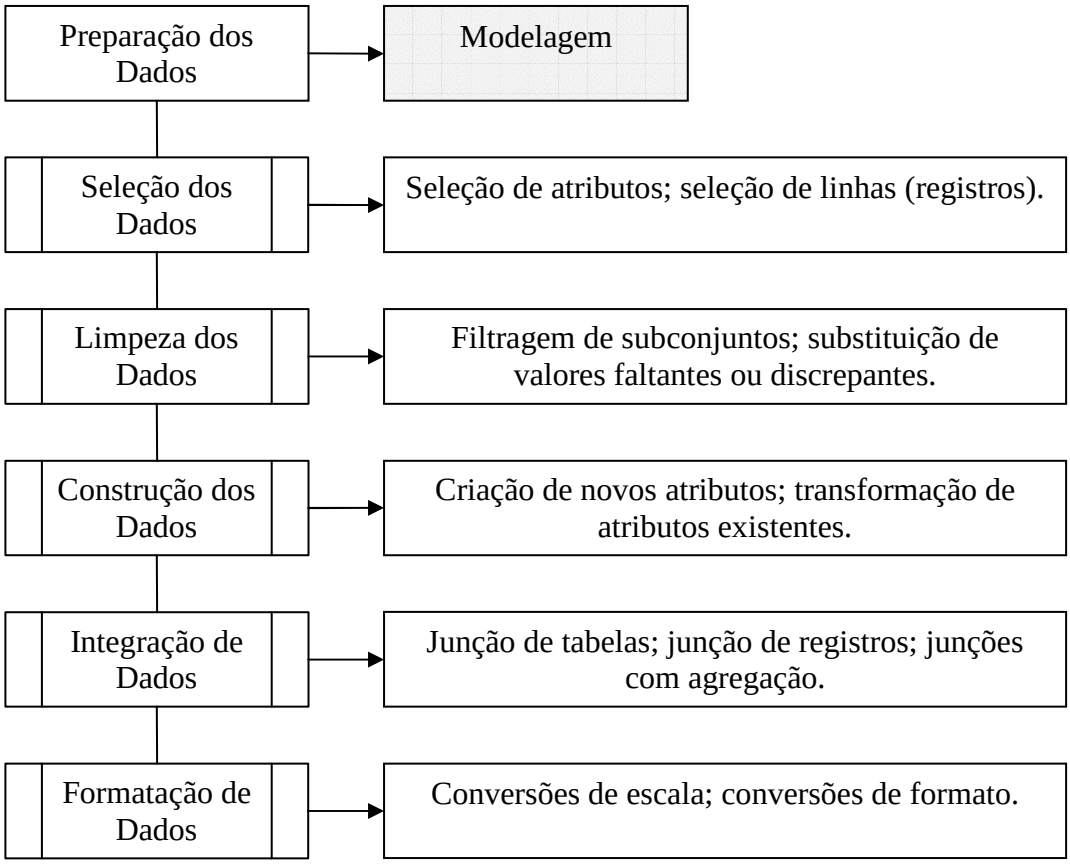


Figura 3.3: Tarefas Genéricas da Fase de Preparação dos Dados

A fase de preparação de dados é usada para adequar os dados à tarefa de mineração que será feita na próxima fase. Tarefas específicas nesta fase são as mais diversas e tratam de transformar os dados obtidos inicialmente em um conjunto com formatação adequada para os algoritmos que serão utilizados na fase de modelagem. A seguir temos o detalhamento das tarefas genéricas.

Seleção dos Dados

É a decisão de quais dados serão analisados no processo de mineração. Nesta tarefa são levadas em conta às restrições técnicas e os tipos de dados adequados, visando adequar tanto atributos, como registros do conjunto de dados.

Alguns exemplos de tarefas específicas são: filtragens de dados através de condições booleanas sobre atributos e valores, eliminação de atributos com tipos de dados inadequados à mineração, etc.

Limpeza dos Dados

Envolve atividades que visam eliminar valores inadequados por serem faltantes, discrepantes ou incoerentes. Esta limpeza pode ser feita pela eliminação dos registros que contém valores inadequados ou a troca dos valores por outros acrescidos com alguma técnica que estime, com algum grau de veracidade, valores coerentes.

Alguns exemplos de tarefas específicas são: tratamento de valores faltantes e tratamento de valores extremos (*outliers*). Estes tratamentos pressupõem a eliminação dos registros com problemas ou a substituição dos valores por outros mais adequados.

Construção dos Dados

São atividades ligadas à criação de novos dados a partir do conjunto de dados existente. Estes novos dados podem ser gerados através da combinação de atributos (também denotada enriquecimento) ou através de cálculos entre atributos ou sobre um atributo específico. Também é possível gerar registros novos a partir de previsões ou preenchimento de lacunas no conjunto de dados inicial.

Algumas tarefas específicas são: criação de campos calculados a partir de atributos do conjunto inicial, geração de novas tabelas a partir de uma consolidação ou transformação de outras tabelas.

Integração de Dados

Esta tarefa representa a junção de conjunto de dados. A união de duas ou mais tabelas para a formação de um novo conjunto de dados. Também pode representar a junção de registros de fontes diferentes, gerando um conjunto de dados único com atributos comuns. A integração de dados é feita principalmente quando temos dados vindos de fontes distintas e que devam ser combinadas para entrarem no processo de modelagem.

Formatação de Dados

Esta tarefa engloba as transformações que visam adequar a forma dos dados para o processo de modelagem. O significado dos dados não é modificado, mas a apresentação (formato) é alterada. Por exemplo, uma data pode ser representada por um número inteiro que indica o número de dias decorridos desde 01/01/1900. Algumas conversões também entram como parte desta tarefa de adequar a forma para os algoritmos de modelagem. Em geral, essas convenções são definidas pela ferramenta de software a ser utilizada na fase de modelagem.

3.3.4 Modelagem

Em geral, como resultado da fase de preparação dos dados, o conjunto de dados é dividido em, pelo menos dois subconjuntos: um para treinamento (geração de modelos) e outro para avaliação (estimar a performance dos modelos gerados). Na fase de modelagem temos a geração de modelos que resultam da aplicação de algoritmos específicos ao conjunto de treinamento. Opcionalmente, pode-se usar um terceiro subconjunto, denotado subconjunto de teste, para realizar as calibrações de parâmetros nos algoritmos, para que se possa chegar a um melhor resultado. Normalmente é possível aplicar mais de um algoritmo para determinado problema de mineração.

Algumas tarefas clássicas de mineração já foram mapeadas e para estes problemas um conjunto de ferramentas já foram testadas e estabelecidas. Este conhecimento permite que o minerador possa, a partir dos objetivos da mineração que deseja realizar, selecionar a qual tarefa pertence o problema específico e prever que tipos de técnicas poderão ser usados e, por fim, direcionar cada fase executando as tarefas que melhor adequem ao conjunto de técnicas selecionadas.

As tarefas genéricas para a modelagem, segundo o modelo CRISP/DM, são:

- selecionar a técnica
- planejar o teste
- criar modelos
- avaliar modelos

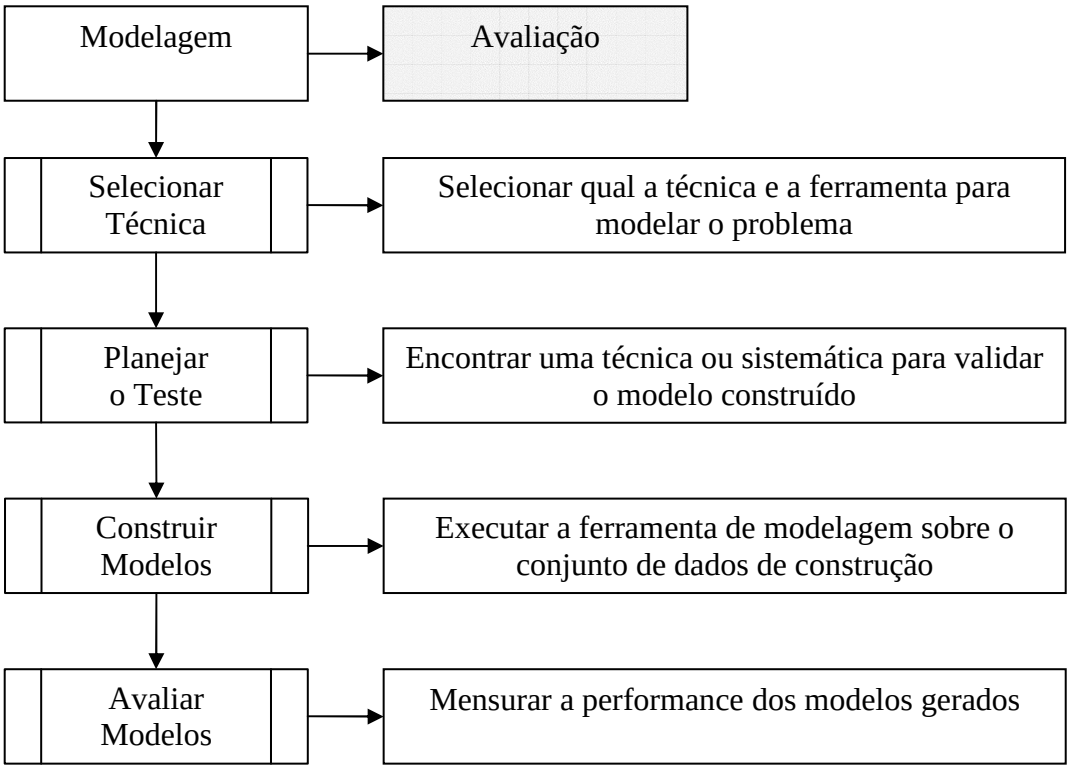


Figura 3.4: Tarefas Genéricas da Fase de Modelagem

Selecionar Técnica

Nesta tarefa deve ser escolhido o tipo de técnica de mineração de dados a será usada. Em muitas tarefas de mineração podem ser usados mais de um tipo de técnica, nestes casos, os processos devem ser feitos para cada técnica.

É importante ressaltar que cada técnica requer certos pré-requisitos, por exemplo, não aceitam valores faltantes, não trabalham com valores categóricos, etc. É importante que os pré-requisitos sejam respeitados para que os resultados sejam coerentes.

Alguns exemplos de tarefas de mineração são: agrupamento, classificação, regras de associação, regressão, sumarização e visualização dentre outras. Para executar determinadas tarefas, na maioria das vezes, estão disponíveis vários algoritmos. Por exemplo, para agrupamento podemos usar o algoritmo K-means ou o algoritmo O-cluster. Nesta tarefa uma instância de processo deve fornecer as informações de que tarefa de mineração será executada, assim como qual algoritmo ou ferramenta será usado para modelar o problema.

Planejar o Teste

Nesta tarefa a preocupação maior é com encontrar uma forma de testar o modelo. Definir uma estratégia ou sistemática que possibilite validar o modelo encontrado. Esta estratégia deve estar alinhada com a técnica escolhida na tarefa anterior.

Um exemplo é separar do conjunto de dados principal um subconjunto de dados de treinamento e avaliação. O primeiro é usado para modelar o problema e construir o modelo. O segundo servirá para validar o modelo criado. Se o modelo requer o ajuste de parâmetros, é usual usarmos um terceiro subconjunto de dados, o subconjunto de testes, para calibrar esses parâmetros. Para determinar os tamanhos desses subconjuntos, se o conjunto de dados original é grande, uma heurística comum é reservar 2/3 dos dados para treinamento e testes, e 1/3 para avaliação. Idealmente esses subconjuntos são obtidos através de uma amostragem aleatória estratificada. Se o conjunto de dados for pequeno, pode-se usar a análise de curva ROC e cálculo da área sob a curva ROC, denominada AUC [Fawcett, 2004]. Este tipo de estratégia é bem comum em tarefas de classificação.

O plano de teste será retomado na última tarefa desta fase, avaliando o modelo criado, a diferença das tarefas genéricas de “planejar o teste” e “avaliar modelos” em relação à fase de Avaliação é que nestas tarefas estamos preocupados somente com o modelo gerado, já na fase de avaliação todos as fases do processo de mineração devem ser consideradas.

Construir Modelos

É a execução da ferramenta de modelagem sobre o conjunto de dados que foi preparado nas fases anteriores. O preparo feito nas fases e tarefas anteriores desta fase é agora consolidado. Esta consolidação é passada para ferramenta de modelagem que irá gerar um modelo.

O resultado desta tarefa é um modelo construído para solucionar um problema. Não constituí uma decisão ou relatório, mas sim um modelo gerado pela ferramenta de modelagem. O modelo é a junção coerente de dados e ferramenta de modelagem. Este modelo necessita ser interpretado inicialmente pelo minerador e posteriormente pelos especialistas de negócio para que possa ser validado.

Avaliar Modelos

Nesta tarefa queremos validar o modelo gerado na etapa anterior com base no plano de teste gerado anteriormente. Como podem ser gerados vários modelos para um mesmo problema de mineração uma sistemática de avaliação dos modelos fornecerá uma medida de acurácia ou qualidade do modelo para que se possam classificar os modelos em um ranking.

Esta avaliação preliminar sobre o modelo pode ser feita pelo minerador, que posteriormente irá discutir os resultados com especialistas de negócio.

Muitas vezes mesmo usando uma técnica específica para gerar o modelo é possível gerar modelos diferentes através da sintonização de parâmetros. Então uma forma de mensurar a qualidade dos modelos gerados é importante para que se possa escolher o modelo que melhor represente um contexto ou modele um problema.

3.3.5 Avaliação

A fase de avaliação é dedicada a revisar o processo de mineração como um todo. De posse do modelo ou modelos gerados é o momento de comparar as performances de cada modelo nas suas várias faces, por exemplo, acuracia, tempo de resposta, resultados. O mais importante é avaliar se o modelo atende os objetivos de mineração colocados no início do processo.

As técnicas de avaliação são diversas e também dependem da tarefa de mineração que esta sendo executada. É importante que a avaliação leve em consideração a opinião do especialista de negócio. Neste momento é importante fazer uma revisão sobre todos os tópicos da mineração, assim como verificar se as expectativas do usuário foram alcançadas até o momento.

Nesta fase a decisão de aplicar o modelo ou não deve ser tomada, assim como determinar os próximos passos, que inclusive podem ser de refazer o processo de mineração parcialmente ou totalmente.

É importante ressaltar que a melhor forma de avaliar um modelo é testá-lo na prática, muitas vezes sobre um conjunto de dados já conhecido ou em uma sistemática pré-determinada.

A seguir temos a lista de tarefas genéricas desta fase segundo o modelo de referência CRISP/DM:

- Avaliar Resultados
- Revisar o Processo
- Determinar os Próximos Passos

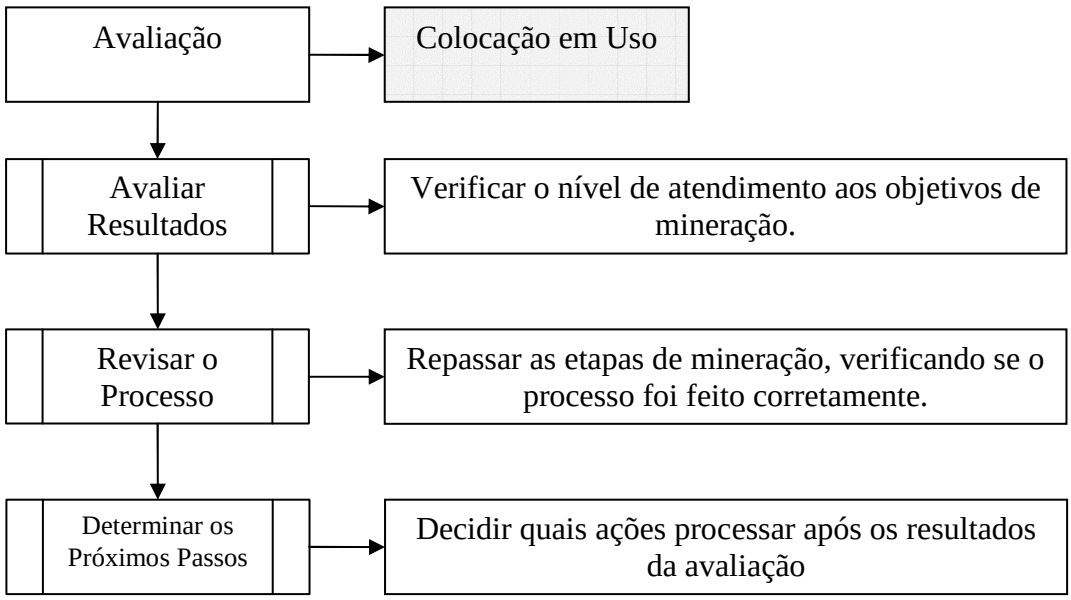


Figura 3.5: Tarefas Genéricas da Fase de Avaliação

Avaliar Resultados

Esta avaliação é diferente da avaliação executada na fase de modelagem. Esta trata do nível de compatibilidade do modelo em relação aos objetivos de negócio. Se existem pontos fracos ou omissos do modelo em relação às expectativas do negócio.

Esta avaliação requer a participação de um especialista que irá determinar em última analise se os resultados da mineração satisfazem as necessidades do negócio. Os resultados devem ser visualizados, analisados e uma decisão sobre sua aprovação deve ser emitida. Direcionamentos para novas pesquisas e estudos também podem resultar da analise.

Revisar o Processo

Nesta tarefa devemos considerar o processo inteiro, repassando as tarefas executadas e se estas estão de acordo com os pré-requisitos das ferramentas e do negócio. Nesta tarefa também deve ser verificado se alguma atividade importante não foi executada.

É uma passada geral no projeto de mineração para que se possa garantir a corretude de execução das tarefas. É possível também encontrar a necessidade de repetição de alguma tarefa para confirmação dos resultados.

Determinar os Próximos Passos

Nesta tarefa será determinado como proceder. Após as avaliações do modelo e a revisão do processo de mineração, já temos subsídios para decidir se o modelo deverá ser colocado em uso, ou se novas iterações de mineração devem ser feitas ou se um novo projeto de mineração é requerido.

As decisões ligadas a esta tarefa são exclusivas do processo de mineração e não decisões de negócio que serão tratadas na fase de colocação em uso. Aqui podemos decidir que a mineração foi satisfatória e que o modelo pode ser disponibilizado em uma aplicação real ou ainda que é necessário promover outras tarefas de mineração para alcançar os objetivos no seu todo.

3.3.6 Colocação em Uso

Esta fase trata de usufruir os resultados da mineração. As formas de aplicação são diversas e irão depender inclusive das técnicas utilizadas. A colocação em uso deve levar em conta o cliente da mineração. Os resultados ou aplicações devem estar em uma forma que o cliente entenda e possa aplicar na solução de seus problemas.

Normalmente a construção do modelo não determina o fim da mineração. Em muitas técnicas um modelo pode ser aplicado sobre novos conjuntos de dados. Uma rotina disponível ao usuário final para que possa aplicar modelos aprovados é uma das formas de colocação em uso de um projeto de mineração.

É importante que seja determinado um plano de manutenção desta fase de colocação em uso, visto que um modelo pode ao longo do tempo perder aderência com a realidade do negócio ou comportamento dos elementos envolvidos no processo. Determinar rotinas de acompanhamento, bem como formas de medição de erros e acertos podem fornecer um indicador importante sobre a necessidade de renovação do projeto de mineração.

É importante determinar como os resultados da mineração serão vistos, por relatório, por aplicação de software, ou outros elementos informatizados. Com base nesta escolha deverão ser tratados problemas diversos de distribuição, segurança e disponibilidade da aplicação.

Sempre que possível é interessante que o processo de mineração seja incorporado também aos processos da empresa, para que esta possa realmente utilizar o projeto de mineração.

A seguir temos uma lista com as tarefas genéricas desta fase segundo o modelo de referência CRISP/DM:

- Planejamento de Implantação
- Planejamento de Manutenção
- Relatório Final
- Revisão do Projeto

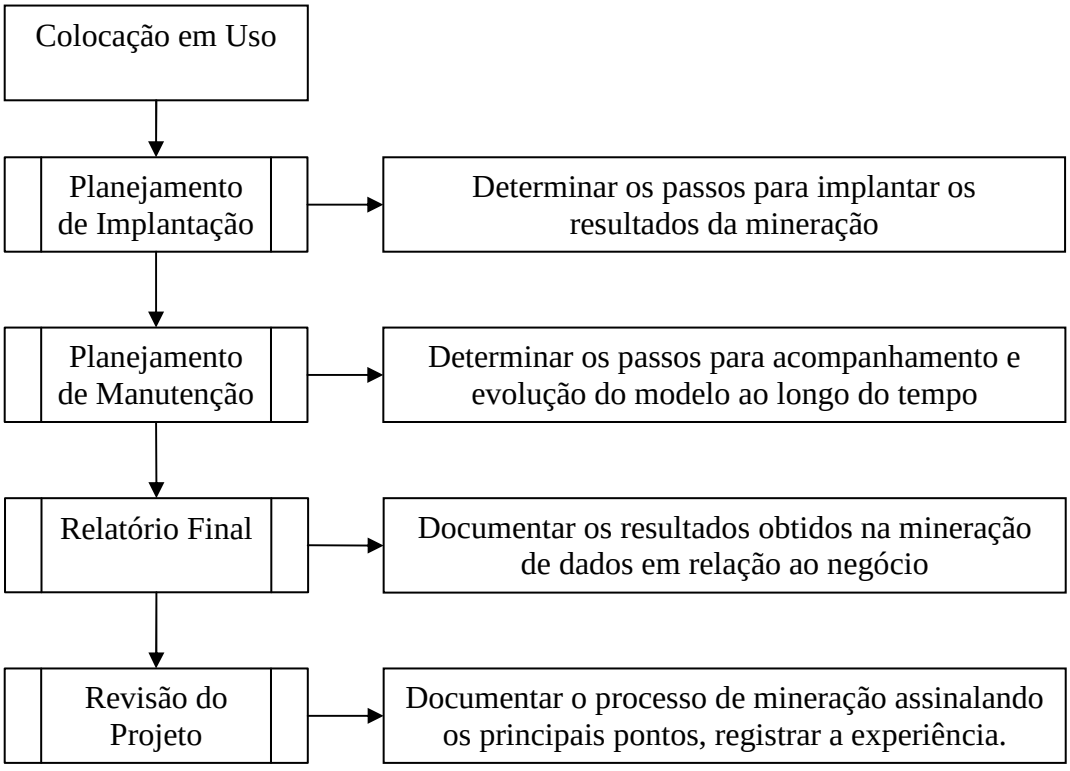


Figura 3.6: Tarefas Genéricas da Fase de Colocação em Uso

Planejamento de Implantação

Nesta tarefa iremos planejar a forma como os resultados da mineração serão aplicados ao domínio de negócio. Os passos para estabelecer a aplicação de um modelo ou resultados de um modelo.

Em termos de informática é possível automatizar algumas atividades de colocação em uso. Aqui não cabe a decisão de colocar em uso, pois isto é feito na fase anterior, mas decidido que será utilizado, os resultados da mineração devem ser aplicados, os pré-requisitos e a forma de aplicação é que são considerados neste planejamento.

Uma lista com os passos para a execução desta fase é um documento importante desta tarefa.

Planejamento de Manutenção

É importante que além de aplicar os resultados da mineração esteja claro uma forma de medir a eficácia da aplicação ao longo do tempo. O contexto e os elementos envolvidos na mineração podem se transformar ao longo do tempo e fazer com que os resultados da mineração não sejam adequados a novo realidade estabelecida.

Este planejamento é um trabalho específico para cada tipo de aplicação, os resultados podem ser usados uma vez para tomada de decisão ou continuamente em um processo integrado. Cada tipo de aplicação requer planos de manutenção diferentes.

Neste planejamento, principalmente de aplicações de longo prazo, decisões de aplicar novos testes ao modelo, refazer alguma fase do projeto de mineração ou mesmo partir para um novo projeto de mineração são listadas como alternativas a respostas menos precisas da aplicação. Note que é importante em um plano de manutenção ter um indicador da qualidade do modelo, se este atende ou não aos requisitos do negócio.

Relatório Final

Esta tarefa documental serve para consolidar as informações sobre os resultados do projeto de mineração. É o documento conclusivo que associa os objetivos iniciais com os resultados alcançados. Pontos fortes e fracos do projeto de mineração são assinalados.

Este relatório final foca principalmente os resultados da mineração. Este é direcionado ao cliente da mineração, demonstrando o trabalho em termos práticos, no que o conhecimento adquirido pode ajudar no negócio.

Revisão do Projeto

Esta tarefa documental serve para registrar a experiência do time no processo de mineração. Muitos elementos são aprendidos durante o processo de mineração. O objetivo aqui é guardar esta experiência para que possa servir de subsidio no futuro para

novos projetos de mineração. Registros de falhas e soluções, aquilo que funcionou conforme esperado e também processos que não deram certo.

Um projeto de mineração como o nome indica tem marcos iniciais e finais. É importante esta ciente que a mineração é um processo recorrente e que a experiência adquirida dentro da empresa, nos processos e pelas pessoas devem ser registrados para que situações similares no futuro possam ser resolvidas com soluções desenvolvidas com esforço anterior.

3.4 Workflow da Metodologia Proposta

A metodologia proposta prevê um processo de mineração iterativo entre as fases e na maioria das vezes seqüencial dentro de uma fase em um modelo ideal, porém na prática uma tarefa deve pode ser executada mais de uma vez e a realimentação do processo deve ser feita sempre que for encontrada uma falha significativa. A seguir temos a Figura 3.7 com um esquema geral da metodologia proposta.

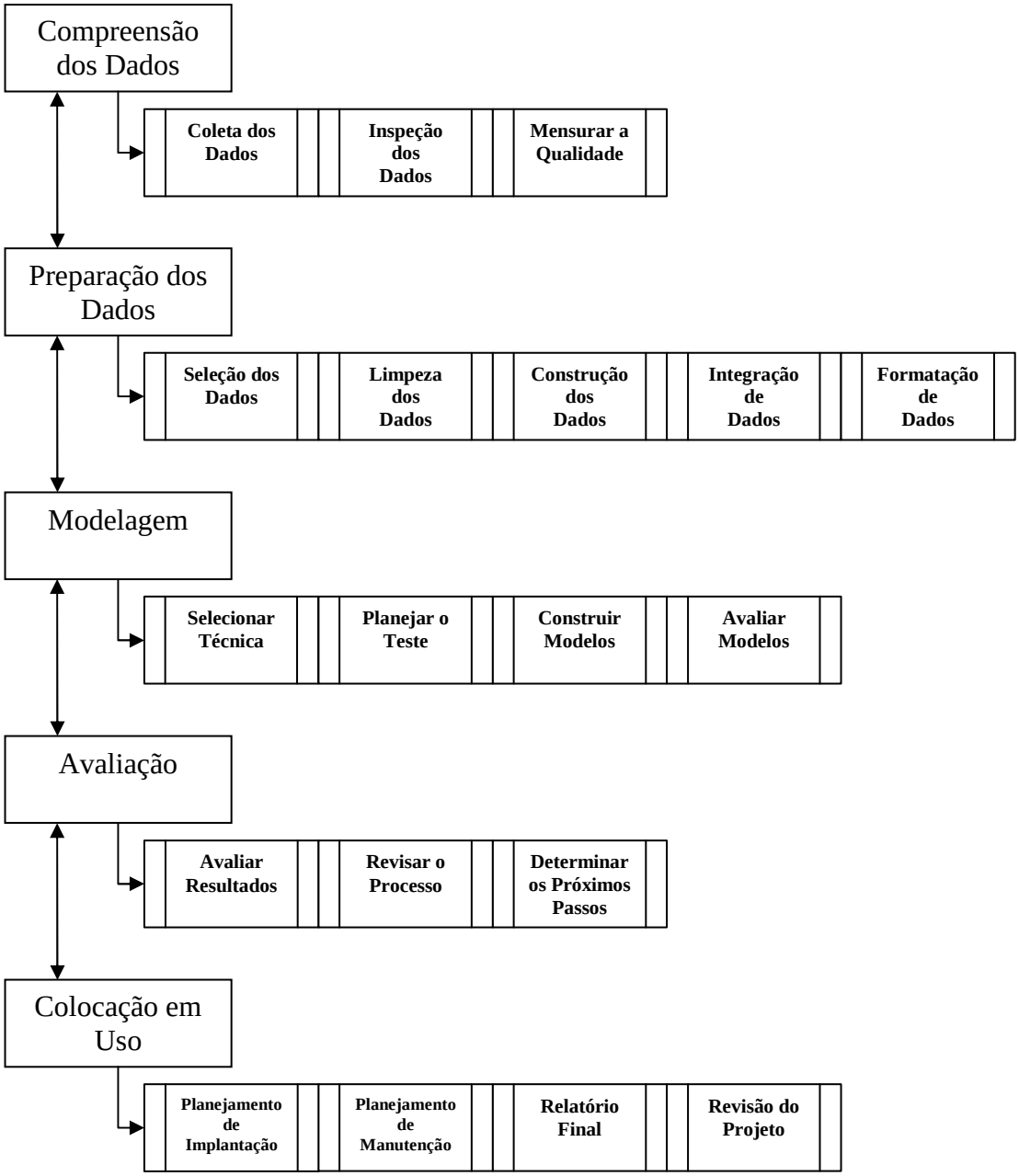


Figura 3.7: Workflow da Metodologia Proposta

Capítulo 4 Ferramenta Implementada

Este capítulo apresenta a ferramenta que materializa a metodologia proposta. Serão abordadas as principais funcionalidades implementadas e forma de utilização da ferramenta HaDog. Na Seção 4.1 são discutidas as necessidades e os objetivos da existência da ferramenta integrada para mineração. Na Seção 4.2 apresentamos a estrutura, dependências e plataforma de desenvolvimento da ferramenta. Na Seção 4.3 temos a descrição da interface da ferramenta. Na Seção 4.4 apresentamos um roteiro de uso tradicional sobre um problema de regras de associação usando o algoritmo de *APriori*. Na Seção 4.5 apresentamos em resumo as principais funcionalidades implementadas. Na Seção 4.6 são apresentados os principais algoritmos utilizados na modelagem. Por fim na Seção 4.7 são apresentadas perspectivas futuras em relação à ferramenta.

4.1 Motivação

Um dos objetivos principais desta pesquisa é fornecer ao não-informata (pesquisadores em germoplasma) a possibilidade de executar um projeto de mineração.

A metodologia proposta no capítulo anterior fornece um fluxo de execução estruturado que serve de roteiro para os especialistas, que apesar de serem leigos em informática, têm um grau de formação compatível com realização de atividades complexas.

O ponto crítico neste contexto são os detalhes de informática: a implementação de algoritmos de computadores e interações com os softwares envolvidos.

O HaDog vem para facilitar a interação entre estes pesquisadores e a execução de projetos de mineração cobrindo boa parte das fases da metodologia proposta. Os principais requisitos de projeto do HaDog são:

- estar integrado com o SIBRARGEN
- estar disponível via Extranet
- ser aderente a metodologia proposta
- ser de fácil utilização
- facilitar a realização das principais tarefas de mineração de dados

4.2 Plataforma

A estruturação do projeto de construção do HaDog foi pautada nos requisitos listados na Seção 4.1. Iremos descrever de forma sintética a estrutura e plataforma deste projeto. Acompanhe a seguir alguns tópicos relevantes neste contexto.

Como Sistema Gerenciador de Banco de Dados temos o Oracle® na versão 10g. Este SGBD é utilizado pelo SIBRARGEN e também servirá de repositório para as informações de projetos de mineração do HaDog.

Como linguagem de desenvolvimento utilizamos Java na versão 1.6, padrão Sun®. Esta linguagem já é utilizada por outros sistemas na instituição em foco, Embrapa e é ferramenta dominada pelo implementador.

Como IDE utilizamos o NetBeans na versão 6. Este é um IDE de uso livre, que tem boa integração com a linguagem e com servidores Web, tais como Tomcat e JBoss, uma característica importante para um projeto essencialmente para ambiente Web.

Utilizamos várias API com destaque para as API: JDBC, Weka e *Oracle Data Mining*.

O servidor Web para desenvolvimento foi o Tomcat na versão 6 da Apache. Para produção estamos utilizando o JBoss versão 4.

Estruturamos o projeto sobre um pequeno *framework* de autoria própria para implementar parte do modelo MVC (*Model-View-Control*), permitindo controle total por parte do implementador. Fizemos esta escolha, pois necessitamos de interações personalizadas, que poderiam torna-se difíceis de implementar em *frameworks* maiores tais como o Struts.

A execução de uma tarefa no HaDog ocorre através de uma *Servlet* que faz o papel de controle (*Control*). Esta lê uma *flag* vinda da Web que determina qual ação proceder. Nesta *Servlet* já é feito todo o controle de segurança, minimizando falhas. A partir da passagem correta pelo esquema de segurança a *Servlet* executa uma classe que implementa as tarefas necessárias (*Model*) e por fim páginas em HTML são processadas por esta mesma *Servlet* introduzindo dados vindos da camada *Model*.

A divisão das páginas e classes foi coerente com as fases de um projeto de mineração, acrescida de necessidades auxiliares, tais como a identificação dos usuários. Procuramos concentrar os esforços de implementação comuns em classes únicas compartilhadas no projeto. Assim como utilizar constantes em parâmetros comuns, tornando a manutenção do projeto mais simples.

O sistema esta disponível na URL http://ashwall.cenargen.embrapa.br/HaDog. Rodando sobre um servidor Linux CentOS 5 com Tomcat 6, JBoss 4 e Java 1.6.

O banco de dados Oracle® está hospedado em um servidor com Windows® Server 2003 Enterprise Edition.

Os servidores estão localizados no CENARGEN (Embrapa Recursos Genéticos e Biotecnologia).

Para utilizar o sistema é necessário que o pesquisador tenha acesso ao SIBRARGEN, especificamente ao módulo BAG (Banco de Germoplasma) com perfil de leitura e escrita, neste caso o usuário automaticamente terá acesso a todas opções do HaDog dentro do conjunto de dados do seu BAG, não interferindo na área de outros usuários.

Como os potencias usuários do HaDog são também usuários do SIBRARGEN utilizamos o mesmo esquema de autenticação e não nos preocupamos em gerenciar os usuários, perfis e permissões. Portanto para utilizar o HaDog é necessário que o usuário seja também um usuário do SIBRARGEN.

A arquitetura funcional do HaDog em uma macro-visão envolve a interface com o SIBRARGEN, uma área própria de trabalho que é subdividida em área para dados e área para modelos e aplicações. A gerência destas áreas é feita pelo HaDog e controlada pelo usuário através dos *wizards*. A Figura 4.1 mostra um esquema macro desta arquitetura.

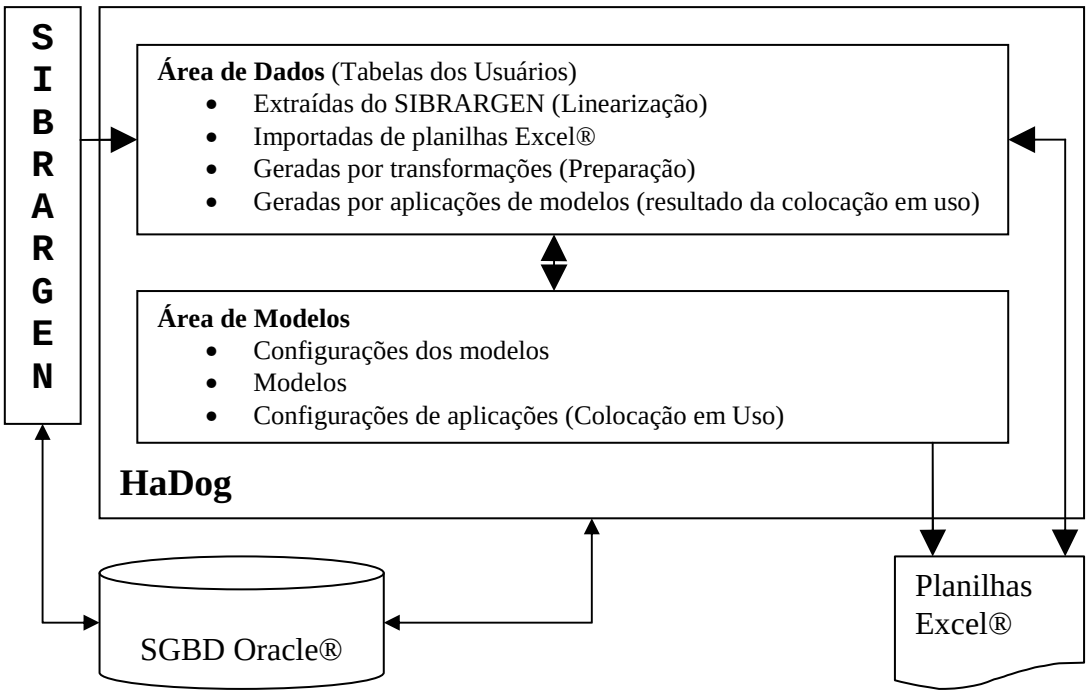


Figura 4.1: Arquitetura Macro do HaDog

4.3 Interface

O projeto é baseado em uma interface de *wizard*, cada tarefa executada dentro da ferramenta é separada em passos que cumulativamente vão fornecendo informações para que determinada ação possa ser executada no final.

A implementação destes *wizards* foi implementada através de *skins*, que são estruturas de páginas Web compartilhadas por mais de um processo. O projeto contém alguns *skins*, porém os três principais são denominados de "wi", "wm" e "wf", representando respectivamente "Wizard Inicial", "Wizard Meio (interior)" e "Wizard Fim".

Estes *skins* são apresentados na Figura 4.2 em mosaico, integralmente para o passo inicial e somente rodapé para o passo intermediário e passo final.

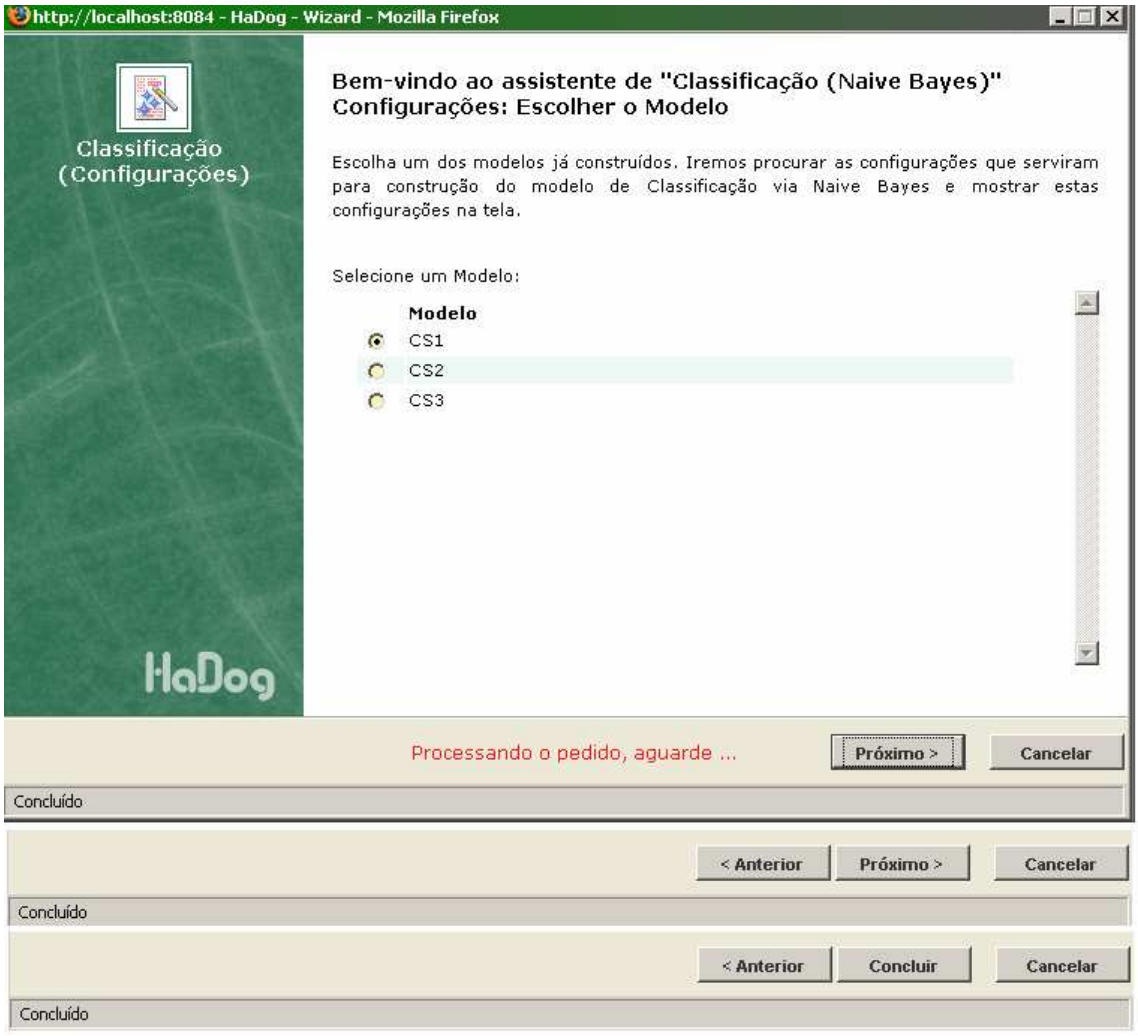


Figura 4.2: Skin de Wizard para o Passo Inicial e Rodapés

A grande maioria dos elementos gráficos de interface é construída usando HTML, que é comum entre os *browsers*. Interações dos elementos HTML e validações em entrada de dados são feitas prioritariamente em *JavaScript* de cliente. Todas as interações são feitas através de *browser*, assim o sendo, o usuário final irá necessitar somente um navegador para uso da ferramenta.

O processamento dos algoritmos de mineração, assim como as operações sobre o SGBD estão concentradas nos servidores. Estamos utilizando a idéia de cliente leve.

Foram utilizados alguns elementos em *Flash* para geração de gráficos. Neste caso necessitando do *plug-in* adequado. Atualmente *browsers* como Mozzila®, Firefox® e Opera® já vem integrados com este *plug-in* e o Internet Explorer® tem atualização automática para o mesmo.

A escolha de uma interface de *wizard* é creditada ao objetivo de tornar mais simples as tarefas de mineração, visto que desta forma o usuário é direcionado passo a passo até alcançar um resultado. A ferramenta atualmente conta com dezenas de *wizards* para as mais diversas tarefas do processo de mineração de dados.

O processo inicial até chegar aos *wizards* desejados tem dois passos descritos a seguir.

Inicialmente deve-se fazer a autenticação no sistema, cada usuário tem acesso somente aos bancos de germoplasma indicados no esquema de segurança do SIBRARGEN. Os usuários e permissões são compartilhados com o HaDog. Como na Figura 4.3.

Figura 4.3: Autenticação no Sistema

Posteriormente deve-se acessar uma das etapas disponíveis no sistema. As etapas são similares às fases da metodologia de mineração. A Figura 4.4 mostra o menu com as etapas disponíveis.

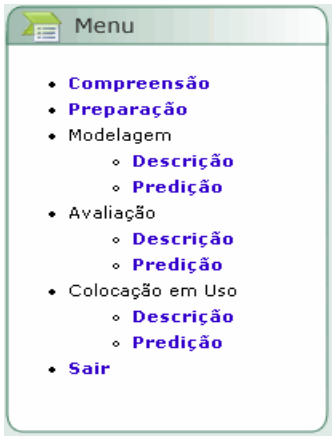


Figura 4.4: Menu com as Etapas (Fases da Mineração)

Acessando uma das etapas o usuário terá um submenu com as tarefas implementadas para a etapa escolhida. Por exemplo, ao acessar “Modelagem > Descrição” aparecerá uma janela como a da Figura 4.5.

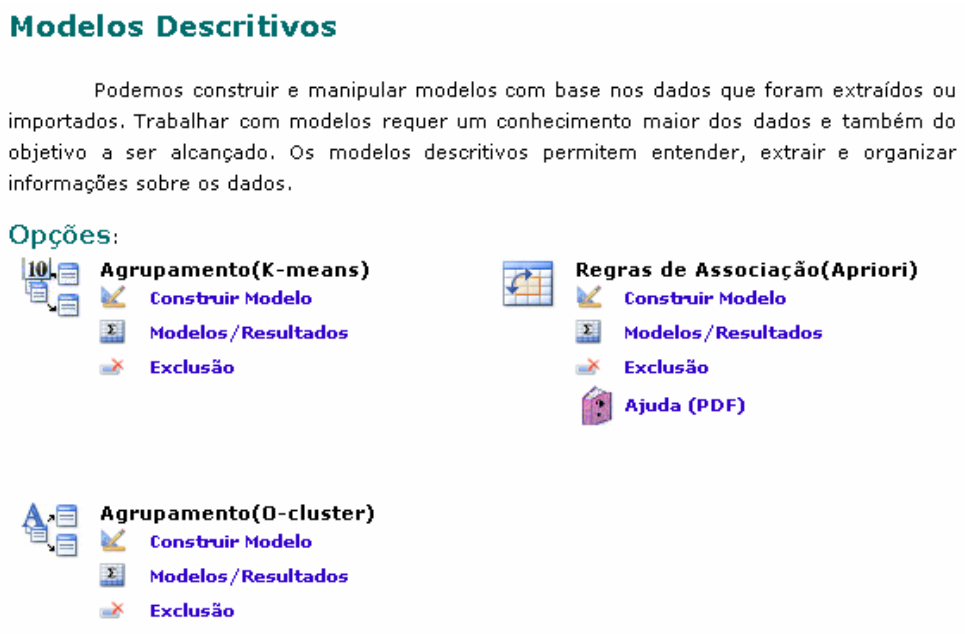


Figura 4.5: Submenu de "Modelagem > Descrição"

A maioria das opções disponíveis em um submenu chama um *wizard* nos moldes descritos no início desta seção.

4.4 Exemplo: Regras de Associação

Nesta seção iremos demonstrar algumas funcionalidades do projeto, através de um exemplo de extração de regras de associação sobre um pequeno conjunto de dados. Estes dados são de acesso livre, inclusive já foram publicados.

As outras funcionalidades do sistema têm funcionamento similar, o que muda são os objetivos da mineração, porém como foi dito a maioria das tarefas são disponibilizadas através *wizards*, sempre com o mesmo comportamento gráfico e de interface.

A atividade de "Regras de Associação" em mineração de dados prediz a probabilidade de co-ocorrência de um determinado conjunto de valores dentro de uma transação. Um caso comum é do carrinho de compras, onde é feita uma analise dos produtos encontrados em cada compra e então é calculada uma probabilidade de itens ocorrerem em conjunto em uma mesma compra. Pode-se dizer que estamos procurando um grau de afinidade entre objetos, relações ou correlações entre itens de um conjunto. No final é determinado com certo grau de certeza um comportamento diante de uma ação.

No HaDog temos implementado o algoritmo de *APriori* para fornecer as regras de associação. Cada regra tem dois indicadores que mensuram a qualidade da regra. O suporte que representa a ocorrência da regra no conjunto total de regras e a confiança que a medida da ocorrência do conseqüente dado o antecedente.

4.4.1 Compreensão dos Dados

A forma comum de obter dados vindos do SIBRARGEN na ferramenta é através da opção de linearização na primeira etapa de compreensão. Acessando a opção "Compreensão > Linearização" poderemos visualizar um cenário como o da Figura 4.6.

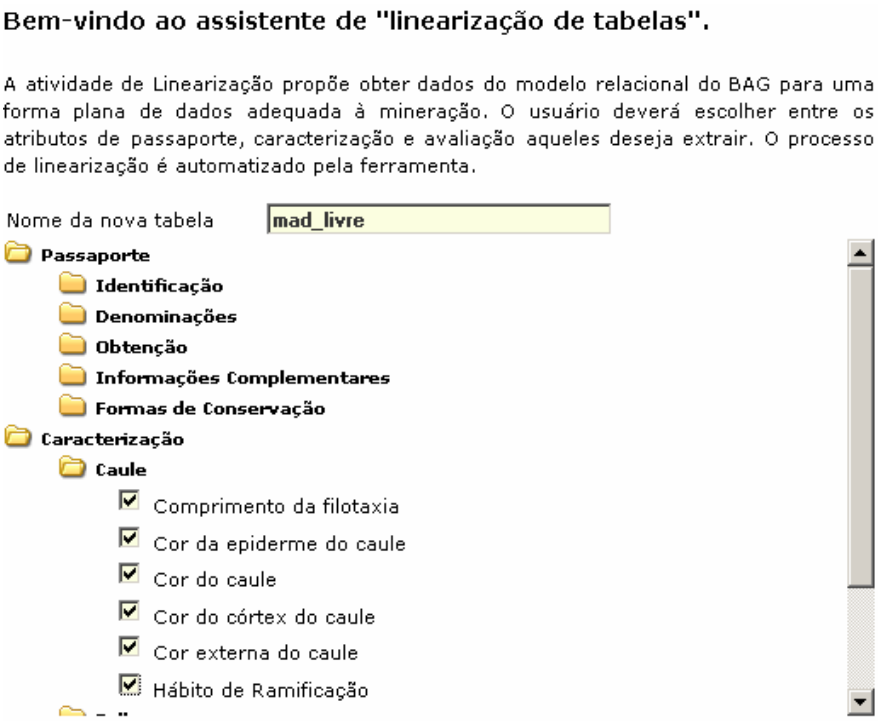


Figura 4.6: Uma Visão da Linearização

Iremos escolher um atributo de identificação do acesso e os atributos de caracterização e avaliação. O algoritmo de linearização irá descobrir as relações de chave primaria e estrangeira entre as tabelas envolvidas e gerar uma linha única em uma nova tabela. No caso de dados de caracterização e avaliação, o algoritmo ainda levará em conta metadados que descrevem os atributos especificamente para cada espécie.

Nesta etapa de compreensão dos dados é importante verificar a qualidade dos dados extraídos. Verificar a existência de valores faltantes entre outros erros. Uma forma é visualizar os dados extraídos. A ferramenta permite uma visualização dos dados e a visualização de um resumo, que iremos mostrar a seguir.

Para visualizar uma sumarização com estatística descritiva de uma tabela deve-se acessar a opção "Compreensão > Visualizar Resumo". Também é possível acessar um histograma da distribuição das categorias. Acompanhe um exemplo na Figura 4.7.

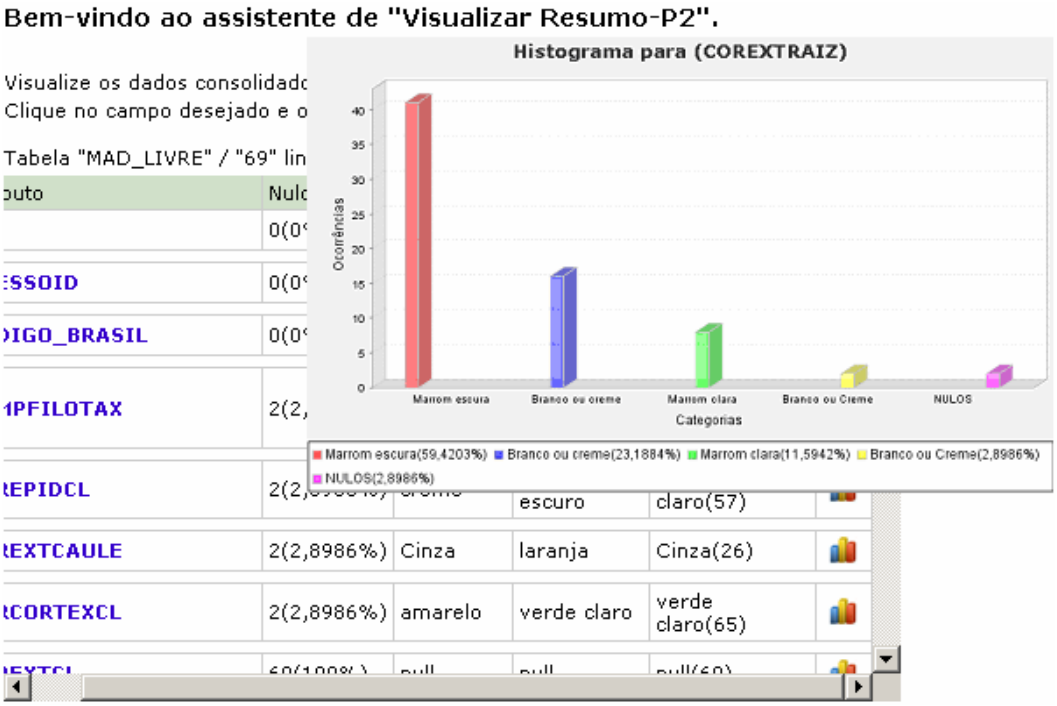


Figura 4.7: Visualização de Resumo

4.4.2 Preparação dos Dados

Notamos que existem valores faltantes que devem ser tratados. Iremos então acessar uma opção da etapa de preparação de dados, “Preparação > Tratamento de Valores Faltantes”.

O primeiro passo é escolher a tabela que iremos tratar, no caso “MAND_LIVRE” conforme foi nomeada na linearização. Depois aparece um resumo com os atributos que contém valores faltantes, neste momento já podemos definir a estratégia de tratamento. E por fim escolhemos o que fazer para cada atributo com valores faltantes. O resultado será uma nova tabela com valores faltantes tratados.

Nossa estratégia será eliminar os atributos que tem muitos valores nulos, por exemplo, acima de cinqüenta por cento. Aqueles que tem uma quantidade menor de valores nulos serão tratados assim: se categórico substitui pela moda, se numérico substitui pela média.

Este tratamento é necessário, pois a maioria dos algoritmos de mineração de dados não trabalha com valores faltantes. Desta forma é importante verificar a qualidade dos dados antes de prosseguir em um projeto de mineração.

O último passo do *wizard* apresentará um quadro semelhante à Figura 4.8.

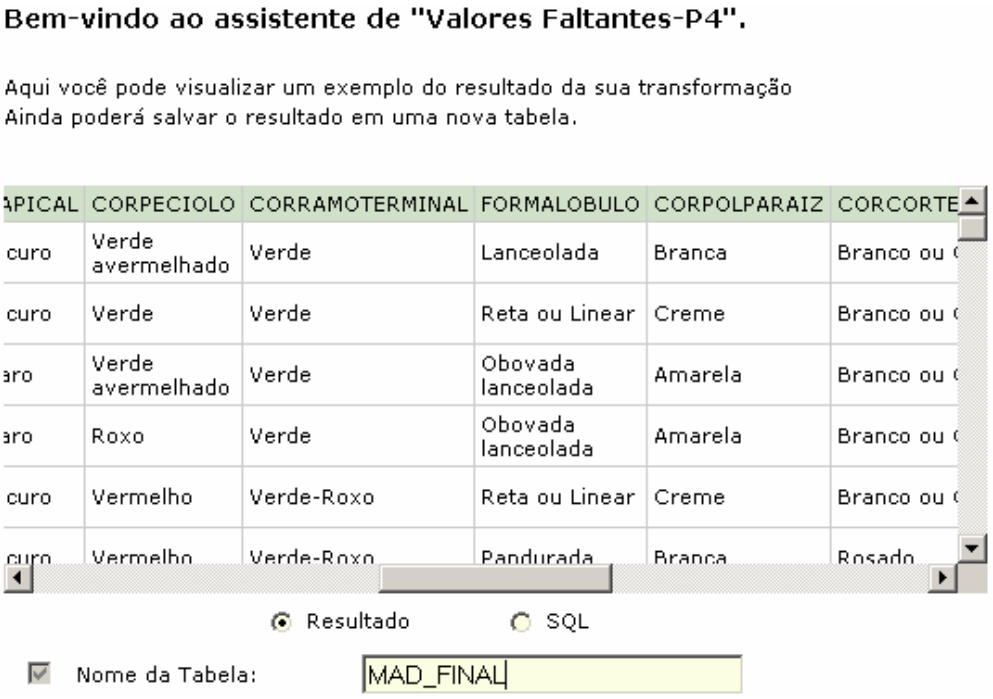


Figura 4.8: Resultado de Tratamento de Valores Faltantes

4.4.3 Modelagem

O conjunto de dados que será usado na criação do modelo de regras de associação será o "MAND_FINAL" resultante da etapa anterior, especificamente do tratamento de valores faltantes.

Um modelo de regras de associação é responsável por explicar dados que estão na base, ou seja, é um modelo descritivo, onde estamos preocupados em descobrir relações implícitas já existentes na base.

Qualquer modelo gerado na ferramenta é armazenado e gerenciado pelo HaDog, então ele funciona como se fosse um arquivo que pode ser reutilizado ou excluído do sistema. O primeiro passo na criação de um modelo é nomeá-lo como na Figura 4.9.

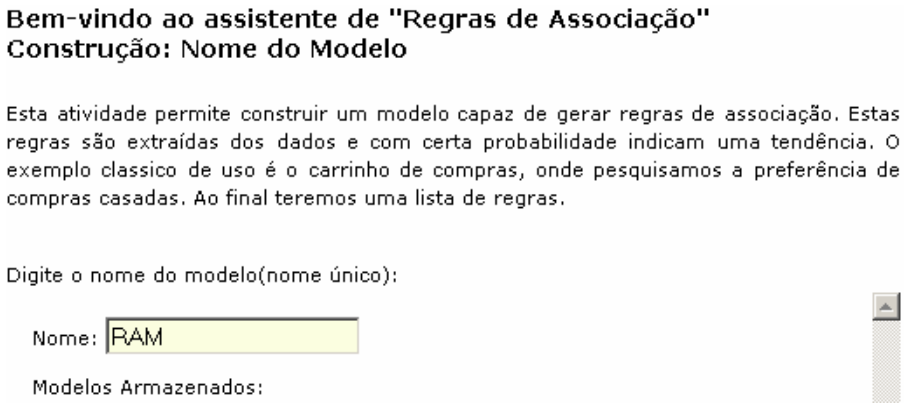


Figura 4.9: Regras de Associação - Nomeando um Modelo

Os nomes, por exemplo, de modelos e tabelas devem ser únicos dentro da ferramenta.

O segundo passo na geração de um modelo de regras de associação é escolher a tabela (conjunto de dados) que será utilizado na construção do modelo. Algumas informações como quantidade de linhas e tamanho médio da linha são retirados das estatísticas de banco de dados, que nem sempre estão disponíveis. Neste exemplo vamos usar a tabela "MAND_FINAL" como é mostrado na Figura 4.10.

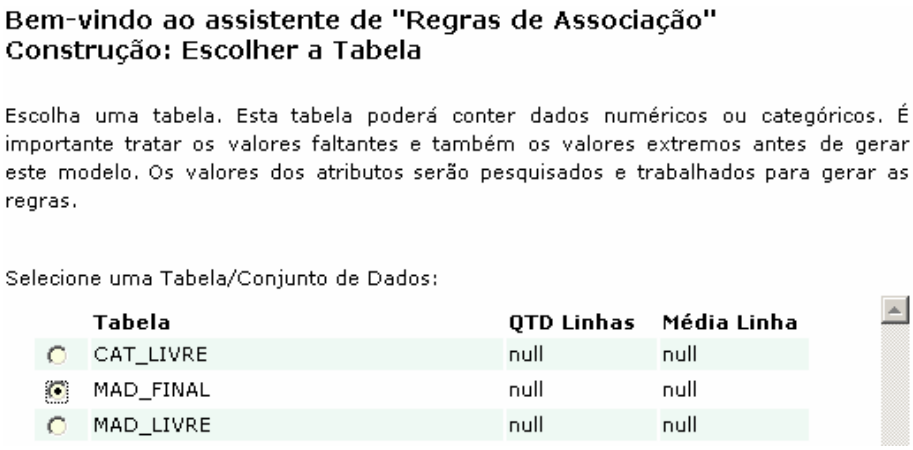


Figura 4.10: Regras de Associação – Escolha de Tabela

O terceiro passo será escolher os atributos que farão parte da mineração. É importante excluir da modelagem os atributos que são identificadores, assim como aqueles que não tem valores distintos. Neste conjunto de dados temos como identificadores: id, acessoid e codigo_brasil. Os atributos país e estado não têm valores distintos. Os atributos dos dois grupos devem ser excluídos (desmarcados) da mineração. Este passo é mostrado na Figura 4.11.

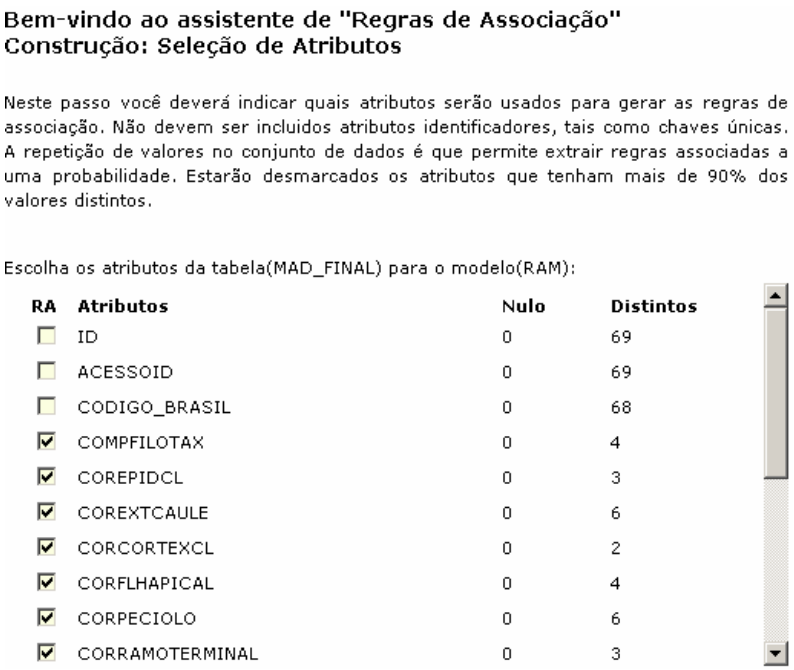


Figura 4.11: Regras de Associação – Escolha de Atributos

No quarto passo temos a configuração ou parametrização do algoritmo *APriori*, podem ser configurados os valores mínimos para suporte e confiança, ou seja, serão relacionadas apenas as regras que tiverem valores mínimos para estes indicadores. Assim como é possível limitar o número de atributos por regra, isso é interessante para não deixar a regra muito complexa, de difícil entendimento. Este passo é mostrado na Figura 4.12.

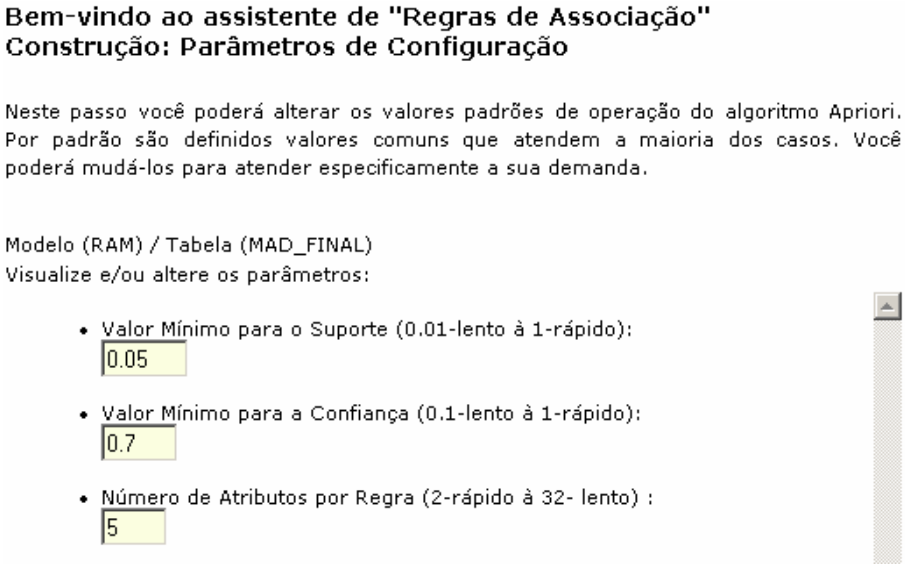


Figura 4.12: Regras de Associação – Parametrização

Após a parametrização o modelo será gerado e armazenado para consulta posterior.

4.4.4 Avaliação

A avaliação de modelos de regra de associação na ferramenta é feita somente através da analise dos indicadores de suporte e confiança.

Além de fornecer para cada regra os valores dos indicadores é possível visualizar um gráfico que sumariza os dados de confiança. Onde cada fatia do gráfico corresponde ao número de regras que tem confiança igual ou acima de uma porcentagem.

A Figura 4.13 mostra o gráfico para o modelo gerado. Note que neste caso específico as regras encontradas pelo modelo têm grau de confiança alto, maior ou igual a noventa por cento.

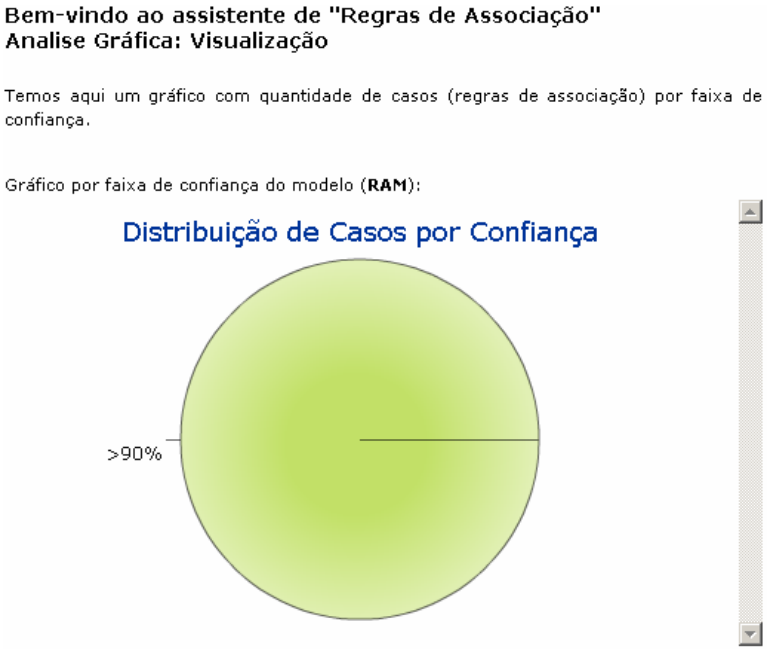


Figura 4.13: Gráfico de Distribuição de Casos por Confiança

4.4.5 Colocação em Uso

Em um problema de regras de associação o especialista deverá analisar as regras e verificar se algumas delas com suporte e confiança altas podem ser utilizadas em proveito do negócio.

A ferramenta permite que sejam mostradas as regras e também exportada para uma planilha Excel® para ser trabalhada fora do ambiente. A seguir temos a Figura 4.14 que mostra parte das regras formadas pelo modelo.

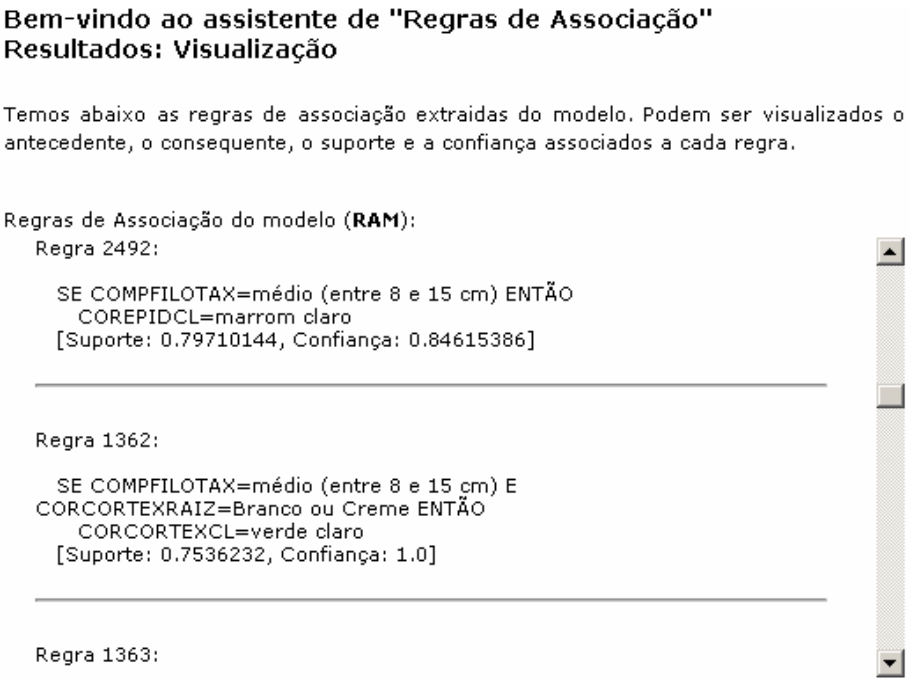


Figura 4.14: Regras Geradas pelo Modelo

4.5 Funcionalidades

O HaDog abrange todas as etapas da metodologia proposta e implementa as tarefas comuns de cada fase. A seguir iremos listar as principais funcionalidades implementadas.

Compreensão de Dados

1. Linearização
2. Exportação e Importação de Planilhas Excel®
3. Visualizações

Preparação de Dados

1. Tratamento de Valores Faltantes
2. Tratamento de Exceções
3. Geração de Tabelas de Construção e Teste
4. Filtragens
5. Agregações
6. Substituição de Valores
7. Campos Calculados
8. Definição de Tipos de Dados: Categórico ou Numérico

Modelagem

1. Descrição
    a. Agrupamento: *K-means*
    b. Agrupamento: *O-cluster*
    c. Regras de Associação: *APriori*
2. Predição
    a. Importância de Atributos: MDL (*Minimum Description Length*)
    b. Classificação: *Naive Bayes*
    c. Regressão: SVM (*Support Vector Machine*) Linear

Avaliação

1. Descrição
    a. Análises Gráficas
    b. Heurísticas
2. Predição
    a. Análises Gráficas
    b. Matriz de Confusão

Colocação em Uso

1. Descrição
    a. Aplicação de Modelos de Agrupamento
    b. Exportações de Resultados
2. Predição
    a. Aplicação de Modelos de Classificação
    b. Aplicação de Modelos de Regressão
    c. Exportação de Resultados

4.6 Algoritmos de Modelagem

A seguir são apresentados os algoritmos utilizados na fase de modelagem. Estes algoritmos são de domínio público e aqui são apresentados de forma resumida, evidenciando os pontos principais e abstraindo os detalhes de implementação.

4.6.1 Algoritmo de *K-means*

A idéia básica do algoritmo de K-means é partir da escolha de n pontos a serem utilizados como estimativa inicial para os centróides. Então entramos em um laço onde examinamos cada ponto a ser agrupado e então este ponto é associado ao centróide mais próximo. Após este laço é atualizado o valor do centróide com base no novo arranjo. O algoritmo para quando nenhum ponto seja associado a outro centróide ou então o número máximo de iterações previstas seja alcançado.

Aspectos importantes na implementação do algoritmo de K-means recaem sobre a escolha da função de similaridade, ou seja, como calcular a distância de um ponto ao centróide definido. Algumas abordagens passam por:

- distância geométrica
- medidas de distância personalizadas
- proximidade ortográfica
- proximidade semântica

Além disso, outro parâmetro importante é o número de *clusters* a gerar, este parâmetro irá determinar o quanto detalhado queremos a separação entre os elementos.

No HaDog podem ser informados três parâmetros: número de *clusters*, número de iterações e índice máximo de tolerância à erro. Os parâmetros do algoritmo visam alterar o grau de acuidade do algoritmo.

Quando aumentamos o número de grupos temos a geração de mais pontos iniciais como centróides, isso no início da organização dos dados em grupos. O número de iterações força o algoritmo a reavaliar o centróide “n” vezes conforme escolhido, isso faz com que a cada iteração um novo calculo de centróide seja efetivado, provavelmente reorganizando melhor os grupos. Por fim o índice de tolerância a erros determina a acuidade do calculo do centróide em relação aos elementos do grupo, quanto menor o índice mais difícil classificar um elemento dentro do grupo. Uma tolerância a erros pequena pode resultar em vários registros não classificados.

Na Figura 4.15 temos de forma abstrata o algoritmo de *K-means*.

function Direct-k-means()

Initialize k prototypes $(w_1, \ldots, w_k)$ such that $w_j = i_l,\ j \in \{1, \ldots, k\},\ l \in \{1, \ldots, n\}$

Each cluster C_j is associated with prototype w_j

Repeat

for each input vector i_l, where $l \in \{1, \ldots, n\}$, *do*

Assign i_l to the cluster C_{j*} with nearest prototype w_{j*} (i.e., $\mid i_l - w_{j*} \mid \ \leq \ \mid i_l - w_j \mid,\ j \in \{1, \ldots, k\}$)

for each cluster C_j, where $j \in \{1, \ldots, k\}$, *do*

Update the prototype w_j to be the centroid of all samples currently in C_j, so that $w_j = \sum_{i_l \in C_j} i_l / \mid C_j \mid$

Compute the error function:

$$E = \sum_{j=1}^{k} \sum_{i_l \in C_j} \mid i_l - w_j \mid^2$$

Until E does not change significantly or cluster membership no longer changes

Figura 4.15: Algoritmo de *K-means* utilizado

Escolhemos implementar o *K*-means por ser um algoritmo consagrado. Este já foi testado em diversas situações e os pesquisadores contribuiram ao longo do tempo da evolução do algoritmo com incrementos de flexibilização e performance.

O pseudocódigo e explicações sobre o uso do algoritmo são facilmente encontrados na Internet, assim como exemplos de utilização na área de mineração de dados.

4.6.2 Algoritmo de *O-cluster*

O algoritmo de *O-cluster* é baseado na construção de histogramas. É um algoritmo alternativo ao *K-means*. Na Figura 4.16 que mostra o fluxograma deste algoritmo.

Apesar da documentação do algoritmo *O-cluster* ser mais restrita que a de *K-means*, visto que tivemos dificuldades na obtenção do pseudocódigo, escolhemos implementá-lo, por representar uma alternativa promissora no agrupamento de elementos, cujos atributos, são de natureza categórica.

Existem adaptações do *K-means* que trabalham com atributos categóricos, porém a implementaçã do *O-cluster* na base de dados Oracle® se mostrou mais adequada. Os cálculos de histogramas e a divisão dos agrupamentos são operações disponíveis dentro da ferramenta de SGBD, diminuindo assim o nosso trabalho de codificação.

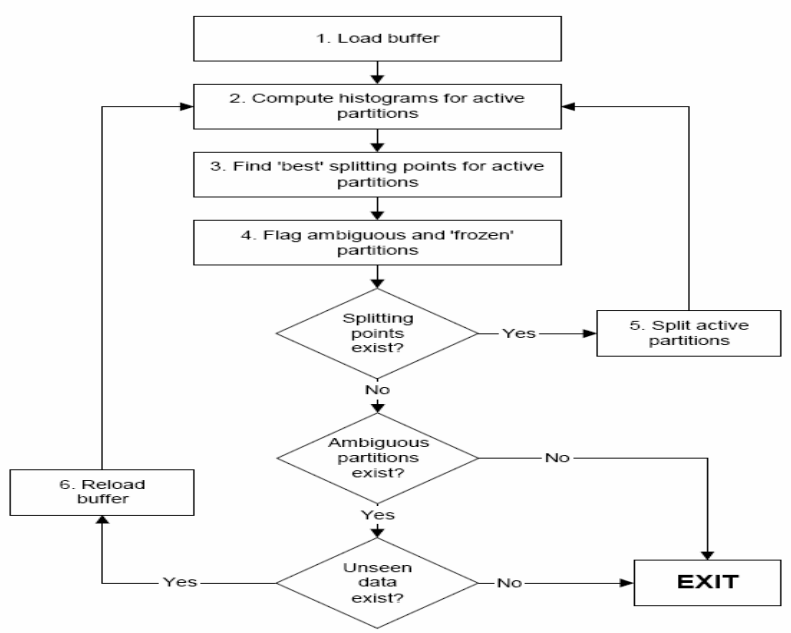


Figura 4.16: Fluxograma do Algoritmo *O-Cluster*

O objetivo do *O-cluster* é determinar áreas de grande densidade de dados, com base nestas informações separar os dados em *clusters*. Utilizam-se projeções de histogramas para encontrar picos e vales. A cada ocorrência de vales dois grupos são criados a partir da separação, isto é mapeado em algoritmo para uma árvore binária.

O que determina se um vale ocorreu ou não é o parâmetro de sensibilidade ao erro do algoritmo, sendo que um valor pequeno irá considerar pequenas variações como vales e, portanto gerar um número maior de *clusters*.

Na Figura 4.17 temos a representação de como este algoritmo trata a distribuição de categorias e como estes pontos de separação são encontrados.

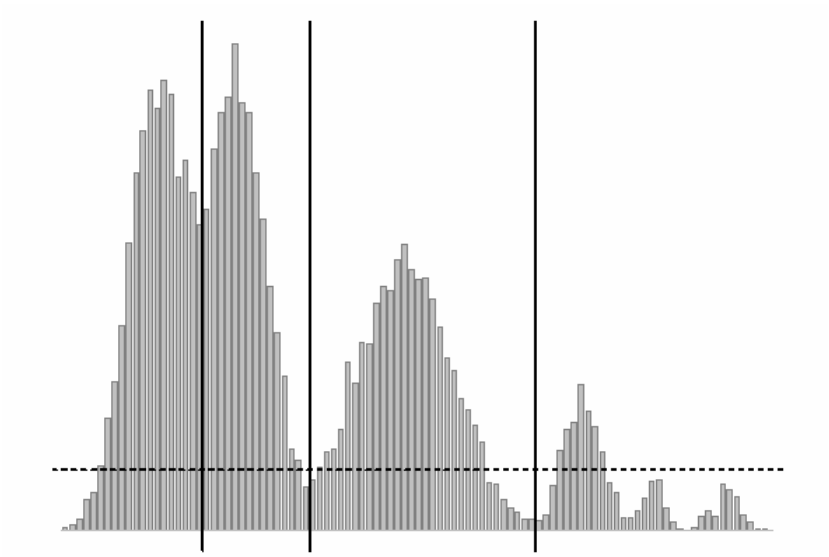


Figura 4.17: Determinação de pontos de vale

Na atividade de *load buffer* é criado uma única partição que irá conter todos os pontos, caso não caibam todos os dados no *buffer* é utilizada uma amostra aleatória do tamanho permitido.

Na atividade de *compute histograms* o objetivo é calcular o valor do histograma para cada valor discreto ou categoria, organizado sob uma determinada ótica, se número pela ordem crescente dos valores, se categórico é feita uma contagem individual para cada categoria.

Na atividade de *find best splitting* são combinados os valores dos histogramas para os valores numéricos e categóricos e as melhores combinações são incorporadas, entende-se por melhor, aqueles elementos que ocorrem maior número de vezes combinado.

Na atividade de *flag ambiguous partitions* é feita uma verificação, utilizando o parâmetro de sensibilidade do algoritmo se o ponto de partição é adequado ou não. Caso seja adequado o ponto é marcado.

Na atividade de *split partitions* é feita a separação das partições que realimentam o algoritmo no passo de *compute histograms*.

4.6.3 Algoritmo de *APriori*

O algoritmo de *APriori* trabalha na questão de co-ocorrência de itens relacionados. Digamos uma implicação da forma $X \Rightarrow Y$, onde X e Y são conjuntos disjuntos de itens de um registro consolidado.

São utilizadas duas métricas principais, a saber, suporte que é a probabilidade que um registro contenha tanto X quanto Y e confiança que é a probabilidade que um registro contenha X dado que contêm Y.

Na definição do problema, visando à solução temos:

seja $I=\{i_1, i_2, i_3, ..., i_n\}$ um conjunto de itens

seja T um conjunto de transações, onde cada transação t é um conjunto de itens, tal que $t \subseteq T$ e cada transação é única.

Sejam $X \subseteq T$ e $Y \subseteq T$ e $X \cap Y = \varnothing$, uma regra de associação é uma implicação da forma $X \Rightarrow Y$.
Desejamos gerar todas as regras que tenham suporte e confiança maiores que os valores mínimos especificados, ou seja, os parâmetros, aqui chamamos de minsup e minconf, respectivamente.

Na primeira etapa iremos determinar todos os conjuntos de itens cujo suporte é maior do que o suporte mínimo especificado. Eles são chamados de conjuntos freqüentes de itens.

Na segunda etapa iremos gerar as regras de associação a partir dos conjuntos freqüentes de itens. Dados conjuntos freqüentes X Y Z e X Y, a regra $Z \Leftarrow X\ Y$ é válida se a razão conf = supp (X Y Z) ÷ supp (X Y) é maior que a confiança mínima.

Algoritmo de Apriori

```
1. F1 = {1-itemsets}
2. for (k=2; Fk-1 ≠ 0; k++) {
3.      Ck = gera_candidatos(Fk-1);
4.      for all transações T in BD
5.              for all subconjuntos t in T
6.                      if (c está_em Ck: c=t) c.count++;
7.      Fk = {c in Ck | c.count >= minsup};
8. }
9. for all Fk, k>2
10.     gera_regras(Fk);
```

```
1. gera_candidatos (Fk-1) {
2.      for (i=1; i<=k-1; i++)
3.              if (p.item[i] ≠ q.item[i])
4.                      return 0; // Não gera candidato
5.      if (p.item[k] < q.item[k]) {
6.              Ck = p ∪ q;
7.      }
8.      for all conjuntos_candidatos c in Ck
9.              for all (k-1)subconjuntos s in c
10.                     if (s não_está_em Fk-1)
11.                             remove c de Ck;
12.     return Ck;
13. }
```

No algoritmo de APriori é fornecido uma interface para três parâmetros: valor mínimo de suporte, valor mínimo de confiança e número máximo de atributos por regra.

O valor mínimo de suporte restringe as regras geradas àquelas que tem um valor mínimo de represetatividade diante do conjunto de dados. O valor fornecido significa a porcentagem de zero a um em relação ao motante existente de corelação entre os itens.

O valor mínimo de confiança restringe as regras geradas àquelas que dado um antecedente à probabilidade que o conseqüente aconteça tenha valor mínimo aquele parametrizado. Note que a existência de um valor grande para confiança não representa que a regra é comum, quem determina uma regra forte e a composição de suporte e confiança.

O número máximo de atributos por regra restringe as regras mais simples e diretas, por exemplo, se o número de atributos for grande, torna o processo de conhecimento dos dados complexo e muitas vezes difícil de serem analisados e aplicados melhoramentos. Deve-se verificar o número de atributos do conjunto de dados inicial e ensaiar com diversos números máximo de atributos. Uma analise sobre as regras geradas em termos dos índices de suporte e confiança poderá auxiliar no processo de avaliação do modelo construído.

O algoritmo de APriori é consagrado em mineração de dados, tendo pseudocódigo de fácil acesso, assim como uma série de casos já documentados de uso. Vários artigos tratando de mineração de dados citam o algoritmo de APriori.

4.6.4 Algoritmo de MDL

Inicialmente o algoritmo de MDL - *Minimum Description Length* foi e é utilizado para compactação de dados. Ele trabalha no intuito de encontrar a menor representação para uma dada informação.

No problema de determinar a importância de um atributo em relação a outro destino consideramos que a menor representação, ou seja, a mais compacta também representa a melhor maneira de explicar ou influenciar o atributo destino.

O modelo de importância de atributos utiliza a MDL considerando cada atributo como um modelo preditivo simples em relação ao atributo destino. Cada valor ou categoria do atributo destino é associada a um índice *i* e para cada um destes valores é calculada a distribuição de propabilidade de ocorrência de valores do atributo considerado em relação ao atributo destino.

Com base na analise anterior dos *n* atributos considerados é construída uma lista em ordem descrescente de probabilidade. No calculo da probabilidade é usada a função logarítmica para suavizar a influência de um atributo em relação ao outro.

No MDL os atributos são os parâmetros mais importantes, inícialmente é necessário determinar qual é o atributo destino, ou seja, aquele que queremos analisar. Esta analise recai sobre o grau de influência que outros atributos tem na definição de valores para o atributo destino.

Entre os atributos que serão utilizados no algoritmo devem ser excluídos atributos que tenham valores diferentes para todas as ocorrências, por exemplo, chaves primárias. Também devem ser excluídos no outro extremo os atributos que não tenham variabilidade necessária, esta analise é individual e só temos a certeza quando todas as ocorrências têm o mesmo valor.

Utilizamos o MDL por ter seu código disponível para uso através de API do Oracle. Apesar do código fonte não ser comentado, existe boa documentação para uso. Alguns exemplos de utilização também estão disponíveis e foram utilizados para agilizar no processo de desenvolvimento do HaDog.

4.6.5 Algoritmo de *Naive Bayes*

O algoritmo de *Naive Bayes* funciona bem para conjuntos relativamente pequeno de dados e atende as necessidades em termos de quantidade dos dados de germoplasma considerados.

O algoritmo é baseado no teorema de Bayes:

P(A | B) = (P(B | A) P(A))/P(B)

Condideramos a probabilidade a priori como estimativa inicial da probabilidade de um certo evento ocorrer e a probabilidade a posteriori é uma informação adicional obtida na revisão da probabilidade a priori para obter a probabilidade a posteriori, previsão.

INPUT: training set T, hold-out set H, initial number of components k_0, and convergence thresholds δ_{EM} and δ_{Add}.

Initialize M with one component.
$k \leftarrow k_0$
repeat
  Add k new mixture components to M, initialized using k random examples from T.
  Remove the k initialization examples from T.
  repeat
    E-step: Fractionally assign examples in T to mixture components, using M.
    M-step: Compute maximum likelihood parameters for M, using the filled-in data.
    If $\log P(H|M)$ is best so far, save M in M_{best}.
    Every 5 cycles, prune low-weight components of M.
  until $\log P(H|M)$ fails to improve by ratio δ_{EM}.
  $M \leftarrow M_{best}$
  Prune low weight components of M.
  $k \leftarrow 2k$
until $\log P(H|M)$ fails to improve by ratio δ_{Add}.
Execute E-step and M-step twice more on M_{best}, using examples from both H and T.
Return M_{best}.

Figura 4.18: Algoritmo *Naive Bayes* utilizado

Considerando o teorema de Bayes e em relação aos parâmetros podemos informar dois limites mínimos. O algoritmo conta o número de casos em que A e B ocorrem em conjunto, isto como uma porcentagem sobre o total de casos, este é o primeiro parâmetro. O segundo parâmetro é a porcentagem em que ocorre A em relação a todos os casos.

Escolhemos implementar o algoritmo de NaiveBayes por constituir uma ferramenta já utilizada em mineração de dados, principalmente em modelos de classificação. Pudemos ter acesso à parte do código fonte na Internet e complementado com boa literatura sobre o pseudocódigo.

Também tivemos fácil acesso a casos que aplicaram o NaiveBayes em problemas de classificação em mineração de dados, o que constituiu fator primordial para determinação dos parâmetros dos algoritmos aplicados em mineração de dados.

4.6.6 Algoritmo de SVM

O núcleo do algoritmo de SVM – *Support Vector Machine* foi utilizado neste trabalho apenas com funções lineares, sendo uma implementação mais simples, apesar de menos precisa em relação a modelos de predição. A estrutura básica do algoritmo encontra-se na Figura 4.19.

Initialize $\alpha_i, \alpha_i^* = 0$
Choose arbitrary working set S_w
repeat
  Compute coupling terms (linear and constant) for S_w

  Solve reduced optimization problem
  Choose new S_w from variables α_i, α_i^* not satisfying the
  KKT conditions
until working set $S_w = \emptyset$

Figura 4.19: Algoritmo SVM com núcleo linear

O calculo no algoritmo de SVM no HaDog pode ser feita através de dois tipos de função: desvio médio padrão ou desvio médio absoluto. São utilizadas no núcleo funções lineares.

Além disso, é possível definir um parâmetro de tolerância a erro que fique entre 0 e 0.1, quanto menos a tolerância a erro mais acurado será o modelo, porém dependendo do conjunto de dados trabalhado não será possível a determinação de um modelo, já que este parâmetro é um ponto de parada.

O algoritmo de SVM constitui um conjunto de algoritmos que podem ser criados a partir de núcleo comum. Este núcleo comum tem seu pseudocódigo disponível para a comunidade.

Escolhemos uma implementação mais simples somente para funções lineares, apesar de não ser adequada a uma grande gama de problemas, atende na nossa perspectiva inicial de fomentar mineração de dados entre os especialistas em recursos genéticos. As estruturas de dados e de programação já foram pensadas para se adequar a um novo conjunto de funções, possibilitando a expansão do algoritmo no futuro.

4.7 Perspectivas

Espera-se que a ferramenta possa ser usada em massa pelos especialistas usuários do SIBRARGEN fornecendo o ferramental necessário para que estes pesquisadores incluam em suas pesquisas técnicas de mineração de dados.

Por ser desenvolvido para ambiente Web, estar organizado de forma estruturada nos moldes do CRISP/DM e ser baseado em *wizards* acreditamos que possa ser usado por pesquisadores em germoplasma distribuídos geograficamente no Brasil.

Esperamos que a ferramenta fomente a mineração de dados no contexto de recursos genéticos e que sirva de incentivo, com os resultados, para que os pesquisadores possam ter um cuidado maior com seus dados, tanto na captação, manutenção e organização dos mesmos.

Como trabalhos futuros esperamos estabilizar a ferramenta e disponibilizá-la a comunidade científica da Embrapa e parceiros.

Em um passo posterior desejamos agregar novos algoritmos de mineração de dados, principalmente na atividade de classificação. E complementarmente agregar funcionalidades na fase de avaliação e colocação em uso conforme as necessidades reais que irão surgir com o uso da ferramenta pela comunidade.

A ferramenta foi construída de forma modular, permitindo que novos módulos sejam agregados. Já vislumbramos a possibilidade de inclusão de novas tarefas de mineração, principalmente nas fases de preparação dos dados. Os primeiros usuários do HaDog já demandam novas funcionalidades a partir da experiência inicial adquirida.

A transformação de dados, a normalização, a importação e exportação de dados possibilitaram aos usuários uma manipulação mais genérica das informações. Estes usuários retiram informações para tratamento em outras ferramentas e também anexam novos dados a base.

Várias sugestões de incrementos na preparação de dados foram feitas, entre elas, a comparação de dados tabelados com dados importados, a expansão da criação de campos calculados, permitindo a geração de expressões mais complexas, a geração de gráficos parametrizados pelos usuários, entre outros.

Capítulo 5 Estudo de Caso: Acessos Representativos

Este capítulo apresenta um estudo de caso em bases de germoplasma que pretende selecionar um conjunto de acessos que melhor represente um grupo. Este grupo é composto de acessos de uma espécie ou mais espécies que são afins. Na Seção 5.1 conceituamos os termos utilizados e contextualizamos o problema. Na Seção 5.2 temos o planejamento do projeto de mineração listando as atividades que serão executadas. Na Seção 5.3 temos a descrição das atividades executadas no HaDog segundo o planejamento feito na seção anterior. Por fim na Seção 5.4 é apresentado de forma sintética os resultados e algumas considerações acerca do projeto de mineração.

5.1 Conceituação e Contextualização do Problema

Existem bases de dados que registram acessos, ou seja, variedades de uma espécie. Um acesso representa a classificação mais detalhada dentro de uma espécie. Colecionar acessos é importante para garantir a perpetuação da diversidade genética, pois é este material que irá permitir as pesquisas em melhoramento.

Para uma dada espécie é possível ter centenas ou até milhares de acessos. Cada um destes acessos possui características próprias que são avaliadas em processos de caracterização e avaliação. Na caracterização são considerados atributos perenes ligados ao genótipo do indivíduo. Na avaliação os atributos estão ligados à relação de como os gens são afetados pelo ambiente, ou seja, é a expressão do fenótipo.

O melhoramento se dá através da combinação de acessos, o que pode ocorrer de várias formas, desde o cruzamento tradicional até as manipulações genéticas. Não se cria gens novos, mas pode-se combinar gens existentes em acessos diferentes para formação de um novo. A possibilidade de combinação é enorme e a pesquisa deve direcionar seus experimentos para aquelas poucas combinações que chegam ao resultado esperado.

Um melhorista, pesquisador que procura combinar acessos para encontrar outro que possa atender a demandas específicas, necessita trabalhar com um determinado conjunto de germoplasma (acessos). A quantidade de acessos pode ser bem grande no início de uma pesquisa se for considerado todo o banco de germoplasma daquela espécie.

Por questões de tempo e economia não é viável que cada melhorista tenha disponível cópia de todos os acessos de um banco de germoplasma de uma dada espécie. Mas no início de uma pesquisa é interessante que esteja disponível a diversidade genética da espécie sendo trabalhada.

Aqui temos um problema de logística: a partir de um estoque central de acessos de uma espécie (banco de germoplasma) queremos distribuir para os melhoristas acessos que representam, com um certo grau de veracidade, a diversidade do banco de germoplasma.

O objetivo principal é diminuir, em um primeiro instante, o número de acessos que devem ser distribuídos, diminuindo assim custos de replicação, manutenção e transporte.

Os dados de caracterização e avaliação nos dão uma idéia detalhada dos atributos que descrevem um acesso. Estes atributos podem ser usados para agrupar os acessos em um determinado número de grupos. Cada grupo representa um grupo diferente em relação ao outro. Como a composição destes grupos é feita com base em atributos que descrevem genótipo e fenótipo, que são expressões dos gens, temos a formação pautada na diversidade genética.

Por questões de viabilidade e economia desejamos selecionar somente alguns acessos do conjunto total. Estes acessos podem ser selecionados a partir dos grupos formados, considerando aqueles que estão mais perto do centro do grupo, ou seja, aqueles que tem uma possibilidade maior de concentrar as características de atributos (regras) que formaram o grupo.

Cabe ressaltar que em um momento mais avançado da pesquisa o melhorista poderá necessitar de um outro tipo de acesso, por exemplo, todos os acessos que participam de um determinado grupo que contém características desejadas e que cujo acesso representativo figurou expoente na pesquisa. Assim podendo fazer o ajuste fino na pesquisa inicial.

5.2 Planejamento da Mineração

Nesta seção iremos planejar o projeto de mineração de dados sob a ótica da metodologia proposta. Iremos detalhar os passos que iremos executar, assim como listar as considerações importantes neste projeto de mineração.

Este projeto de mineração irá derivar uma sistemática para auxiliar no processo decisório sobre quais acessos distribuir, quando existe a necessidade de iniciar uma pesquisa com determinada espécie. Cada subseção tratará de uma fase do projeto de mineração que passará por: compreensão dos dados, preparação dos dados, modelagem, avaliação e colocação em uso.

5.2.1 Compreensão dos Dados

Os dados utilizados para mineração serão primordialmente os dados de caracterização e avaliação, pois estes representam a expressão visível da variabilidade genética. Estes dados podem ser obtidos diretamente da base de dados de germoplasma.

Na coleta dos dados devemos considerar que os atributos que serão utilizados na mineração são por natureza diversos, dependendo da espécie considerada. Além disso, o especialista poderá utilizar alguns atributos em detrimento de outros, conforme sua necessidade ou interesse. Esta escolha inicial caracterizará o conjunto de dados que estará disponível para mineração nas próximas fases.

Uma analise preliminar dos dados obtidos na coleta deve ser feita para verificar a possibilidade de continuidade do projeto. É importante verificar se estão completos e se para aquela espécie os acessos foram caracterizados e avaliados nos atributos de interesse. Se não forem, é necessário alimentar o SIBRARGEN para tornar possível o projeto de mineração para a espécie em questão. Caso estes dados estejam semipreenchidos ou com pequenos erros é interessante anotar as inconsistências e se possível tratá-las na próxima fase.

5.2.2 Preparação dos Dados

Os dados coletados na fase anterior e avaliados sobre a ótica da quantidade e qualidade podem requerer alguma transformação para deixá-los compatíveis com a tarefa de mineração que desejamos executar, neste caso um agrupamento. As anotações de inconsistências da fase anterior devem ser tratadas nesta fase.

Um pré-passo importante é verificar o tipo de dado de cada atributo. O SIBRARGEN define os atributos de caracterização e avaliação como categóricos, mesmo que estes contenham dados numéricos. Assim, após extrair dados das bases do SIBRARGEN, é necessário redefinir os tipos dos atributos, em categórico ou numérico conforme a origem.

Outras transformações necessárias podem ser o tratamento de valores faltantes, assim como limpeza de valores extremos, provavelmente ocasionados por erros de digitação. Caso seja verificado que não é possível transformar os dados em um conjunto confiável de dados é necessário abortar o projeto de mineração e retomar a alimentação do SIBRARGEN. Em caso contrário, o conjunto de dados derivado do pré-processamento será utilizado na construção do modelo de agrupamento.

5.2.3 Modelagem

De posse do conjunto final de dados poderemos aprender um modelo. Já temos determinada à tarefa de mineração, que é um agrupamento, passamos agora a escolher a ferramenta (algoritmo) que desejamos utilizar. Podemos escolher o algoritmo *K-means* ou *O-cluster*. O segundo foi preparado para atuar com atributos categóricos ou numéricos na implementação feita.

O modelo construído gera regras condição-ação do tipo se-então. Estas regras são avaliadas pelos indicadores suporte e confiança. O indicador suporte representa a probabilidade de ocorrência da regra em relação ao conjunto total de regras. A confiança representa a probabilidade condicional do conseqüente, dado o antecedente da regra.

O valor de suporte indica a freqüência relativa da regra no conjunto de dados. O valor de confiança representa a certeza de que se a condição for satisfeita então o conseqüente ocorre no conjunto de dados. Os valores de suporte e confiança estão expressos entre zero e um, inclusive.

Como um modelo de agrupamento é do tipo descritivo, estamos entendendo e explicando os dados. Neste caso podemos dispor de algumas facilidades de visualização, assim como métricas baseadas em heurísticas para determinar a qualidade do modelo de agrupamento.

No momento da criação do modelo é importante passar ao algoritmo apenas os atributos que são de interesse no processo de agrupamento.

Atributos identificadores, tais como chave primária ou chave única, devem ser excluídos da escolha de atributos para mineração. Já que estes representam apenas a necessidade do modelo relacional em individualizar cada linha e não tem um

significado no contexto do negócio. Exemplos: identificador do acesso ou código do acesso no Brasil.

Atributos com apenas um valor também não devem ser escolhidos para mineração, visto que não é possível diferenciar um acesso do outro por um atributo onde todos os acessos têm o mesmo valor. O agrupamento é um processo também de diferenciação, separação em grupos. Exemplo: Cor da folha se todas tiverem o mesmo valor.

Outra consideração é em relação ao número de grupos que deve ser gerado. Este número é variável e irá depender da necessidade do especialista. É possível que sejam gerados vários modelos com número de grupos variáveis e depois feita uma avaliação com base em visualização de regras, resultados e heurística para eleger um dos modelos.

5.2.4 Avaliação

Nesta fase iremos nos preocupar em mensurar a qualidade dos modelos gerados. Em uma tarefa de mineração de agrupamento temos alternativas empíricas.

Devemos visualizar as regras e conferir os valores para os indicadores de suporte e confiança. Regras com confiança alta indicam um bom grau de acerto do conseqüente se o antecedente ocorrer. Se as regras que formam os grupos tiverem confiança alta é um indicativo de uma boa separação entre os grupos. Porém o especialista deve estar atento para a regra formada, pois esta pode englobar tantos acessos que não atenderia a necessidade de diversificação do problema. Neste caso um modelo com maior número de grupos pode ser a solução.

Usando o algoritmo de *K-means* é possível determinar o número de grupos do modelo, já no algoritmo de *O-cluster* aumentando o parâmetro de sensibilidade do algoritmo pode se obter uma quantidade maior de grupos. Caso necessário pode-se retornar a fase de modelagem e construir novos modelos com parâmetros diferentes.

5.2.5 Colocação em Uso

Nesta fase os resultados da mineração serão usados para auxiliar na escolha de quais acessos deverão ser distribuídos se necessitarmos fornecer a diversidade genética do Banco de Germoplasma.

Existem duas formas principais de visualizar os resultados de uma mineração de agrupamento: separar os elementos (linhas) pelas regras de formação dos grupos ou reclassificar os elementos conforme o grau de aderência em relação ao centróide dos grupos formados.

Separando os elementos pelas regras de formação teremos um subconjunto de dados com todos os elementos do grupo, porém sem um indicador do quanto o elemento é aderente ao centróide.

Já no segundo caso aplicamos os dados sobre o modelo escolhido. O resultado será um novo conjunto de dados onde cada linha terá um indicativo (probabilidade) do elemento estar contido no grupo indicado.

5.3 Execução do Projeto de Mineração

Na seção anterior descrevemos uma sistemática sob a luz da metodologia proposta para descoberta de acessos representativos de um determinada espécie ou espécies afins. Nesta seção iremos aplicar o projeto sobre dados da espécie *Manihot esculenta Crantz*, popularmente conhecida como mandioca. Vale lembrar que a sistemática é similar para aplicação em outras espécies.

Utilizaremos a ferramenta HaDog para nos auxiliar no processo de mineração. Dividiremos as subseções nas opções encontradas na ferramenta, que são aderentes a metodologia proposta.

5.3.1 Compreensão

Iremos trabalhar sobre um conjunto de dados composto por 1168 acessos de Mandioca. Estes acessos contêm dados de passaporte e caracterização. Os dados de passaporte estão preenchidos para os atributos obrigatórios, os outros atributos podem ou não conter dados. Para os atributos de caracterização os dados estão preenchidos.

Para obter os dados de mandioca iremos entrar na ferramenta HaDog e na opção Compreensão iremos escolher a subopção Linearização. Nesta subopção conseguimos extrair da base do SIBRARGEN os dados para mineração. Queremos obter os seguintes dados:

- Dado de passaporte com o objetivo de identificar o acesso para posterior colocação em uso: código no Brasil.
- Dados de caracterização: aqueles atributos que são de interesse por parte do especialista (biólogos ou agrônomos)

A seguir temos uma figura que demonstra a subopção de linearização:

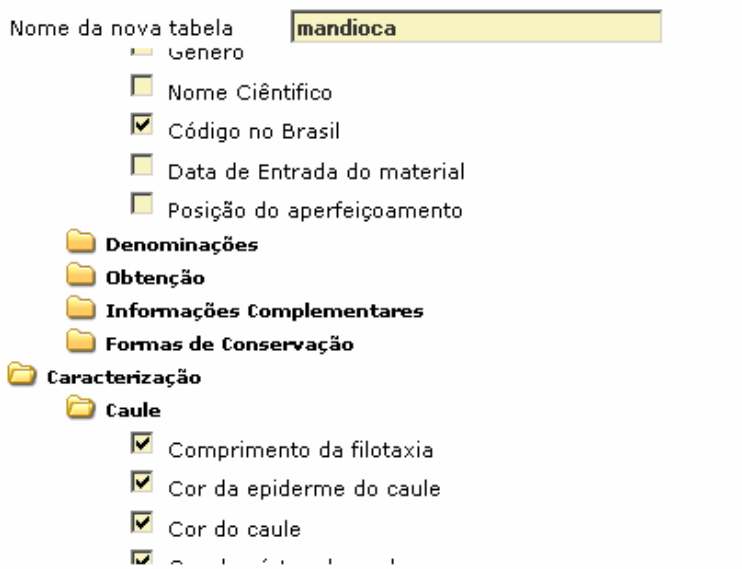


Figura 5.1: Linearização – Escolha dos Atributos

Conforme indicação do especialista foram escolhidos todos os atributos de caracterização sobre o caule e dois atributos para raiz e folha. O resultado da execução da Linearização pode ser visualizado na subopção “Visualização de Dados”. Abaixo temos a Figura 5.2 com parte da visão gerada:

Conjunto de Dados: MANDIOCA

CODIGO_BRASIL	COMPFILOTAX	COREPIDCL	COREXTCAULE	CORCORTEXCL	COREXTCL	HABITORAMIFICACAO
BRA-002097	curto (menor que 8 cm)	marrom escuro	Dourado	amarelo	marrom escuro	Indiviso
BRA-002143	curto (menor que 8 cm)	marrom claro	laranja	amarelo	cinza	Tricotômico
BRA-002186	médio (entre 8 e 15 cm)	marrom escuro	Prateado	verde claro	laranja	Indiviso
BRA-002224	longo (maior que 15 cm)	marrom escuro	Marrom escuro	amarelo	verde amarelado	Indiviso
BRA-002275	curto (menor que 8 cm)	laranja	laranja	verde escuro	cinza	Tricotômico

Figura 5.2: Visualização – Parte dos Dados de Mandioca Linearizados

Outra visão possível dos dados é obtida através da subopção “Visualizar Resumo”. Nesta opção temos uma primeira sumarização dos dados com uma estatística descritiva de valores nulos, maior, menor, media ou moda. É também possível acessar um histograma com dados da distribuição das categorias na população total.

Tabela "MANDIOCA" / "1168" linhas:

Atributo	Nulos	Menor	Maior	Média/Comum	Hist.
ID	0(0%)	1	1168	584,5	
ACESSOID	0(0%)	173348	174515	173.931,5	
CODIGO_BRASIL	0(0%)	BRA-000027	BRA-109592	BRA-001856(2)	
COMPFILOTAX	0(0%)	curto (menor que 8 cm)	médio (entre 8 e 15 cm)	médio (entre 8 e 15 cm)(421)	

Funções	Valores
Menor Comprimento	22
Maior Comprimento	23
Média Comprimento	22,6807
Valores Distintos	2

COREPIDCL	0(0%)	creme	marrom	marrom claro(307)	

Figura 5.3: Visualizar Resumo – Estatística Descritiva

Podemos observar que os atributos extraídos de caracterização para mandioca são categóricos. O Código no Brasil é o identificador único dentro do conjunto considerado. Também observamos pela sumarização anterior que não temos valores faltantes, desta forma dispensando a etapa de preparação de dados.

5.3.2 Preparação

Para este conjunto de dados não serão necessárias operações de transformação. Em outros conjuntos de dados pode haver necessidade de tratamento de valores faltantes, identificado na fase anterior.

O tratamento de valores extremos também é dispensável, visto que a natureza dos dados é categórica e dentro do SIBRARGEN é tabelada, minimizando erros de digitação. Além disso, analisando os histogramas dos atributos verificamos que as categorias têm uma distribuição uniforme, não existindo valores raros.

Não existem dados de interesse do tipo numérico, desta forma uma definição de dados também não é necessária, já que estão corretamente definidos como categóricos.

Não identificamos neste conjunto de dados outras necessidades de transformação.

5.3.3 Modelagem

Iremos criar modelos de agrupamento. Na ferramenta HaDog temos disponíveis dois tipos de algoritmos *K-means* e *O-cluster*. O algoritmo *O-cluster* pode trabalhar com atributos categóricos, que compõe o conjunto de dados deste caso. Iremos criar cinco modelos todos utilizando como entrada os atributos de caracterização, o que iremos modificar é o parâmetro de sensibilidade, que indica em última analise a quantidade de cluster que serão geradas.

Chamaremos os modelos de mr1, mr2, mr3, mr4 e mr5. Para cada um destes modelos iremos escolher todos os atributos de caracterização e deixando desmarcados os outros conforme figura a seguir:

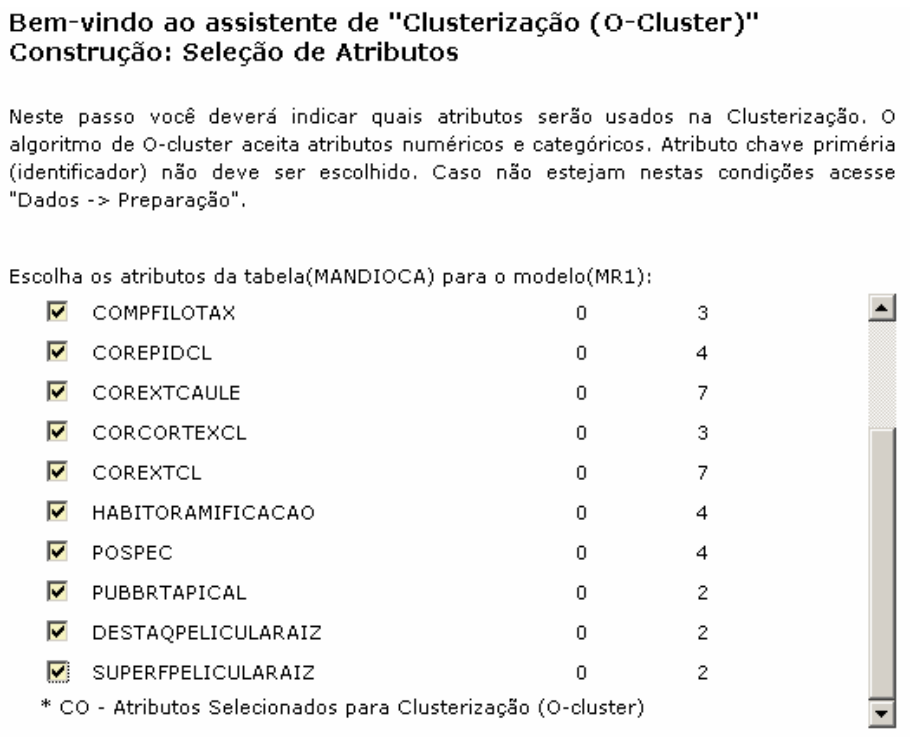


Figura 5.4: Modelo O-cluster - Atributos

Em um modelo de agrupamento não devemos incluir os atributos identificadores, assim como atributos que também não contém valores distintos. Os atributos de interesse, atributos de caracterização, não são identificadores e todos contêm valores distintos como verificamos na fase de compreensão dos dados.

Na parte de parametrização do modelo iremos modificar os parâmetros de sensibilidade e o número máximo de grupos (*cluster*). A idéia é modelar os cinco grupos com número cada vez maior de grupos e depois avaliar para podemos escolher aquele que melhor atenda as necessidades de encontrar um conjunto de acessos representativos.

No algoritmo de *O-cluster* é possível aumentar os números de grupos, aumentando o valor da sensibilidade, este parâmetro é um número real entre 0 e 1, inclusive, iremos usar cinco valores um para cada modelo gerado. Para cada um dos valores de sensibilidade teremos um número real de grupos gerados.

Iremos parametrizar o algoritmo de *O-cluster* com sensibilidade variada e com número máximo de *cluster* no topo, ou seja, o valor 64. Como mostrado na figura a seguir:

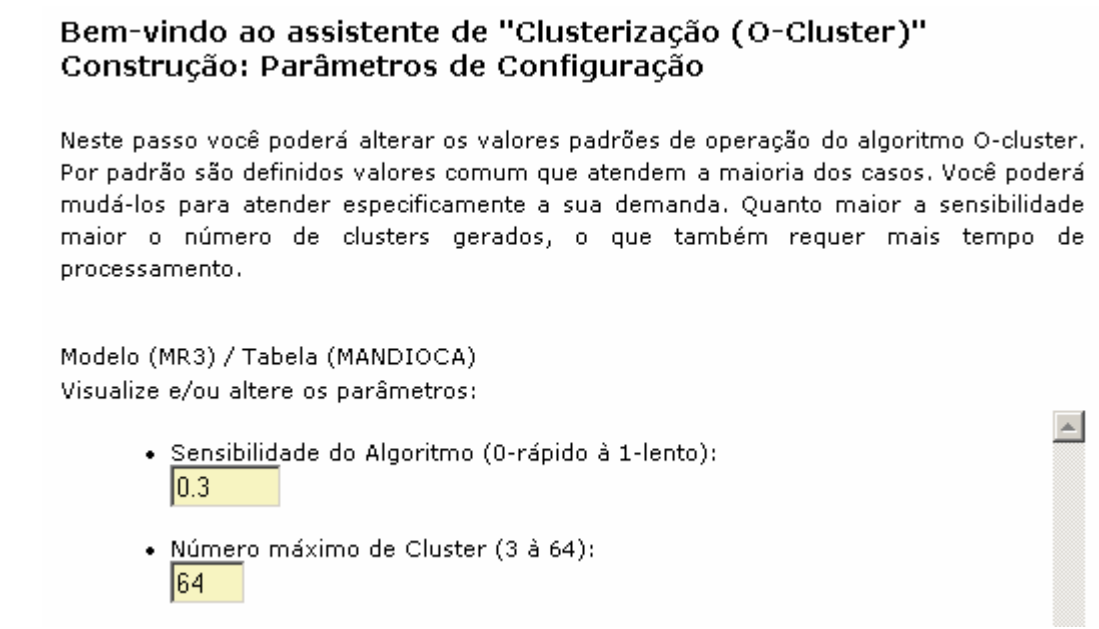


Figura 5.5: Modelo O-cluster - Parametrização

A tabela a seguir mostra os dados de sensibilidade e número de grupos formados, verificamos que aumentando a sensibilidade do algoritmo o número de grupos cresce.

Tabela 5.1: Valores de Sensibilidade e Número de Grupos

Modelo	**MR1**	**MR2**	**MR3**	**MR4**	**MR5**
Sensibilidade	0,3	0,5	0,7	0,9	1,0
Grupos	9	12	18	42	52

5.3.4 Avaliação

Retomando o problema inicial, desejamos formar um grupo de acessos que possa representar o todo, no sentindo da diversidade genética. O especialista deverá ter em mente que o número de acessos que deseja disponibilizar é um bom indicador do número de grupos que pretende formar, por exemplo, se o especialista deseja disponibilizar 45 amostras é interessante trabalhar com os modelos MR4 ou MR5. Idealmente dispondo de um acesso por grupo gerado ou próximo disso. Porém é necessário avaliar a qualidade dos modelos gerados, antes de escolher um para aplicação.

Uma forma disponibilizada na ferramenta HaDog é a visualização em gráfico dos grupos formados em termos da quantidade de acessos por grupo. Também é fornecida uma série de informações sobre o grupo, entre elas o suporte e a confiança. Com esta visualização é possível verificar se os grupos tiveram ou não distribuição uniforme e se as regras formadas são confiáveis.

A seguir temos a analise gráfica para o modelo MR2, os detalhes da janela de mensagem são obtidos clicando-se na fatia do grupo desejado.

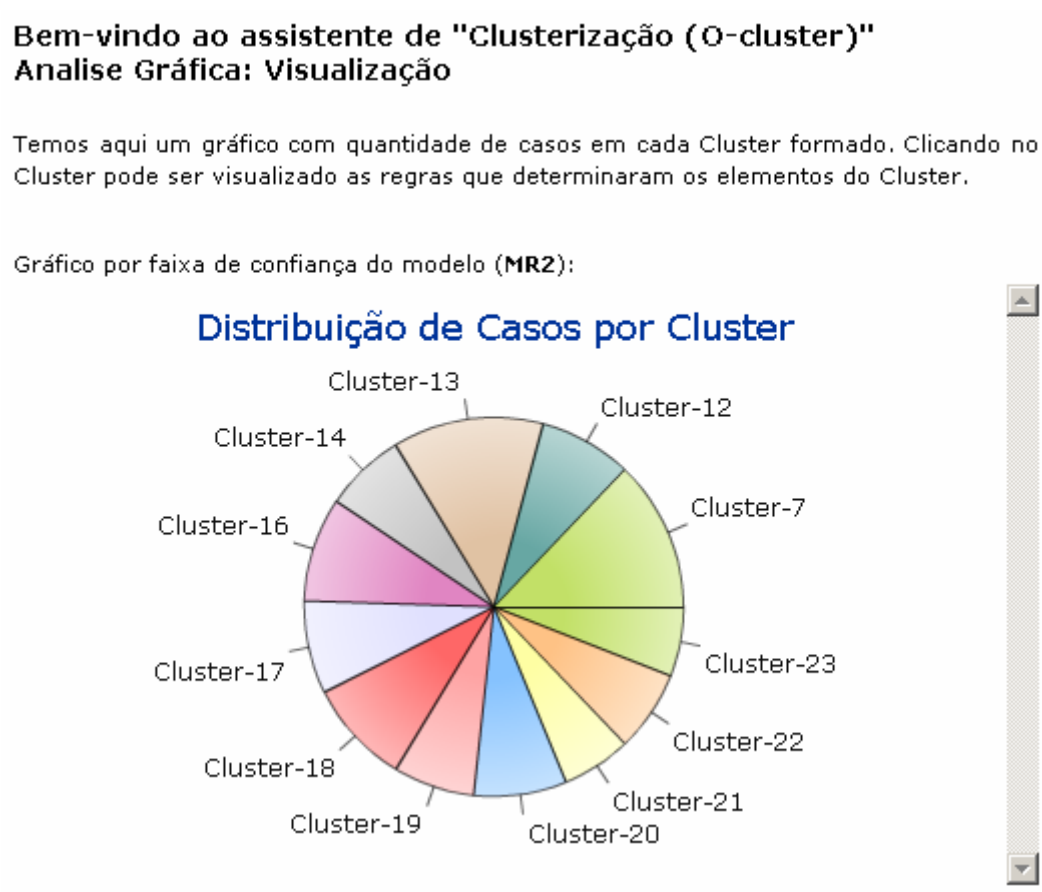


Figura 5.6: Avaliação – Gráfico de Distribuição por Cluster

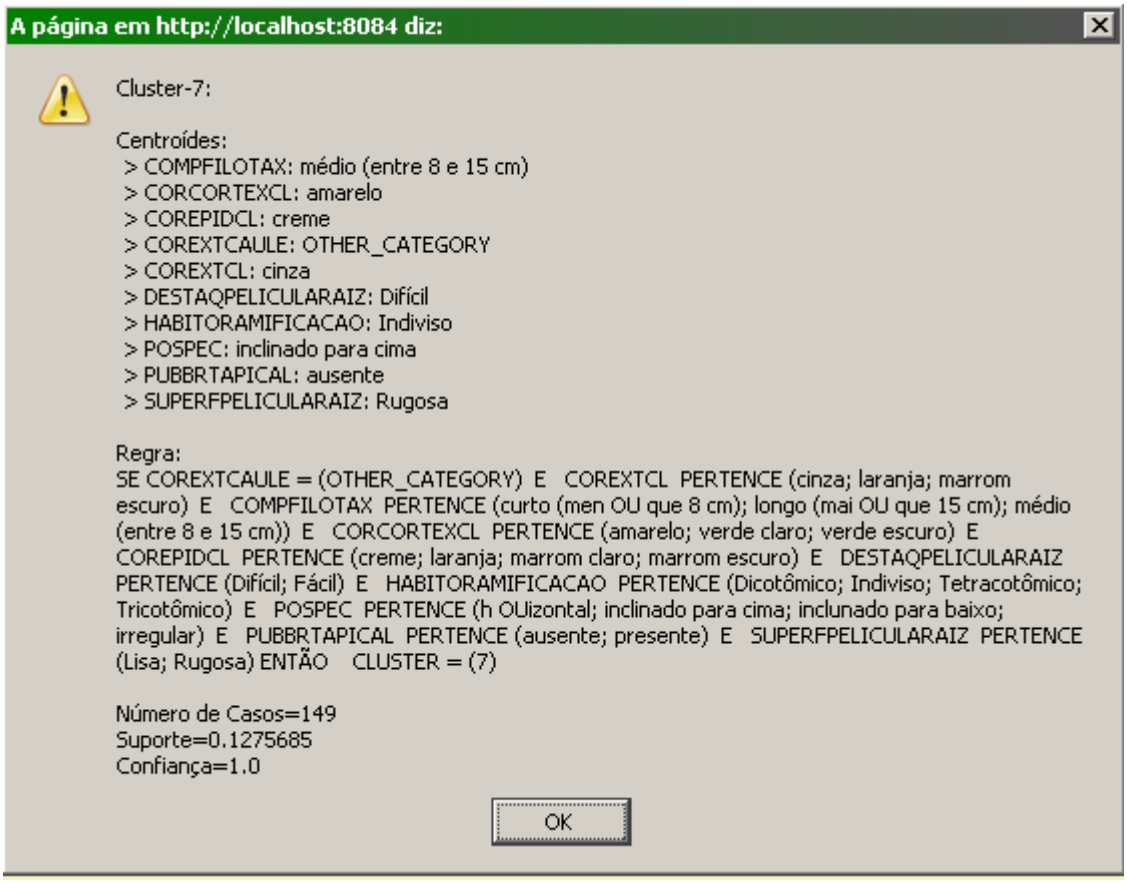


Figura 5.7: Avaliação – Detalhes de um Grupo

A outra opção de avaliação é um relatório sumarizado sobre o modelo gerado. Este relatório contém informações sobre todos os grupos e médias calculadas sobre os índices de suporte, confiança e número de elementos. Veja a seguir o relatório para o modelo MR2:

Bem-vindo ao assistente de "Clusterização (O-cluster)"
Relatório do Modelo: Visualização

Temos aqui um relatório sumarizado de um modelo de clusterização. Este relatório é trabalhado sobre as seguintes informações: quantidade de elementos, confiança e suporte.

Relatório do Modelo (**MR2**):

Numero de Clusters: 12

Variável	Média	Desvio Padrão
QTD.Elementos	97	27,5659
Média de Suporte:	0,0833	0,0236
Média de Confiança:	1	0

Cluster Folha	Qtd.Elementos	Suporte	Confiança
23	69	0,0591	1
22	79	0,0676	1
21	75	0,0642	1
20	89	0,0762	1
19	82	0,0702	1
18	102	0,0873	1
17	99	0,0848	1
16	100	0,0856	1
14	79	0,0676	1

Figura 5.8: Avaliação – Relatório do Modelo

Coletando dados dos relatórios dos modelos chegamos à tabela a seguir, que contém uma sumarização dos dados de avaliação dos cinco modelos gerados.

Tabela 5.2: Acessos Representativos - Dados de Avaliação

Modelo	Grupos	Média (Desvio Padrão)		
		QTD. Elementos	Suporte	Confiança
MR1	9	129 (30,52)	0,11 (0,02)	1,0 (0,0)
MR2	12	97 (27,56)	0,08 (0,02)	1,0 (0,0)
MR3	18	64 (13,83)	0,05 (0,01)	1,0 (0,0)
MR4	42	27 (5,32)	0,02 (0,00)	1,0 (0,0)
MR5	52	22 (4,35)	0,01 (0,00)	1,0 (0,0)

Analisando a tabela anterior verificamos que a distribuição dos elementos vai ficando cada vez mais uniforme com o aumento da sensibilidade e por conseqüência do número de grupos.

A confiança na regra de formação do grupo é grande, ou seja, caso o antecedente ocorra o conseqüente sempre ocorre. Neste caso, se um elemento tiver as características delimitadas no antecedente ele com certeza fará parte de um grupo específico. Portanto, não existem intersecções entre os grupos formados, podemos dizer que os modelos estão bem construídos.

Por estas analises deixamos a critério do especialista escolher o modelo, retornando ao parágrafo inicial desta subseção, a escolha pode ser feita com base no número de acessos que o especialista deseja disponibilizar, escolhendo o modelo cujo número de grupos seja próximo do número de acessos.

5.3.5 Colocação em Uso

Como citado estamos trabalhando com um conjunto padrão de amostras, estamos querendo selecionar 45 acessos. Para esta quantidade de acessos é adequado usamos os modelos MR4 ou MR5.

Iremos aplicar os dois modelos ao conjunto de dados e fazer uma comparação dos resultados obtidos. O especialista deve analisar os resultados e então escolher a aplicação que melhor atender aos objetivos de mineração.

Na ferramenta HaDog iremos aplicar os modelos ao conjunto de dados inicial que chamamos de “mandioca”. E fazer uma analise gráfica das duas aplicações.

Na aplicação iremos escolher todos os atributos, inclusive aqueles que não foram utilizados na criação dos modelos. Isto é importante para identificar os acessos dentro dos grupos.

A seguir temos uma figura em mosaico, mostrando a aplicação do modelo MR4 sobre o conjunto de dados de “mandioca”. São necessários quatro passos até chegar ao resultado final que será um novo conjunto de dados. Este será acrescido de dois atributos um número de grupo e da probabilidade da linha esta contida no grupo. Este mesmo processo deverá ser feito para o modelo MR5, que também é adequado ao nosso estudo de caso.

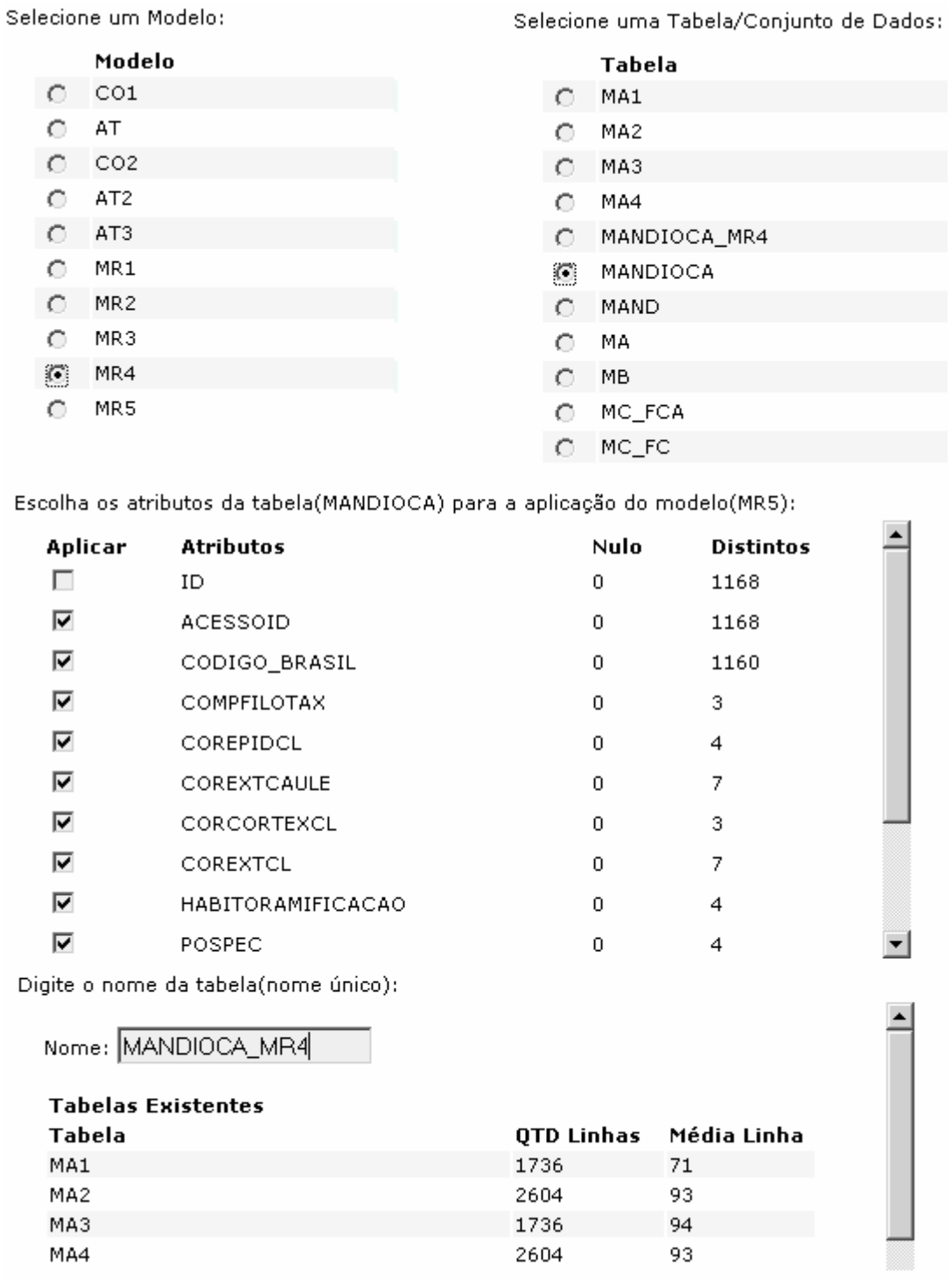


Figura 5.9: Colocação em Uso

Note que aplicamos o modelo ao mesmo conjunto de dados usado para gerá-lo, isso não impede que seja usado um novo conjunto de dados derivado de novas entradas de dados no SIBRARGEN.

Após a aplicação do modelo podemos visualizar os resultados gerados em duas novas tabelas “MANDIOCA_MR4” e “MANDIOCA_MR5”, resultantes da aplicação dos modelos “MR4” e “MR5” ao conjunto de dados “MANDIOCA”.

Escolhemos a subopção “Exportação Parcial” de “Colocação em Uso” para podemos configurar a quantidade de acessos que desejamos ter ao final da exportação. Os passos para exportação de uma aplicação de agrupamento no HaDog são:

- Escolha do modelo aplicado;
- Parametrização da saída, inclusive com o número de elementos desejado;
- Gravação da planilha.

Mostramos estes passos para o modelo MR4. Inicialmente devemos escolher qual modelo aplicado queremos exportar, só estarão presentes nas listagens os modelos que tiverem sido aplicados.

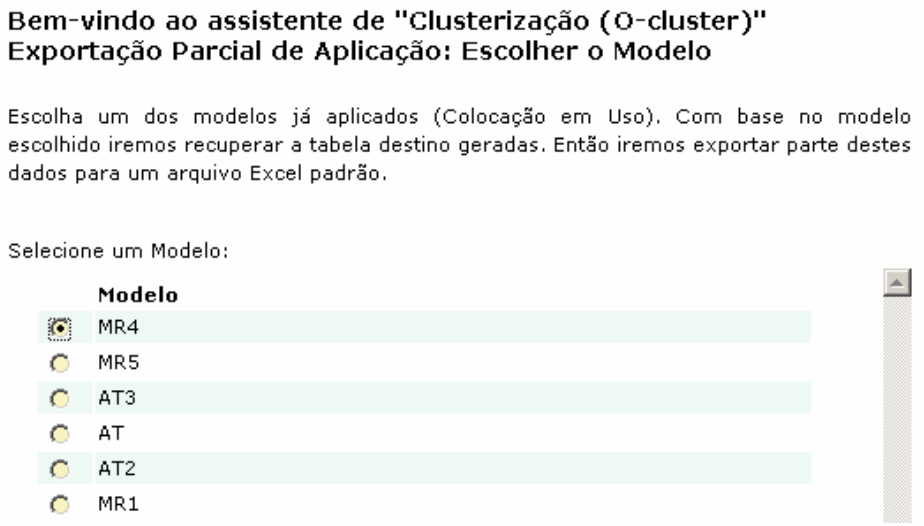


Figura 5.10: Exportação de Aplicação de Agrupamento - Modelo

Em seguida é mostrada uma janela para parametrização da exportação. Isso ocorre visto que escolhemos uma exportação parcial, nesta podemos determinar quantos elementos são desejados. A exportação irá considerar elementos distribuídos de forma uniforme entre os grupos, por exemplo, se o modelo gerou cinco grupos e desejamos exportar sete elementos teremos dois elementos dos dois primeiros grupos e um elemento dos demais.

A seguir iremos limitar o número de elementos em 45 o número previsto no nosso estudo de caso. Este valor é determinado pelo especialista conforma a demanda que o mesmo teve.

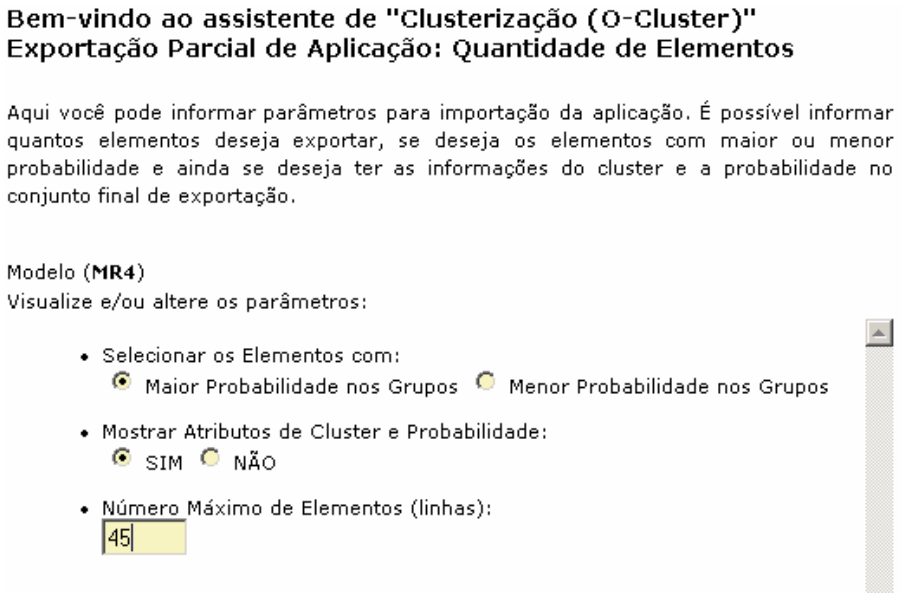


Figura 5.11: Exportação de Aplicação de Agrupamento - Parametrização

A primeira opção diz respeito ao tipo de elementos que queremos selecionar, elementos com valor de probabilidade alta ou baixa. Já a segunda opção é com relação ao conteúdo da planilha, se desejamos ou não que os atributos de processamento da aplicação (*cluster* e probabilidade) sejam incluídos na saída. E a última opção permite limitar o número máximo de linhas que serão exportadas. Aqui estamos chamando cada linha de elemento.

O resultado final da exportação é uma planilha no formato Excel® com os dados parciais da aplicação do modelo, conforme figura abaixo:

	A	B	C	D	E	F	G	H
1	O_NUM_CLUSTER	O_PROBABILIDADE	ACESSOID	CODIGO_BRASIL	COMPFILOTAX	COREPIDCL	COREXTCAULE	CORCORTEXCL
2	25	1	173.997	BRA-057754	médio (entre 8 e 15 cm)	marrom claro	Verde amarelado	verde escuro
3	30	1	173.553	BRA-098876	curto (menor que 8 cm)	marrom claro	Marrom escuro	verde claro
4	35	1	173.499	BRA-000671	curto (menor que 8 cm)	marrom escuro	Marrom escuro	amarelo
5	41	1	173.448	BRA-000957	médio (entre 8 e 15 cm)	marrom claro	Marrom escuro	verde claro
6	42	1	174.276	BRA-006939	longo (maior que 15 cm)	marrom claro	Marrom claro	amarelo
7	45	1	174.294	BRA-007129	longo (maior que 15 cm)	creme	Cinza	verde escuro
8	46	1	174.257	BRA-008061	longo (maior que 15 cm)	marrom escuro	Dourado	amarelo
9	48	1	173.412	BRA-001783	longo (maior que 15 cm)	creme	Prateado	verde claro
10	50	1	173.952	BRA-058084	médio (entre 8 e 15 cm)	marrom escuro	Cinza	verde claro
11	51	1	173.929	BRA-058602	curto (menor que 8 cm)	marrom claro	Cinza	verde escuro
12	52	1	173.916	BRA-058488	médio (entre 8 e 15 cm)	marrom escuro	Cinza	verde escuro
13	53	1	174.258	BRA-008079	longo (maior que 15 cm)	creme	Cinza	verde escuro
14	54	1	174.327	BRA-006661	longo (maior que 15 cm)	marrom escuro	Verde amarelado	amarelo
15	55	1	173.642	BRA-081663	curto (menor que 8 cm)	laranja	Verde amarelado	verde escuro
16	56	1	173.604	BRA-082015	médio (entre 8 e 15 cm)	creme	Dourado	amarelo
17	57	1	173.497	BRA-000655	curto (menor que 8 cm)	creme	Prateado	amarelo
18	58	1	174.320	BRA-006602	longo (maior que 15 cm)	marrom escuro	Marrom escuro	verde claro
19	59	1	173.681	BRA-081353	longo (maior que 15 cm)	laranja	Marrom escuro	verde claro

Figura 5.12: Exportação de Aplicação de Agrupamento - Planilha

Este resultado vai de encontro ao objetivo inicial do projeto de mineração. Que era listar uma determinada quantidade de acessos representativos, ou seja, aqueles que tiverem maior diversidade genética.

Usamos como indicador de diversidade os atributos de caracterização, que expressam o genótipo do acesso.

Utilizamos um algoritmo de agrupamento para gerar grupos em quantidade próxima ao número de acessos desejados. Este tipo de algoritmo nos permite formar grupos naturais

que caso o indicador de confiança seja alto, nos diz que a separação entre estes grupos é boa. Foi o que ocorreu neste estudo específico.

Nos modelos com valor de sensibilidade maior tivemos por conseqüência um número maior de grupos e estes tiveram uma distribuição de elementos mais uniforme, também por esta razão os modelos MR4 e MR5 foram escolhidos para aplicação. Dado que se for retirado um elemento de cada grupo teremos um representante igualitário.

Tanto o modelo MR4, quanto o MR5 possuem os requisitos para serem aplicados. Eles têm grau de confiança alto e o número de grupos gerados próximo do número de elementos desejados. O especialista preferiu o modelo MR4, criando uma heurística de escolher o modelo com maior número de grupos possível menor ou igual ao número de elementos desejados.

5.4 Considerações Finais

Como síntese dos resultados alcançados temos a derivação de uma sistemática para encontrar acessos representativos de uma espécie através do uso da ferramenta HaDog e baseado na metodologia proposta. Temos o seguinte algoritmo:

1. Linearizar os dados do Banco de Germoplasma considerado os módulos de passaporte (atributos de identificação), caracterização (Genótipo) e avaliação (fenótipo);

2. Verificar se o conjunto de dados extraído é adequado à mineração, se existem dados suficientes de caracterização ou avaliação, caso contrário abortar o projeto e retornar ao SIBRARGEN;

3. Tratar os valores faltantes (*missing values*) se existirem;

4. Tratar os valores extremos (*outliers*) se existirem;

5. Gerar modelos de agrupamento com os atributos de caracterização e avaliação, baseado em *K-means* se os dados forem numéricos e em *O-cluster* se forem categóricos; Estes modelos devem ter números variáveis de grupos, idealmente alguns próximos ao número de acessos que desejamos intercambiar;

6. Avaliar os modelos em termos dos indicadores de suporte e confiança. O suporte deve ser o mais uniforme e a confiança deve ser alta.

7. Selecionar um ou mais modelos para aplicação conforme avaliação;

8. Aplicar o modelo ou modelos selecionados ao conjunto de dados inicial, ou se for uma segunda aplicação aos novos dados;

9. Extrair os dados da aplicação, limitando a saída ao número de acessos desejados.

Como o ferramental necessário para exploração dos dados neste tipo de problema estão disponíveis no HaDog o especialista que tem acesso ao SIBRARGEN também terá acesso a estas ferramentas, pois é utilizado o mesmo esquema de autenticação.

A introdução de novos atributos, especialmente aqueles derivados de novas técnicas em genética possibilitará uma separação ainda mais precisa dos grupos segundo especialistas. A regra é quanto mais detalhados forem os atributos de caracterização e avaliação mais precisa será a mineração de dados.

A sistemática pode ser usada em outros bancos de germoplasma existentes no SIBRARGEN. Já que as entradas e saídas são similares, independente da espécie considerada.

Neste mesmo processo onde o curador, responsável pelo Banco de Germoplasma, fornece amostras a um melhorista, podemos ter um segundo momento, onde ao invés de selecionar acessos de grupos diferentes, os melhoristas irão requerer do curador acessos que sejam similares a alguns acessos que tiveram papel de destaque na pesquisa inicial.

Para esta tarefa pode ser usado o mesmo conjunto resultado da aplicação. A subopção "Exportação Completa" gera uma planilha com todos os acessos classificados por grupos.

Com esta planilha podem-se encontrar os acessos de destaque, descobrir quais grupos pertencem e então selecionar os outros acessos que fazem parte do grupo para promover o intercâmbio entre o curador e o melhorista.

Capítulo 6 Estudo de Caso: Coleta Direcionada

Este capítulo apresenta um outro estudo de caso em bases de germoplasma que pretende extrair da base de dados informações sumarizadas sobre locais de coleta de acessos com determinadas características desejadas. O objetivo é fornecer informações para que novas coletas sejam direcionadas a lugares geográficos mais propensos a abrigarem os acessos procurados. O processo de mineração será similar ao estudo de caso anterior. Na Seção 6.1 conceituamos os termos utilizados e contextualizamos o problema. Na Seção 6.2 temos o planejamento do projeto de mineração listando as atividades que serão executadas. Na Seção 6.3 temos a descrição das atividades executadas no HaDog segundo o planejamento feito na seção anterior. Por fim na Seção 6.4 apresentamos os resultados e considerações acerca do projeto de mineração.

6.1 Conceituação e Contextualização do Problema

Um banco de germoplasma é composto acessos, que são subvariedades específicas, normalmente de uma mesma espécie ou espécies afins. Uma das formas de aumentar o número de acessos de um banco de germoplasma é através da coleta.

Estas coletas ocorrem em expedições organizadas por instituições de pesquisas, como a Embrapa, e são direcionadas a locais onde por experiência ocorrem acessos da espécie desejada. Além de aumentar o número de acessos e por conseqüência a diversidade genética, a coleta pode servir para aquisição de acessos com características específicas.

Neste último sentindo a coleta pode se tornar uma tarefa aleatória, já que diante do cenário geográfico os acessos são encontrados ao acaso. Uma expedição com objetivos específicos podem demandar muito tempo e recursos senão for corretamente direcionada. Apesar da atividade de coleta envolver necessariamente o fator acaso é possível melhorar o planejamento com base em dados de acessos que já foram coletados.

Muitos acessos obtidos via processo de coleta contém atributos de localização geográfica. São três informações que consolidam a localização de onde o acesso foi encontrado: a latitude, a longitude e a altitude. Sendo importantes para a localização no plano cartesiano os dois primeiros atributos.

Os dados de localização geográfica normalmente são fornecidos quando o acesso foi obtido via coleta. A saber, temos outras formas de obtenção de um acesso, por exemplo, via processo de cruzamento, biotecnológico ou outro.

Os acessos obtidos recentemente têm dados de localização geográfica mais confiável, principalmente com o advento do GPS (*General Positioning System*), que tornou possíveis anotações de localização com erro mínimo. Porém mesmo os dados mais antigos são adequados para o tipo de descoberta que pretendemos alcançar.

Desejamos trabalhar no caso em que se pretende coletar acessos com determinadas características. Entende-se, característica, certo valor de atributo dos módulos de caracterização ou avaliação.

Com a premissa de objetivar coletar acessos específicos desejamos fornecer informações de que região ou regiões é mais provável encontrá-los. Entende-se aqui região como sendo um retângulo no espaço geográfico, formado por duas posições cartesianas, ou seja, dois pares de latitude e longitude.

A estratégia é obter os dados a partir da base do SIBRARGEN. Iremos fazer uma operação de linearização sobre os principais dados de passaporte e também sobre os atributos de interesse nos módulos de caracterização e avaliação.

A operação de linearização nos módulos de caracterização e avaliação é especial, pois requer o uso de metadados antes de obter os valores dos atributos. Esta operação esta automatizada no HaDog.

Após a linearização partiremos para algumas operações de preparação dos dados necessária por questões de natureza e qualidade dos dados de interesse e também por conta da necessidade de filtrar somente os acessos com as características desejadas.

De posse do conjunto final de dados podemos emitir um relatório preliminar. Este forneceria uma visão geral, com os dados dos acessos selecionados pelas características desejadas e onde foram encontrados.

É também possível trabalhar o conjunto de dados através de agrupamento, por exemplo, com o algoritmo de *K-means* poderemos gerar grupos com acessos afins em relação à localização. Grupos densos, ou seja aqueles em que os acessos têm valores próximos ao centróide seriam tratados com ênfase, pois indicam uma região onde foram encontradas grandes quantidades de acessos.

6.2 Planejamento da Mineração

Nesta seção iremos planejar o projeto de mineração de dados sob a ótica da metodologia proposta. Iremos como no estudo de caso anterior detalhar os passos que iremos executar, assim como listar as considerações importantes neste projeto de mineração de coleta direcionada.

O objetivo deste projeto de mineração é derivar uma sistemática capaz de auxiliar no processo decisório de coleta, no que diz respeito a direcionar qual rota ou regiões explorar quando a meta é buscar acessos com determinadas características. Cada subseção tratará de uma fase do projeto de mineração que passará por: compreensão dos dados, preparação dos dados, modelagem, avaliação e colocação em uso.

6.2.1 Compreensão dos Dados

Os dados utilizados para mineração serão adquiridos dos três módulos tratados, a saber, passaporte, caracterização e avaliação. Do passaporte é necessário obter os dados de identificação e de localização geográfica do acesso, assim como o tipo de obtenção (coleta, melhoramento ou procedimento biotecnológico). De caracterização e avaliação é necessário obter os atributos de interesse, ou seja, aqueles que determinam as características desejadas pelo coletor.

Esta pré-seleção de dados já é um filtro inicial, pois consideramos apenas os atributos que serão utilizados no projeto de mineração. Esta pré-seleção pode ser feita no processo de linearização automatizado no HaDog.

Como no estudo de caso anterior é necessário fazer uma vistoria sobre os dados obtidos via linearização, verificando a possibilidade de continuidade do projeto. É necessário primeiro verificar se os dados de interesse estão presentes, se existe informações sobre coleta, tanto do tipo de obtenção, quanto da localização geográfica, caso contrário o projeto de mineração é inviável.

6.2.2 Preparação dos Dados

Neste estudo de caso esta fase é mais trabalhada. Várias tarefas devem ser feitas para que os dados estejam preparados. Da fase anterior foram obtidos os dados de um determinado banco de germoplasma, os atributos já foram selecionados e uma vistoria inicial indicou que é possível continuar o processo de mineração.

Inicialmente devemos filtrar os dados selecionando as linhas que nos interessam. Devemos executar as seguintes filtragens:

- Do conjunto extraído devemos selecionar somente aqueles que são de coleta, ou seja, forma de obtenção igual à coleta (“COLE”).

- Sendo de coleta é necessário selecionar os acessos que tem as características procuradas. Neste passo dever ser feito o filtro sobre os atributos de caracterização e avaliação.

Após a filtragem será gerado um subconjunto de dados, este deve ser submetido à tarefa de tratamento de valores faltantes (*missing values*), se houverem. Escolhemos como estratégia eliminar as linhas onde tiverem valores faltantes para os dados de localização geográfica. Neste contexto outras estratégias não são adequadas levando em conta a semântica destes dados.

Caso o subconjunto de dados resultante seja grande, por exemplo, com mais de uma centena de acessos selecionados, pode-se aplicar a tarefa de tratamento de valores extremos (*outliers*). Estes dados podem ter sido originados de erros de digitação no SIBRARGEN e mesmo que não sejam erros, como o conjunto é relativamente grande podemos desprezá-los, já que o objetivo é indicar regiões que contenham um grande número de acessos e não a localização específica de um acesso.

A partir do subconjunto de dados gerado na fase de preparação de dados podemos proceder em duas linhas. Caso o subconjunto seja pequeno um relatório com os dados obtidos é a melhor forma para subsidiar o coletor, no caso de um subconjunto maior de dados podemos modelar com agrupamento tentando encontrar grupos densos, que significariam uma quantidade maior de acessos próximos de uma mesma região.

6.2.3 Modelagem

No caso do subconjunto de dados derivado da fase de preparação conter uma quantidade grande de acessos podemos processar os dados através da tarefa de agrupamento. O objetivo será encontrar grupos mais densos, ou seja, aqueles cujos elementos do grupo estejam mais próximos ao centróide.

O algoritmo de *K-means* é mais indicado para este tipo de processamento, visto que nossa métrica de quão bom é um grupo esta intimamente ligada à média de proximidade do centróide, podemos alternativamente usar o algoritmo *O-cluster*.

Usaremos os atributos de localização geográfica para alimentar o algoritmo e gerar o modelo. Depois podemos aplicar o modelo escolhido ao subconjunto com todos os atributos.

Na construção do modelo os atributos de interesse serão a latitude e a longitude. Gerado o modelo podemos verificar a composição do modelo, os indicadores de suporte e confiança. Se o grupo tem valor alto para suporte indica que muitos acessos fazem parte deste grupo, analisando a regra de formação do grupo poderemos definir uma região onde foram encontrados muitos acessos.

Se possível gere modelos com um número de *clusters* compatíveis com o número de municípios onde já foram feitas as coletas da espécie em estudo. E depois varie para outros modelos com número de *clusters* próximos. Esta heurística é proposta, visto que as coletas normalmente são direcionadas para determinado município.

Esta experimentação permitirá que os vários modelos e aplicações sejam analisados, para determinação de um par modelo-aplicação que possa descrever o problema.

6.2.4 Avaliação

O objetivo maior neste projeto de mineração é conhecer os dados que já estão na base, relacionando um novo conhecimento, que é regiões onde é provável encontrar acessos de determinado tipo.

Com base no objetivo anterior a avaliação dos modelos deve ser feita considerando um par modelo-aplicação. É importante verificar duas variáveis em relação ao grupo formado:

- O número de acessos
- O grau de densidade

É valido lembrar que não nos interessa um grupo que tenha alta densidade, mas um número pequeno de acessos. O contrário também é verdadeiro, um modelo com poucos grupos formados podem ter um número de acessos grande, mas a sua densidade não será boa.

Uma métrica de avaliação deve considerar as duas variáveis. Optamos por usar uma fórmula em que as latitudes e longitudes extremas são utilizadas para calcular a densidade do grupo.

Os vários pares modelos-aplicações gerados podem ser submetidos a esta analise e então selecionado para colocação em uso o par com maiores valores para a métrica calculada anteriormente.

6.2.5 Colocação em Uso

A colocação em uso neste caso é uma listagem com uma sumarização dos dados do par modelo-aplicação. Queremos fornecer uma visão ordenada, da mais densa para a menos densa, de regiões onde encontramos mais acessos. Esta listagem servirá de subsidio para tomada de decisão sobre qual rota de coleta tomar, dado que o objetivo é encontrar acessos com determinadas características.

6.3 Execução do Projeto de Mineração

Na seção anterior foi descrita uma sistemática levando em consideração a metodologia proposta, visando auxiliar no processo decisório de escolha de locais de coleta. Nesta seção a atividade chamada de coleta direcionada será aplicada sobre dados da espécie *Manihot esculenta Crantz*, popularmente conhecida como mandioca. Vale lembrar que a sistemática é similar para aplicação em outras espécies.

Utilizaremos a ferramenta HaDog para nos auxiliar no processo de mineração. Dividiremos as subseções nas opções encontradas na ferramenta, que são aderentes a metodologia proposta.

6.3.1 Compreensão

Iremos trabalhar sobre o mesmo conjunto de dados usado no estudo de caso anterior, que é composto por 1168 acessos de Mandioca. Estes acessos contêm dados de passaporte e caracterização. Os dados de passaporte estão preenchidos para os atributos obrigatórios, os atributos de localização geográfica (latitude e longitude) estão parcialmente preenchidos. Para os atributos de caracterização os dados estão preenchidos.

Para obter os dados necessários de mandioca iremos entrar na ferramenta HaDog e acessar a opção compreensão, depois escolher a subopção Linearização. Nesta subopção iremos da base do SIBRARGEN os dados para mineração. Queremos obter os seguintes dados:

- Dado de passaporte com o objetivo de identificar o acesso para posterior colocação em uso: Código no Brasil.
- Dados de passaporte de localização geográfica: Latitude e Longitude.
- Dados de caracterização: aqueles atributos que são de interesse por parte do especialista (biólogos ou agrônomos)

A seguir temos uma figura que demonstra a subopção de linearização:

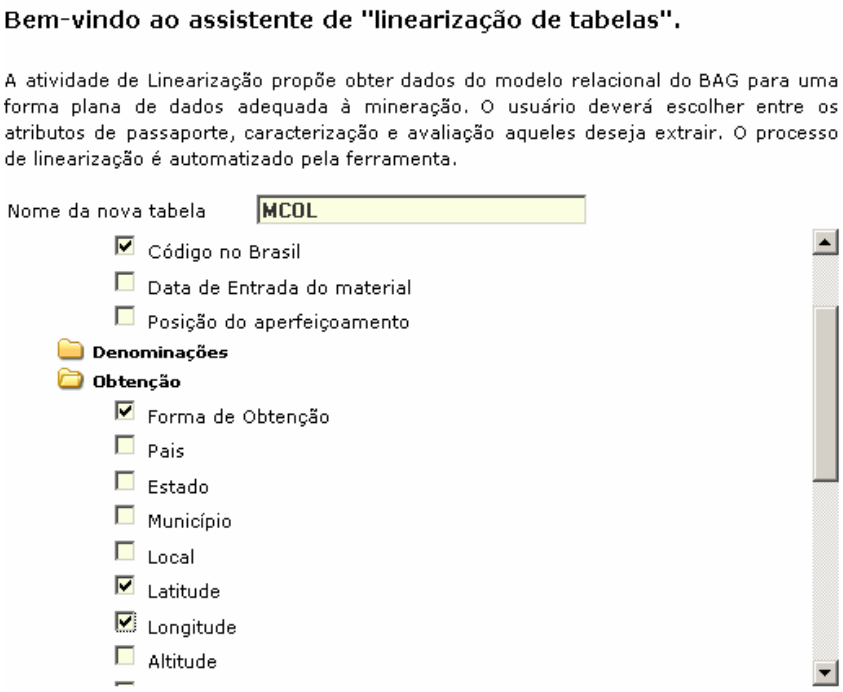


Figura 6.1: Linearização – Escolha dos Atributos

Além dos dados já descritos de passaporte foram escolhidos para linearização, por indicação do especialista, todos os atributos de caracterização para raiz da planta. O resultado da execução da Linearização pode ser visualizado na subopção "Visualização de Dados". Abaixo temos a Figura 6.2 com parte da visão gerada:

Conjunto de Dados: MCOL

CODIGO_BRASIL	FORMA_OBTENCAO	LATITUDE_DEGREE	LONGITUDE_DEGREE	CORPELICULARAIZ	CORPOLPARAIZ	CORCORTEXRAIZ
BRA-002127	COLE	0	0	Marron Escuro	Rosada	Roxa
BRA-002135	COLE	3,1333	58,4333	Marron Escuro	Rosada	Branco ou Creme
BRA-002143	COLE	0	0	Creme	Creme	Branco ou Creme
BRA-002208	MELH	22,85	43,7667	Marrom Claro	Branca	Roxa
BRA-002224	MELH	3,4333	77,5	Marrom Claro	Rosada	Branco ou Creme
BRA-002275	COLE	0	0	Marron Escuro	Creme	Roxa
BRA-002321	COLE	3,7833	39,2667	Marrom Claro	Branca	Branco ou Creme
BRA-002348	COLE	0	0	Marrom Claro	Rosada	Roxa
BRA-002356	COLE	0	0	Marron Escuro	Amarela	Rosado
BRA-002429	COLE	0	0	Creme	Creme	Branco ou Creme

Figura 6.2: Visualização – Parte dos Dados de Mandioca Linearizados

Outra visão possível dos dados é obtida através da subopção "Visualizar Resumo". Nesta subopção temos uma primeira sumarização dos dados com uma estatística descritiva de valores nulos, maior, menor, media ou moda. É também possível acessar um histograma com dados da distribuição das categorias na população total.

Os valores zero para os dados de localização geográfica devem ser considerados como faltantes segundo o especialista.

Tabela "MCOL" / "1168" linhas:

CODIGO_BRASIL	0(0%)	BRA-000027	BRA-109592	BRA-001856(2)
FORMA_OBTENCAO	41(3,5103%)	COLE	MELH	COLE(974)
LATITUDE_DEGREE	318(27,226%)	0,0333	29,5833	10,142
LONGITUDE_DEGREE	318(27,226%)	28,9167	77,65	43,994
CORPELICULARAIZ	0(0%)	Creme	Marron Escuro	Creme(397)
CORPOLPARAIZ	0(0%)	Amarela	Rosada	Rosada(310)
CORCORTEXRAIZ	0(0%)	Amarelo	Roxa	Amarelo(308)
COREXTRAIZ	0(0%)	Amarela	Marrom escura	Marrom clara(303)
DESTAQPELICULARAIZ	0(0%)	Difícil	Fácil	Difícil(615)

Figura 6.3: Visualizar Resumo – Estatística Descritiva

Podemos observar que os atributos extraídos de caracterização são para mandioca categóricos. O Código no Brasil é o identificador único dentro do conjunto considerado. Também observamos pela sumarização anterior que temos valores faltantes para os dados de localização geográfica, isto deverá ser tratado na etapa de preparação de dados.

6.3.2 Preparação

Na atividade de coleta direcionada a preparação dos dados é muito explorada. Além do tratamento de dados que não estão em conformidade com a atividade de mineração de dados é necessário fazer uma filtragem para selecionar o subconjunto de dados de interesse.

Caso o subconjunto de dados resultante seja pequeno não há necessidade de continuar a mineração. Deve-se neste caso gerar uma listagem com os dados resultantes desta etapa. Esta listagem será analisada visualmente pelo especialista.

Filtragem de Dados

Existem dados faltantes para os atributos de localização geográfica. Não é possível inferir um valor razoável para estes atributos segundo o especialista. No caso de não existir dados suficientes será necessário reconsiderar a voltar no processo de documentação e preencher os dados no SIBRARGEN para então trabalhar no HaDog.

Portanto, a estratégia será eliminar as linhas (elementos) que contenham valores faltantes para os atributos de localização geográfica.

Uma forma alternativa de eliminar os valores faltantes é filtrando o conjunto de dados e selecionando somente aqueles que tem valor.

É também importante selecionar somente os acessos que são de coleta (COLE), pois este é o foco do estudo. A seguir temos as condições que comporão o filtro que será executado neste estudo de caso, os três primeiros são comuns para qualquer atividade de coleta direcionada, as outras foram propostas, no sentido de exemplificação pelo especialista, que tentou selecionar mandiocas que normalmente tem bom paladar.

- O atributo forma de obtenção (forma_obtencao) deve ser “COLE”;
- A latitude (latitude_degree) deve ser maior que zero;
- A longitude (longitude_degree) deve ser maior que zero;
- Cor da película da raiz (corpelicularaiz) deve ser creme ou marrom claro;
- Cor do córtex da raiz (corcortexraiz) deve ser amarelo;
- Cor da poupa da raiz (corpolparaiz) deve ser amarela;
- Superfície da película da raiz (superfpelicularaiz) deve ser lisa;

A execução desta filtragem ocorre em quatro passos na subopção “Filtragem Interativa de Dados”. No primeiro selecionamos a tabela conforma a Figura 6.4.

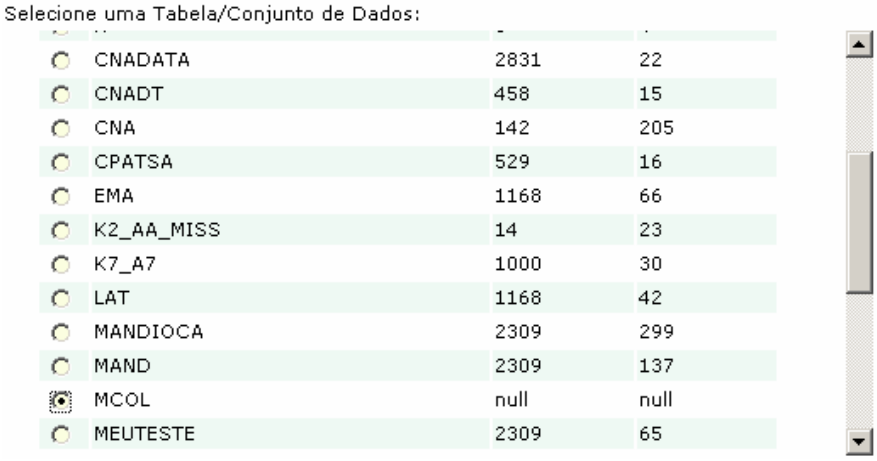


Figura 6.4: Filtragem Interativa – Escolha da Tabela

O segundo passo é sobre a montagem do filtro. É possível escolher mais de uma condição para o filtro. As condições podem ser unidas através dos operadores lógicos “e” ou “ou”. Para atributos categóricos é possível escolher mais de uma categória. Já para os atributos numéricos é possível selecionar pelos operadores de comparação básicos.

Ainda para os categóricos é apresentada uma caixa de escoha com as categórias encontradas na base e o número de ocorrências da mesma.

A Figura 6.5 mostra a filtragem para este estudo de caso.

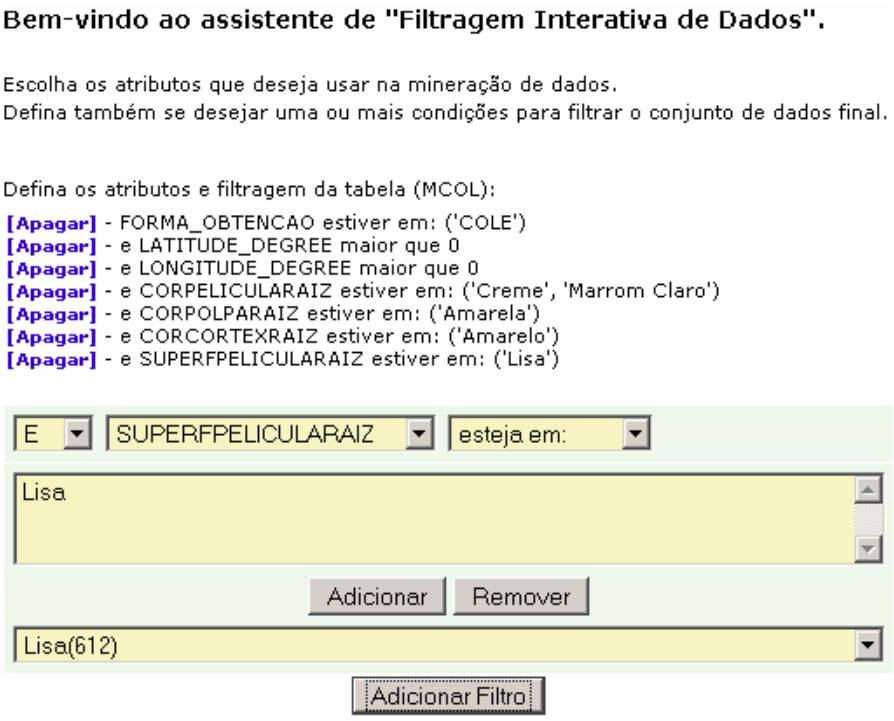
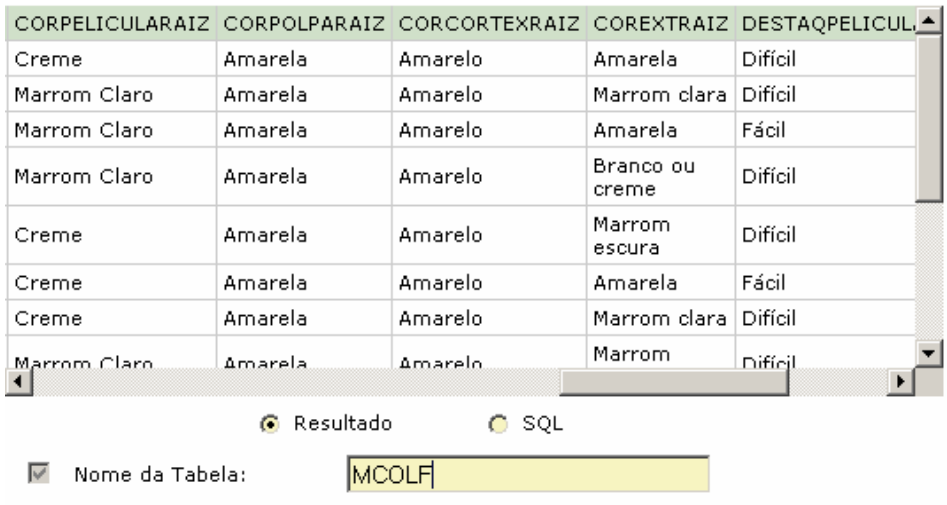


Figura 6.5: Filtragem Interativa – Montagem do Filtro

No terceiro passo temos a visualização dos dados selecionados após a filtragem. Dependendo das condições este conjunto de dados pode ser pequeno e passível de analise visual pelo especialista, não necessitando de continuidade no processo de mineração. No caso de um conjunto maior será possível aplicar a tarefa de agrupamento conforme será descrito na Subseção 6.3.3. A Figura 6.6 mostra o resultado da aplicação do filtro.



CORPELICULARAIZ	CORPOLPARAIZ	CORCORTEXRAIZ	COREXTRAIZ	DESTAQPELICUL
Creme	Amarela	Amarelo	Amarela	Difícil
Marrom Claro	Amarela	Amarelo	Marrom clara	Difícil
Marrom Claro	Amarela	Amarelo	Amarela	Fácil
Marrom Claro	Amarela	Amarelo	Branco ou creme	Difícil
Creme	Amarela	Amarelo	Marrom escura	Difícil
Creme	Amarela	Amarelo	Amarela	Fácil
Creme	Amarela	Amarelo	Marrom clara	Difícil
Marrom Claro	Amarela	Amarelo	Marrom	Difícil

Figura 6.6: Filtragem Interativa – Resultado da Filtragem

No quarto passo salvamos os dados resultantes do filtro em uma nova tabela que agora conterá somente o subconjunto desejado. Esta tabela pode ser exportata para Excel® através da opção “Compreensão > Exportação de Dados”.

Após obter o novo conjunto de dados retornamos a etapa anterior. Isto porque desejamos verificar se é necessário continuar o processo de mineração.

De volta a etapa de compreensão iremos analisar o novo conjunto de dados através da opção “Visualizar Resumo”. A Figura 6.7 mostra um resumo dos dados.

Tabela "MCOLF" / "13" linhas:

Atributo	Nulos	Menor	Maior	Média/Comum	Hist.
ID	0(0%)	169	1.154	714,6923	
ACESSOID	0(0%)	173.667	174.442	174.062,9231	
CODIGO_BRASIL	0(0%)	BRA-005681	BRA-081175	BRA-005681(1)	
FORMA_OBTENCAO	0(0%)	COLE	COLE	COLE(13)	
LATITUDE_DEGREE	0(0%)	3,7833	20,1	11,1128	
LONGITUDE_DEGREE	0(0%)	35,2	44,4167	38,6679	
CORPELICULARAIZ	0(0%)	Creme	Marrom Claro	Marrom Claro(7)	
CORPOLPARAIZ	0(0%)	Amarela	Amarela	Amarela(13)	
CORCORTEXRAIZ	0(0%)	Amarelo	Amarelo	Amarelo(13)	

Figura 6.7: Visualizando Resumo do Resultado da Filtragem

O atributo “forma_obtencao” para este problema deve ser sempre “COLE”, pois indica que selecionamos somente os acessos obtidos via processo de coleta.

Um segundo ponto de destaque é o número de linhas selecionado, bastante reduzido neste estudo de caso, encontramos apenas treze acessos com as características desejadas.

Segundo o especialista esta será uma realidade em muitos casos, pois a coleta é motivada quando existe a necessidade de aumentar a variabilidade genética do banco de germoplasma, justamente naquelas características que são mais raras.

Com um número reduzido de acessos uma analise visual do conjunto de dados resultante é suficiente para atender a demanda, se for necessário é possível extrair os dados para uma planilha em Excel®. A Figura 6.8 mostra a listagem completa dos dados filtrados.

Conjunto de Dados: MCOLF

CODIGO_BRASIL	FORMA_OBTENCAO	LATITUDE_DEGREE	LONGITUDE_DEGREE	CORPELICULARAIZ	CORPOLPARAIZ	CORCORTEXRAIZ	COREXTRAIZ	DESTAQPELICULARAIZ	SUPERFPELICULARAIZ
BRA-080900	COLE	7,7833	39,9333	Creme	Amarela	Amarelo	Amarela	Difícil	Lisa
BRA-080985	COLE	7,8667	38,7667	Marrom Claro	Amarela	Amarelo	Marrom clara	Difícil	Lisa
BRA-074799	COLE	20,1	35,2	Marrom Claro	Amarela	Amarelo	Amarela	Fácil	Lisa
BRA-006047	COLE	6,1833	35,8	Marrom Claro	Amarela	Amarelo	Branco ou creme	Difícil	Lisa
BRA-006971	COLE	11,25	37,6167	Creme	Amarela	Amarelo	Marrom escura	Difícil	Lisa
BRA-005991	COLE	12,6667	39,1	Creme	Amarela	Amarelo	Amarela	Fácil	Lisa
BRA-072044	COLE	12,2333	44,4167	Creme	Amarela	Amarelo	Marrom clara	Difícil	Lisa
BRA-056651	COLE	18,7	39,85	Marrom Claro	Amarela	Amarelo	Marrom escura	Difícil	Lisa
BRA-007447	COLE	12,6667	39,1	Marrom Claro	Amarela	Amarelo	Marrom clara	Difícil	Lisa
BRA-005681	COLE	12,6667	39,1	Creme	Amarela	Amarelo	Branco ou creme	Difícil	Lisa
BRA-007170	COLE	10,6833	37,4167	Creme	Amarela	Amarelo	Amarela	Fácil	Lisa
BRA-056341	COLE	3,7833	39,2667	Marrom Claro	Amarela	Amarelo	Marrom escura	Difícil	Lisa
BRA-081175	COLE	7,8833	37,1167	Marrom Claro	Amarela	Amarelo	Branco ou creme	Fácil	Lisa

Figura 6.8: Resultado da Filtragem

Uma outra forma de abordar o problema é considerar o conjunto maior de informações, ou seja, estaremos interessados em buscar regiões com maior incidência de acessos daquela espécie. Esta abordagem também é possível com resultados de filtragens para espécies cujo número de acessos é maior que de mandioca, por exemplo: *Zea mays L.* (milho), *Phaseolus vulgaris L.* (feijão) e *Oryza sativa L.* (arroz).

Para continuidade do processo de mineração e demonstração da sistemática proposta neste estudo de caso iremos considerar o conjunto total de acessos de mandioca que são de coleta e que tem dados válidos para localização geográfica. O resumo deste conjunto de dados pode ser visto na Figura 6.9.

Tabela "MCOLC" / "707" linhas:

Atributo	Nulos	Menor	Maior	Média/Comum	Hist.
ID	0(0%)	2	1.166	575,355	
ACESSOID	0(0%)	173.349	174.515	173.949,3508	
CODIGO_BRASIL	0(0%)	BRA-000035	BRA-109592	BRA-006599(2)	
FORMA_OBTENCAO	0(0%)	COLE	COLE	COLE(707)	
LATITUDE_DEGREE	0(0%)	0,0333	29,5833	9,9401	
LONGITUDE_DEGREE	0(0%)	28,9167	64,7833	40,8005	
CORPELICULARAIZ	0(0%)	Creme	Marron Escuro	Creme(246)	
CORPOLPARAIZ	0(0%)	Amarela	Rosada	Rosada(201)	
CORCORTEXRAIZ	0(0%)	Amarelo	Roxa	Amarelo(192)	

Figura 6.9: Resumo dos Dados de Coleta de Mandioca Válidos

Se tentarmos fazer transformações sobre este conjunto de dados procurando determinar regiões onde ocorre mais acessos teremos uma visão difícil de analisar. Isto ocorre porque a quantidade de registros é grande.

Por exemplo, se criarmos uma nova tabela através da subopção “Preparação > Campo Calculado” arrendodando os valores de localização geográfica e depois aplicarmos uma agregação de dados teremos a visão da Figura 6.10.

COUNT_CODIGO_BRASIL	EXP_1	EXP_2
1	0	51
11	1	47
2	1	49
2	2	50
1	2	51
2	2	55
1	2	61
4	3	44
4	3	45
8	3	57

Figura 6.10: Tentativa de Descrição dos Dados por Sumarização

Na Figura 6.10 é mostrado o número de ocorrências de acessos (COUNT_CODIGO_BRASIL) na latitude “EXP_1” e na longitute “EXP_2”.

O resultado anterior derivado de tarefas de preparação de dados não é suficiente para explicar os dados da forma que desejamos. Primeiro não é possível combinar as duas variáveis de forma satisfatória para gerar a sumarização e segundo o SQL faz suas operações internas de forma booleana deixando a agregação rígida.

Uma opção é modelar usando a técnica de agrupamento. Iremos utilizar o algoritmo de *K-means* conforme explicado no planejamento.

6.3.3 Modelagem

Iremos criar modelos de agrupamento usando o algoritmo *K-means*. Através de agregações constatamos que existem 28 municípios onde foram coletados acessos de mandioca iremos criar três modelos todos utilizando como entrada os atributos de localização geográfica, o que iremos modificar é o número de *clusters*.

Chamaremos os modelos de mc1, mc2 e mc3. Para cada um destes modelos iremos escolher os atributos de localização geográfica e deixando desmarcados os outros conforme figura a seguir:

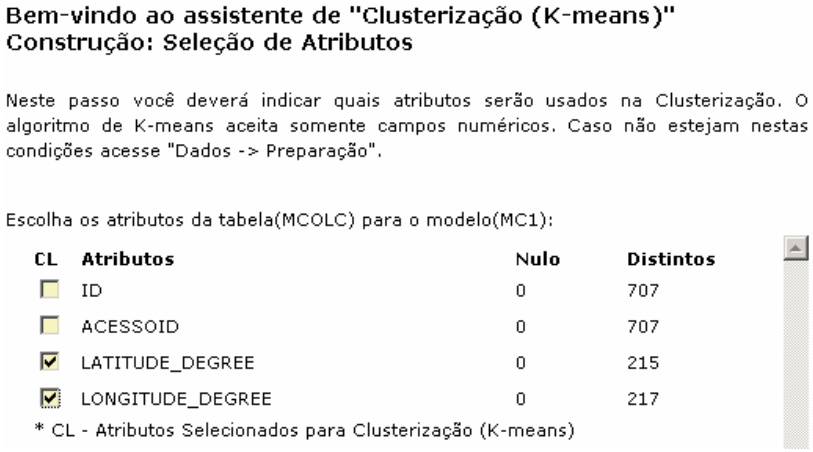


Figura 6.11: Modelo K-means - Atributos

Em um modelo de agrupamento não devemos incluir os atributos identificadores, assim como atributos que não contém valores distintos. Os atributos de interesse, atributos de localização geográfica, não são identificadores e os dois contêm valores distintos como verificamos na fase de compreensão dos dados.

Na parte de parametrização do modelo iremos modificar o parâmetro de número de c*lusters* para 24, 28 e 32 para os modelos mc1, mc2 e mc3 . A idéia é gerar um número de grupos próximos ao número de municípios visitados nas coletas.

No algoritmo de *k-means* é possível determinar exatamente o número de grupos que será gerado. Iremos deixar os outros parâmetros com os valores padrões.

Os ajustes nos outros parâmetros podem aumentar ou diminuir a sensibilidade do algoritmo, fazendo com que os elementos da intersecção de grupos sejam classificados em grupos diferentes conforme a parametrização. Aqui desejamos obter informações no nível macro, não necessitando deste ajuste fino.

6.3.4 Avaliação

Retomando o problema inicial, desejamos informações sobre uma região onde foram encontrados mais acessos de uma dada espécie.

No caso de um conjunto pequeno de dados derivado de uma filtragem não avançamos na modelagem e por isso não temos tarefas na avaliação.

Em caso de um conjunto maior de acessos podemos avaliar os modelos gerados através de um indicador já discutido no planejamento da mineração.

Calcular a area segundo nossa heurística:

$A = (Lat_{max} - Lat_{min}) + (Long_{max} - Long_{min})$

Se Area (A) menor que 1 então A=1

Calcular a avaliação para o grupo:

Ag = Ne / A, onde:

Ag: Métrica de Avaliação de um grupo
Ne: Número de elementos
A: Area

Calcular a média das avaliações do grupo:

Métrica = Média(Ag)

A comparação entre modelos pode ser feita com base na média dos valores de MA, já que cada grupo formado em um modelo terá um valor de MA. A Tabela 6.1 contém estas médias para os três modelos gerados.

Tabela 6.1: Coleta Direcionada - Dados para Avaliação

Modelos	Número de Grupos	Média da Métrica
mc1	24	6.62
mc2	28	6.52
mc3	32	8.47

Com base na métrica iremos escolher o modelo “mc3” para colocação em uso. Neste caso deverá existir uma sumarização da planilha exportada da aplicação.

6.3.5 Colocação em Uso

O HaDog no estágio atual de desenvolvimento permite a exportação de dados de uma aplicação de modelo. O usuário neste estudo de caso poderá analisar via Excel® os grupos do par modelo-aplicação “mc3”. Importante salientar que os dados do centróide dos grupos com maior número de elementos é um bom indicativo de região para encontrar novos acessos.

6.4 Considerações Finais

Como síntese dos resultados alcançados temos a criação de uma sistemática para encontrar regiões onde determinados acessos já ocorreram, dando subsídio para direcionar uma coleta. A sistemática é resumida no seguinte algoritmo:

1. Linearizar os dados do banco de germoplasma considerado os módulos de passaporte (atributos de identificação), caracterização (genótipo) e avaliação (fenótipo). Sempre incluir os dados de localização geográfica do passaporte;

2. Verificar se o conjunto de dados extraído é adequado à mineração, principalmente se existem dados de localização geográfica e se os mesmos são válidos;

3. Tratar os valores faltantes (*missing values*) se existirem;

4. Tratar os valores extremos (*outliers*) se existirem;

5. Fazer a filtragem dos dados pelos atributos de interesse;

6. Se o conjunto de dados for pequeno exportar o resultado da filtragem e finalizar o processo, caso contrário continuar;

7. Gerar modelos com os atributos de localização geográfica utilizando o algoritmo de *K-means.* A experimentação pode começar com modelos que tenham número de grupos próximo ao número de municípios onde já foram feitas coletas;

8. Avaliar os modelos em termos dos grupos formados, levando em conta o número de elementos e a densidade do grupo;

9. Extrair os dados da aplicação de forma completa e analisar os dados de centróides e números de elementos dos grupos formados.

A sistemática pode ser usada em outros bancos de germoplasma existentes no SIBRARGEN. Já que as entradas e saídas são similares, independente da espécie considerada.

É importante salientar que uma visualização gráfica em plotador de pontos como o ArcView® é um recurso interessante a ser usado sobre o conjunto final de posições de latitude e longitude derivados do projeto de mineração.

Capítulo 7 Conclusão

Esse capítulo descreve as conclusões e resultados obtidos e as linhas de trabalho futuro.

Para melhor contextualização, é apresentada a motivação, uma breve recapitulação do objetivo geral e das linhas de ação adotadas nesta pesquisa. Logo após, são apresentados os resultados obtidos e as contribuições. Também são explicitadas as limitações da solução proposta, assim como trabalhos futuros.

7.1 Motivação e Objetivos

Atualmente existe um sistema na Embrapa que capta informações de recursos genéticos denominado SIBRARGEN. Este sistema possui informações sobre germoplasma manipulado na empresa. Esta base tem um grande potencial de pesquisa em termos de mineração de dados que ainda não foi explorado.

Esta pesquisa visou documentar uma metodologia de mineração de dados aplicável a bases de germoplasma, contemplando as principais fases de um projeto de mineração e permitindo que especialistas da área (biólogos e agrônomos) apliquem técnicas de mineração com facilidade através de um ferramental Web intuitivo.

7.2 Estratégia Adotada

A metodologia proposta é baseada em CRISP/DM e abarca cinco das seis fases do CRISP/DM: compreensão dos dados, preparação dos dados, modelagem, avaliação e colocação em uso. A fase de entendimento do negócio foi suprimida porque o público alvo é constituído por especialistas do domínio.

Para facilitar o processo de mineração foi desenvolvida uma ferramenta Web denominada HaDog que materializa a metodologia proposta, contemplando as fases abordadas. As tarefas de cada fase foram materializadas em *wizards* para conduzir o usuário leigo em informática a executar atividades de mineração.

Como validação experimental da metodologia proposta foram executados dois estudos de casos sobre bases de germoplasma de *Manihot esculenta Crantz* (mandioca). O primeiro quer selecionar acessos representativos de uma espécie e o segundo pretende direcionar coletas de acessos no campo.

7.3 Resultados Obtidos e Contribuições

A prática e manipulação de dados de recursos genéticos pelos especialistas é incrementada com a adoção do HaDog, isto se constitui em uma contribuição para a comunidade científica em recursos genéticos, na medida em que a disseminação e utilização destes dados também é incrementada.

Outro ponto de contribuição é que os conceitos de documentação, implícitos nos atributos dos módulos de passaporte, caracterização e avaliação são expostos pela ferramenta.

Uma grande dificuldade em documentação de recursos genéticos é a padronização, que pode ser vista sob os aspectos de unificação de unidades de medida, utilização de nomenclatura única, tabelamento de dados possíveis para determinados atributos, conceituação de atributos. O HaDog permite o contato mais próximo de pesquisador com esta realidade, disseminando estas idéias, compartilhando dados e unificando atributos que são usados para caracterizar e avaliar as espécies vegetais.

Outra contribuição é alcançada através dos estudos de caso apresentados, que derivaram sistemáticas que podem ser aplicadas aos outros bancos de germoplasma, além do de mandioca, ou seja, foi possível determinar um macro-algoritmo que pode ser utilizado para outras espécies.

Este macro-algoritmo foi posto a prova, visto que os estudos de casos foram realizados com a participação do Curador de mandioca, pesquisador do CENARGEN, o qual validou, com base na experiência dele, os resultados obtidos.

Da ferramenta implementada já estão em uso pela comunidade as duas primeiras etapas de compreensão e preparação de dados, que envolvem automatização de atividades de captação, exportação e transformação de dados.

As exportações de dados para planilhas eletrônicas eram executadas pelos analistas de informática, demandando tempo que poderia ser usado em outras atividades. Agora os pesquisadores executam suas demandas e têm resposta imediata. Com base em informações extraídas do sistema de ordens de serviço foram solicitadas 32 extrações em setembro, 41 em outubro, 20 em novembro e apenas 3 em dezembro de 2007. Os módulos de compreensão e preparação dos dados do HaDog foi disponibilizado a partir de novembro.

Pelo levantamento bibliografico realizado e por informação de pesquisadores da área, a metodologia proposta constitui a primeira metodologia de mineração de dados em bases de germoplasma, contribuindo para a futura normatização e padronização do processo de mineração de dados neste domínio de conhecimento.

A implementação do HaDog constitui uma contribuição tecnológica neste domínio de recursos genéticos, por facilitar a realização das tarefas de mineração de bases de germoplasma em ambiente Web disponível diretamente ao usuário final. Pela especificidade dos usuários finais, em geral pesquisadores com formação em biologia ou agronomia, que não possuem familiaridade com a área de mineração de dados ou com programação, enfim o uso do HaDog contribui para introduzir a mineração de dados na área de recursos genéticos da Embrapa e dos parceiros.

7.4 Limitações e Trabalhos Futuros

O HaDog esta preparado para interagir com o SIBRARGEN somente para os módulos de passaporte, caracterização e avaliação. Em um momento posterior são importantes as integrações com outros módulos: intercâmbio e conservação.

Após a verificação do potencial de uso das tarefas de preparação de dados foram mapeadas novas atividades que podem ser automatizadas, tais como validação de dados em relação a tabelas existentes, conversões de dados, normalização de valores, entre outras.

Novos algoritmos podem ser disponibilizados na fase de modelagem. A escolha desses algoritmos está condicionada aos interesses específicos dos pesquisadores em germoplasma. Por exemplo, a tarefa de classificação foi implementada com o algoritmo *Naive Bayes*. Outros classificadores, baseados em algoritmos de árvore de decisão, poderão ser implementados aumentando a possibilidade de encontrar um modelo mais adequado.

As fases de avaliação e colocação em uso deverão receber ajustes e incrementos conforme novos estudos de caso forem surgindo. Idéias como área sobre curva ROC já foram citadas para compor o ferramental de avaliação. Na colocação em uso, podem ser disponibilizadas novas formas de formatação e filtragem das saídas das aplicações – tais como especificar as regras de associação que devem ser exibidas em função da indicação de um conjunto de atributos –, assim como permitir a aplicação dos modelos de predição em casos específicos. Atualmente, os modelos aprendidos podem ser aplicados apenas a tabelas contendo diversos casos de interesse.

Os algoritmos e telas envolvidas nos modelos de predição foram testados somente sob o ponto de vista da corretidute de implementação do algoritmo. Como trabalho futura queremos encontrar um estudo de caso real no contexto de recursos genéticos que possa validar na prática os modelos de predição implementados.

Bibliografia

Abras, G.; Ballarin, V. L. (2005). **A Weighted k-Means Algorithm Applied to Brain Tissue Classification**. Signal Processing Laboratory, School of Engineering, University Nacional de Mar del Plata.

Alsabti, K.; Ranka, S.; Singh, V. (2000). **An Efficient k-Means Clustering Algorithm**. University of Florida.

Brefeld, U. & Scheffer, T. (2005). **AUC Maximizing Support Vector Learning**. Proceedings of the ICML Workshop on ROC Analysis in Machine Learning.

Boley, D.L. (1998). **Principal Direction Divisive Partitioning**. Data Mining and Knowledge Discovery, v.2, n.4, p.325-344.

Carvalho, L. A. V. (2005). **Datamining**. Rio de Janeiro: Editora Ciência Moderna LTDA.

Chen, S.; Jeong, K. (2007). **Support Vector Regression Based GARCH Model with Application to Forecasting Volatility of Financial Returns**. China Center for Economic Studies, Fudan University.

Cheung, Y. (2003). **k-Means: A New Generalized k-Means Clustering Algorithm**. Department of Computer Science, Hong Kong Baptist University.

Chu, W.; Keerthi, S. S. (2006). **New Approaches to Support Vector Ordinal Regression**. Yahoo! Research Labs.

Daskalaki, S.; Kopanas, I. & Avouris, N. (2006). **Evaluation of Classifiers for an Uneven Class Distribution Problem**. Applied Artificial Intelligence. v.20, p.381-417

Faber, V. (1994). **Clustering and the Continuous k-Means Algorithm**. Los Alamos Science. Nº 22, p. 138-144.

Faiad, M.G.R.; Salomão, A.N.; Ferreira, F.R.P.; Gondim, M.T.P; Wetzel, M.M.V.S.; Mendes, R.A .; Goes, M. de. (1998). **Manual de procedimentos para conservação de germoplasma semente em longo prazo na Embrapa**, Brasília: Embrapa. P. 21. (Embrapa Recursos Genéticos e Biotecnologia. Documento, 30).

Faraoun, K. M.; Boukelif A. (2006). **Neural Networks Learning Improvement Using k-Means Clustering Algorithm to Destect Network Intrusions**. Département d´électronique, Djillali Liabès University.

Fayyad, U.M. (1997). **Editorial: Data Mining and Knowledge Discovery**. v.1 p.5-10.

Fayyad, U.M. (2004). (Editor). **Special Issue on Learning from Imbalanced Data** Sets. ACM SIGKDD Explorations. v.6.

Fernández, M. C.; Menasalvas, E.; Marbán, O.; Peña, J. M.; Millán, S. (2001). **Minimal Decision Rules Base don the Apriori Algorithm**. International Journal Application Math Computer Science. Vol.11, Nº 3, p. 691-704.

Ferri, C.; Flach, P. & Hernández-Orallo, J.H. (2002). **Learning Decision Trees using the Area under the ROC curve**. In C.S.A. Hoffman, editor, Nineteenth International Conference on Machine Learning (ICML'2002). Morgan Kaufmann Publishers. p.139–146.

Gama, J. & Brazdil, P. (2000) **Cascade Generalization**. Machine Learning. v.41 n.3 p.315-343.

Grünwald, P. (2005). **A Tutorial Introduction to the Minimum Description Length Principle**. Centrum Voor Wiskunde en Informatica.

Gunn, S. R. (1998). **Support Vector Machines for Classification and Regression**. University of Southampton.

Hart, P. E. (1968). **The Condensed Nearest Neighbor Rule**. IEEE Transactions on Information Theory IT-14. p.515–516.

Havold, J. (2005). **Naïve Bayes Spam Filtering Using Word-Position-Based Attributes**. Department of Computer Science, Lind University.

He, Z; Xu, X. & Deng, S. (2002). **Squeezer: an Efficient Algorithm for Clustering Categorical Data**. Journal of Computer Science and Technology. v.17, n.5, p.611-625

Hiragi, O. G.; Costa, S. R. I. (2001) **BAG - Banco de Germoplasma**. In: SIMPOSIO DE RECURSOS GENETICOS PARA AMERICA LATINA E CARIBE - SIGERALC, 3., Londrina. Recursos Genéticos: conservar para a vida - anais. Londrina: [s.n.].

Japkowicz, N. (2002). **Supervised Learning with Unsupervised Output Separation**. In Proceedings of the IASTED International Conference on Artificial Intelligence and Soft Computing (ASC). p.321-325.

Ladeira, M; Vieira, M.H.P; Prado, H.A; Noivo, R.M & Castanheira, D.B.S (2005). **UnBMiner - Ferramenta Aberta Para Mineração de Dados**. *Revista Tecnologia da Informação*, Brasília-DF, v.5, n.1, p.45-63.

Langley, P.; Iba, W. & Thompson, K. (1992). **An Analysis of Bayesian Classifiers**. In Proceedings of the 10th National Conference on Artificial Intelligence. AAAI Press and MIT Press. p.223-228.

Lopes, A. M. (2006). **O valor dos Recursos Genéticos**, Brasília: Embrapa, p. 31. (Embrapa Recursos Genéticos e Biotecnologia. Documento, 56).

Lowd, D.;Domingos, P. (2003). **Naive Bayes Models for Probability Estimation**. Department of Computer Science and Engineering, University of Washington.

MacQueen, J.B. (1967). **Some Methods for Classification and Analysis of Multivariate Observations**. Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press, v.1, p.281-297.

Mauser, A.; Bezrukov, I.; Deselaers, T. Keysers, D. (2004). **Predicting Customer Behavior Using Naïve Bayes and Maximum Entropy**. Lehrstuhl für Informatik VI, Computer Science Department RWTH Aachen University.

Merz, C.J. & Murphy, P.M. (1998) **UCI Repository of Machine Learning Datasets**. http://www.ics.uci.edu/~mlearn/MLRepository.html. (Acesso em 10/01/2008).

Milenova, B. L.; Campos, M. M. (2001). **Clustering Large Databases with Numeric and Nominal Values Using Orthogonal Projections**. Oracle Corporation.

Mitchell, T. (1997). **Machine Learning**. New York. McGraw Hill

Nickerson, A.; Japkowicz, N. & Millos, E. (2001). **Using Unsupervised Learning to Guide Resampling in Imbalanced Data Sets**. In Proceedings of the 8th International Workshop on AI and Statistics. Key West. p.261-65.

Oliveira, G.L. & Neto, M.G.M. (2004). **ExperText: Uma Ferramenta de Combinação de Múltiplos Classificadores Naive Bayes**. Anales de la 4ª Jornadas Iberoamericanas de Ingeniería de Software e Ingeniería de Conocimiento. Madrid. v.1, p.317-32.

Oracle Corporation (2003). **Oracle Database: Data Warehousing Guide, 10g Release 1**.

Oracle Corporation (2003). **Oracle Database: SQL Reference, 10g Release 1**.

Oracle Corporation (2003). **Oracle Data Mining: Administrator´s Guide, 10g Release 1**.

Oracle Corporation (2003). **Oracle Data Mining: Application Developer´s Guide, 10g Release 1**.

Oracle Corporation (2003). **Oracle Data Mining: Concepts, 10g Release 1**.

Rakesh, A.; Srikant, R. (2000). **Fast Algorithms for Mining Association Rules**. IBM Almaden Research Center.

Rish, I.; Hellerstein;J. Thathachar, J. (1998). **An Analysis of Data Characteristics that Affect Naïve Bayes Performance**. IBM T.J. Watson Research Center.

Romão, W. Niederauer, C. A. P.; Martins, A. Tcholajian, A. Pacheco, R. C. S.; Barcia, R. M. (2001). **Extração de Regras de Associação em C&T: Algoritmo APriori**. Programa de Pós-Graduação em Engenharia de Produção. Universidade Federal de Santa Catarina.

Shenzhi Li; Belapurkar, A. P.; Xiaoning Y., Dilsizian M. J.; Pottenger, W. M.; Ganiz, M. C.; Janneck, C. D. (2004). **Higher Order Apriori**. Lehigh University Department of Computer Science and Engineering 19 Memorial Drive West.

Smola, A. J.; Schölkopf, B. (2004). **A Tutorial on Support Vector Regression**. Kluwer Academic Publishers. Statistics and Computing 14, p. 199-222.

SPSS Inc.; NCR Systems Engineering Copenhagen & DaimlerChrysler AG (1999). **CRISP-DM 1.0 – Step-by-step Data Mining Guide**. SPSS & CRISP-DM Consortium. (Disponível em w. Acesso em 26/04/2006).

Vaidya, J.; Clifton, C. (2004). **Privacy Preserving Naïve Bayes Classifier for Vertically Partitioned Data**. University of Zürich.

Van Rijsbergen, C. J. (1979). **Information Retrieval**. 2ª Edição, London, Butterworths.

Webb,G.I.; Boughton, J.R.; Wang, Z. (2004). **Not So Naïve Bayes: Aggregating One-Dependence Estimators**. School of Computer Science and Software Engineering.

Wilson, D.R. & Martinez, T.R. (2000). **Reduction Techniques for Exemplar-Based Learning Algorithms**. Machine Learning. v.38, n.3, p 257-286.

Wu, C.; H. J.; Lee, D. (2004). **Travel Time Prediction with Support Vector Regression**. IEEE Trasactions on Intelligent Transportation Systems 5.4, p. 276-281.

Zhang, J.; Jin, R.; Yang, Y.; Hauptmann, A. G. (2003). **Modified Logistic Regression: An Approximation to SVM and Its Applications in Large-Scale Text Categorization**. School of Computer Science, Carnegie Mellon University.

Zhu, X. (2001). **Text categorization with Naive Bayes Classifiers**. International Journal Application Math Computer Science. Vol.11, Nº 3, p. 714-718.

Apêndice A Modelo Relacional do SIBRARGEN

Este apêndice serve para mostrar parte da estrutura do banco de dados do SIBRARGEN. Serão mostradas a estrutura dos módulos de passaporte, caracterização e avaliação. Na Seção A.1 temos modelos entidade-relacionamento simplificados. Na Seção A.2 temos a definição das principais tabelas com os atributos e tipos correspondentes.

A.1 Modelos Entidade-Relacionamento

A seguir temos os modelos entidade-relacionamento simplificados para os módulos de passaporte e caracterização-avaliação. As figuras foram retiradas da ferramenta ERWin®. Na Figura A.1 temos o modelo para passaporte e na Figura A.2 temos o modelo para caracterização-avaliação.

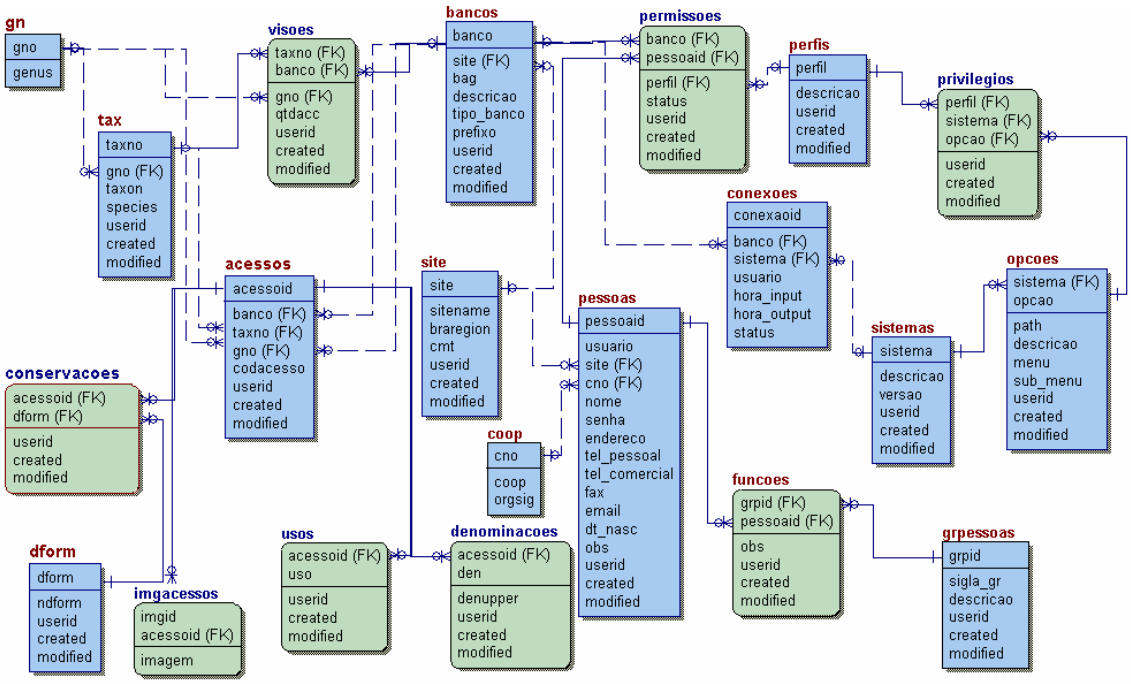


Figura A.1: MER Simplificado do Módulo de Passaporte

A tabela principal é “acessos” que é ó foco do sistema. Neste modelo também temos as tabelas do esquema de autenticação e privilégios dos usuários que são as tabelas à direita na figura.

Em azul temos as tabelas fortes e em verde temos tabelas associativas. O atributo ou os atributos que estão separados na primeira parte da caixa forma a chave primária da tabela. A marca “(FK)” indica que o atributo é uma chave estrangeira.

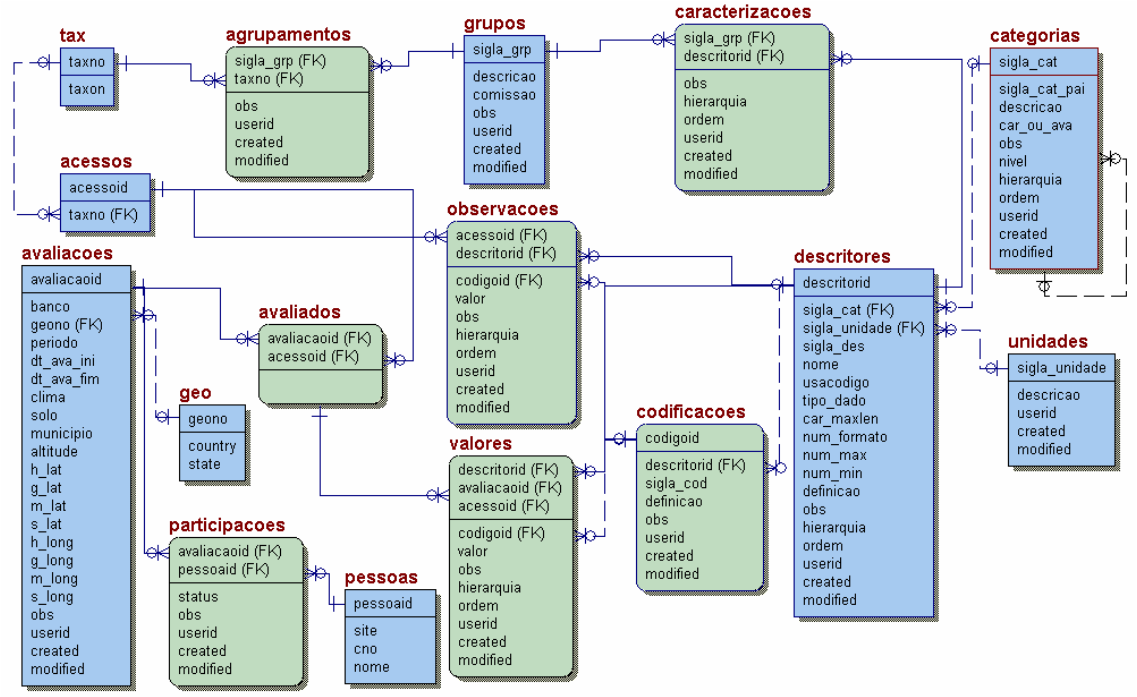


Figura A.2: MER Simplificado dos Módulos de Caracterização e Avaliação

Nas tabelas “descritores” e “codificacoes” temos os principais metadados. Em “descritores” temos as definições dos atributos para uma dada espécie. Em “codificacoes” temos as classes para cada atributo categórico.

Os dados de caracterização e avaliação são armazenados pelo sistema nas tabelas “observacoes” e “valores”, respectivamente.

A.2 Principais Tabelas

Nas figuras a seguir temos a descrição de três das principais tabelas dos módulos de passaporte, caracterização e avaliação. Esta definição foi retirada do produto SQL Developer da Oracle®.

Na Figura A.3 temos a tabela de acessos que contém os atributos de identificação, assim como seu histórico de origem e a características comuns aos acessos, independente da espécie considerada.

Column Name	Data Type	Nullable	Data Default
ACESSOID	NUMBER(8,0)	No	(null)
BANCO	VARCHAR2(20 BYTE)	No	(null)
CODACESSO	NUMBER(8,0)	Yes	(null)
ACID	NUMBER(8,0)	Yes	(null)
GNO	NUMBER(8,0)	No	(null)
TAXNO	NUMBER(8,0)	No	(null)
DTRECEB	DATE	Yes	(null)
LPROCED	NUMBER(8,0)	Yes	(null)
IPROCED	NUMBER(8,0)	Yes	(null)
ACIMPT	VARCHAR2(10 BYTE)	Yes	(null)
FOBTEN	VARCHAR2(4 BYTE)	Yes	(null)
LOBTEN	NUMBER(8,0)	Yes	(null)
MOBTEN	VARCHAR2(60 BYTE)	Yes	(null)
ALTOBTEN	NUMBER(5,0)	Yes	(null)
GLAT	NUMBER(3,0)	Yes	(null)
MLAT	NUMBER(2,0)	Yes	(null)
SLAT	NUMBER(2,0)	Yes	(null)
HLAT	VARCHAR2(1 BYTE)	Yes	(null)
GLONG	NUMBER(3,0)	Yes	(null)
MLONG	NUMBER(2,0)	Yes	(null)
SLONG	NUMBER(2,0)	Yes	(null)
HLONG	VARCHAR2(1 BYTE)	Yes	(null)
METODO	VARCHAR2(4 BYTE)	Yes	(null)
GENEALOGIA	VARCHAR2(240 BYTE)	Yes	(null)
COLBASE	VARCHAR2(1 BYTE)	Yes	(null)
CORE	VARCHAR2(1 BYTE)	Yes	(null)
DISPACC	VARCHAR2(1 BYTE)	Yes	(null)
ACREST	VARCHAR2(1 BYTE)	Yes	(null)
MORF	VARCHAR2(1 BYTE)	Yes	(null)
REPRO	VARCHAR2(1 BYTE)	Yes	(null)
GENETICA	VARCHAR2(1 BYTE)	Yes	(null)
NCARACT	VARCHAR2(1 BYTE)	Yes	(null)
OUTRACARFLAG	VARCHAR2(1 BYTE)	Yes	(null)
OUTRACAR	VARCHAR2(50 BYTE)	Yes	(null)
OUTRAAVAFLAG	VARCHAR2(1 BYTE)	Yes	(null)
OUTRAAVA	VARCHAR2(50 BYTE)	Yes	(null)
RENDIM	VARCHAR2(1 BYTE)	Yes	(null)
FBIOTICO	VARCHAR2(1 BYTE)	Yes	(null)
FABIOTICO	VARCHAR2(1 BYTE)	Yes	(null)
NAVAL	VARCHAR2(1 BYTE)	Yes	(null)
OBS	VARCHAR2(500 BYTE)	Yes	(null)
USERID	VARCHAR2(10 BYTE)	No	substr(USER,1,10)
CREATED	DATE	No	SYSDATE
MODIFIED	DATE	Yes	(null)

Figura A.3: Estrutura da Tabela de Acessos

Na Figura A.4 temos a tabela de descritores que armazena parte dos metadados utilizados para tornar genérico o sistema, ou seja, que possam ser definidos um conjunto diferente de atributos para cada espécie.

Column Name	Data Type	Nullable	Data Default
DESCRITORID	NUMBER(8,0)	No	(null)
SIGLA_CAT	VARCHAR2(20 BYTE)	No	(null)
SIGLA_DES	VARCHAR2(20 BYTE)	No	(null)
NOME	VARCHAR2(80 BYTE)	No	(null)
DEFINICAO	VARCHAR2(500 BYTE)	Yes	(null)
TIPO_DADO	VARCHAR2(1 BYTE)	Yes	(null)
SIGLA_UNIDADE	VARCHAR2(10 BYTE)	Yes	(null)
CAR_MAXLEN	NUMBER(2,0)	Yes	(null)
NUM_FORMATO	VARCHAR2(20 BYTE)	Yes	(null)
NUM_MIN	NUMBER	Yes	(null)
NUM_MAX	NUMBER	Yes	(null)
USACODIGO	VARCHAR2(1 BYTE)	No	(null)
OBS	VARCHAR2(500 BYTE)	Yes	(null)
USERID	VARCHAR2(10 BYTE)	No	substr(USER,1,10)
CREATED	DATE	No	SYSDATE
MODIFIED	DATE	Yes	(null)
HIERARQUIA	NUMBER	Yes	(null)
ORDEM	NUMBER	Yes	(null)
SIGLA_GRP	VARCHAR2(50 BYTE)	Yes	(null)

Figura A.4: Estrutura da Tabela de Descritores

Na Figura A.5 temos a tabela de observações que registram os dados de caracterização dos acessos. O SIBRARGEN utiliza os metadados de “descritores” e “codificacoes” especialmente para validar as entradas de dados na tabela de “observacoes”.

Column Name	Data Type	Nullable	Data Default
ACESSOID	NUMBER(8,0)	No	(null)
DESCRITORID	NUMBER(8,0)	No	(null)
CODIGOID	NUMBER(8,0)	Yes	(null)
VALOR	VARCHAR2(500 BYTE)	Yes	(null)
OBS	VARCHAR2(200 BYTE)	Yes	(null)
USERID	VARCHAR2(10 BYTE)	No	substr(USER,1,10)
CREATED	DATE	No	SYSDATE
MODIFIED	DATE	Yes	(null)
HIERARQUIA	NUMBER	Yes	(null)
ORDEM	NUMBER	Yes	(null)

Figura A.5: Estrutura da Tabela de Observações