UNIVERSIDADE FEDERAL DO PARANÁ

RENAN TANABE

COMPARAÇÃO DE MÉTODOS DE OTIMIZAÇÃO COLETIVA APLICADOS À MINERAÇÃO DE DADOS

CURITIBA

2016

RENAN TANABE

COMPARAÇÃO DE MÉTODOS DE OTIMIZAÇÃO COLETIVA APLICADOS À MINERAÇÃO DE DADOS

Trabalho apresentado como requisito parcial à obtenção do grau de Bacharel em Gestão da Informação no curso de graduação em Gestão da Informação, Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná.

Orientadora: Prof.ª Dra. Denise FukumiTsunoda

CURITIBA

2016

RESUMO

O presente trabalho compara métodos de otimização por inteligência coletiva aplicados à mineração de dados. Estes métodos quando aplicados à mineração de dados contribuem para a gestão da informação, visto a quantidade do volume de dados são geradas e armazenadas atualmente. O problema de pesquisa delimita-se a partir da questão de como comparar os métodos de otimização coletiva. Visto aausência de conteúdo acadêmico na área, o interesse do aluno pelo tema e contribuição para o Curso de Gestão da Informação, decidiu-se pela a realização do trabalho. Para responder ao problema, o trabalho contou com o aporte téorico de inteligência artificial e aprendizagem de máquinas, mineração de dados, otimização coletiva, métodos de otimização coletiva, método de colônia de formigas, método de enxame de partículas, método de cultura de bactérias, método de colônia artificiais de abelhas e método de cardume. Para a realização da comparação dos métodos foram selecionadas três bases de dados, sendo uma numérica, uma nominal e uma mista e processadas nas ferramentas de mineração de dados AntMiner e Weka. Utilizou-se de procedimentos metodológicos de pesquisa exploratória descritiva, método de classificação J48, método de colônias de formigas e método de selecionador de atributos (PSOSearch). A partir dos critérios estabelecidos, houve a análise e a comparação dos resultados.

Palavras chave: Métodos de otimização por inteligência coletiva, inteligência artificial, mineração de dados.

ABSTRACT

The present work has as general objective compare methods of optimization by collective intelligence applied to data mining. The collective optimization methods applied to data mining contribute to the information management, because the amount of data volume is generated and stored today. The research problem is delimited from the question of how to compare collective optimization methods. Considering the absence of academic content in the area, the student interest in the subject and contribution to the Information Management Course, it was decided to carry out the work. In order to answer the problem, the work counted on the theoretical contribution of artificial intelligence and machine learning, data mining, collective optimization, collective optimization methods, ant colony method, particle swarm method, bacterial culture method, artificial colony of bees method and fisch school method. In order to carry out the comparison of the methods, three databases were selected: one numerical, one nominal and one mixed and were processed in AntMiner and Weka data mining tools. It were used methods of descriptive exploratory research,J48 classification method, ant colony method and attribute selector (PSOSearch). From the established criteria the results were analyzed and compared.

Key words: Optimization by collective intelligence method, artificial intelligence, data mining.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

ACO - Ant Colony Optimization

ARFF - Attribute-Relation File Format

BFO - Bacterial Foraging Optimization

CSV - Comma Separated Values

IA - Inteligência Artificial

KDD - Knowledge Discovery in Databases

PSO - Particle Swarm Optimization

SUMÁRIO

1 INTRODUÇÃO

Em decorrência do crescimento de volume de dados em informações em formato digital, aliado à facilidade de compartilhamento destes pelos mecanismos disponibilizados pela *Word Wide Web*, o conhecimento de ferramentas e técnicas de organização e recuperação de dados, informações e conhecimentos se apresenta como um dos grandes desafios da atualidade. Neste contexto, a inteligência artificial vem sendo utilizada como alternativa para realização destas atividades, por exemplo, a aplicação em sistemas de informações.

Da mesma forma que diversas outras abordagens relacionadas à tecnologia, a utilização de inteligência artificialem organizações ainda está iniciando mas não faltam áreas em que já é possível visualizar o futuro em que será “normal” ver as máquinas inteligentes cooperando com o ser humano. Segundo o Estadão[1] (2016) diversos investimentos estão sendo realizados e afirmam que no ano de 2015 foram investidos um total de US$ 8,5 bilhões. De acordo com a pesquisa realizada pelo portal Insider Pro[2], grandes empresas como: Facebook, Apple, Google e Microsoft, apostam e investem muito nesta área.

Para o futuro, imagina-se que a inteligência artificial tenha aplicações especializadas tais como a redução de congestionamentos, com a chegada de carros autônomos. Outra área que é promissora para o futuro são as aplicações de IA na na robótica, onde é esperado que os robôs não só realizem trabalhos manuais, mas também façam“companhia” para uma pessoa[3].

Ao aprofundar o tema de inteligência artificial, pode-se relacionar com os métodos de otimização coletiva. Tais métodos foram propostos, poissão baseados no comportamento coletivo de sistemas auto-organizados, ditribuídos, flexíveis e dinâmicos (SERAPIÃO, 2009).Os métodos podem ser aplicados em diversas áreas, por exemplo: o controle de trajetória de robôs (CURKOVIC; JERBIC, 2007) que utiliza o método de colônia de abelhas, otimização de máquinas (KASHAN; KARIMI, 2009) aplicando o método de otimização por enxame de partículas, processamento

[1] Disponível em: <http://link.estadao.com.br/noticias/inovacao,inteligencia-artificial-caminha-a-passos-lentos,10000048402>. Acesso em: 13 jun. 2016.
[2] Disponível em: <https://insider.pro/pt/article/80392/>. Acesso em: 13 jun. 2016.
[3]Disponível em: <http://www.dca.fee.unicamp.br/~gudwin/ftp/publications/Dincom05_Gudwin>. Acesso em: 13 jun. 2016.

de imagens (MAITRA; CHARTTERJEE, 2008) através do método de otimização de cultura de bactérias, bioinformática (CHAN; FREITAS, 2006), no qual utilizam otimização por colônia de formigas.

Segundo Millonas (1994), as principais propriedades de escolha dos métodos de otimização coletiva, deve-se à:

a) aproximidade: os agentes são capazes de interagir;
b) qualidade: os agentes são capazes de avaliar seus comportamentos;
c) diversidade: permite ao sistema reagir a situações não-esperadas,
d) estalibidade: nem todas as alterações ambientais devem modificar o comportamento de um agente;
e) adaptabilidade: capacidade de adaptação a diversas condições ambientais.

Assim, esta pesquisa analisará os conceitos que permeiam os métodos de otimização que serão escolhidos para estudo, e na sequência realizará a comparação destes de forma empírica em bases de dados. É importante ressaltar que não será avaliada a complexidade algorítmica dos métodos.

1.1 PROBLEMATIZAÇÃO E QUESTÃO DE PESQUISA

Diante do cenário da inteligência artificial e inteligência coletiva, observou-se a oportunidade de estudar os métodos de otimização coletiva.

A definição da problematização para o presente trabalho, pode-se dar partir do momento em que os usuários dos métodos de otimização coletiva necessitem saber qual o método pode ser mais adequado para ser utilizado em determinadas situações, e neste momento deparam-se com a falta de trabalhos existentes que contenha tal tema.

Diante do exposto é possível definir a questão de pesquisa como: comocomparar osmétodos de otimização por inteligência coletiva aplicados à mineração de dados.

1.2 OBJETIVOS

O presente trabalho tem como objetivo geral comparar métodos de otimização por inteligência coletiva aplicados à mineração de dados.

Derivados do objetivo geral, os objetivos específicos definidos são:

- definir os métodos e ferramentas que serão objetos de comparação;
- escolher as bases de dados que serão utilizadas nos experimentos para comparação;
- submeter as bases às ferramentas / métodose registrar os resultados.

1.3 JUSTIFICATIVA

Os principais motivos para a escolha do tema proposto são a ausência de conteúdo acadêmico, o interesse pelo tema do pesquisador, visto que é um tema interdisciplinar, com argumentos relacionados totalmente com os conceitos de informação e tecnologia da informação, além de pode ser explorado por diferentes áreas do conhecimento.

A importância deste estudo se dá a partir do momento que a inteligência coletiva é uma maneira de valorização das competências individuais, onde se propõe a colocar em sinergia os indivíduos através do uso de tecnologias, a fim de que eles possam se juntar para que haja o compartilhamento da inteligência e conhecimento.

Outro fator importante para realização do trabalho foi o resultado da pesquisa pelo termo em inglês "*comparison methodology swarm optimization*" em bases de científicas. Na base Biblioteca Digital Brasileira de Teses e Dissertações (BDTD)[4], foi pesquisado o termo e retornou três resultados em que as abordagens eram voltadas Absorção de Corantes em Carvão Ativado, Otimização para Problemas de Roteirizarão de Veículos e Algoritmos para Formação de Grupos na Aprendizagem Colaborativa. Também foi realizada a pesquisa na base cientifica *Web Of Science*

[4] Disponível em: <http://bdtd.ibict.br/vufind/Search/Results?lookfor=comparison+methodology+swarm+optimization&type=AllFields>. Acesso em: 30 mai. 2016.

(WOS)[5], com refinamento de pesquisa pelo termo "*computer science*" retornando 64 resultados, porém os resultados abordando temas como: Otimização para Controle de Sensores, Otimização Baseada na Exploração do Espaço e Otimização em Eletromagnetismo. E por fim foi pesquisado utilizando a base de Periódicos CAPES[6], com refinamento de pesquisa pelo termo "*swarm intelligence*", e foram encontrados 20 trabalhos, porém com aplicações em Turbinas de Vento, Engenharia e Difusão de Umidade. Todos os resultados encontrados não condizem com o tema da pesquisa.

Pode-se colocar também como motivação para a realização do trabalho, a contribuição para o Curso de Gestão da Informação que, até o presente momento, não teve trabalhos de conclusão de curso abordando tal temática.

1.4 DELIMITAÇÃO DA PESQUISA

A presente pesquisa tem como proposta o estudo da otimização coletiva e alguns de seus métodos, pois, dentre os diversos propostos na literatura de base, apenas os que atenderem aos critérios estabelecidos neste estudo serão detalhados e comparados.

Sendo um trabalho de conclusão de curso de graduação em Gestão da Informação, o foco da comparação não será sob o ponto de vista de complexidade algorítmica.

1.5 ESTRUTURA DO DOCUMENTO

Este documento está organizado em cinco seções, sendo a primeira esta introdução. A segunda, todo o referencial teórico. A terceira seção os procedimentos metodológicos que foram utilizados neste trabalho, a quarta seção de análise de resultados e por fim, a quinta seção, as considerações finais do trabalho.

---

[5] Disponível em: <https://apps.webofknowledge.com/Search.do?product=UA&SID=3FxHQvucpSorjkCslWv&search_mode=GeneralSearch&prID=ba8a0070-fb19-468b-b0f3-26e3f275935d>. Acesso em: 30 mai. 2016

[6] Disponível em: <http://www-periodicos-capes-gov br.ez22.periodicos.capes.gov.br/index.php?option=com_pmetabusca&mn=88&smn=88&type=m&metalib=aHR0cDovL21scGx1cy5ob3N0ZWQuZXhsaWJyaXNncm91cC5jb20vcHJpbW9fbGlicmFyeS9saWJ3ZWIvYWN0aW9uL3NlYXJjaC5kbz9kc2NudD0wJmZyYmc9JnNjcC5zY3BzPXByaW1vX2NlbnRyYWxfbXVsdGlwbGVfZmUmdGFiPWRlZmF1bHRfdGFiJmN0PXNlYXJjaCZtb2RlPUJhc2ljJmR1bT10cnVlJmluZHg9MSZmbj1zZWFyY2gmdmlkPUNBUEVT&buscaRapidaTermo=comparison+methodology+swarm+optimization>. Acesso em: 30 mai. 2016.

2 REFERENCIAL TEÓRICO

A finalidade deste capítulo é apresentar abordagens teóricas que norteiam este trabalho e que auxiliam na orientação e no suporte ao que será proposto a seguir. Será referenciada a inteligência artificial e aprendizagem de máquina, inteligência coletiva, métodos de otimização coletiva, e por fim alguns métodos: colônia de formigas, exame de partículas, cultura de bactérias, colônia de abelhas e cardume.

2.1 INTELIGÊNCIA ARTIFICIAL E APRENDIZAGEM DE MÁQUINA

Os primeiros rastros da inteligência artificial nasceram na década de 40, sendo marcada pela II Guerra Mundial, o que acabou concluindo que se havia uma necessidade de projetar uma tecnologia para análise de balística, quebra de códigos e cálculos para projetar a bomba atômica. Desta forma, nascia os primeiros grandes projetos de construção de computadores.

Para Barr e Feigenbaum (1981) a inteligência artificial é parcela da ciência da informação que é responsável por projetar e criar sistemas computacionais inteligentes, isto é, sistemas que possuem características associadas com a inteligência no comportamento humano – por exemplo, raciocínio e resolução de problemas.

Segundo Charniak e McDermott (1987) afirmam que inteligência artificial é a aprendizagem das escolas por meio da utilização de modelos computacionais. E Kurzweil (1990) complementa, inteligência artificial é o conhecimento de construir máquinas que exercem funções, no qual exigem inteligência no momento em que são executadas por pessoas.

De acordo com Luger & Stubblefield (1993), inteligência artificial é o conjunto de adversidades e métodos estudados pelos cientistas de inteligência artificial.

A grande dificuldade da inteligência artificial é o entendimento da maneira que os seres humanos raciocinam, com o objetivo de aperfeiçoar os raciocínios através

de um conjunto de processos computacionais, com a finalidade de criar implicações da lógica matemática que reproduza o comportamento do cérebro humano.

Em relação à aprendizagem de máquina, Bishop (2007) afirma que pertence a uma área de inteligência artificial, com o objetivo de criar técnicas computacionais nos procedimentos de aprendizado.

Segundo Simon (1983), aprendizado pode ser visto como uma modificação num sistema que aprimore sua performance na segunda vez que for repetido a mesma atividade ou outra atividade da mesma população.

O aprendizado de máquina tem por sua finalidade a criação de teorias computacionais com o foco no desenvolvimento do conhecimento artificial. Softwares criados a partir desta tecnologia contêm a aspectos de tomar decisões baseando-se no aprendizado prévio acumulado por meio do contato com o ambiente.

2.2 MINERAÇÃO DE DADOS

A mineração de dados surgiu em meados da década de 1990, como área de pesquisa e aplicação independente, porém suas origens na matemática, estatística e computação são muito anteriores a este período.

Segundo Castro & Ferrari (2016) a mineração de dados explora uma base de dados utilizando algoritmos adequados para obter conhecimento. A mineração de dados é parte integrante de um processo amplo, conhecido como descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases* ou KDD).

De acordo com Thomé (2002) o KDD é realizado pelas seguintes etapas:

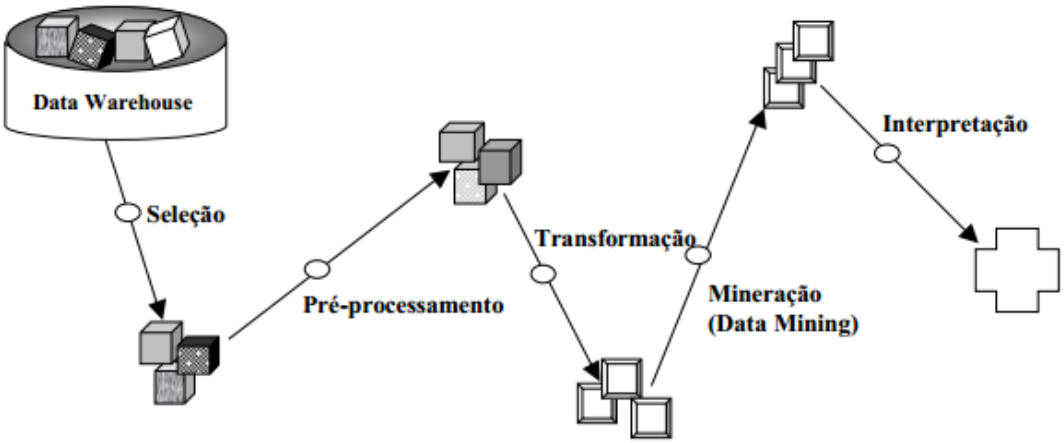
- seleção: etapa responsável em analisar e selecionar os dados a serem utilizados na busca por padrões e novos conhecimentos;
- pré-processamento: preparação e tratamentos dos dados que serão utilizados posteriormente pelos algoritmos. Nesta etapa os dados inválidos consistentes e redundantes devem ser identificados e descartados;
- transformação: aplicação de algum tipo de transformação linear ou até mesmo não linear nos dados, se necessário. Geralmente nesta fase são

aplicadas técnicas de redução de dimensionalidade e de projeção de dados;

- mineração: a etapa consiste em buscar encontrar padrões por meio de aplicação de técnicas e suas respectivas heurísticas;
- interpretação: etapa em que os dados minerados são analisados e interpretados para a geração de conhecimento e utilização dos resultados em benefício do negócio.

Na Figura 1, as etapas podem ser compreendidas de uma melhor forma.

FIGURA 1 - ETAPAS DO KDD



FONTE: Thomé (2002)

De acordo com Castro & Ferrari (2016)as funcionalidades da mineração de dados são divididas da seguinte forma:

- análise descritiva de dados: as análises descritivas permitem uma sumarização e compreensão dos objetos da base e seus atributos;

- predição – classificação e regressão: predição refere-se à construção e ao uso de um modelo para avaliar a classe de um objeto não rotulado ou para estimar o valor de um ou mais atributos de dado objeto. A classificação e regressão fazem parte dos dois principais tipos de problemas de predição, no qual a classificação é utilizada para valores discretos, e a regressão é utilizada para valores contínuos;

- análise de grupos: agrupamento *(clustering)* é o processo de particionar ou segmentar um conjunto de objetos em grupos de objetos similares. O agrupamento considera dados de entrada não rotulados. No processo de agrupamento os objetos são agrupados com o objetivo de maximizar a distância interclasse e minimizar a distância intraclasse. O agrupamento pode ser definido como uma coleção de objetos similares uns aos outros e dissimilares aos objetos pertencentes aos outros agrupamentos;

- associação: corresponde à descoberta das regras de associação que apresentam valores de atributos que ocorrem concomitantemente em uma base de dados;

- detecção de anomalias: as anomalias são objetos que não seguem o comportamento ou não possuem a característica comum dos dados ou de um modelo que os representem. As anomalias podem ser detectadas de diversas formas, incluindo métodos estatísticos que assumem uma distribuição ou modelo de probabilidade dos dados, ou medidas de distância por meio das quais objetos substancialmente distantes dos demais são considerados como anomalias.

2.3 OTIMIZAÇÃO COLETIVA

O comportamento social de diferentes espécies tem despertado a curiosidade de pesquisadores de diversas áreas nas últimas décadas. O que caracteriza estas sociedades é o fato de muitas tarefas e atividades são executadas em conjunto buscando um determinado objetivo que auxilie toda população.

Otimização coletiva foi introduzida em 1989 e descreve a habilidade de variadas espécies de animais que sem um líder, atuam de maneira a ampliar suas chance de sobrevivência. Este tipo de comportamento coletivo é estudado nos dias de hoje nos ambientes naturais de espécies como: colônias de formigas e abelhas, bando de pássaros, cupins, crescimento de bactérias, cardumes e até mesmo multidões de seres humanos. O motivo de estudo de comunidade de animais é pelo

fato de descreverem comportamentos complexos e sofisticados somente quando estão em grupo (PINHEIRO, 2008).

A otimização coletiva é parte de um sistema onde o comportamento coletivo de indivíduos relacionam-se localmente dentro do seu ambiente, possibilita o surgimento de um modelo funcional e coerente. A otimização coletiva fornece uma base de maneira permitir a exploração coletiva ou distribuída, para solução de problemas sem um controle centralizado ou necessidade de um modelo global (MOEDINGER, 2005).

Neste assunto, Dorigo *et al.* (1999) cita nos anos 50 foi proposto o conceito de *Stigmergy,* no qual mostrou que a informação origina-se do ambiente local e o trabalho em crescimento consegue conduzir as atividades individuais. Tal processo permite uma coordenação quase ideal do trabalho coletivo e nos da impressão que a colônia está seguindo um plano bem determinado.

Para exemplificar a otimização coletiva nos dias de hoje, pode-se citar o *Google* no qual a otimização coletiva da web é usada para definir a importância da página. Outro exemplo para ser citado, é o *Wikipedia*, onde na medida que permite a edição coletiva de verbetes e links hipertextuais (MILLER, 2007).

2.4 MÉTODOS DE OTIMIZAÇÃO COLETIVA

No estudo da otimização coletiva, pode-se perceber variados tipos de métodos. A seguir será exposto os métodos escolhidos para estudo neste trabalho.

2.4.1 Colônia de formigas

O método de otimização por meio de colônia de formigas (*Ant Colony Optimization – ACO*), foi proposto por Marco Dorigo, em 1992, baseado no estudo do comportamento das formigas.

De acordo com Ribeiro (2013) o método foi inspirado na observação de formigas na natureza e foram realizados experimentos com formigas reais. Nestes experimentos observou-se que diferentes espécies de formigas depositavam uma substância denominada de feromônio por onde passam, quando estão em busca de uma fonte de alimento. O feromônio funciona como uma fonte de comunicação de

uma formiga com as demais que também estão a procura de uma fonte de alimento. A presença do feromônio influencia na escolha do caminho a ser trilhado pelas outras formigas, ou seja, tendem a seguir caminho com maior concentração de feromônio. O feromônio depositado faz um caminho que permite às formigas encontrarem boas fontes de alimento que haviam sido encontradas anteriormente pelas outras formigas.

Deneubourg (1990) realizou um experimento utilizando formigas reais, onde foi colocado um ninho de formigas em um aquário com uma fonte de alimentos na outra ponta. Para caminhar até esse alimento foram desenvolvidos duas escolhas de caminho, sendo que um maior que o outro. Quando se iniciou o experimento, cada formiga seguiu um caminho exploratório aleatório. Como as formigas que escolheram o menor caminho trilhavam o caminho mais rapidamente que as outras, elas depositavam uma maior quantidade de feromônio nesse caminho escolhido em relação ao outro em um mesmo intervalo de tempo. Então, em um determinado momento a intensidade do feromônio no caminho mais curto estava tão elevado que quase todas as formigas trilhavam por este caminho.

2.4.2 Enxame de partículas

A otimização de enxame de partículas (*Particle Swarm Optimization – PSO*) foi proposto em 1995 por James Kennedy e Russel Elberhart para resolver problemas de interesse público. O método surgiu de experiências que modelam o convívio e comportamento social considerado em muitas espécies de pássaros, cardumes, chegado a até analisar o comportamento do ser humano.

Segundo Serapião (2009) uma teoria sócio-cognitiva simples está trás da PSO. Cada indivíduo de uma população tem suas experiências e é apto de ponderar a qualidade dessas experiências. Da forma que os indivíduos possuem características de ser social, eles também tem conhecimentos sobre como os seus pares se comportam. Os dois tipos de informação tem o sentido de aprendizagem individual (cognitiva) e transmissão cultural (social). Portanto a estatística de que um indivíduo toma uma certa decisão será uma função de desempenho no passado e do desempenho de alguns de seus vizinhos.

Kennedy *et al.*(2001) utilizou três princípios para reunir o processo de adaptação cultural:

- avaliar: indivíduos tem habilidade de perceber o ambiente de forma prever seu próprio comportamento;
- comparar: indivíduos usam um aos outros como referência comparativa;
- imitar: a imitação acontece de forma centralizada em grupos sociais humanos, e é relevante para aquisição e manutenção dos conhecimentos mentais.

De acordo com Serapião (2009), a PSO assim como outras abordagens de inteligência coletiva, está baseada em uma população de indivíduos aptos de relacionar entre si e o ambiente. Levando em consideração as propriedades de autoavaliação, comparação e imitação, os indivíduos são aptos de enfrentar umnúmero de possíveis situações que o ambiente lhes apresenta. Os comportamentos globais serão resultados emergentes dessas interações.

2.4.3 Cultura de bactérias

O método de otimização por meio de cultura de bactérias (*Bacterial Foraging Optimization* – BFO) começou a ser estudado por Kelvin Passino em 2001, no qual foi baseado nas estratégias de localização, manipulação e ingestão de alimentos de células da bactéria *Escherichia Coli*.

Tal método é composto na suposição de que os seres vivos buscam conseguir nutrientes de uma maneira em que o gasto de energia seja o menor possível, e a absorção de energia seja a maior possível. Desta maneira busca-se uma tentativa de maximizar a absorção de energia por unidade de tempo gasta na busca (MOEDINGER, 2005).

Segundo Passino (2002), essa maximização oferece fontes de nutrientes para sobreviver e energia extra para demais tarefas relevantes, como por exemplo, lutar, fugir, reproduzir e abrigar-se. As tarefas de *foraging* são variadas entre as espécies.

Para os herbívoros, o alimento é achado facilmente, porém o consumo é em grande quantidade. Já nas espécies dos carnívoros encontrar os alimentos é uma tarefa mais difícil, mas não é necessário em grande quantidade (PASSINO, 2002)..

O ambiente controla o teste padrão dos nutrientes que estão acessíveis e as dificuldade em adquiri-los, mostrando que as características necessárias para o sucesso na procura dependem quem busca o alimento. Os nutrientes para variedades animais estão em locais específicos (lagos, campos, árvores), nos casos o *foraging* integra tais locais, e no momento desta procura, resolver se o atual local satisfaz ou se deve procurar outro local com mais fontes de alimentos. Normalmente, estes locais são encontrados logo na sequência, e os fatores de risco e esforço são considerados na decisão se irão ou não procurar outro local de alimentação. (PASSINO, 2002).

De acordo com Stephens e Krebs (1986), uma teoria possível teoria de otimização motivada em colônia de bactérias formula o problema de *foraging* como obstáculo de otimização, com base no comportamento da bactéria *Escherichia Coli* que fazem parte de nosso intestino. Neste caso, métodos computacionais ou analíticos podem ser utilizados para formular políticas de *foraging,* que específica as variáveis já citadas e como as decisões devem ser tomadas.

2.4.4 Colônias artificiais de abelhas

A otimização coletiva por meio de colônias artificiais de abelhas (*Bee System - BS*) foi proposto por Karaboga em 2005, com a finalidade de solucionar problemas de otimização baseando-se no comportamento inteligente das abelhas na natureza. As colônias de abelhas são estudadas devido à sua alta capacidade organizacional.

As abelhas são seres que se organizam e realizam uma inteligência coletiva no qual ampliam o desempenho no ambiente em que convivem. As colônias tem três características de particular interesse: auto-organização, adaptação e divisão do trabalho (AKAY; KARABOGA, 2010).

Alocadas em colônias, quando na procura de alimentos, existem três diferentes tipos de comportamento: trabalhadoras, exploradoras e oportunistas. As abelhas trabalhadoras são aquelas que estão organizadas em alguma fonte de néctar perto da colmeia, essas abelhas levam o néctar colhido e informações sobre a quantidade de néctar da fonte onde ela está localizada. Essas informações são repassadas na área chamada área de dança. Nesta área, as abelhas oportunistas observam as abelhas trabalhadoras e tomam a decisão sobre qual fonte querem

visitar. Esta decisão é tomada levando em consideração em função da proximidade e da quantidade de néctar. As abelhas exploradores são aquelas que fazem buscas aleatórias nos arredores da colmeia para encontrar novas fontes de néctar. Dessa maneira, as colônias de abelha, por meio da interação das abelhas trabalhadores e oportunistas, desenvolvem uma inteligência coletiva que otimiza a busca de alimentos (ANDRADE; CUNHA, 2012)

2.4.5 Cardume

O método inspirado por meio do comportamento coletivo de um cardume de peixes foi proposto por Bastos Filho e Lima Neto em 2009.

Na natureza, diversas espécies vivem em conjunto, com o objetivo de aumentar suas chances de sobrevivência. Nos cardumes de peixes, isto pode funcionar com o objetivo de proteção mútua em relação aos predadores e a procura coletiva de alimentos (LACERDA, 2012).

Lacerda (2012) também cita comportamentos que serviram de inspiração natural para a elaboração do método.

- alimentação: os indivíduos se alimentam para sobrevivência, com o objetivo do crescimento saudável e capacidade de reprodução;
- nado: observando o processo de nado, no qual acontece de forma ordenada e organizada, permitindo que o cardume realize buscas por fontes de alimento em seu local de procura;
- reprodução: caso exista a reprodução, pode-se concluir que o processo de buscam está sendo bem executado. A reprodução possui como objetivo executar uma procura mais detalhada em bons locais encontrados pelo cardume, permitindo assim a sobrevivência dos mais adaptados e excluindo dos menos sucedidos.

Tendo sido apresentandos os principais conceitos relacionados ao tema pesquisado, a próxima seção apresenta os procedimentos metodológicos adotados.

3 PROCEDIMENTOS METODOLÓGICOS

Do ponto de vista metodológico, a pesquisa é caracterizada como pesquisa exploratória. Para Gil (2002), a pesquisa exploratória tem como objetivo demonstrar uma maior familiaridade com o problema proposto, com vistas a tornar o problema mais nítido ou construir hipóteses. Estas pesquisas têm como objetivo principal o aprimoramento de ideias ou intuições. Em grande quantidade dos casos, essas pesquisas envolvem levantamento bibliográfico, entrevistas com pessoas que possuem experiências práticas com o problema de pesquisa e análise de exemplos que incite a compreensão (SELLTIZ *et al*., 1967).

Segundo Polit e Hugler (1987) a pesquisa exploratória é um estudo preliminar para desenvolver ou refinar hipóteses, ou para testar e definir os métodos de coleta de dados. As razões para o envolvimento dos pesquisadores em uma pesquisa exploratória são que o investigador é simplesmente curioso e quer um entendimento do fenômeno de interesse, isso acontece quando não existem estudos satisfatórios para serem identificados ou trata-se de uma nova área de estudo. Ainda, os estudos exploratórios são conduzidos para estimar a viabilidade e o custo de realização de uma investigação mais precisa de um projeto do mesmo tema.

Para iniciar os estudos primeiramente foi feito um levantamento da literatura pertinente sobre o tema proposto para se definir quais métodos de otimização seriam estudados neste projeto de pesquisa.

A partir disto, foi realizada uma pesquisa na base cientifica *Web Of Science*[7], onde foi realizada a busca pelo termo "*methodology swarm optimization"* onde foram encontrados 611 resultados e por diversas vezes os métodos de otimização coletiva que apareceram com frequência foram os por meio de formigas (*Ant Colony Optimization*), enxame de partículas (*Particle Swarm Optimization*), cultura de bactérias (*Bacterial Foraging Optimization*) e colônia artificiais de abelhas (*Bee System*). Para a escolha do método de otimização de cardume de peixes, também

[7] Disponível em: <https://apps.webofknowledge.com/summary.do?product=UA&parentProduct=UA&search_mode=GeneralSearch&qid=4&SID=3FxHQvucpSorjkCslWv&page=1&action=changePageSize&pageSize=50>. Acesso em: 10 jun. 2016.

foi pesquisado na base *Web Of Science*[8], e nenhum resultado foi encontrado. Devido a isso, este método também foi escolhido.

Para a concepção do instrumento de comparação dos métodos, foi realizada uma busca na literatura pertinente sobre comparação de métodos de otimização. Os critérios encontrados e definidos estão descrito na sequência deste trabalho, na seção 3.2.

Foram utilizadas três bases de dados com tipo de atributos diferentes (numérica, nominal e mista) disponíveis na UCI – Machine Learning Repository[9], um repositório online de bases de dados. A ferramenta AntMiner e Weka (explicadas nas seções 3.1.7 e 3.1.8, respectivamente) foram utilizadas o processamento dos dados que auxiliando na comparação dos métodos definidos para estudo.

3.1 MATERIAIS E MÉTODOS

Este tópico foi reservado para apresentação de materiais e métodos utilizados na realização do trabalho.

3.1.1 Bases de dados

Esta seção apresenta as três bases de dados utilizadas para análise nesta pesquisa, com os seguintes critérios: numérica, nominal e mista (numérica e nominal). Na escolha pelas bases, procurou-se o equilíbrio entre quantidade de instâncias e atributos. A base *Soybean* difere-se na quantidade de registros de instâncias e atributos em comparação com as bases *Wine* e *Hepatitis*, pois não houve sucesso em utilizar a base *Lymphography* (base com quantidade de instâncias e atributos “similares” as bases *Wine* e *Hepatitis*), visto que a base está protegida pela UCI e, mesmo após contato por e-mail (conforme instruções no site), não houve resposta.

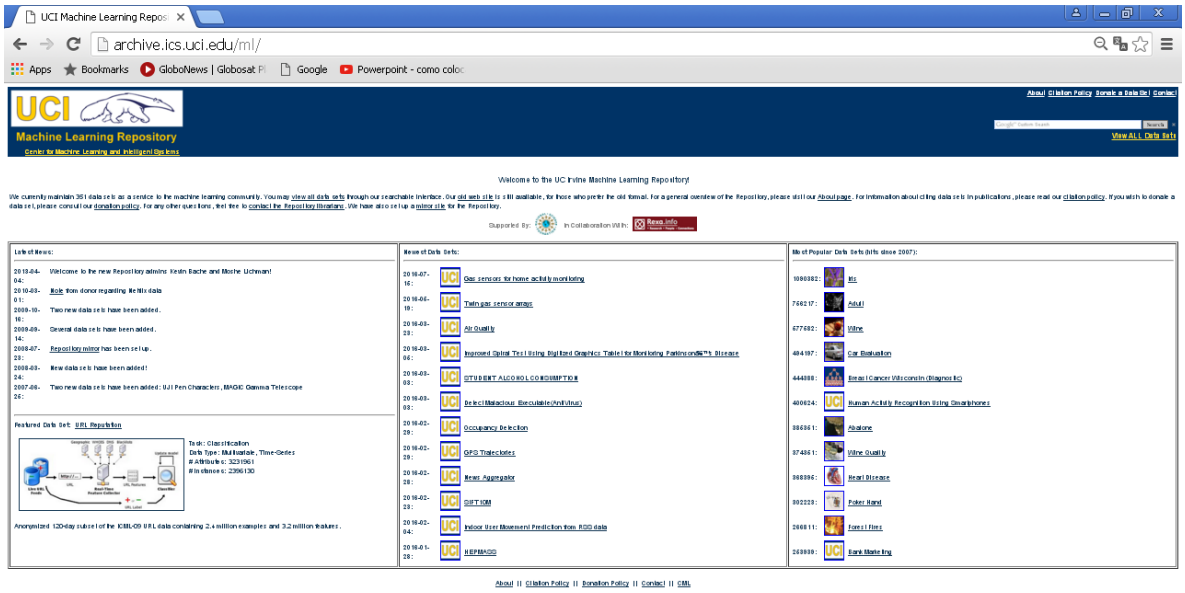
3.1.2 Repositório UCI

A UCI Machine Learning Repository é uma coleção de base de dados que são utilizadas para análise empírica dos algoritmos de aprendizagem de máquina. Criado em 1997 por David Aha, o repositório é utilizado pela comunidade como fonte

[8] Disponível em: <https://apps.webofknowledge.com/UA_GeneralSearch_input.do?SID=3FxHQvucpSorjkCslWv&product=UA&search_mode=GeneralSearch&errorQid=6#searchErrorMessage>. Acesso em:10 jun. 2016.
[9] Disponível em: <http://archive.ics.uci.edu/ml/>. Acesso em: 14 jun. 2016.

primária de conjunto de dados de aprendizado de máquina. A versão atual do site foi projetada em 2007 por Arthur Assunção, David Newman e também contou com a colaboração da Universidade de Massachusetts.

FIGURA 2-INTERFACE DO UCI



FONTE: UCI (2016)

Na UCI é possível realizar buscas de bases utilizando filtros como:

- tipo de atributos: numérico, nominal e mista (numérico e nominal);
- tipo de dados: multivariados, univariada, sequencial e texto;
- área: ciências biológicas, físicas e sociais, engenharia, negócios e jogos;
- quantidade de atributos: menos que 10, 10 à 100 e maior que 100;
- quantidade de instâncias: menos que 10, 10 à 100 e maior que 100;
- tipo de formato: matriz e não matriz.

Para fins desta pesquisa o filtro utilizado para a escolha das bases foi a do tipo de atributos, pois foi selecionada uma base com tipo de atributo numérico, nominal e mista.

3.1.3 Primeira base de dados: *Soybean data set*

A base *Soybean Data Set*[10] é uma base de dados de doenças provenientes da soja. Criada por R.S. Michalski e R.L. Chilausky as principais características da base são definidas conforme abaixo:

- tipo de atributos: nominal;
- tipo de dados: multivariada;
- área: ciência biológicas;
- quantidade de atributos: 36;
- quantidade de instâncias: 683;
- data de doação: 11/07/1998.

3.1.4 Segunda base de dados: *Wine data set*

A base *Wine Data Set*[11] é uma base de dados que utilizam análises químicas para determinar a origem do vinho. Criada por M. Forina, as principais características da base são definidas conforme abaixo:

- tipo de atributos: numérica
- tipo de dados: multivariada;
- área: física;
- quantidade de atributos: 13;
- quantidade de instâncias: 178;
- data de doação: 01/07/1991.

3.1.5 Terceira base de dados: *Hepatitis data set*

A base *Hepatitis Data Set*[12] é uma base de dados que reúne um conjunto de informações sobre hepatite. Criada por G.Gong as principais características da base são definidas conforme abaixo:

- tipo de atributos: mista (nominal e numérica);
- tipo de dados: multivariada;
- área: ciências biológicas;

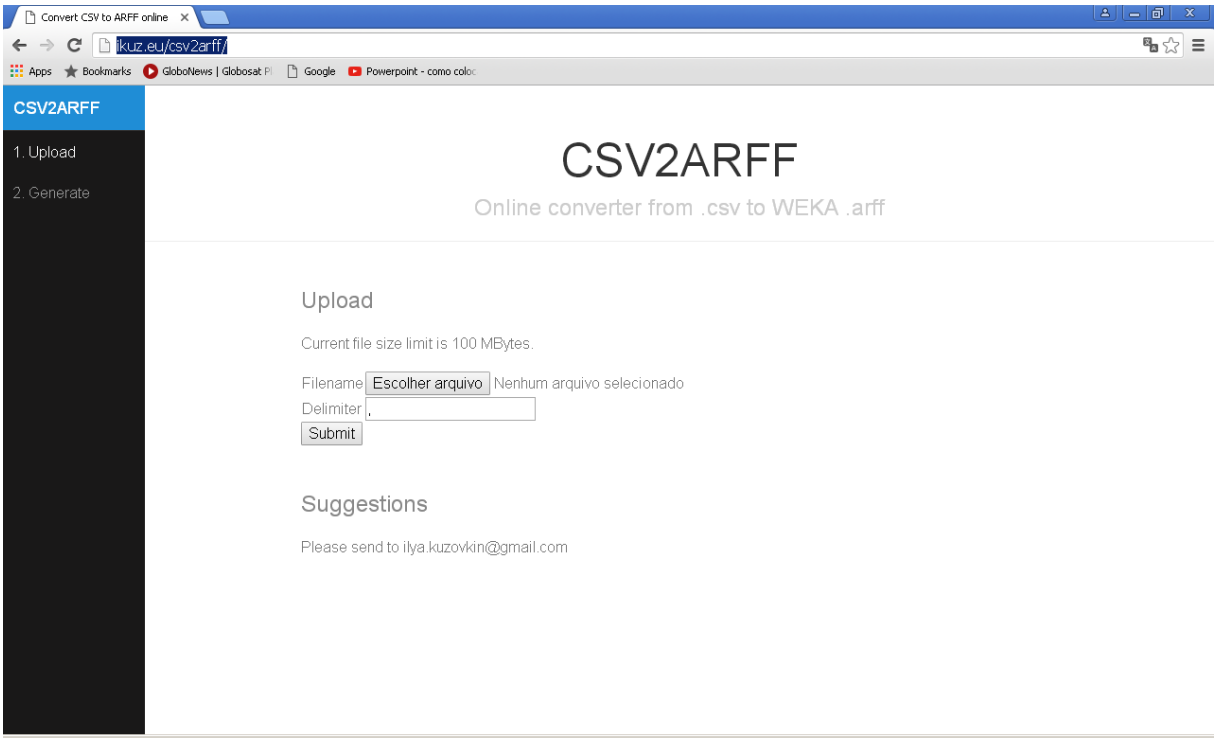
[10] Disponível em: <http://archive.ics.uci.edu/ml/datasets/Soybean+%28Large%29>. Acesso em: 08 out. 2016.
[11] Disponível em: <http://archive.ics.uci.edu/ml/datasets/Wine>. Acesso em: 08 out. 2016.
[12] Disponível em: <http://archive.ics.uci.edu/ml/datasets/Hepatitis>. Acesso em: 08 out. 2016.

- quantidade de atributos: 19;
- quantidade de instâncias: 155;
- data de doação: 01/11/1988.

3.1.6 Conversor CSV para ARFF

Para carregar as bases nas ferramentas Ant Miner e Weka, foi necessário a conversão da extensão das bases de .CSV para .ARFF (Attribute Relation File Format). Para isso, utilizou-se o CSV2ARFF[13]. O CSV2ARFF é uma ferramenta gratuita e para ter acesso basta ter conexão com a internet.

FIGURA 3-INTERFACEDO CSV2ARFF



FONTE: CSV2ARFF (2016)

Para a conversão foi necessário a escolher a base em formato .CSV, escolher o delimitador (vírgula), e realizar a conversão.

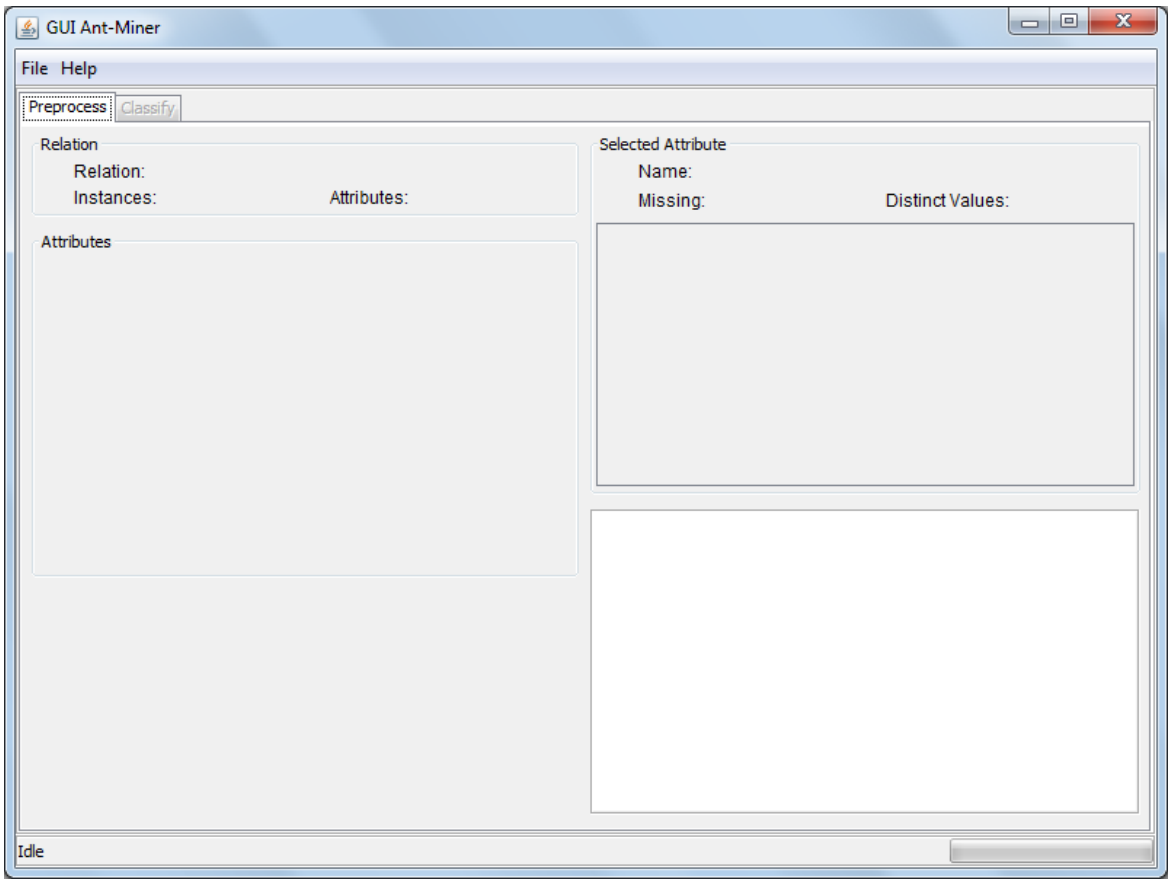
3.1.7 Ant Miner

O Ant Miner[14]é um software criado pelos pesquisadores Fernando Meyer e Rafael Stubs Parpinelli. É uma ferramenta de mineração de dados baseados em colônia de formigas. O software é gratuito para pesquisas e ensino. Desenvolvido

[13] Disponível em: <http://ikuz.eu/csv2arff/>. Acesso em 08 out. 2016.
[14] Disponível em: <http://www.aco-metaheuristic.org/aco-code/>. Acesso em: 28 ago. 2016.

em linguagem Java, é compatível com os principais sistemas operacionais e possui uma interface gráfica “amigável” ao usuário.

FIGURA 4-INTERFACE DO ANT MINER



FONTE: ANT MINER (2016)

O formato de entrada dos dados deve ser com a extensão .ARFF. O software somente aceita bases com atributos nominais. É possível compreender os resultados da ferramenta por meio de regras que são geradas após a execução de uma base de dados.

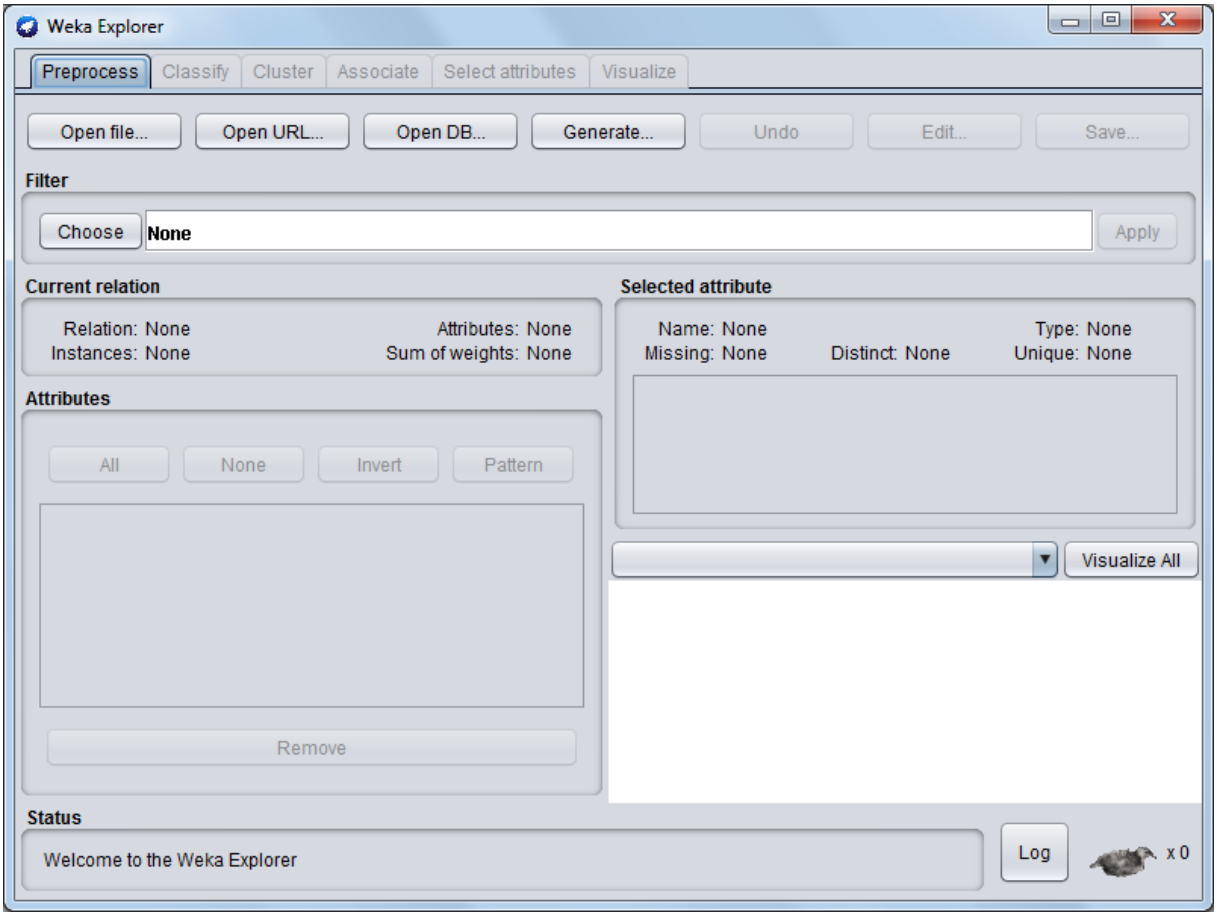
3.1.8 Weka

O Waikato Environment for Knowledge Analysis (Weka[15]) é um pacote desenvolvido pela Universidade de Waikato, em 1993, com o intuito de agregar algoritmos para mineração de dados na área de Inteligência Artificial. O software é licenciado pela General Public License (GPL) sendo, assim, possível a alteração do seu código-fonte. Weka é elaborado em linguagem Java. O formato de entrada dos dados na ferramenta é a extensão .ARFF.

[15] Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>. Acesso em: 17 set. 2016.

Possui uma série de heurísticas para mineração de dados relacionadas à classificação, regressão, agrupamento, regras de associação e visualização, entre elas: NaiveBayes, Linear Regression, IB1, Bagging, LogistBoot, Part, Ridor, ID3 e LMT.

FIGURA 5-INTERFACE DO WEKA



FONTE: WEKA (2016)

Para fins desta pesquisa foi utilizada a versão 3.8 do WEKA.

3.1.9 Fish School Search (Vanilla Version)

O Fish School Search (Vanilla Version)[16] é um software gratuito, desenvolvido por Carmelo Filho, Fernando Neto, Antônio Nascimento e Marília Lima, pela Universidade de Pernambuco. A ferramenta é baseada no método de cardumes de peixes.

[16] Disponível em: < http://www.fbln.pro.br/fss/versions.htm>. Acesso em: 17 set. 2016.

Não houve possibilidade de utilizar a ferramenta, pois para iniciar era necessário compilar o programa no Eclipse[17],ao compilar o programa não houve o retorno do Eclipse. Ao entrar em contato via e-mail com os criadores do software, a explicação que foi dada, requeria conhecimentos em avançados em programação e linguagem Java (no qual não são aprendidas ao decorrer do curso de Gestão da Informação), onde era necessário passar as informações das bases para o programa em forma de código, diferentemente do AntMiner e Weka, onde já existe uma interface, e também era necessário configurar indivíduo e como representar em formato de peixes.

3.1.10 J48

O J48 é uma implementação do algoritmo C4.5, desenvolvido por Quilan Ross em 1993. Segundo Librelotto e Mozzaquatro (2013), o J48 tem a finalidade de gerar uma árvore de decisão baseada em um conjunto de dados de treinamento, sendo este modelo usado para classificar as instâncias no conjunto de teste.

Um dos aspectos para a utilização do método J48 em mineração de dados é que o mesmo se mostra adequado para os procedimentos, envolvendo dados qualitativos contínuos e discretos presentes nas bases de dados. O J48 é considerado o que apresenta o melhor resultado na montagem das árvores de decisão (LIBRELOTTO; MOZZAQUATRO, 2013).

Segundo Witten e Frank (2005), para a montagem da árvore de decisão o J48 utiliza a abordagem de dividir para conquistar, no qual um problema complexo é decomposto em subproblemas mais simples, aplicando recursivamente a mesma estratégia a cada subproblema, dividindo o espaço definido pelos atributos em subespaços, associando-se a eles uma classe.

O método J48 será executado na ferramenta Weka e os resultados são visualizados por meio da árvore de decisão.

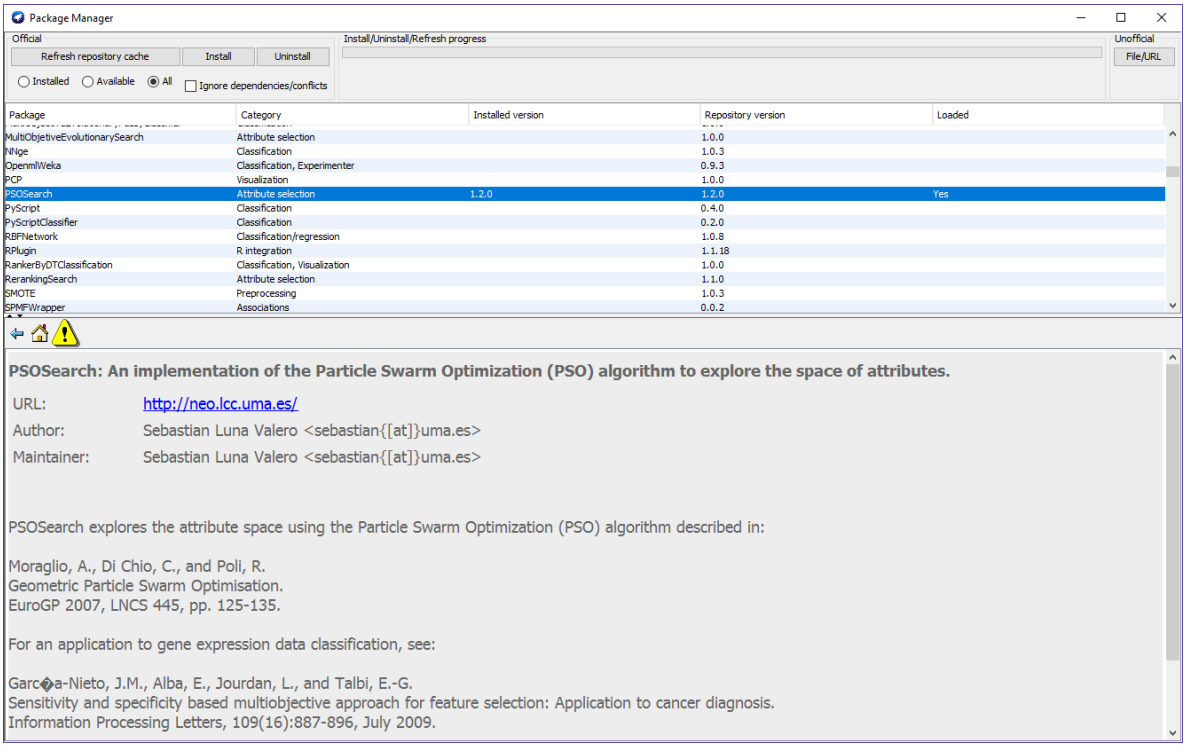
3.1.11 Weka – PSOSearch

No Weka existe um gerenciador de pacotes, e após uma busca neste gerenciador foi encontrado o pacote PSOSearch criado por Sebastian Luna Valero.

[17] Disponível em: <https://www.ibm.com/developerworks/br/library/os-eclipse-platform/>. Acesso em: 17 set. 2016.

O pacote é referente a uma implementação do algoritmo de enxame de partículas para seleção de atributos. Após a busca no gerenciador é necessário a baixar e instalar (Figura 5).

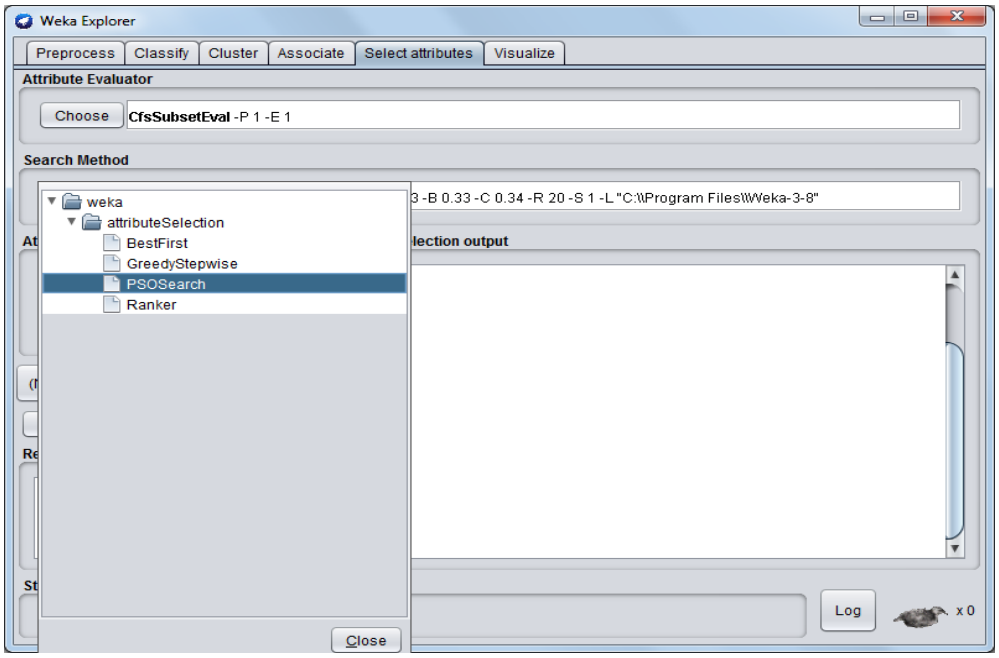
FIGURA 6- INTERFACE DO WEKA PSOSEARCH



FONTE: WEKA (2016)

Após a instalação, os PSOSearch fica disponível na aba “Select Attributes” (Figura 6).

FIGURA 7- PSOSEARCH INSTALADO NO WEKA



FONTE: WEKA (2016)

Uma vez que o PSOSearch disponível na ferramenta Weka não é um método de classificação, mas um para seleção de atributos. Apesar de ter sido utilizado, não apresenta os mesmos resultados que o AntMiner, J48 e outros.

3.1.12 Configuração de Máquina

As configurações básicas do computador utilizado para processamento das três bases de dados nas ferramentas são:

- sistema operacional: Microsoft Windows 7;
- processador: Intel Celeron 1.86 GHz;
- memória RAM: 2,49 GB.

3.2 CRITÉRIOS UTILIZADOS NA COMPARAÇÃO

Assim, para fins de comparação, foram realizadas doze tarefas de minerações:

a) três bases de dados com AntMiner com atributos originais;
b) três bases de dados com Weka (J48) com atributos originais;
c) três bases de dados com AntMiner com atributos selecionados pelo PSOSearch;

d) três bases de dados com Weka (J48) com atributos selecionados pelo PSOSearch.

Todos os testes foram realizados com validação cruzada de 10 (dez) partições. Os resultados estão apresentados na próxima seção.

Segundo Ramisch (2009) os critérios que podem ser observados para comparação são:

- tempo de processamento;
- taxa de acerto e taxa de erro;
- compreensibilidade dos padrões descobertos.

De acordo com Nascimento e Cruz (2013), outros critérios que podem ser utilizados para comparação são:

- matriz de confusão para que se possa verificar em quais situações o método acerta / erra com maior frequência;
- tamanho das árvores considerando-se número de folhas e tamanho da árvore.

Portanto para a presente pesquisa, os critérios acima serão utilizados para comparação dos resultados, no qual serão expostos na seção 4.

4 ANÁLISE DE RESULTADOS

Conforme explicado na seção de procedimentos metodológicos, foram testadas doze atividades de mineração de dados. Nesta seção, os resultados foram agrupados em:

a) mineração com atributos originais;

b) mineração com atributos selecionados pelo PSOSearch;

Ao final desta seção é apresentada uma comparação que contempla todos os experimentos realizados.

4.1 MINERAÇÃO DE DADOS COM ATRIBUTOS ORIGINAIS

Os resultados apresentados nos tópicos 4.1.1, 4.1.2 e 4.1.3 foram com base na mineração de dados com atributos originais das bases.

4.1.1 *Soybean data set*

Ao executar a base *Soybean* (com 683 instâncias e 36 atributos) no Weka utilizando o método J48 e validação cruzada de 10 (dez) partições, obtiveram-se os principais resultados: tempo de processamento de 0,12 segundos, instâncias corretamente classificadas 625 (91,50%) e instâncias incorretamente classificadas 58 (8,49 %).

A partir destes resultados o Weka apresentou a matriz de confusão da FIGURA 8, no qual é possível verificar que as classes que apresentaram maior número de classificações corretas foram: *phytophthora-rot*, *brown-spot* e *alternarialeaf-spot.*

FIGURA 8- MATRIZ DE CONFUSÃO *SOYBEAN* ATRIBUTOS ORIGINAIS

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  s   <-- classified as
 19  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  a = diaporthe-stem-canker
  0 20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  b = charcoal-rot
  1  0 19  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  c = rhizoctonia-root-rot
  0  0  0 87  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0 |  d = phytophthora-rot
  0  0  0  0 44  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  e = brown-stem-rot
  0  0  0  0  0 20  0  0  0  0  0  0  0  0  0  0  0  0  0 |  f = powdery-mildew
  0  0  0  0  0  0 20  0  0  0  0  0  0  0  0  0  0  0  0 |  g = downy-mildew
  0  0  0  0  0  0  0 85  0  0  0  0  2  1  4  0  0  0  0 |  h = brown-spot
  0  0  0  0  0  0  0  0 20  0  0  0  0  0  0  0  0  0  0 |  i = bacterial-blight
  0  0  0  0  0  0  0  0  1 19  0  0  0  0  0  0  0  0  0 |  j = bacterial-pustule
  0  0  0  0  0  0  0  0  0  0 20  0  0  0  0  0  0  0  0 |  k = purple-seed-stain
  0  0  0  4  0  0  0  0  0  0  0 40  0  0  0  0  0  0  0 |  l = anthracnose
  0  0  0  0  0  0  0  3  0  0  0  0 14  0  3  0  0  0  0 |  m = phyllosticta-leaf-spot
  0  0  0  0  0  0  0  1  0  0  0  0  0 85  5  0  0  0  0 |  n = alternarialeaf-spot
  0  0  0  0  0  0  0  3  0  0  0  0  1 20 67  0  0  0  0 |  o = frog-eye-leaf-spot
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 15  0  0  0 |  p = diaporthe-pod-&-stem-blight
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 14  0  0 |  q = cyst-nematode
  0  0  0  1  0  0  0  0  0  0  0  1  0  0  0  0  0 14  0 |  r = 2-4-d-injury
  0  0  0  1  0  0  0  0  0  0  0  1  0  0  0  1  0  2  3 |  s = herbicide-injury
```

FONTE: WEKA (2016)

O Weka também apresentou a árvore do método J48 com número de folhas igual a 61 e tamanho da árvore igual 93. A visualização gráfica da árvore fica comprometida pelo número de folhas (61) e largura da mesma (93). Assim, o APÊNDICE 1 apresenta a árvore em formato texto.

Na ferramenta AntMiner utilizando o método da colônia de formigas e validação cruzada de 10 (dez) partições, obtiveram-se como principais resultados: tempo de processamento de 174 segundos, instâncias corretamente classificadas 600 (87,90%), instâncias incorretamente classificadas 83 (12,10%)e 234 regras. A ferramenta AntMiner não apresenta a matriz de confusão ao final do processamento.

4.1.2 *Wine data set*

Na execução da base *Soybean* (com 178 instâncias e 14 atributos) no Weka utilizando o método J48 e validação cruzada de 10 (dez) partições, obtiveram-se os principais resultados: tempo de processamento de 0,05 segundos, instâncias corretamente classificadas 167 (93,82%) e instâncias incorretamente classificadas 11 (6,17%).

A partir destes resultados o Weka apresentou a seguinte matriz de confusão (FIGURA 9), onde é possível verificar que 58 instâncias foram classificadas corretamente para classe 1, 67 instâncias para a classe 2 e 42 instâncias para a classe 3.

FIGURA 9- MATRIZ DE CONFUSÃO *WINE* ATRIBUTOS ORIGINAIS

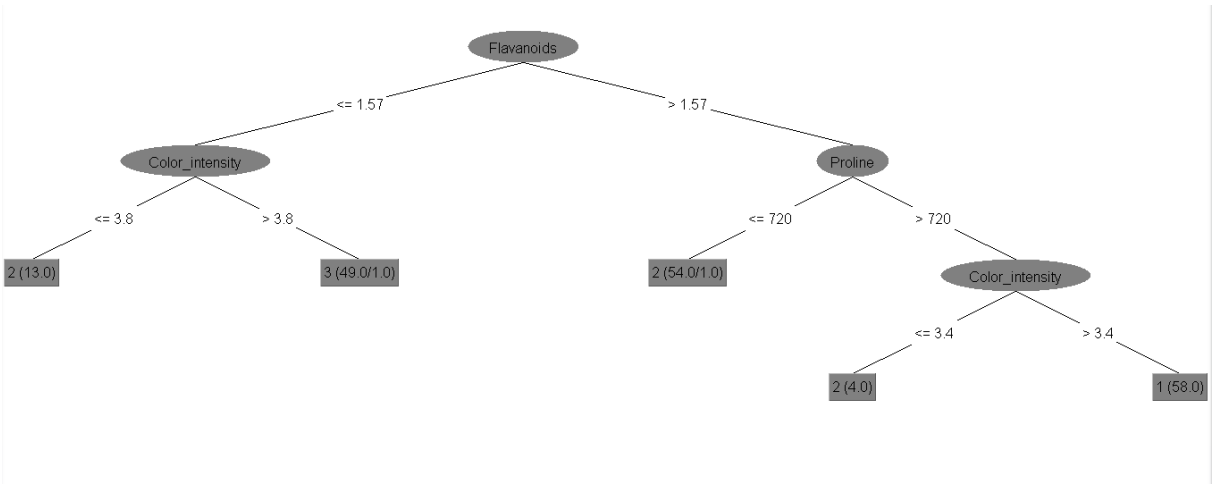
```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 58  1  0 |  a = 1
  3 67  1 |  b = 2
  1  5 42 |  c = 3
```

FONTE: WEKA (2016)

O Weka também apresentou a árvore do método J48 com número de folhas igual a 5 e tamanho da árvore igual 9. Abaixo a pode-se visualizar a árvore.

Flavanoids<= 1.57
|Color_intensity<= 3.8: 2 (13.0)
|Color_intensity> 3.8: 3 (49.0/1.0)
Flavanoids> 1.57
|Proline<= 720: 2 (54.0/1.0)
|Proline> 720
|| Color_intensity<= 3.4: 2 (4.0)
||Color_intensity> 3.4: 1 (58.0)

FIGURA 10- ÁRVORE DE DECISÃO *WINE* - ATRIBUTOS ORIGINAIS



FONTE: WEKA (2016)

Não foi possível executar a base no AntMiner, pois a ferramenta somente aceita bases com atributos nominais.

4.1.3 *Hepatitis data set*

Ao executar a base *Hepatitis* (com 155 instâncias e 20 atributos) no Weka utilizando o método J48 e validação cruzada de 10 (dez) partições, obtiveram-se os principais resultados: tempo de processamento de 0,14 segundos, instâncias corretamente classificadas 130 (83,87%) e instâncias incorretamente classificadas 25 (16,12%).

A partir destes resultados o Weka gerou a seguinte matriz de confusão (FIGURA 11), no qual é possível verificar que 14 instâncias foram classificadas corretamente para a classe *die* e 116 para classe *live.*

FIGURA 11 - MATRIZ DE CONFUSÃO *HEPATITIS* ATRIBUTOS ORIGINAIS

```
=== Confusion Matrix ===

  a   b   <-- classified as
 14  18 |   a = DIE
  7 116 |   b = LIVE
```

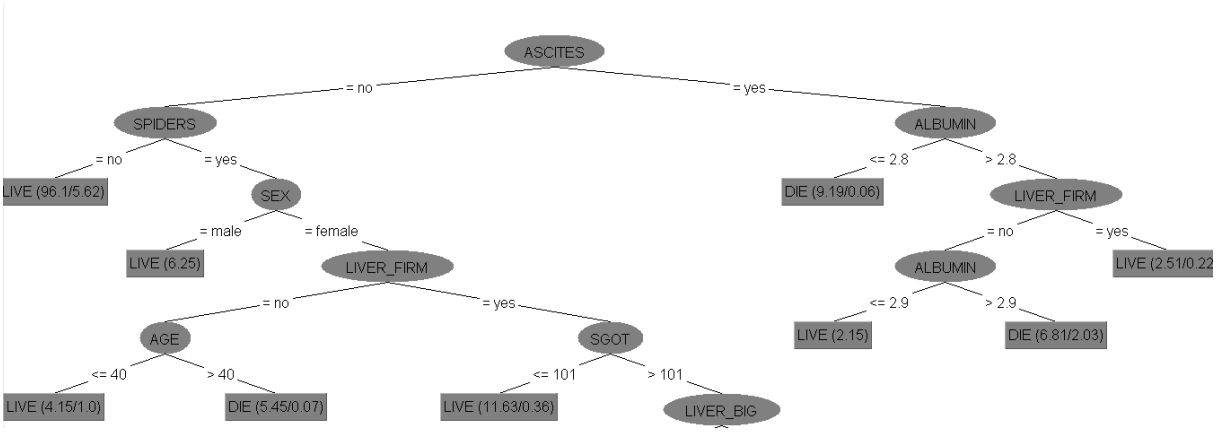
FONTE: WEKA (2016)

O Weka também apresentou a árvore do método J48 com número de folhas igual a 11 e tamanho da árvore igual 21. A Figura 12 apresenta a árvore resultante pelo J48.

```
ASCITES = no
|SPIDERS = no: LIVE (96.1/5.62)
|SPIDERS = yes
|   |   SEX = male: LIVE (6.25)
|   |   SEX = female
|   |   |   LIVER_FIRM = no
|   |   |   |   AGE <= 40: LIVE (4.15/1.0)
|   |   |   |   AGE > 40: DIE (5.45/0.07)
|   |   |   LIVER_FIRM = yes
|   |   |   |   SGOT <= 101: LIVE (11.63/0.36)
|   |   |   |   SGOT > 101
|   |   |   |   |   LIVER_BIG = no: DIE (3.23/0.08)
|   |   |   |   |   LIVER_BIG = yes: LIVE (7.54/2.36)
ASCITES = yes
|   ALBUMIN <= 2.8: DIE (9.19/0.06)
|   ALBUMIN > 2.8
```

```
|   |   LIVER_FIRM = no
|   |   |   ALBUMIN <= 2.9: LIVE (2.15)
|   |   |   ALBUMIN > 2.9: DIE (6.81/2.03)
|   |   LIVER_FIRM = yes: LIVE (2.51/0.22)
```

FIGURA 12 - ÁRVORE DE DECISÃO *HEPATITIS* - ATRIBUTOS ORIGINAIS



FONTE: WEKA (2016)

Não foi possível executar a base no AntMiner, pois a ferramenta somente aceita bases com atributos nominais.

4.2 MINERAÇÃO DE DADOS COM ATRIBUTOS SELECIONADOS PELO *PSOSEARCH*

Os resultados que serão apresentados a seguir foram com base na mineração de dados com atributos selecionados pelo *PSOSearch*.

O *PSOSearch* selecionou os seguintes atributos para bases:

Soybean Data Set

O método selecionou os seguintes atributos na base *Soybean*, dentre eles:

- *date;*
- *precip;*
- *temp;*
- *area-damaged;*
- *seed-tmt;*

- *germination;*
- *plant-growth;*
- *leaves;*
- *leafspots-halo;*
- *leafspots-marg;*
- *leafspot-size;*
- *leaf-malf;*
- *leaf-mild;*
- *stem-cankers;*
- *canker-lesion;*
- *fruiting-bodies;*
- *external-decay;*
- *int-discolor;*
- *fruit-pods;*
- *fruit-spots;*
- *seed-discolor;*
- *roots;*
- *class.*

Wine Data Set

O método selecionou os seguintes atributos na base *Wine*, dentre eles:

- *class;*
- *alcohol;*
- *alkalinity of ash;*
- *magnesium;*
- *total phenols;*
- *nonflavanoid phenols.*

Hepatitis Data Set

O método selecionou os seguintes atributos na base *Hepatitis*, dentre eles:

- *age;*
- *sex;*

- *fatigue;*
- *spiders;*
- *ascites;*
- *varices;*
- *bilirubin;*
- *albumin;*
- *protime;*
- *histology;*
- *class.*

4.2.1 *Soybean data set*

Ao executar a base *Soybean* (com 683 instâncias e 23 atributos) no Weka utilizando o método J48 e validação cruzada de 10 (dez) partições, obtiveram-se os principais resultados: tempo de processamento de 0,01 segundos, instâncias corretamente classificadas 622 (91,06%) e instâncias incorretamente classificadas 61 (8,93%).

A partir destes resultados o Weka apresentou a matriz de confusão da FIGURA 8, no qual é possível verificar que as classes que apresentaram maior número de classificações corretas foram: *phytophthora-rot, brown-spot* e*alternarialeaf-spot.*

FIGURA 13 - MATRIZ DE CONFUSÃO *SOYBEAN* ATRIBUTOS SELECIONADOS

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  s   <-- classified as
 20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  a = diaporthe-stem-canker
  0 20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  b = charcoal-rot
  1  0 19  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  c = rhizoctonia-root-rot
  0  0  0 87  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0 |  d = phytophthora-rot
  0  0  0  0 44  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  e = brown-stem-rot
  0  0  0  0  0 20  0  0  0  0  0  0  0  0  0  0  0  0  0 |  f = powdery-mildew
  0  0  0  0  0  0 20  0  0  0  0  0  0  0  0  0  0  0  0 |  g = downy-mildew
  0  0  0  0  0  0  0 84  0  0  0  0  2  1  5  0  0  0  0 |  h = brown-spot
  0  0  0  0  0  0  0  0 20  0  0  0  0  0  0  0  0  0  0 |  i = bacterial-blight
  0  0  0  0  0  0  0  0  3 17  0  0  0  0  0  0  0  0  0 |  j = bacterial-pustule
  0  0  0  0  0  0  0  0  0  0 20  0  0  0  0  0  0  0  0 |  k = purple-seed-stain
  1  0  0  5  0  0  0  0  0  0  0 38  0  0  0  0  0  0  0 |  l = anthracnose
  0  0  0  0  0  0  0  5  0  0  0  0 12  0  3  0  0  0  0 |  m = phyllosticta-leaf-spot
  0  0  0  0  0  0  0  0  0  0  0  0  0 86  5  0  0  0  0 |  n = alternarialeaf-spot
  0  0  0  0  0  0  0  3  0  0  0  0  0 15 73  0  0  0  0 |  o = frog-eye-leaf-spot
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 15  0  0  0 |  p = diaporthe-pod-&-stem-blight
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 14  0  0 |  q = cyst-nematode
  0  0  0  6  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0 |  r = 2-4-d-injury
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  4  3 |  s = herbicide-injury
```

FONTE: WEKA (2016).

O Weka também apresentou a árvore do método J48 com número de folhas igual a 74 e tamanho da árvore igual 106. O APÊNDICE 2 apresenta a árvore em formato texto.

Na ferramenta AntMiner utilizando o método da colônia de formigas, validação cruzada de 10 (dez) partições e a base com os 22 atributos selecionados pelo AntMiner, obtiveram-se como principais resultados: tempo de processamento de 67 segundos, 666 (97,54%) instâncias corretamente classificadas,17 (2,46%) instâncias incorretamente classificadas e 84 regras. A ferramenta AntMiner não apresenta a matriz de confusão ao final do processamento.

4.2.2 *Wine data set*

Ao executar a base *Wine* (com 178 instâncias e 6 atributos) no Weka utilizando o método J48 e validação cruzada de 10 (dez) partições, obtiveram-se os principais resultados: tempo de processamento de 0,01 segundos, instâncias corretamente classificadas 143 (80,33%) e instâncias incorretamente classificadas 35 (19,66 %).

A partir destes resultados o Weka gerou a seguinte matriz de confusão (FIGURA 14), no qual pode-se verificar que a classe 1 e 2 tiveram 55 instâncias classificadas corretamente e classe 3, 33 instâncias.

FIGURA 14 - MATRIZ DE CONFUSÃO *WINE* ATRIBUTOS SELECIONADOS

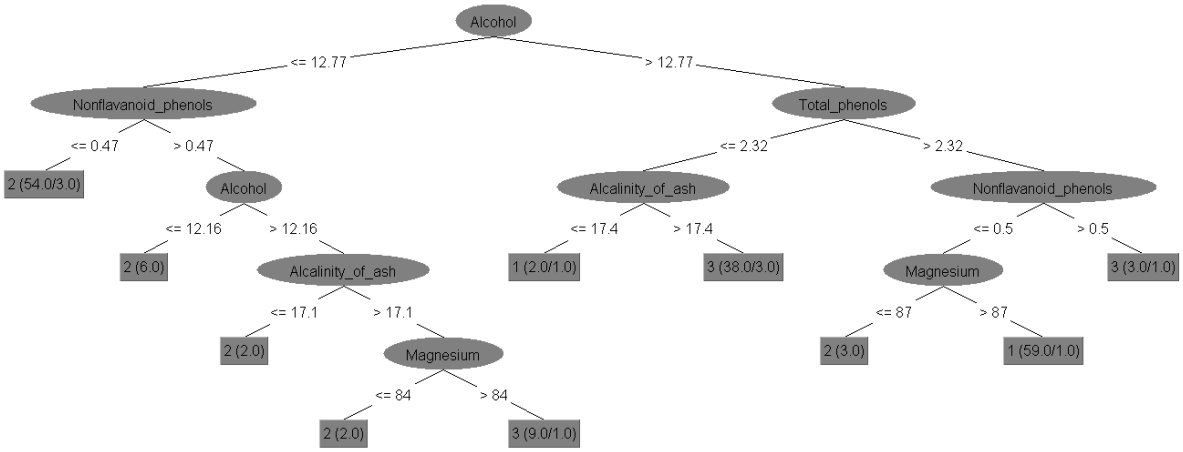
```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 55  0  4 |  a = 1
  2 55 14 |  b = 2
  1 14 33 |  c = 3
```

FONTE: WEKA (2016).

O Weka também apresentou a árvore do método J48 com número de folhas igual a 10 e tamanho da árvore igual 19.Abaixo a pode-se visualizar a árvore.

```
Alcohol<= 12.77
|Nonflavanoid_phenols<= 0.47: 2 (54.0/3.0)
|Nonflavanoid_phenols> 0.47
||   Alcohol<= 12.16: 2 (6.0)
||   Alcohol> 12.16
||   |   Alcalinity_of_ash<= 17.1: 2 (2.0)
|   |   |   Alcalinity_of_ash> 17.1
|   |   |   |   Magnesium <= 84: 2 (2.0)
|   |   |   |   Magnesium > 84: 3 (9.0/1.0)
Alcohol > 12.77
|   Total_phenols<= 2.32
|   |   Alcalinity_of_ash<= 17.4: 1 (2.0/1.0)
|   |   Alcalinity_of_ash> 17.4: 3 (38.0/3.0)
|   Total_phenols> 2.32
|   |   Nonflavanoid_phenols<= 0.5
|   |   |   Magnesium <= 87: 2 (3.0)
|   |   |   Magnesium > 87: 1 (59.0/1.0)
|   |   Nonflavanoid_phenols> 0.5: 3 (3.0/1.0)
```

FIGURA 15 - ÁRVORE DE DECISÃO *WINE* – ATRIBUTOS SELECIONADOS



FONTE: WEKA (2016)

Não foi possível executar a base no AntMiner, pois a ferramenta somente aceita bases com atributos nominais.

4.2.3 *Hepatitis data set*

Ao executar a base *Hepatitis* (com 155 instâncias e 11 atributos) no Weka utilizando o método J48 e validação cruzada de 10 (dez) partições, obtiveram-se os principais resultados: tempo de processamento de 0,03 segundos, instâncias corretamente classificadas 129 (83,22%) e instâncias incorretamente classificadas 26 (16,77%).

A partir destes resultados o Weka gerou a seguinte matriz de confusão (FIGURA 16), onde é possível verificar que 12 instâncias foram classificadas corretamente para *die* e 117 para *live.*

FIGURA 16 - MATRIZ DE CONFUSÃO *HEPATITIS* ATRIBUTOS SELECIONADOS

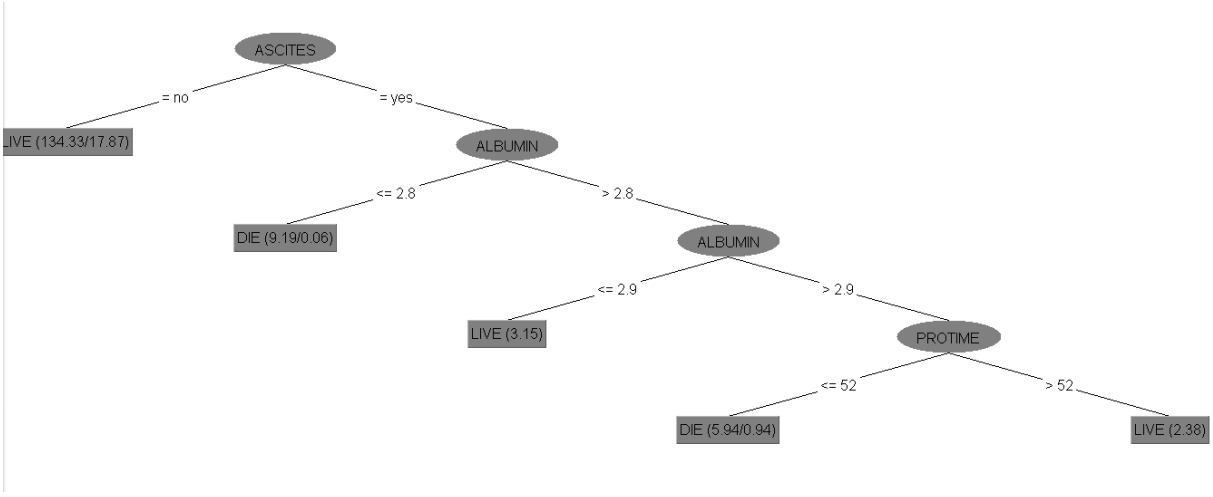
```
=== Confusion Matrix ===

  a   b   <-- classified as
 12  20 |   a = DIE
  6 117 |   b = LIVE
```

FONTE: WEKA (2016).

O Weka também apresentou a árvore do método J48 com número de folhas igual a 5 e tamanho da árvore igual 9.

```
ASCITES = no: LIVE (134.33/17.87)
ASCITES = yes
|   ALBUMIN <= 2.8: DIE (9.19/0.06)
|   ALBUMIN > 2.8
|   |   ALBUMIN <= 2.9: LIVE (3.15)
|   |   ALBUMIN > 2.9
|   |   |   PROTIME <= 52: DIE (5.94/0.94)
|   |   |   PROTIME > 52: LIVE (2.38)
```

FIGURA 17 - ÁRVORE DE DECISÃO - *HEPATITIS* - ATRIBUTOS SELECIONADOS



FONTE: WEKA (2016).

Não foi possível executar a base no AntMiner, pois a ferramenta somente aceita bases com atributos nominais.

4.3 COMPARAÇÃO DOS RESULTADOS

A base de dados *Soybean*com atributos originais e a base de dados *Soybean*comatributos selecionados pelo PSOSearch foram processadas nas ferramentas Weka e AntMiner,a partir dos resultados pode-se perceber que o tempo mais rápido de processamento ocorreu na base com atributos selecionados com tempo de resposta de 0,01 segundos utilizando o método J48. Em relação às instâncias classificadas corretamente a base com os atributos selecionados executada na ferramenta AntMiner classificou 666 instâncias corretamente. Em comparação com as árvores de decisões geradas pelo Weka utilizando o método J48, a árvore que é mais fácil de ser entendida é a base com os atributos originais.Em relação às regras geradas pelo AntMiner, a base com atributos originais teve 234 regras contra 84 regras geradas na base com atributos selecionados pelo AntMiner.

QUADRO 1 - COMPARAÇÃO *SOYBEAN*

SOYBEAN DATA SET	ATRIBUTOS ORIGINAIS J48 - WEKA	ATRIBUTOS ORIGINAIS - ANT MINER	ATRIBUTOS SELECIONADOS J48 - WEKA	ATRIBUTOS SELECIONADOS - ANT MINER
Tempo de Processamento	0,12 segundos	174 segundos	0,01 segundos	67 segundos
Instâncias corretamente classificadas	625 (91,50 %)	600 (87,90%)	622 (91,06 %)	666 (97,54 %)
Instâncias incorretamente classificadas	58 (8,49 %)	83 (12,10%)	61 (8,93 %)	17 (2,46 %)
Número de Folhas da Árvore de	61	Não Aplicável	74	Não Aplicável

Decisão				
Tamanho da Árvore de Decisão	93	Não Aplicável	106	Não Aplicável
Número de Regras	Não Aplicável	234	Não Aplicável	84

FONTE: O AUTOR (2016).

Na comparação dos resultados na base *Wine,* no critério tempo de processamento a base com atributos selecionados pelo PSOSearch foi processada mais rápida do que a base com atributos originais (0,05 segundos x 0,01 segundos), porém em relação a taxa de instâncias corretamente classificadas, a base com atributos originais teve um melhor resultado (167 instâncias x 143 instâncias). Para os critérios relacionados à árvore de decisão, a árvore gerada pela base com atributos originais é menor em comparação da árvore gerada pelos atributos selecionados, demonstrando maior facilidade para leitura e compreensão da árvore de decisão.

QUADRO 2 - COMPARAÇÃO *WINE*

WINE DATA SET	ATRIBUTOS ORIGINAIS –J48 –WEKA	ATRIBUTOS SELECIONADOS PSO SEARCH –J48 –WEKA
Tempo de Processamento	0,05 segundos	0,01 segundos
Instâncias corretamente classificadas	167 (93,82%)	143 (80,33%)
Instâncias incorretamente classificadas	11 (6,17%)	35 (19,66 %)
Número de Folhas da Árvore de Decisão	5	10
Tamanho da Árvore de Decisão	9	19

FONTE: O AUTOR (2016).

Por fim, na base *Hepatits* no critério de tempo de processamento pode-se perceber que a base com atributos selecionados pelo PSOSearch teve uma resposta mais rápida, porém a taxa de instâncias corretamente classificadas foi menor do que a base com os atributos originais (129 instâncias x 130 instâncias). Em relação às árvores de decisões, a gerada pela base dos atributos selecionados é menor nos critérios de número de folhas e tamanho da árvore, o que demonstra também maior facilidade para compreensão da árvore. O quadro abaixo auxilia na visualização da comparação.

QUADRO 3 - COMPARAÇÃO *HEPATITIS*

HEPATATIS DATA SET	ATRIBUTOS ORIGINAIS -J48 -	ATRIBUTOS SELECIONADOS PSO SEARCH -J48 -WEKA

	WEKA	
Tempo de Processamento	0,14 segundos	0,03 segundos
Instâncias corretamente classificadas	130 (83,87%)	129 (83,22%)
Instâncias incorretamente classificadas	25 (16,12%)	26 (16,77%)
Número de Folhas da Árvore de Decisão	11	5
Tamanho da Árvore de Decisão	21	9

FONTE: O AUTOR (2016).

Após todas as análises dos resultados e comparações realizadas, é possível concluir que:

- *Soybean Data Set:* em relação ao critério “tempo de processamento” a base com atributos selecionados pelo PSOSearch e executada no Weka utilizando o método J48 apresentou o melhor resultado, porém a taxa de instâncias classificadas corretamente foi inferior à base com atributos selecionados pelo PSOSearch e executada no AntMiner. Conclui-se também que melhor compreensibilidade na visualização dos resultados é nas bases que foram testadas utilizando o método das colônias de formigas, visto que elas são apresentadas por meio de regras e não através de árvore de decisão, como é apresentado no método J40, e quanto maior a base, maior será a árvore de decisão e consequentementemaior será a complexidade para entendimento;

- *Wine Data Set:* em relação ao tempo de processamento a base com atributos selecionados pelo PSOSearch e executada no Weka utilizando o método J48 apresentou o melhor resultado, porém em relação as instâncias corretamente classificadas e a melhor compreensibilidade dos resultados, visto que o tamanho da árvore de decisão é menor, a base com atributos originais e processada no Weka utilizando o método J48, se destaca;

- *Hepatitis Data Set:* em relação às instâncias corretamente classificadas destaca-se a base com atributos originais e processada no Weka utilizando o método J48, porém relação ao tempo de processamento e melhor compreensibilidade dos resultados, visto que o tamanho da árvore de decisão é menor, destaca-se a base com atributos selecionados pelo PSOSearch e executada no Weka utilizando o método J48.

5 CONSIDERAÇÕES FINAIS

Após todo o exposto do presente trabalho, pode-se verificar que o objetivo geral “comparar métodos de otimização por inteligência coletiva aplicados à mineração de dados” foi atingindo a partir da comparação do método de colônia de formigas com o método J48, por meio da utilização de bases de dados e executadas nas ferramentas Weka e AntMiner e o método de enxame de partículas foi testado a partir da utilização de um método de seleção de atributos no Weka (PSOSearch). Ambos os resultados foram registrados e comparados e estão presentes na seção 4.

Em relação aos objetivos específicos que foram definidos para este trabalho, pode-se perceber que para definição dos métodos e ferramentas que foram objetos de comparação,foram selecionados os métodos J48 e PSOSearch e as ferramentas selecionadas foram o AntMiner e Weka.

Referente à escolha das bases de dados que foram utilizadas nos experimentos para comparação, houve a seleção de 3 (três) bases de dados no repositório UCI. Os critérios para a escolha foram os tipos de atributos das bases, portanto, foi selecionada uma base com atributos numéricos (*Wine Data Set),* nominal (*Soybean Data Set*) e mista (*Hepatitis Data Set*). Por fim, as bases foram submetidas às ferramentas / métodos e os resultados foram registrados. Desta maneira, concluiu-se que a partir dos resultados apresentados, a ferramenta AntMiner conseguiu classificar a maior quantidade de instâncias corretamente, utilizando o métodos da colônia de formigas e a base *Soybean* com atributos selecionados pelo PSOSearch.

Em relação ao tempo de processamento, é possível verificar que nas três bases de dados com atributos selecionados pelo PSOSearch e executada no Weka utilizando o método J48, tiveram o melhor desempenho em relação a este critério.

Como foi exposto na delimitação de pesquisa, o trabalho não analisou a complexidade dos algoritmos dos métodos, portanto é importante ressaltar que as comparações realizadas podem não trazer uma notória contribuição para a ciência da computação.

Como sugestões para trabalhos futuros que possam ser realizados poderiam ser comparados os outros métodos citados neste trabalho, utilizando suas ferramentas para que se possa concluir qual método é o mais adequado a um problema específico. Também se sugere que outras bases de dados sejam

utilizadas, por exemplo, com tamanhos e atributos diferentes dos testados nesta pesquisa.

Por fim, seguindo o raciocínio apresentado na justificativa deste trabalho, pode-se observar que ainda existe falta de trabalhos acadêmicos na área dos métodos de otimização coletiva aplicados à mineração de dados, especialmente em língua portuguesa, apresentando-se como uma oportunidade para pesquisas futuras.

REFERÊNCIAS

AKAY B.,KARABOGA D. **A modified artificial bee colony algorithm for real-parameter optimization**.Information Science, 2010.

ANDRADEL. A. G., CUNHA, B. **Algoritmo de colônia artificial de abelhas para um problema de clusterização capacitado**. In Anais do XLIII Simpósio Brasileiro de Pesquisa Operacional. Ubatuba. SP. 2012.

BARR, Avron; FEIGENBAUM, Edward. **The handbook of artificial intelligence.** California: HeurisTech Press, 1981. 1 v.

BISHOP, Christopher. **Pattern recognition and machine learning.** Cambridge: Springer, 2007.

CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes. **Introdução à mineração de dados.**São Paulo: Saraiva, 2016.

CHAN, A.; FREITAS, A.A. **A new ant colony algorithm for multi-label classification with applications in bioinformatics**.Proceedings of the 8th annual conference on Genetic and evolutionary computation (GECCO ’06), pp. 27–34. 2006.

CHARNIAK, Eugene; MCDERMOTT, Drew V.. **Artificial intelligence programming.** 2. ed. Seattle: Psychology Press, 1987.

CONDUTA, Bruno; MAGRIN, Diego. **Aprendizagem de máquina.** 2010. 19 f. Dissertação (Mestrado), Universidade Estadual de Campinas, Limeira, 2010.

CURKOVIC, P.; JERBIC, B. **Honey-bees optimization algorithm applied to path planning problem**. International Journal of Simulation Modelling, Vol. 6, No. 3, pp. 154–164. 2007.

DENEUBOURG, J.-L.; ARON, S; Goss, S.; PASTEELS, J.-M.**The self-organizing exploratory pattern of the argentine ant**.Journal of Insect Behav., Vol. 3, pp. 159–168, 1990.

DORIGO, M.; DI CARO, G. & GAMBARDELLA, L. M. **Ant algorithms for discrete optimization**.Artificial Life, 5,2, p. 137-172, 1999.

ESTADÃO. **Inteligência artificial caminha a passos lentos.** 2016. Disponível em: <http://link.estadao.com.br/noticias/inovacao,inteligencia-artificial-caminha-a-passos-lentos,10000048402>. Acesso em: 13 jun. 2016.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa.** 4. ed. São Paulo: Atlas, 2002. 176 p.

GUDWIN, Ricardo Ribeiro. **Novas fronteiras na inteligência artificial e na robótica**. Bauru. 2005.

PRO, Insider. **Inteligência artificial:** 7 Empresas que estão a fazer maravilhas. Disponívelem: <https://insider.pro/pt/article/80392/>. Acessoem: 13 jun. 2016.

KASHAN, A.H.; KARIMI, B. **A discrete particle swarm optimization algorithm for scheduling parallel machines.** Computers and Industrial Engineering, Vol. 56,No. 1, pp. 216–223. 2009.

KENNEDY, J.; EBERHART, R.C.; SHI., Y. **Swarm intelligence**, San Francisco: Morgan Kaufmann/ Academic Press. 2001

KURZWEIL, Ray. **The age of intelligent machines.** Nova York: The Mit Press, 1990.

LACERDA, Marcelo Gomes Pereira de. **Uma nova heurística de segregação de cardumes para otimização multi-solução de problemas multimodais.** 2012. 84 f. TCC (Graduação) - Curso de Engenharia de Computação, Universidade de Pernambuco, Pernambuco, 2012.

LIBRELOTTO, Solange Rubert; MOZZAQUATRO, PatriciaMariotto. Análise dos algoritmos de mineração j48 e apriori aplicados na detecção de indicadores da qualidade de vida e saúde. **Revista Interdisciplinar de Ensino Pesquisa e Extensão**, Santa Maria, v. 1, n. 1, p.26-37, 2013.

LIU, Y.; PASSINO, K.M. **Biomimicry of social foraging bacteria for distributed optimization: models, principles and emergent behaviors.**Journal of OptimizationTheories and Applications, Vol. 115, No. 3, pp. 603– 628, 2002.

LUGER, George; STUBBLEFIELD, William. **Artificial intelligence**: structures and strategies for complex problem solving. Pensilvânia: Addison Wesley Longman, 1993.

MAITRA, M.; CHATTERJEE, A.**A novel technique for multilevel optimal magnetic resonance brain image thresholding using bacterial foraging.** Journal of the International Measurement Confederation, Vol. 41, No.10, pp. 1124–1134. 2008.

MILLER, P. **Teoria dos enxames**.NationalGeographic Brasil. Ano 7, n. 88, p. 36-57, Julho, 2007

MILLONAS, M. M. **Swarms, phase transitions, and collective intelligence**. In C.G. Langton (Ed.), Artificial Life III, pp. 417–445, 1994.

MOEDINGER, Luis Henrique. **Algoritmos evolutivos e inteligência coletiva aplicados a problemas de otimização não-linear com restrições:** fundamentos e estudo comparativo**.** 2005. 89 f. Dissertação (Mestrado) - Curso de Engenharia de Produção e Sistemas, Pontifícia Universidade Católica do Paraná, Curitiba, 2005.

NASCIMENTO, Lídice Cabral; CRUZ, Carla Bernadete Madureira. Mineração de dados e adaptação de modelos de classificação de cobertura e uso da terra para imagem Worldview2. In: Anais xvi simpósio brasileiro de sensoriamento remoto, 16., 2013, Foz da Iguaçu. **Mineração de dados e adaptação de modelos de**

classificação de cobertura e uso da terra para imagem worldview2. Foz do Iguaçu: Inpe, 2013. p. 2345 - 2352.

PASSINO, K.M. **Biomimicry of bacterial foraging for distributed optimization and control.**IEEE Control Systems Magazine, Vol. 22, No. 3, pp. 52–67. 2002

PINHEIRO, Constance Ribeiro de Souza. **Revisão de métodos de inteligência coletiva para a solução de problemas logísticos na distribuição eficiente de suprimentos**. Enegep**,** Rio de Janeiro, out. 2008.

POLIT, D. F.; HUNGLER, B. P. **Nursing research:** principles and methods. 3a ed. Philadelphia: J. B. Lippincott, 1987.

RAMISCH, Carlos. **Trabalho prático de mineração de dados:** Algoritmos de aprendizado para avaliação de carros.2009.

RIBEIRO, Sergio Ferreira. **Avaliando algoritmos de otimização baseados em colônia de formigas utilizando métricas de redes complexas.** 2013. 40 f. TCC (Graduação) - Curso de Engenharia de Computação, Universidade de Pernambuco, Pernambuco, 2013.

SELLTIZ, Claire et ai**. Métodos de pesquisa nas relações sociais**. São Paulo: Herder, 1967.

SERAPIÃO, Adriane Beatriz de Souza. Fundamentos de otimização por inteligência de enxames: uma visão geral. **Revista Controle & Automação,** São Paulo, v. 20, n. 3, p.271-304, ago. 2009.

SIMON, H.A. **Why should machines learn?** In Michalski et al.,1983.

STEPHENS, D.; KREBS; J. **Foraging theory**. Princeton, NJ: Princeton Univ. Press, 1986.

THOMÉ, Antônio C. G. Data Warehouse, Data Mining. **Redes neurais –** uma ferramenta para kdd e data mining**.**2002.

WITTEN, I. H.; FRANK, E. **Data mining:** practical machine learning tools andtechniques**.** 2 ed. San Francisco: Morgan KaufmannPublishers, 2005.

APÊNDICE 1 – ÁRVORE DE DECISÃO DA BASE *SOYBEAN* ATRIBUTOS ORIGINAIS.

```
leafspot-size = lt-1/8
|   canker-lesion = dna
|   |   leafspots-marg = w-s-marg
|   |   |   seed-size = norm: bacterial-blight (21.0/1.0)
|   |   |   seed-size = lt-norm: bacterial-pustule (3.23/1.23)
|   |   leafspots-marg = no-w-s-marg: bacterial-pustule (17.91/0.91)
|   |   leafspots-marg = dna: bacterial-blight (0.0)
|   canker-lesion = brown: bacterial-blight (0.0)
|   canker-lesion = dk-brown-blk: phytophthora-rot (4.78/0.1)
|   canker-lesion = tan: purple-seed-stain (11.23/0.23)
leafspot-size = gt-1/8
|   roots = norm
|   |   mold-growth = absent
|   |   |   fruit-spots = absent
|   |   |   |   leaf-malf = absent
|   |   |   |   |   fruiting-bodies = absent
|   |   |   |   |   |   date = april: brown-spot (5.0)
|   |   |   |   |   |   date = may: brown-spot (24.0/1.0)
|   |   |   |   |   |   date = june
|   |   |   |   |   |   |   precip = lt-norm: phyllosticta-leaf-spot (4.0)
|   |   |   |   |   |   |   precip = norm: brown-spot (5.0/2.0)
|   |   |   |   |   |   |   precip = gt-norm: brown-spot (21.0)
|   |   |   |   |   |   date = july
|   |   |   |   |   |   |   precip = lt-norm: phyllosticta-leaf-spot (1.0)
|   |   |   |   |   |   |   precip = norm: phyllosticta-leaf-spot (2.0)
|   |   |   |   |   |   |   precip = gt-norm: frog-eye-leaf-spot (11.0/5.0)
|   |   |   |   |   |   date = august
|   |   |   |   |   |   |   leaf-shread = absent
|   |   |   |   |   |   |   |   seed-tmt = none: alternarialeaf-spot (16.0/4.0)
|   |   |   |   |   |   |   |   seed-tmt = fungicide
|   |   |   |   |   |   |   |   |   plant-stand = normal: frog-eye-leaf-spot (6.0)
|   |   |   |   |   |   |   |   |   plant-stand = lt-normal: alternarialeaf-spot (5.0/1.0)
|   |   |   |   |   |   |   |   seed-tmt = other: frog-eye-leaf-spot (3.0)
|   |   |   |   |   |   |   leaf-shread = present: alternarialeaf-spot (2.0)
|   |   |   |   |   |   date = september
|   |   |   |   |   |   |   stem = norm: alternarialeaf-spot (44.0/4.0)
|   |   |   |   |   |   |   stem = abnorm: frog-eye-leaf-spot (2.0)
|   |   |   |   |   |   date = october: alternarialeaf-spot (31.0/1.0)
|   |   |   |   |   fruiting-bodies = present: brown-spot (34.0)
|   |   |   |   leaf-malf = present: phyllosticta-leaf-spot (10.0)
|   |   |   fruit-spots = colored
|   |   |   |   fruit-pods = norm: brown-spot (2.0)
|   |   |   |   fruit-pods = diseased: frog-eye-leaf-spot (62.0)
|   |   |   |   fruit-pods = few-present: frog-eye-leaf-spot (0.0)
|   |   |   |   fruit-pods = dna: frog-eye-leaf-spot (0.0)
|   |   |   fruit-spots = brown-w/blk-specks
```

```
|   |   |   crop-hist = diff-lst-year: brown-spot (0.0)
|   |   |   crop-hist = same-lst-yr: brown-spot (2.0)
|   |   |   crop-hist = same-lst-two-yrs: brown-spot (0.0)
|   |   |   crop-hist = same-lst-sev-yrs: frog-eye-leaf-spot (2.0)
|   |   |  fruit-spots = distort: brown-spot (0.0)
|   |   |  fruit-spots = dna: brown-stem-rot (9.0)
|   |  mold-growth = present
|   |   |  leaves = norm: diaporthe-pod-&-stem-blight (7.25)
|   |   |  leaves = abnorm: downy-mildew (20.0)
|  roots = rotted
|   |  area-damaged = scattered: herbicide-injury (1.1/0.1)
|   |  area-damaged = low-areas: phytophthora-rot (30.03)
|   |  area-damaged = upper-areas: phytophthora-rot (0.0)
|   |  area-damaged = whole-field: herbicide-injury (3.66/0.66)
|  roots = galls-cysts: cyst-nematode (7.81/0.17)
leafspot-size = dna
|  int-discolor = none
|   |  leaves = norm
|   |   |  stem-cankers = absent
|   |   |   |  canker-lesion = dna: diaporthe-pod-&-stem-blight (5.53)
|   |   |   |  canker-lesion = brown: purple-seed-stain (0.0)
|   |   |   |  canker-lesion = dk-brown-blk: purple-seed-stain (0.0)
|   |   |   |  canker-lesion = tan: purple-seed-stain (9.0)
|   |   |  stem-cankers = below-soil: rhizoctonia-root-rot (19.0)
|   |   |  stem-cankers = above-soil: anthracnose (0.0)
|   |   |  stem-cankers = above-sec-nde: anthracnose (24.0)
|   |  leaves = abnorm
|   |   |  stem = norm
|   |   |   |  plant-growth = norm: powdery-mildew (22.0/2.0)
|   |   |   |  plant-growth = abnorm: cyst-nematode (4.3/0.39)
|   |   |  stem = abnorm
|   |   |   |  plant-stand = normal
|   |   |   |   |  leaf-malf = absent
|   |   |   |   |   |  seed = norm: diaporthe-stem-canker (21.0/1.0)
|   |   |   |   |   |  seed = abnorm: anthracnose (9.0)
|   |   |   |   |  leaf-malf = present: 2-4-d-injury (3.0)
|   |   |   |  plant-stand = lt-normal
|   |   |   |   |  fruiting-bodies = absent: phytophthora-rot (50.16/7.61)
|   |   |   |   |  fruiting-bodies = present
|   |   |   |   |   |  roots = norm: anthracnose (11.0/1.0)
|   |   |   |   |   |  roots = rotted: phytophthora-rot (12.89/2.15)
|   |   |   |   |   |  roots = galls-cysts: phytophthora-rot (0.0)
|  int-discolor = brown
|   |  leaf-malf = absent: brown-stem-rot (35.73/0.73)
|   |  leaf-malf = present: 2-4-d-injury (3.15/0.68)
|  int-discolor = black: charcoal-rot (22.22/2.22)
```

APÊNDICE 2 – ÁRVORE DE DECISÃO DA BASE *SOYBEAN* ATRIBUTOS SELECIONADOS PELO PSOSEARCH.

```
leafspot-size = lt-1/8
|   canker-lesion = dna
|   |   leafspots-marg = w-s-marg
|   |   |   roots = norm: bacterial-blight (21.66/1.66)
|   |   |   roots = rotted: bacterial-pustule (2.06/0.06)
|   |   |   roots = galls-cysts: cyst-nematode (0.51/0.02)
|   |   leafspots-marg = no-w-s-marg: bacterial-pustule (17.91/0.91)
|   |   leafspots-marg = dna: bacterial-blight (0.0)
|   canker-lesion = brown: bacterial-blight (0.0)
|   canker-lesion = dk-brown-blk: phytophthora-rot (4.78/0.1)
|   canker-lesion = tan: purple-seed-stain (11.23/0.23)
leafspot-size = gt-1/8
|   roots = norm
|   |   leaf-mild = absent
|   |   |   fruit-spots = absent
|   |   |   |   fruiting-bodies = absent
|   |   |   |   |   leaf-malf = absent
|   |   |   |   |   |   date = april: brown-spot (5.0)
|   |   |   |   |   |   date = may: brown-spot (24.0/1.0)
|   |   |   |   |   |   date = june
|   |   |   |   |   |   |   precip = lt-norm: phyllosticta-leaf-spot (4.0)
|   |   |   |   |   |   |   precip = norm: brown-spot (5.0/2.0)
|   |   |   |   |   |   |   precip = gt-norm: brown-spot (21.0)
|   |   |   |   |   |   date = july
|   |   |   |   |   |   |   precip = lt-norm: phyllosticta-leaf-spot (1.0)
|   |   |   |   |   |   |   precip = norm: phyllosticta-leaf-spot (2.0)
|   |   |   |   |   |   |   precip = gt-norm: frog-eye-leaf-spot (11.0/5.0)
|   |   |   |   |   |   date = august
|   |   |   |   |   |   |   seed-tmt = none: alternarialeaf-spot (17.0/4.0)
|   |   |   |   |   |   |   seed-tmt = fungicide
|   |   |   |   |   |   |   |   germination = 90-100: frog-eye-leaf-spot (5.0)
|   |   |   |   |   |   |   |   germination = 80-89: alternarialeaf-spot (2.0/1.0)
|   |   |   |   |   |   |   |   germination = lt-80: alternarialeaf-spot (5.0/1.0)
|   |   |   |   |   |   |   seed-tmt = other: frog-eye-leaf-spot (3.0)
|   |   |   |   |   |   date = september
|   |   |   |   |   |   |   stem-cankers = absent: alternarialeaf-spot (44.0/4.0)
|   |   |   |   |   |   |   stem-cankers = below-soil: alternarialeaf-spot (0.0)
|   |   |   |   |   |   |   stem-cankers = above-soil: alternarialeaf-spot (0.0)
|   |   |   |   |   |   |   stem-cankers = above-sec-nde: frog-eye-leaf-spot (2.0)
|   |   |   |   |   |   date = october: alternarialeaf-spot (31.0/1.0)
|   |   |   |   |   leaf-malf = present: phyllosticta-leaf-spot (10.0)
|   |   |   |   fruiting-bodies = present: brown-spot (34.0)
|   |   |   fruit-spots = colored
|   |   |   |   fruit-pods = norm: brown-spot (2.0)
```

```
|   |   |   fruit-pods = diseased: frog-eye-leaf-spot (62.0)
|   |   |   fruit-pods = few-present: frog-eye-leaf-spot (0.0)
|   |   |   fruit-pods = dna: frog-eye-leaf-spot (0.0)
|   |   |  fruit-spots = brown-w/blk-specks
|   |   |   leaves = norm: diaporthe-pod-&-stem-blight (6.8)
|   |   |   leaves = abnorm
|   |   |   |  canker-lesion = dna: brown-spot (0.0)
|   |   |   |  canker-lesion = brown: brown-spot (2.0)
|   |   |   |  canker-lesion = dk-brown-blk: frog-eye-leaf-spot (2.0)
|   |   |   |  canker-lesion = tan: brown-spot (0.0)
|   |   |  fruit-spots = distort: brown-spot (0.0)
|   |   |  fruit-spots = dna: brown-stem-rot (9.0)
|   |  leaf-mild = upper-surf: brown-spot (0.0)
|   |  leaf-mild = lower-surf: downy-mildew (20.45/0.45)
|  roots = rotted
|   |  area-damaged = scattered: herbicide-injury (1.1/0.1)
|   |  area-damaged = low-areas: phytophthora-rot (30.03)
|   |  area-damaged = upper-areas: phytophthora-rot (0.0)
|   |  area-damaged = whole-field: herbicide-injury (3.66/0.66)
|  roots = galls-cysts: cyst-nematode (7.81/0.17)
leafspot-size = dna
|  int-discolor = none
|   |  leaves = norm
|   |   |  stem-cankers = absent
|   |   |   |  canker-lesion = dna: diaporthe-pod-&-stem-blight (5.53)
|   |   |   |  canker-lesion = brown: purple-seed-stain (0.0)
|   |   |   |  canker-lesion = dk-brown-blk: purple-seed-stain (0.0)
|   |   |   |  canker-lesion = tan: purple-seed-stain (9.0)
|   |   |  stem-cankers = below-soil: rhizoctonia-root-rot (19.0)
|   |   |  stem-cankers = above-soil: anthracnose (0.0)
|   |   |  stem-cankers = above-sec-nde: anthracnose (24.0)
|   |  leaves = abnorm
|   |   |  canker-lesion = dna
|   |   |   |  plant-growth = norm: powdery-mildew (22.0/2.0)
|   |   |   |  plant-growth = abnorm: diaporthe-stem-canker (13.0/3.0)
|   |   |  canker-lesion = brown
|   |   |   |  leaf-malf = absent
|   |   |   |   |  fruit-pods = norm
|   |   |   |   |   |  fruiting-bodies = absent: anthracnose (4.0)
|   |   |   |   |   |  fruiting-bodies = present: diaporthe-stem-canker (10.0)
|   |   |   |   |  fruit-pods = diseased: anthracnose (6.0)
|   |   |   |   |  fruit-pods = few-present: cyst-nematode (0.63)
|   |   |   |   |  fruit-pods = dna: rhizoctonia-root-rot (1.0)
|   |   |   |  leaf-malf = present: 2-4-d-injury (2.87/0.64)
|   |   |  canker-lesion = dk-brown-blk
|   |   |   |  area-damaged = scattered: 2-4-d-injury (2.95/1.26)
|   |   |   |  area-damaged = low-areas
|   |   |   |   |  fruit-pods = norm: phytophthora-rot (2.4/1.08)
```

```
|   |   |   |   |   fruit-pods = diseased
|   |   |   |   |   |   roots = norm: anthracnose (4.14/0.14)
|   |   |   |   |   |   roots = rotted: phytophthora-rot (5.46/0.18)
|   |   |   |   |   |   roots = galls-cysts: phytophthora-rot (0.0)
|   |   |   |   |   fruit-pods = few-present: phytophthora-rot (2.97/1.34)
|   |   |   |   |   fruit-pods = dna: phytophthora-rot (45.59/1.52)
|   |   |   |   area-damaged = upper-areas: 2-4-d-injury (2.62/0.93)
|   |   |   |   area-damaged = whole-field: anthracnose (7.72/2.72)
|   |   |   canker-lesion = tan: phytophthora-rot (0.0)
|   int-discolor = brown
|   |   leaf-malf = absent: brown-stem-rot (35.73/0.73)
|   |   leaf-malf = present: 2-4-d-injury (3.15/0.68)
|   int-discolor = black: charcoal-rot (22.22/2.22)
```