UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE CIÊNCIAS EXATAS E DA TERRA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA E MEIO AMBIENTE

DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS DE MONITORAMENTO AMBIENTAL PARA AVALIAÇÃO DA QUALIDADE DA ÁGUA

INARA APARECIDA FERRER SILVA

ORIENTADOR: PROF. DR. PETER ZEILHOFER

Cuiabá- MT, junho/2007

UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE CIÊNCIAS EXATAS E DA TERRA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA E MEIO AMBIENTE

DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS DE MONITORAMENTO AMBIENTAL PARA AVALIAÇÃO DA QUALIDADE DA ÁGUA

INARA APARECIDA FERRER SILVA

Dissertação apresentada ao Programa de Pós-graduação em Física e Meio Ambiente da Universidade Federal de Mato Grosso, como parte dos requisitos para obtenção do título de Mestre em Física e Meio Ambiente.

ORIENTADOR: PROF. DR. PETER ZEILHOFER

Cuiabá- MT, Junho/2007

DEDICATÓRIA

Ao bom Deus pela vida, aos meus pais pelo infinito amor e ajuda nesta caminhada, ao meu esposo pelo companheirismo e paciência, a minha irmã Ivana que me incentivou em todos os sentidos, aos meus filhos Giovanni e Enzo que me fazem renascer a cada dia.

AGRADECIMENTOS

Ao Prof. Dr. Peter Zeilhofer, pela orientação, incentivo, paciência, confiança e grande amizade, itens que possibilitaram a realização deste trabalho, a minha gratidão.

Ao Prof. José de Souza Nogueira por permitir através da coordenação deste curso de mestrado o meu ingresso e realização.

Aos especialistas Prof. Márcio Mecca, Marizete Caovilla, Andreza Thiesen, Liliana Zeilhofer, que responderam `a nossa classificação com dedicação, item primordial para a realização deste trabalho, meus sinceros agradecimentos.

A todos os professores do Programa de Mestrado em Física e Meio Ambiente da Universidade Federal de Mato Grosso pela dedicação a este curso e transmissão de conhecimento.

Ao meu pai, pela força moral, pelo apoio nas horas difíceis, pela palavra certa nas horas incertas, meu eterno amor.

`A minha mãe, que não mediu esforços para me ajudar com meus filhos, a minha eterna gratidão e amor.

`A minha irmã que pelos incentivos e forças nas horas em que elas desaparecem.

`A Soilce pela gentileza em nos atender, pela amizade e carinho de sempre.

Aos funcionários e técnicos da UFMT que proporcionaram o ambiente propício para as aulas e desenvolvimento deste trabalho.

A todos os colegas de mestrado, pela amizade, pela palavra amiga, pela ajuda, toda a minha gratidão.

Ao Centro Federal de Educação Tecnológica, pelo apoio e liberação de 50% da carga horária.

Aos colegas de trabalho do CEFET que assumiram os 50% de carga horária para que eu pudesse realizar este trabalho.

`A minha chefe Danusa Balthazar Gonçalves pela compreensão e apoio para a realização deste trabalho, dedico o sucesso desta realização profissional, meus sinceros agradecimentos.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS

IA- Inteligência Artificial

KDD- Knowledge Discovery in Databases

DM- Data Mining

BD- Banco de Dados

SIBAC- Sistema de Monitoramento Ambiental da Bacia do Rio Cuiabá

SGBD- Sistema Gerenciador de Banco de Dados

CF- Coliforme Termotolerante

CT- *Eschrichia Coli*

NTK- Nitrogênio Total Kjheldal

P- Fósforo

OD- Oxigênio Dissolvido

DBO- Demanda Bioquímica de Oxigênio

pH- Potencial Hidrogeniônico

ST- Sólidos Dissolvidos Totais

DQO- Demanda Química de Oxigênio

Turb- Turbidez

CPRH- Agência Estadual do Meio Ambiente do Estado de Pernambuco

RESUMO

SILVA, I.A.F. ***Descoberta de Conhecimento em Base de Dados de Monitoramento Ambiental para Avaliação da Qualidade da Água.*** Cuiabá, 2007. 134 p. Dissertação (Mestrado) – Programa de Pós-Graduação em Física e Meio Ambiente, Universidade Federal de Mato Grosso.

Com o sucessivo crescimento econômico e populacional, aumenta a demanda pelo uso dos recursos hídricos, intensificam-se os conflitos entre os diversos usos e a degradação da qualidade dos corpos da água. Assim entidades e órgãos governamentais mantêm bases de dados de qualidade de água para subsidiar o monitoramento ambiental. Porém, o volume crescente desses dados, associado a limitação humana em analisá-los e relacioná-los em sua totalidade, faz com que a interpretação conclusiva da situação real da qualidade da água de uma bacia seja uma tarefa difícil. Em conseqüência disso, a necessidade de se desenvolver novas ferramentas e técnicas de extração de conhecimento a partir de dados armazenados também vem crescendo, juntamente com a necessidade de se criar instrumentos automatizados para avaliar a evolução da qualidade das águas em seus estados atuais. Assim, o objetivo deste trabalho é propor a descoberta de conhecimento em bases de dados de monitoramento ambiental, utilizando técnicas de mineração de dados (Data Mining) para descobrir regras que representem níveis de qualidade da água nos seus estados atuais. Em um estudo de caso utilizando a base de dados de monitoramento ambiental da bacia do Rio Cuiabá (SIBAC), no Estado de Mato Grosso, obteve-se uma classificação de especialistas do domínio para quatro usos (Balneabilidade, Abastecimento, Irrigação e Manutenção dos Ciclos Biogeoquímicos Naturais) nos níveis de qualidade (ótimo, bom, regular, ruim e péssimo). Posteriormente as amostras foram submetidas a um software de Mineração de Dados, oito algoritmos de classificação foram utilizados para desenvolvimento de regras de avaliação automatizada. Nas classificações obtidas, destacaram-se os algoritmos PART, JRIP e OneR com níveis de acerto entre 37.7 e 74.5 %, mostrando uma possível contribuição das técnicas avaliadas para a sintetização e interpretação de dados presentes em bases de dados de qualidade da água.

Palavras-chave: mineração de dados, qualidade da água, monitoramento ambiental, base de dados.

ABSTRACT

SILVA, I.A.F. ***Knowledge Discovery in Databases of monitored ambient to Evaluation of the Quality of the Water.*** Cuiabá, 2007. 134p. Dissertação (Mestrado) – Programa de Pós-Graduação em Física e Meio Ambiente, Universidade Federal de Mato Grosso.

The successive economic and population growth has increased the demand for the use of the hybrid resources, the conflicts between the diverse uses and the degradation of the quality of the bodies of the water have been intensified. Thus, entities and governmental organizations have kept databases about the quality of water in order to subsidize the environmental supervision. However, the increasing volume of these data associated the limitation of human resources for analyzing and relating them to the totality, makes the conclusive interpretation of the real situation about the quality of the basin water a difficult task. In consequence, the necessity of developing new tools and extraction techniques of knowledge from stored data have been increasing with the necessity of creating automatized instruments to evaluate the evolution of the quality of waters in their current phases. Consequently the objective of this study is to consider the knowledge discovery in databases of monitored ambient by means of techniques of data mining to find out rules that represent levels of quality of the water in their current phases. In a case study using the database of monitored ambient of the basin of Cuiaba river (SIBAC), in the State of Mato Grosso, in Brazil. A classification of specialists was obtained. It presents the domain for four uses (Balneability, Supplying, Irrigation and Maintenance of the Natural Biogeochemical Cycles) in the quality levels (excellent, good, regular, bad and very bad). Afterward the samples had been submitted to a Data Mining Software, eight algorithms of classification had been used for development of automatized rules evaluation. In the obtained classifications, the algorithms PART, JRIP and OneR with levels of rightness between 37.7 and 74.5 % had been distinguished, revealing to a possible contribution of the techniques evaluated for the sintetization and interpretation of data gifts in databases of quality of the water.

Key words: data mining, quality of the water, environmental checking, data base.

1- INTRODUÇÃO

A água é um bem que não se produz, é um recurso que a natureza recicla através do seu ciclo hidrológico. Observa-se que a medida que as regiões se desenvolvem, mais intenso é o uso dos recursos hídricos, maior o potencial de conflitos entre usos e maiores os riscos de degradação da qualidade dos corpos d'água. Assim a preocupação com a água é crescente, ela precisa ser gerida como um bem escasso, de alto valor econômico, não só nos seus aspectos quantitativos, mas, principalmente em relação a sua qualidade, causando prejuízos e restrições nos seus usos múltiplos.

Os problemas relativos a qualidade da água envolvem um espectro bastante amplo dentro das áreas de estudo hidroambiental e na determinação das potenciais fontes de contaminação resultantes de: disposições inadequadas de resíduos sólidos e líquidos, de natureza doméstica e industrial, alterações provocadas por empreendimento para geração de energia, resfriamento de águas de termoelétricas, agricultura e criação de animais. Todas essas ações de origem antrópica que ocorrem na bacia causam impactos que se inter-relacionam com os processos naturais que ocorrem na bacia.

A questão da qualidade ganhou evidência com a sanção da Lei Federal N° 9.433, de 8 de janeiro de 1997, que instituiu a Política Nacional de Recursos Hídricos, tendo como um dos seus fundamentos gerir tais recursos, proporcionando uso múltiplo, em consonância com objetivos que assegurem a atual e as futuras gerações a necessária disponibilidade de água, em padrões de qualidade adequados aos respectivos usos. Isto demonstra a preocupação com a integração da gestão dos aspectos de qualidade e quantidade, onde uma das ações principais é a "integração da gestão de recursos hídricos com a gestão ambiental".

O CONSELHO NACIONAL DO MEIO AMBIENTE-CONAMA N° 357, de 17 de março de 2005, dispõe sobre a classificação dos corpos d'água e diretrizes ambientais para o seu enquadramento, bem como estabelece as condições e padrões

de lançamento de efluentes , e dá outras providências. Segundo o CONAMA 357 as águas são tratadas como: águas doces, águas salobras, águas salinas, ambiente lêntico e ambiente lótico. As águas doces, salobras e salinas do Território Nacional são classificadas, segundo a qualidade requerida para os seus usos preponderantes, em treze classes de qualidade. As águas doces, objeto do nosso estudo de caso na Bacia do Rio Cuiabá, podem ser classificadas em: classe especial, classe 1, classe 2, classe 3 e classe 4. O rio Cuiabá, é classificado como um rio de classe 2, onde suas águas podem ser destinadas:

- Ao abastecimento para consumo humano, após tratamento convencional;
- A proteção de comunidades aquáticas;
- A recreação de contato primário, tais como natação, esqui aquático e mergulho, conforme Resolução CONAMA 274, de 2000;
- A irrigação de hortaliças, plantas frutíferas e de parques, jardins, campos de esporte e lazer, com os quais o público possa vir a ter contato direto; e,
- A aqüicultura e `a atividade de pesca.

Os padrões de qualidade das águas determinados nesta Resolução estabelecem limites individuais para cada substância. O conjunto de parâmetros de qualidade de água selecionado para subsidiar a proposta de enquadramento deverá ser monitorado periodicamente pelo Poder Público.

Sendo o enquadramento de um manancial nas classes do CONAMA 357/2005 comumente um ato estático, persiste a necessidade de se criar instrumentos para avaliar a evolução da qualidade das águas. Neste contexto, este trabalho propõe a descoberta de conhecimento para avaliação das condições atuais de qualidade da água dos mananciais da bacia do Rio Cuiabá, em sua grande maioria enquadrada na classe 2 do CONAMA, visando o estabelecimento de critérios mais restritivos para os usos dos recursos hídricos na bacia do rio Cuiabá.

Para monitorar a qualidade das bacias hidrográficas, universidades e órgãos governamentais mantém registros das coletas realizadas *in loco* em bancos de dados. Normalmente estes registros indicam o trecho do rio onde foi feita determinada coleta, o mês e o ano da coleta, os elementos químicos, físico-químicos e biológicos

presentes e suas respectivas concentrações. O volume de dados armazenados nessas bases cresce rapidamente, elas contêm informações valiosas, porém, a recuperação dessas informações não é de forma direta, muita informação e conhecimento úteis podem estar sendo desperdiçados, ficando ocultos dentro das bases de dados. Usualmente, estas informações não estão disponíveis devido a falta de ferramentas adequadas para sua extração, associada a limitação humana de analisar extensas bases de dados e extrair relações entre elas.

Assim o aumento no volume das bases de dados de monitoramento ambiental associado a demanda crescente por conhecimento novo, voltado a decisões estratégicas, tem provocado o interesse crescente em descobrir conhecimentos relevantes nessas bases de dados, especialmente em bases de monitoramento ambiental

Neste contexto, a descoberta de conhecimento em bases de dados de monitoramento ambiental, utilizando técnicas de Data Mining, para avaliar a qualidade da água pode ser uma ferramenta importante para o processo de tomada de decisão realizados por órgãos e gestores de recursos hídricos na avaliação qualitativa dos mesmos. Este processo de "busca" de conhecimento implícito em bases de dados é uma área da Inteligência Artificial, chamada de KDD (Descoberta de Conhecimento em Bases de Dados). KDD pode ser visto ainda como o processo da descoberta de novas correlações, padrões e tendências significativas por meio da análise minuciosa de grandes conjuntos de dados. É um processo genérico de descoberta de conhecimento que inclui três fases: pré-processamento, mineração de dados e pós-processamento do conhecimento obtido. A mineração de dados (Data Mining) é a principal fase do processo de KDD, pode ser usada sob vários segmentos que armazenam informações, aplicando algoritmos nesses segmentos. As ferramentas de Data Mining identificam todas as correlações existentes nessas fontes de dados, assim é possível extrair o conhecimento sucinto e com maior facilidade para subsidiar a tomada de decisão.

Considerando que o objetivo deste trabalho é a descoberta de conhecimento não trivial em bases de dados de monitoramento ambiental para avaliação da qualidade da água, este trabalho tem a intenção de realizar:

- Um estudo sobre o processo de descoberta de conhecimento (KDD) em

bases de dados, especificamente as de monitoramento ambiental;

- Acessar padrões dos dados de uma base de Monitoramento Ambiental – SIBAC (Sistema Integrado de Monitoramento Ambiental da Bacia do Rio Cuiabá) utilizando técnicas de mineração de dados para descobrir conhecimento dos níveis de qualidade da água, no contexto de seus possíveis usos. Explicitar este conhecimento através de modelos de regras de classificação e previsão destes padrões na Bacia do rio Cuiabá.
- Avaliar o desempenho de diversos algoritmos de classificação presentes no software Weka para o problema em estudo, visando uma futura utilização destes em um Sistema Especialista para avaliação automatizada da base de dados SIBAC e de saídas de modelos de qualidade de água.

Assim, o propósito deste trabalho é responder a seguinte questão: Como extrair o conhecimento implícito em bases de dados de monitoramento ambiental e com isto melhorar o processo de tomada de decisão para avaliação qualitativa da gestão de recursos hídricos?

2- REVISÃO DE LITERATURA

2.1- A ÁGUA

NALINI (2003) diz que o mais valioso recurso natural da humanidade é a água, sem ela não há vida.

> *O conjunto das atividades humanas cada vez mais diversificado, associado ao crescimento demográfico, vem exigindo atenção maior às necessidades de uso de água para as mais diversas finalidades. Essas necessidades cobram seus tributos tanto em termos quantitativos como qualitativos, e se evidenciam principalmente em regiões com características de maior desenvolvimento urbano, industrial e agrícola .(PHILLIPI JR apud MANCUSO & SANTOS, 2003).*

> *O Brasil detém 8% das reservas de água doce de todo o planeta. De toda a água da Terra, só 0,2% pode ser aproveitada. 97,3% da água concentra-se nos oceanos, 2,07% nas geleiras e calotas polares e somente 0,63% de água doce, própria para o uso. No Brasil 80% da água disponível está na Bacia Amazônica, onde vivem apenas 7% da população. Os 20% restantes distribuem-se desigualmente pelo país, segundo dados de 1997 do CPRM - Serviço Geológico do Brasil. A região Sudeste é aquela em que os rios se encontram mais comprometidos. Enquanto que de 92 a 95% das casas do país recebem água potável, apenas 35% delas possuem coleta de esgotos. Pior ainda, 16% dos dejetos das zonas urbanas são tratados. Os outros 84% são despejados diretamente nos rios.(NALINI, 2003)*

A falta d'agua tornará o produto mais valioso, suas taxas consequentemente serão aumentadas e a disponibilidade do produto será controlada. Hoje cobra-se pelo tratamento, encanamento, ou seja pelos serviços que a companhia de saneamento presta a população. As tarifas são até irrisórias se pensarmos na raridade do bem e nos problemas que estamos enfrentando. Muito em breve teremos a água sendo cobrada como bem econômico justamente pela escassez do produto.

> *Cobrar pela água in natura não eliminará o problema. Mas o atenuará. Em seguida virá a legislação impondo a obrigatoriedade de equipamentos que disciplinem a vazão máxima para as casas e empresas. Depois, normas para a utilização de eletrodomésticos, aparelhos hidráulicos e encanamentos de controle no fornecimento e de vazão intermitente, ou reduzida. Planos para reutilização da água e para o seu uso racional, (NALINI,2003).*

Segundo a Unicef (Fundo das Nações Unidas para a Infância), menos da metade da população mundial tem acesso à água potável. A irrigação corresponde a 73% do consumo de água, 21% vai para a indústria e apenas 6% destina-se ao consumo doméstico. Um bilhão e 200 milhões de pessoas (35% da população mundial) não têm acesso a água tratada. Um bilhão e oitocentos milhões de pessoas (43% da população mundial) não contam com serviços adequados de saneamento básico, ou seja, são dez milhões de pessoas que morrem anualmente em decorrência de doenças intestinais transmitidas pela água.

De acordo com a Política Nacional de Recursos Hídricos instituída pela Lei n° 9.433 de 08.01.97, o intuito da cobrança pela utilização dos recursos hídricos é reconhecer a água como bem econômico e dar ao usuário uma indicação de seu real valor. Além disso, o propósito é incentivar o uso racional e sustentável da água e obter recursos financeiros para o financiamento dos programas e intervenções contemplados nos planos de recursos hídricos. A fixação dos valores a serem cobrados pela utilização dos recursos hídricos considerará três fases: a captação, a extração e derivação, a diluição, transporte e assimilação de efluentes e outros usos que alterem o regime, a quantidade ou a qualidade da água existente em um corpo d'água.

Portanto o uso da água como bem econômico deverá ser controlado por um instrumento jurídico que discipline seu uso e manutenção para garantir o consumo das populações e outras atividades econômicas.

2.2- ASPECTOS INSTITUCIONAIS DE GERENCIAMENTO DE RECURSOS HÍDRICOS

A Lei Federal 9.433, de 08.01.97, institui a Política Nacional de Recursos Hídricos, cria o Sistema Nacional de Gerenciamento de Recursos Hídricos, regulamenta o inciso XIX do art. 21 da Constituição Federal.

> *A Política Nacional de Recursos Hídricos baseia-se nos seguintes fundamentos: a água é um bem de domínio publico; a água é um recurso natural limitado, dotado de valor econômico; em situações de escassez, o uso prioritário dos recursos hídricos e o consumo humano e a dessedentação de animais; a gestão de recursos hídricos deve sempre proporcionar o uso múltiplo das águas; a bacia hidrográfica e a unidade territorial para a implementação da Política Nacional de Recursos Hídricos e a atuação do Sistema Nacional de Gerenciamento de Recursos Hídricos; a gestão de recursos hídricos deve ser descentralizada e contar com a participação do Poder Publico, dos usuários e das comunidades.(*REBOUÇAS *et al.*,1999)

Também estabelece as seguintes diretrizes gerais de ação: a gestão sistemática de recursos hídricos, sem dissociação dos aspectos de qualidade e quantidade; a adequação da gestão de recursos hídricos as diversidades físicas, bióticas, demográficas, econômicas, sociais e culturais das diversas regiões do Pais; a integração da gestão de recursos hídricos com a gestão ambiental; a articulação do planejamento de recursos hídricos com o dos setores usuários e com os planejamentos regional, estadual e nacional; a articulação da gestão de recursos hídricos com a do uso do solo; a integração da gestão das bacias hidrográficas com a dos sistemas estuarinos e zonas costeiras.

Possui como instrumentos: os Planos de Recursos Hídricos; o enquadramento dos corpos de água em classes, segundo os usos preponderantes da água; a outorga dos direitos de uso de recursos hídricos; a cobrança pelo uso de recursos hídricos; a compensação a municípios e o Sistema de Informações sobre recursos hídricos.

Outorga de direito de uso de recursos hídricos é o ato administrativo mediante o qual o poder público outorgante faculta ao outorgado o uso de determinado recurso hídrico, e por prazo determinado, nos termos e nas condições expressas no ato de outorga. Prevista na Lei 9.433/97 como um dos instrumentos da Política Nacional de Recursos Hídricos com o objetivo de assegurar o controle quantitativo e qualitativo da água e o efetivo exercício dos direitos de acesso a este bem.

Conforme a Política Nacional de Recursos Hídricos estão sujeitos a outorga pelo Poder Público os direitos dos seguintes usos de recursos hídricos: derivação ou captação de parcela da água existente em um corpo de água para consumo final, inclusive abastecimento público, ou insumo de processo produtivo; extração de água de aqüífero subterrâneo para consumo final ou insumo de processo produtivo; lançamento em corpo d'água de esgotos e demais resíduos líquidos ou gasosos, tratados ou não, com fim de sua diluição, transporte ou disposição final; aproveitamento dos potenciais hidrelétricos; outros usos que alterem o regime, a quantidade ou a qualidade da água existente em um corpo de água.

Conforme HORA (2001), embora a Lei n° 9.433/97 reconheça a outorga como um instrumento de controle na distribuição de recursos hídricos de competência do Poder Público, estabeleceu alguns quesitos que tornam esse controle participativo, isso para levar em consideração aspectos tradicionais e valores culturais das comunidades que utilizam a bacia. Assim, as prioridades de uso na concessão da outorga, serão definidas pelos Planos de Recursos Hídricos aprovados pelos comitês de Bacias Hidrográficas. Fica estabelecida também ao Conselho Nacional de Recursos Hídricos a competência para estabelecer diretrizes gerais para este instrumento de gestão.

No Estado de Mato Grosso foi instalado o Conselho Estadual de Recursos Hídricos - CEHIDRO em 06 de marco de 2002, através do decreto de regulamentação n° 3.952.

Atualmente, o instrumento de gestão de recursos hídricos "Outorga de Direitos de Uso da Água" está em fase inicial no Estado. Para tanto, foram publicados no DOE do dia 06/06/07 o Decreto n° 336 que regulamenta a outorga de direitos de uso dos

recursos hídricos. A Resolução n° 12 do CEHIDRO é que estabelece critérios técnicos para outorga de captações de águas superficiais de domínio do Estado e a Portaria n° 04 dispõe sobre os procedimentos para tal. Além disso, estuda-se qual a UPG (Unidade de Planejamento e Gerenciamento de Bacia) que dará início a emissão da Outorga, com excessão da outorga para geração de energia hidrelétrica (a qual necessita da solicitação do pedido de outorga, independente da UPG). Com relação ao Plano Estadual de Bacias neste estado, atualmente está em fase de discussão e apresentação para a sociedade do Diagnóstico realizado.

Segundo PROENÇA *et al.(*2004), outro instrumento de gerenciamento de recursos hídricos da Lei n. 9.433/97 é a implementação do enquadramento dos corpos de água em classes, isso demanda o conhecimento da qualidade das águas a serem geridas e das influências ambientais e antrópicas que possam alterá-la. Dessa forma é possível a utilização das normas de qualidade das águas, garantindo os padrões para os usos múltiplos desejados pela comunidade, preservando os aspectos qualitativos para a vida aquática e demais usos.

Assim a medida que as regiões crescem e se desenvolvem, maior é o uso dos recursos hídricos, maiores são os conflitos e riscos ambientais da qualidade dos corpos d' água.

2.3- FONTES DE POLUIÇÃO DA ÁGUA

REBOUÇAS *et al.(*1999) diz que a qualidade dos corpos dágua dependem dos ambientes naturais e antrópicos de onde se originam, circulam , percolam ou ficam estocadas. Os principais problemas de escassez de água que ameaçam a sobrevivência das populações e da vida na Terra, são causados pelo crescimento desordenado das demandas e pelos processos de degradação da qualidade dos corpos d'água, atingindo valores nunca imaginados, a partir da década de 50.

Segundo LIMA (2001), os conceitos de qualidade e poluição estão interligados. A poluição decorre de uma mudança na qualidade física, química, radiológica ou biológica do ar, água ou solo causada pelo homem ou por outra atividade que pode ser prejudicial ao uso do recurso.

HOLT *apud* LIMA (2001) diz que a industrialização e urbanização, juntamente com

a intensificação das atividades agrícolas, têm resultado no aumento da demanda pela água. Em conseqüência disto também aumentam a contribuição de contaminantes nos corpos d'água. As maiores rotas de contaminação são ocasionadas por emissões diretas e indiretas dos esgotos tratados e não-tratados (produtos tóxicos e metais pesados), escoamento e deposição atmosférica e por lixiviação do solo.

LIMA (2001) diz que as variedades de poluentes lançados nos corpos d'água podem ser agrupadas em duas classes:pontual e difusa. Os resíduos domésticos e industriais constituem o grupo das fontes pontuais por se restringirem a um simples ponto de lançamento, facilitando o sistema de coleta através de canais ou rede. Em geral a fonte de poluição pontual pode ser reduzida através tratamento apropriado para posterior lançamento. A poluição difusa, caracteriza-se por apresentar múltiplos pontos de descarga resultantes do escoamento em áreas urbanas e ou agrícolas e ocorrem durante os períodos de chuva, atingindo concentrações bastante elevadas dos poluentes. A redução dessas fontes requer mudanças nas práticas de uso da terra e na melhoria de programas de educação ambiental.

Assim, o índice de poluição não é determinado pela intensidade dos poluentes, mas pela capacidade de assimilação dos corpos d' água.

2.4- PADRÕES AMBIENTAIS E VARIÁVEIS DA QUALIDADE DA ÁGUA

Segundo a Organização Mundial da Saúde-OMS, água potável é aquela que apresenta aspecto límpido e transparente, não apresenta cheiro ou gosto objetáveis, não contém nenhum tipo de microrganismo que possa causar doença; e não contém nenhuma substância em concentrações que possam causar qualquer prejuízo a saúde (padrão de potabilidade para as águas destinadas ao abastecimento humano).

No Brasil, o uso da água doce para consumo humano está sujeito a fatores específicos de qualidade que são definidos pelos Padrões de Potabilidade. Os Padrões de Potabilidade são definidos pelo Ministério da Saúde (BRASIL,2000), através da Portaria n.518 de 26/03/2004. Esses valores são valores máximos permitidos (VMP) de concentração para uma série de substâncias e componentes presentes na água.

Segundo BLUM *apud* MANCUSO & SANTOS (2003) o fato de estabelecer a qualidade da água a seu uso visa simplificar e tornar mais objetiva a avaliação, mas tem limitações que devem ser consideradas. Por exemplo, quando se trata de avaliar a qualidade de uma água para consumo humano, deve-se levar em conta que não são ainda suficientemente conhecidos os efeitos sobre a saúde provocados pela presença de várias substâncias químicas, especialmente compostos orgânicos sintéticos. Ou seja, não se dispõe de padrões de potabilidade para todos os possíveis constituintes de uma água. Além disso, não se conhecem suficientemente os efeitos da associação de duas ou mais substâncias (efeitos sinérgicos), nem estão definidos métodos de análise para identificação e quantificação de outras.

Conforme NASCIMENTO (1998) os padrões são utilizados para a proteção da qualidade da água, de forma a assegurar seus usos previstos. A ABNT (NBR 9896/87) preconiza que os padrões de qualidade são constituídos por um conjunto de parâmetros e respectivos limites, e são estabelecidos com base em critérios científicos que avaliam risco para um dado indivíduo e o dano causado pela exposição a uma dose de determinado poluente. Um critério científico é uma quantidade limite fixada para um determinado parâmetro que, estando dentro dos limites máximos (ou mínimos), conforme a natureza do constituinte, protegerá os usos desejados para um determinado corpo d'agua, dentro de um grau de segurança.

Assim REBOUÇAS *et al*. (1999) diz que a escassez quantitativa é um fator limitante ao desenvolvimento, enquanto que a qualitativa traz problemas sérios a saúde pública, `a economia e ao meio ambiente. Da mesma forma os métodos sofisticados de tratamento podem criar problemas complexos que afetam a qualidade do ambiente e a saúde pública.

A Legislação Brasileira também estabelece padrões de qualidade para águas superficiais, através do CONAMA n°357 de 17 de março de 2005 .

De acordo com REBOUÇAS *et al.(*1999) na avaliação da qualidade da água considera-se a composição de uma amostra que são constituídas por características químicas, microbiológicas e físicas, observando limites estabelecidos pela Resolução CONAMA n°357 de 17 de março de 2005 e o respectivo objetivo.

O enquadramento nas classes da CONAMA, entretanto é estático, precisa de uma avaliação constante. No estado de Pernambuco criou-se, por exemplo, um sistema de

avaliação periódico. A Agência Estadual de Meio Ambiente e Recursos Hídricos do Estado de Pernambuco- CPRH, considerando a real situação das águas nas estações monitoradas, e a classificação das águas interiores segundo os usos preponderantes, estabelecida no Decreto Estadual 7.269/81, adota, na determinação da qualidade dos corpos d' água, a classificação constante do Quadro 1 .

Quadro 1-Avaliação da qualidade das águas de modo a atender seu uso mais restritivo dentro da classe

CLASSIFICAÇÃO	DESCRIÇÃO
Não comprometida	Enquadram-se, nesta categoria, os corpos de água que apresentam condições de qualidade de água compatíveis com os limites estabelecidos para a classe especial das águas doces, salinas e salobras e classe 1 das águas doces (Resolução CONAMA n° 357/05). Estes corpos d'água apresentam qualidade da água ótima, com níveis desprezíveis de poluição.
Pouco comprometida	Enquadram-se, nesta categoria, os corpos de água que apresentam condições de qualidade de água compatíveis com os limites estabelecidos para a classe 2 das águas doces e a classe 1 das águas salinas e salobras (Resolução CONAMA n° 357/05). Estes corpos d'água apresentam qualidade da água boa, com níveis baixos de poluição.
Moderadamente comprometida	Enquadram-se, nesta categoria, os corpos de água que apresentam condições de qualidade de água compatíveis com os limites para a classe 3 das águas doces e a classe 2 das águas salinas e salobras (Resolução CONAMA n° 357/05). Estes corpos d'água apresentam qualidade da água regular, com níveis aceitáveis de poluição.
Poluída	Enquadram-se, nesta categoria, os corpos de água que apresentam condições de qualidade de água compatíveis com os limites estabelecidos para a classe 4 das águas doces e a classe 3 das águas salinas e salobras (Resolução CONAMA n° 357/05). Estes corpos d'água apresentam qualidade da água ruim, com poluição acima dos limites aceitáveis.
Muito Poluída	Enquadram-se, nesta categoria, os corpos de água que não se enquadram em nenhuma das classes acima estabelecida. Estes corpos

	d'água apresentam qualidade da água péssima, com poluição muito elevada.
Não Monitorada	

Fonte: www.cprh.pe.gov.br

Observa-se que um trecho da bacia hidrográfica com 1005 Mg/L de Coliformes Termotolerantes em 80% de 6 amostras e outro com 4.000 Mg/L Coliformes Termotolerantes em 80% de 6 amostras, pertencem a mesma classe 3, segundo a Resolução CONAMA 357, e verifica-se que os investimentos para recuperar esses trechos são financeiramente diferentes. Assim percebe-se que o CONAMA 357/05 generaliza os casos dentro de suas classes pela própria característica estática que possui, um nível de detalhamento na qualidade para estabelecer avaliações periódicas e o estado de qualidade atual do corpo hídrico pode ser uma alternativa viável, como mostra o exemplo adotado pela CPRH (Quadro 1).

2.4.1 - Variáveis da Qualidade da Água

O levantamento de qualquer sistema ambiental depende fundamentalmente da escolha dos parâmetros representativos de seu "status" por ocasião do momento da amostragem. Esses parâmetros e seus padrões de qualidade da água só têm sentido se são determinados em função dos seus usos preponderantes atuais e futuros.

Assim, neste item falaremos dos parâmetros utilizados para avaliar qualidade da água e seus limites estabelecidos pela Resolução CONAMA N° 357 (fixa valores de diferentes classes para um corpo receptor). Em seguida, faremos uma exposição das variáveis físico-químicas e biológicas para identificar a qualidade da água utilizadas neste estudo, são elas: Turbidez, OD, DBO, DQO, NTK, P, pH, Sólidos Dissolvidos Totais, Coliformes (Termotolerantes e *Eschrichia Coli).*

2.4.1.1- pH

O termo pH (potencial hidrogeniônico) é usado universalmente para expressar o grau de acidez ou basicidade de uma solução, ou seja, é o modo de expressar a concentração de íons de hidrogênio nessa solução. A escala de pH é constituída de

uma série de números variando de 0 a 14, os quais denotam vários graus de acidez ou alcalinidade. Valores abaixo de 7 e próximos de zero indicam aumento de acidez, enquanto valores de 7 a 14 indicam aumento da basicidade. Para ESTEVES (1998) as medidas de pH são de extrema utilidade e complexas de se interpretar pelo grande número de fatores que podem influenciá-lo. Às águas superficiais possuem um pH entre 4 e 9. Naturalmente, nesses casos, o pH reflete o tipo de solo por onde a água percorre. Em lagoas com grande população de algas, nos dias ensolarados, o pH pode subir muito, chegando a 9 ou até mais. Isso porque as algas, ao realizarem fotossíntese, retiram muito gás carbônico, que é a principal fonte natural de acidez da água. Geralmente um pH muito ácido ou muito alcalino está associado à presença de despejos industriais.

2.4.1.2- Oxigênio dissolvido

BRANCO (1986) diz que a determinação do oxigênio dissolvido é de fundamental importância para avaliar as condições naturais da água e detectar impactos ambientais como eutrofização e poluição orgânica. Do ponto de vista ecológico, o oxigênio dissolvido é uma variável extremamente importante, pois é necessário para a respiração da maioria dos organismos que habitam o meio aquático. Geralmente o oxigênio dissolvido se reduz ou desaparece, quando a água recebe grandes quantidades de substâncias orgânicas biodegradáveis encontradas, por exemplo, no esgoto doméstico, em certos resíduos industriais, vinhoto, e outros. Os resíduos orgânicos despejados nos corpos d'água são decompostos por microorganismos que se utilizam do oxigênio na respiração. A morte de peixes em rios poluídos se deve, portanto, à ausência de oxigênio e não à presença de substâncias tóxicas. Na ausência de oxigênio temos a geração de maus odores.

2.4.1.3- Demanda bioquímica do oxigênio (DBO) e demanda química do oxigênio (DQO)

LIMA (2001) diz que a Demanda Bioquímica de Oxigênio (DBO), exprime o valor da poluição produzida por matéria orgânica oxidável biologicamente, corresponde à quantidade de oxigênio que é consumida pelos microorganismos do esgoto ou águas poluídas, na oxidação biológica, a DBO5 é convencionalmente usada, considera a

medida de 5 dias, incubada a 20 graus de temperatura associada a fração biodegradável dos componentes carbonáceos. Essa demanda pode ser suficientemente grande, para consumir todo o oxigênio dissolvido da água, o que condiciona a morte de todos os organismos aeróbios de respiração subaquática. O teste de Demanda Química de Oxigênio (DQO) é a medida da quantidade de oxigênio consumido pela oxidação química de substâncias orgânicas presentes na águas (SPERLING,1996).

2.4.1.4 - Sólidos

É o material particulado não dissolvido, encontrado suspenso no corpo d'água, composto por substâncias inorgânicas e orgânicas, incluindo-se aí os organismos planctônicos (fito e zooplâncton). Sua principal influência é na diminuição na transparência da água, impedindo a penetração da luz. BRANCO (1986) afirma que todos os contaminantes com exceção dos gases dissolvidos contribuem para a carga de sólidos, os que possuem características físicas(suspensos e dissolvidos) e químicas(orgânicos e inorgânicos). O mesmo autor ainda complementa que os sólidos voláteis representam uma estimativa orgânica nos sólidos, ao passo que os sólidos fixos caracterizam a presença de matéria inorgânica ou mineral.

2.4.1.5 - Compostos de Nitrogênio

As águas naturais, em geral, contêm nitratos em solução e, além disso, principalmente tratando-se de águas que recebem esgotos, podem conter quantidades variáveis de compostos mais complexos, ou menos oxidados, tais como: compostos orgânicos quaternários, amônia e nitritos. Em geral, a presença destes denuncia a existência de poluição recente, uma vez que essas substâncias são oxidadas rapidamente na água, graças principalmente à presença de bactérias nitrificantes. MOTA (1995) diz que o nitrogênio orgânico e amônia estão associados a efluentes e águas recém-poluídas. Com o passar do tempo, o nitrogênio orgânico é convertido em nitrogênio amoniacal e, posteriormente, se possuem condições aeróbias , a oxidação da amônia acontece transformando nitrito em nitrato, indicadores de estágio de poluição remota.

2.4.1.6 - Fósforo

Apresenta-se na água nas formas orgânica e inorgânica, na forma natural (tais como dissolução de compostos do solo e decomposição da matéria orgânica) e de origem antropogênica (despejos domésticos, esgotos, detergentes, inseticidas fertilizantes). FEITOSA *et al*. (1997) dizem que devido a ação de microrganismos, a concentração de fósforo pode ser baixa em águas com menor índice de poluição (< que 0,5mg/l) e em águas poluídas valores acima de 1,0 mg/l. Os compostos de fósforo são um dos fatores limitantes à vida dos organismos aquáticos e o seu controle, em um corpo d'água, é de importância fundamental no controle ecológico das algas.

2.4.1.7 – Coliformes

Conforme CANHAMERO (2006), a qualidade de uma água de abastecimento é avaliada usando **organismos indicadores.** A probabilidade de existência das doenças na água passadas a ela por fezes do indivíduos doentes, se faz por contagem de microorganismos não patogênicos, produzidos em grande numero no intestino, sendo uma referência(um indicador). Os organismos usados como referência pertencem a um grupo de bactérias chamados **Coliformes** dividido em três sub-grupos: **coliformes totais**, **coliformes fecais e estreptococos fecais.** De acordo com o mesmo autor, os **Coliformes Totais (CT)** reúnem um grande número de bactérias, entre elas a ***Eschrichia Coli*[1]*,*** de origem exclusivamente fecal e que dificilmente se multiplica fora do trato intestinal. O problema é que outras bactérias dos gêneros ***Citrobacter, Eriterobacter*** e ***Klebsiella***, igualmente identificadas pelas técnicas laboratoriais como coliformes totais, podem existir no solo e nos vegetais. Desta forma, não é possível afirmar categoricamente que uma amostra de água com resultado positivo para coliformes totais tenha entrado em contato com fezes. Os **Coliformes Fecais** pertencem a esse subgrupo os microorganismos que aparecem exclusivamente no trato intestinal. Em laboratório, a diferença entre coliformes totais e fecais é feita através da temperatura (os coliformes fecais continuam vivos mesmo a 44ºC, enquanto os coliformes totais têm crescimento a 35ºC). Sua identificação na

[1] No CONAMA 357/2005 Coliforme Total é referenciada como *Eschrichia Coli*

água permite afirmar que houve presença de matéria fecal, embora não exclusivamente humana.

BRANCO (1986) enfatiza que a presença de coliformes na água indica a presença de fezes e, portanto, a potencialidade da veiculação de seres patogênicos.

2.4.1.8- Turbidez

LIMA (2001) afirma que a presença de partículas em suspensão, que causam a turbidez, ou de substâncias em solução, relativas `a cor, pode concorrer para o agravamento da poluição. A turbidez limita a penetração de raios solares, restringindo a realização da fotossíntese que, por sua vez, reduz a reposição do oxigênio.

2.5- TOMADA DE DECISÃO E DESCOBERTA DE CONHECIMENTO (KDD) EM BASES DE DADOS

A capacidade de um indivíduo ou empresa de tomar decisões é frequentemente associada ao conhecimento que ela possui. Um dos problemas dos analistas de informação é a transformação de dados em informação relevante para a tomada de decisão. Através de técnicas para exploração de dados, pode-se desenvolver aplicações que venham a extrair, das bases de dados de empresas e instituições acadêmicas, informações críticas, com o objetivo de subsidiar o processo decisório de uma organização.

REZENDE (2003) diz que os conceitos de dados, informação e conhecimento estão relacionados entre si. Antes de estabelecer a ligação desses conceitos com as tecnologias utilizadas para seu registro e armazenamento, é preciso descrever a distinção entre dado, informação e conhecimento .

MOTTA (2004) diz que os dados são elementos puros, quantificáveis sobre determinado evento; a informação é um contexto, uma referência, um parâmetro para analisar dados e o conhecimento é a habilidade de se criar um modelo mental que

descreva o evento e indique as decisões a tomar (permitindo a compreensão, análise e síntese para a tomada de decisão inteligente).

Para REZENDE (2003) dado além de ser um elemento puro, quantificável sobre um determinado evento, são fatos, números, texto ou qualquer outra mídia que possa ser processada pelo computador. A informação é o dado analisado e contextualizado, envolve a interpretação de um conjunto de dados, é constituída por padrões, associações ou relações que todos os dados juntos podem proporcionar. Esta pode gerar conhecimento que auxilie na análise de padrões históricos para conseguir uma previsão de dados futuros. Uma decisão é o uso explícito de um conhecimento.

Segundo NAVEGA (2002), podemos identificar vários níveis de informação em função da capacidade de entendimento desses conceitos, Figura 1. Assim, a tradicional pirâmide da informação mostra um aumento do valor da informação ao subir de nível, `a medida em que aumenta o grau de conhecimento sobre a relação dos dados.

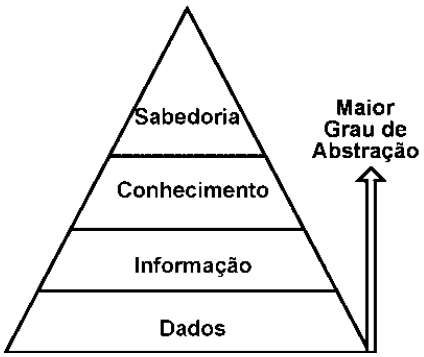

Figura 1- Valor potencial da informação (Fonte: www.inteliiwise.com/navega)

Quanto maior o grau de abstração da informação, ou seja, maior o conhecimento dessa relação, melhor o processo de tomada de decisão.

As empresas e instituições acadêmicas interagem com o meio ambiente a todo o momento. Essa interação tem gerado grandes volumes de dados que são armazenados em suas bases de dados e podem auxiliar na tomada de decisão.

Na Figura 2 podemos verificar a tradução da Figura 1 para a base de dados de uma empresa. Em geral, o valor da informação nas bases de dados, para apoiar a tomada de decisão, aumenta a partir da base da pirâmide. Uma decisão baseada em dados nas camadas mais baixas (nível operacional), onde há tipicamente milhões de registros de dados ainda não-estruturados, não possui valor de conhecimento, pois tem-se pouco entendimento sobre eles. Já a decisão apoiada em dados altamente resumidos nas camadas superiores da pirâmide, onde já passaram por ferramentas de organização e estruturação da informação, tem probabilidade de alto valor estratégico para tomada de decisão nas empresas. Assim, como mostra a Figura 2, o processo de tomada de decisão é fundamentado nos níveis tático (gerencial) e estratégico (direção), onde os dados passaram por técnicas de descoberta de conhecimento, descobrindo novos conceitos e relações entre eles e com isto facilitando o processo decisório.

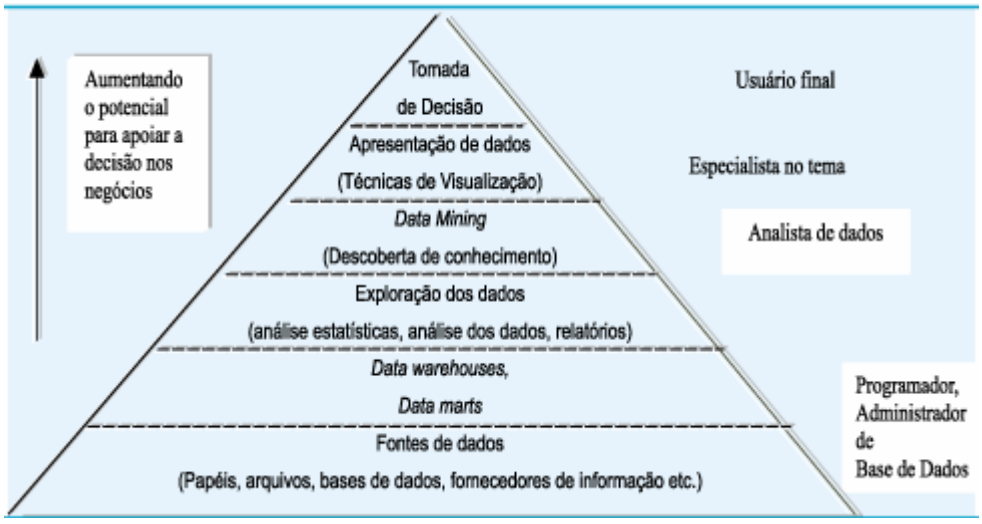

Figura 2- Evolução do valor estratégico das bases de dados (Cabena et al & Tyson apud Quoniam)

Esse processo de transformação de dados em conhecimento em bases de dados, para melhorar o processo decisório utiliza-se da técnica de data mining (principal fase do processo de descoberta de conhecimento), que localiza padrões de informação em bases de dados através do processo de generalização ou indução, descobrindo novos conhecimentos em bases de dados. As ferramentas de *Data Mining* identificam todas

as possibilidades de correlações existentes nas fontes de dados sendo portanto uma fase do processo de descoberta de conhecimento.

2.5.1– Sistemas Inteligentes

Entende-se como inteligente um sistema computacional que obtém e usa o conhecimento para resolver problemas. Sistemas Inteligentes têm aplicação nos mais diversos setores, incluindo: energético, econômico/comercial, seguros, telecomunicações, mercado de capitais, industrial, meio-ambiente e medicina.

De acordo com MOTTA (2005) a obtenção e a utilização do conhecimento são realizadas através de um sistema inteligente ou de um ser humano como forma de apoio a tomada de decisão.

Segundo MOTTA (2005) a cronologia da era do conhecimento obedece ao Quadro 02.

Quadro 2 - Cronologia da era do conhecimento

ÉPOCA	VALOR	DESAFIO
Até 80	Dado	Sistemas de processamento de dados
Entre décadas de 80 e 90	Informação	Sistemas de informação
Atualmente	Conhecimento	**Sistemas inteligentes**

O Quadro 2 mostra que até os anos 80 preocupava-se somente com o armazenamento de dados e com a extração eficiente dos mesmos. Segundo BARRETO (1999), nos anos 60, acreditava-se que os computadores eram enciclopédias ambulantes, nesta época desenvolveram-se o conceito de bases de dados, o computador servia apenas para memorizar e extrair dados. Entre os anos 80 e 90, segundo REZENDE (2003), o desafio foi migrar os dados para as informações através dos Sistemas de Informação que tinham por objetivo a análise e organização da informação para melhorar o processo de tomada de decisão. A partir dos anos 90 o desafio era criar sistemas que pudessem representar e processar conhecimento em resposta `as muitas necessidades dos indivíduos. Assim, atualmente busca-se o desenvolvimento de sistemas inteligentes para auxiliar gestores na difícil tarefa de decidir.

BARRETO (1999) diz que Sistema de Informação e Inteligência Artificial são tópicos intimamente ligados. SI é um problema que engloba aquisição,

armazenamento e recuperação da informação. IA é uma metodologia de resolver problemas em geral, cuja solução requeira inteligência. Assim a junção das duas áreas é perfeita se a aquisição, armazenamento e recuperação da informação caracterizar o uso de inteligência.

Através de técnicas específicas, os computadores além de armazenar e recuperar informações, combinam dados e geram novos conhecimentos. Este fato faz o computador comportar-se como inteligente, ou seja, ele tem habilidades inteligentes.

Para descobrir este novo conhecimento e se tornar um sistema inteligente, temos uma área de estudo na IA , chamada KDD (Knowlegde Discovery Databases) que se preocupa com obtenção de conhecimento útil e interessante , muitas vezes implícito em bases de dados.

2.5.2- Descoberta de Conhecimento (KDD) em Bases de Dados

Com os avanços tecnológicos, os sistemas de armazenamento de dados e monitoramento de fenômenos observados têm acumulado grandes volumes de dados que ocultam conhecimento para o ser humano. Assim, o "KDD - Knowledge Discovery in Databases" (Descoberta de Conhecimento em Banco de Dados) é uma área da IA que se preocupa em descobrir conhecimento oculto, importante para empresas e órgãos governamentais através do uso de algoritmos eficientes de mineração de dados.

> *"O KDD (knowledge Discovery in Databases) pode ser visto como o processo da descoberta de novas correlações, padrões e tendências significativas por meio da análise minuciosa de grandes conjuntos de dados estocados. Este processo se vale de tecnologias de reconhecimento utilizando padrões e técnicas estatísticas e matemáticas. O Data Mining é uma das técnicas utilizada para a realização de KDD. (NORTON, 1999).*

Segundo FAYYAD *et al*. (1996) a descoberta de conhecimento em bases de dados, KDD, é vista como algo mais amplo, e o termo Data Mining (mineração ou

garimpagem de dados) como um componente que trabalha com os métodos de descoberta do conhecimento.

LACERDA & SOUZA (2004) afirmam que, devido ao imenso volume de dados, várias informações acabam ficando escondidas, sendo formadas implicitamente por relações entre campos, formando padrões imperceptíveis a "olho nu". O ciclo iterativo de descoberta do conhecimento em base de dados é mostrado na Figura03.

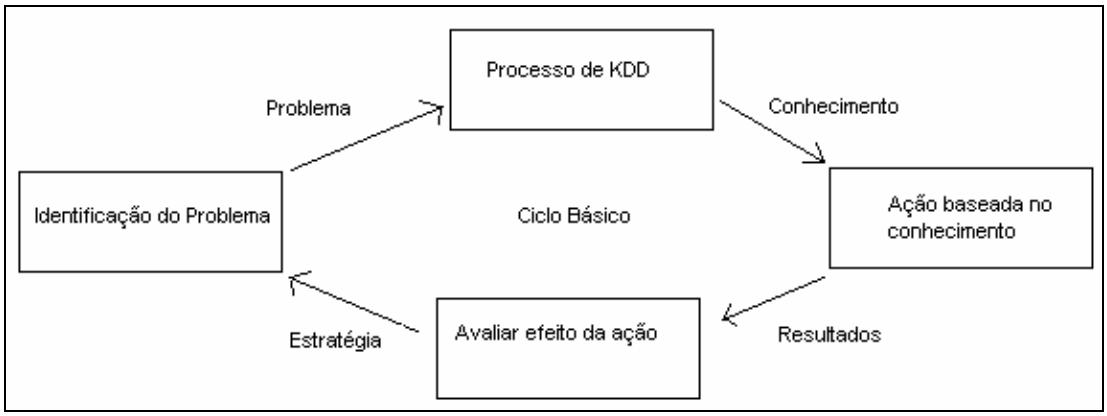

Figura 3 - Ciclo básico de descoberta de conhecimento em BD. (Barreto, 1999)

De acordo com BARRETO (1999) as principais tecnologias usadas em KDD são:

- Organização dos dados (*Data Warehousing*);
- Bancos de dados distribuídos (quando os dados estão distribuídos em diferentes plataformas);
- IA e Sistemas Especialistas;
- Redes Neurais e seus paradigmas de aprendizado supervisionado e não supervisionado;
- Interfaces amigáveis incluindo realidade virtual.

LACERDA & SOUZA (2004) afirmam que a natureza do descobrimento é tanto iterativa quanto interativa. A iteratividade está no fato do processo ser realizado em etapas seqüenciais de maneira que seja possível voltar `as etapas anteriores criando ligações entre elas. Isso traz muitas possibilidades de seqüências, de forma que os passos possam ser repetidos o número de vezes que se façam necessários. O usuário pode dar continuidade ao processo ou transformar uma etapa alterando um atributo,

por exemplo. Assim o usuário é responsável por várias tomadas de decisão durante o ciclo, na modelagem das informações, na escolha do algoritmo a se usado e sob os objetivos a ser seguidos, garantindo a interatividade no processo.

De acordo com MOTTA (2005) existem inúmeras áreas de aplicação de KDD: bancária (aprovação de crédito), ciências e medicina (descoberta de hipóteses, diagnóstico, classificação, predição), comerciais (segmentação, localização de consumidores, identificação de hábitos de consumo), engenharia (simulação e análise, reconhecimento de padrões , processamento de sinais e planejamento), financeira(apoio para investimentos, controle de carteira de ações), gerencial (tomada de decisão, gerenciamento de documentos), internet (ferramentas de busca, navegação, extração de dados), manufatura (modelagem e controle de processos, controle de qualidade, alocação de recursos), segurança (detecção de bombas, icedbergs e fraudes), etc. Recentemente, temos as bases de dados Monitoramento ambiental , utilizando KDD para diagnóstico e prevenção de danos ambientais.

Portanto, o processo de Extração de Conhecimento em Bases de Dados (KDD) tem o objetivo de encontrar conhecimento a partir de um conjunto de dados para ser utilizado em processo decisório. Esta técnica, poderá auxiliar gestores de empresas e órgãos governamentais na tomada de decisão, uma vez que obtém conhecimento novo, que não está explícito na base de dados.

2.5.2.1- O Processo da Descoberta de Conhecimento

A descoberta do conhecimento em bases de dados envolve três etapas principais (Figura 4).

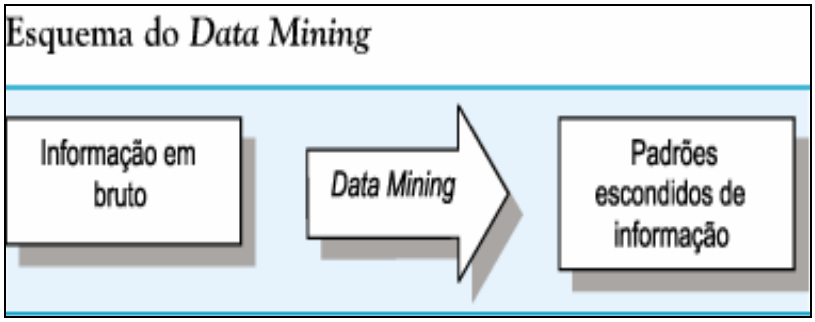

Figura 4 – Etapas da descoberta do conhecimento. (Quoniam *et al*., 2001)

1) **A primeira fase** que contempla a limpeza e preparação dos dados (pré-processamento);
2) **A segunda** que compreende a mineração para descoberta de padrões;
3) **A terceira** que compreende a avaliação e interpretação do resultado obtido.

De acordo com NAVEGA (2002) um ciclo completo de descoberta de conhecimento com Data mining pode ser representado conforme Figura 5. A partir de fontes de dados (bancos de dados, relatórios, logs de acesso, transações, etc) efetua-se uma limpeza (consistência, preenchimento de informações, remoção de ruídos e redundâncias, etc). Assim surgem os repositórios organizados (Data Warehouses). Após isso temos a utilização das ferramentas de Data Mining e a última fase de avaliação ou interpretação do resultado onde um analista refina e conduz o processo até que valiosos padrões apareçam. Assim a hierarquia começa em instâncias elementares (normalmente volumosas) e termina em um ponto relativamente concentrado e mais valioso.

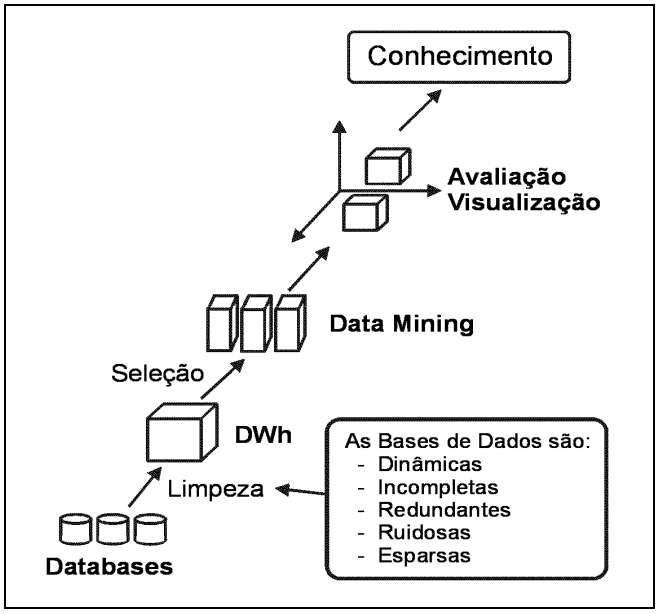

Figura 5 - Fases da descoberta de conhecimento (Fonte: www.inteliiwise.com/navega)

2.5.2.1.1-Pré-Processamento

Segundo QUONIAM *et al.* (2001) nesta fase os dados são coletados, armazenados e passam por uma limpeza na base de dados. Para ter sucesso nesta fase é preciso conhecer a base de dados e entender os dados para que na limpeza e preparação não haja duplicação de conteúdo através de erros de digitação, abreviações, valores omissos, entre outros.

Para LACERDA & SOUZA (2004) o pré-processamento é dividido em sub-etapas:

- **Entendimento do domínio da aplicação**: entender objetivos e metas para posteriormente verificar se o conhecimento descoberto é útil e identificar as transformações ocorridas na base de dados antes do inicio da aplicação.
- **Seleção dos Dados**: Após o entendimento do problema analisar quais variáveis são relevantes para solucioná-lo, ou obter algum resultado. Essa escolha deve ser feita sob o escopo do problema.
- **Limpeza e pré-processamento**: após extração dos dados da base pode ser que ocorram inconsistências, como dados errados, campos

sem preenchimento, etc. É necessário a remoção desses ruídos ou erros para o bom funcionamento do modelo.

- **Transformação**: redução e projeção dos dados considerando conjunto de atributos úteis.

Após esta fase de pré-processamento segue a aplicação da mineração sob os dados.

2.5.2.1.2-Mineração de Dados (Data Mining)

São técnicas de descoberta de conhecimento em bases de dados que não conseguimos visualizar através das consultas realizadas no banco de dados. A mineração de dados trabalhará na busca por padrões e relacionamentos entre os dados a fim de facilitar o entendimento.

> *As ferramentas de Data Mining identificam todas as possibilidades de correlações existentes nas fontes de dados. Através de técnicas para exploração de dados pode-se desenvolver aplicações que venham a extrair, dos bancos de dados informações críticas , com o objetivo de subsidiar plenamente o processo decisório de uma organização, (QUONIAM,2001).*

A definição de Mineração de Dados aceita por diversos autores e pesquisadores foram elaborados por Fayyad *et al. (*1996) e diz: “Extração de Conhecimento de Base de dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados”. Segundo REZENDE (2003) analisando individualmente cada palavra desta definição é melhor para compreendê-la:

- **Dados:** conjunto de fatos presentes em um repositório. Ex: Valores dos campos de um registro de vendas em uma base qualquer.
- **Padrões:** denota alguma abstração de um subconjunto dos dados em alguma linguagem descritiva de conceitos.

- **Processo:** a extração do conhecimento envolve várias etapas como preparação dos dados, busca por padrões e avaliação do conhecimento.
- **Válidos:** os padrões descobertos devem possuir algum grau de certeza, devem satisfazer funções ou limiares que garantam que os exemplos cobertos e os casos relacionados ao padrão sejam aceitáveis.
- **Novos:** um padrão deve fornecer novas informações sobre os dados. Pode ser medido por comparações entre as alterações ocorridas e os dados anteriores.
- **Úteis:** os padrões descobertos devem ser incorporados para serem utilizados.
- **Compreensíveis:** um dos objetivos da MD é encontrar padrões descritos em alguma linguagem que possa ser compreendida pelos usuários permitindo uma análise profunda dos dados.
- **Conhecimento:** o conhecimento é definido em função do seu domínio de aplicação, utilidade, originalidade e compreensão.

LACERDA & SOUZA (2004) dizem que a principal característica do data mining é a aplicação de algoritmos ao dados pré-processados. As sub-etapas da mineração são:

1. **Escolha da tarefa de mineração** (as ferramentas possuem vários tipos): regressão,classificação, associação, sumarização, dependência, agrupamento, etc.
2. **Determinar dentro da tarefa escolhida o algoritmo** que seja mais adequada para o problema em questão.
3. **Aplicar o algoritmo de mineração sob os dados e variáveis** na busca de padrões.

Assim KDD ou descoberta de conhecimento é o conjunto de tarefas que permite detectar conhecimentos implícitos em bases de dados. Data Mining é uma técnica usada em uma das fases do processo de KDD onde o resultado pode ser utilizado para tomada decisão.

2.5.2.1.3-Pós-Processamento

Após as etapas anteriores pode-se verificar se houve mudança nos dados da base de dados e aí passa-se para outra fase de interpretação dos resultados obtidos.
Segundo LACERDA & SOUZA (2004), no pós-processamento a principal função é usar as descobertas úteis. Suas sub-etapas são:

- Interpretação de padrões: avaliar se os padrões descobertos terão alguma utilidade ou geram algum conhecimento. Servem de suporte a tomada de decisão.
- Consolidação dos dados: verificar e utilizar o novo conhecimento sobre os dados através de ferramentas de visualização.

2.5.2.2- Metas da Mineração

O ciclo de descoberta de conhecimento envolve o usuário desde o início de sua fase, ou seja, na preparação dos dados até a obtenção dos resultados. O usuário precisa ter em mente os objetivos que deverão ser alcançados com a fase de mineração de dados, isto porque para alcançar estes objetivos é preciso escolher um algoritmo que atenderá as necessidades do problema a ser resolvido. Outra razão para isto é que existem inúmeros conhecimentos que se pode retirar de uma base de dados. Dentre eles, descoberta de associações, descoberta de regras de previsão, hierarquias de classificação, descoberta de padrões seqüenciais, descoberta de padrões em séries temporais, categorização e segmentação (ALVARES,2000).
Assim faz-se necessário a definição clara do objetivo do problema para o bom desempenho da técnica.
A etapa de mineração de dados possui duas metas (ou atividades) principais (FAYYAD,1996): previsão e descrição.

- **Previsão**: serve para antecipar os valores de variáveis conhecidas ou analisar um possível valor para uma variável com o passar do tempo, ou seja, determina as chances de uma ação ocorrer.
- **Descrição**: é a busca por uma descrição padrão dos dados para posterior entendimento e melhor compreensão dos usuários.

Existem diversas tarefas para se alcançar as metas de previsão e descrição e ainda vários algoritmos estão disponíveis para resolver cada tarefa. Assim depende do objetivo do problema (ou da meta da mineração) a escolha da tarefa ideal para solucioná-lo.

2.5.2.3 -Tarefas da Mineração

Tarefa é no contexto da mineração um tipo de problema de descoberta de conhecimento a ser solucionado, Figura 6. Podem ser de:

- **Classificação**: constrói um modelo de forma que possa ser aplicado a dados não classificados a fim de caracterizá-los em classes, assim é possível antecipar tendências futuras;
- **Estimativa ou regressão**: usada para definir um valor para alguma variável contínua e desconhecida;
- **Associação**: LACERDA & SOUZA (2004) dizem que a associação determina grupos de itens que tendem a ocorrer ao mesmo tempo, na mesma transação, gerando-se uma grande quantidade de regras.
- **Agrupamento, segmentação ou clusterização**: processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos
- **Sumarização**: busca métodos para encontrar uma descrição compacta para um subconjunto de dados. Usada para análise exploratória de dados e criação de relatórios de maneira automática (FAYYAD,1996).
- **Desvio**: tem por objetivo descobrir um conjunto de valores que não seguem padrões definidos, é necessário adotar padrões antecipadamente. Pode-se usar esta tarefa para identificar fraudes baseadas em elementos que estão fora de padrões ou são exceções `a regra (FELDENS,1997).
- **Dependência**: Verifica a existência de um modelo que descreva a dependência entre variáveis.

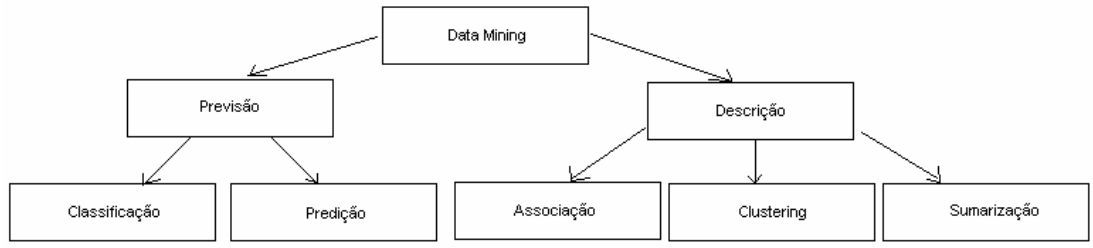

Figura 6 - Tarefas da mineração de dados

2.5.2.4- Métodos ou Técnicas de Mineração

De acordo com DIAS (2001) as técnicas (métodos) de mineração de dados podem ser aplicadas as tarefas. Elas correspondem a técnica computacional implementada por determinado algoritmo para resolver uma tarefa.

HARRISON (1998) dia que não há uma técnica que resolva todos os problemas de mineração de dados. Existem diferentes métodos para diferentes propósitos, cada um com suas vantagens e desvantagens. A familiaridade com as técnicas é necessária para facilitar a escolha de uma tarefa de mineração conforme o problema a ser tratado. Assim para ser executada, cada tarefa possui várias técnicas de mineração envolvidas na busca por padrões ocultos de dados, naturalmente umas mais indicadas que outras para o problema em questão.

LACERDA & SOUZA (2004) dizem que a técnica (método) de regra de associação resolve problemas de associação. A técnica (método) de árvore de decisão resolve problemas de classificação e regressão. As demais técnicas resolvem problemas de classificação e segmentação.

As técnicas (ou métodos) de mineração de dados normalmente usadas são (DIAS, 2001):

Regras de associação: estabelece uma correlação estatística entre atributos de dados e conjunto de dados;

Arvore de decisão: hierarquização dos dados, baseando-se em estágios de decisão (nós) e na separação de classes e subconjuntos. É um bom método quando o objetivo do data mining é a classificação de dados ou predição de saídas, usado para categorizar dados de arquivos.

Raciocínio baseado em casos (MBR): baseado no método do vizinho mais próximo, combina e compara atributos par estabelecer hierarquia de semelhança.

Algoritmos genéticos: métodos gerais de busca e otimização, inspirados na Teoria da Evolução, onde cada nova geração, soluções melhores têm mais chances de ter "descendentes";

Redes neurais artificiais: modelos inspirados na fisiologia do cérebro, na qual o conhecimento é fruto do mapa das conexões neurais e dos pesos dessas conexões. É uma boa maneira de se predizer regras e classificá-las. Não funcionam muito bem com muitas entradas e muitos dados.

2.5.2.5 – Tipo de aprendizado em Bases de Dados

O objetivo de um banco de dados é armazenar os dados com segurança e recuperá-los de forma eficiente conforme necessidade do usuário . Muitas vezes a informação necessária pode não estar contida explicitamente na base de dados, podendo ser inferida. HAL *apud* BOGORNY (2000) salienta duas técnicas para a inferência dos dados: dedução e indução. A dedução é uma conseqüência lógica das informações contidas na base de dados, a informação é extraída da base através de operadores lógicos e comandos dos próprios SGBD. A indução é uma técnica de inferência que generaliza as informações contidas na base de dados. Nela a base é percorrida em busca de padrões ou regularidades, que são combinações de valores para certos atributos que compartilham características comuns. Cada regularidade forma uma regra, prevendo o valor de um atributo com base em outros atributos. O aprendizado indutivo consiste na criação de um modelo, onde os objetos e eventos são agrupados em classes. Para cada classe é criado um conjunto de regras.

AVI *apud* BOGORNY afirma que existem duas técnicas de aprendizado indutivo:

Aprendizado supervisionado ou de exemplos: nesta técnica são fornecidos classes e exemplos de cada classe ao sistema, o qual precisa encontrar a descrição (propriedades comuns nos exemplos) de cada classe. Existe neste aprendizado um professor que guia o processo. Esse professor é o conhecimento prévio dos conceitos ou classes que estão descritas pelo conjunto de exemplos de treinamento.

Aprendizado não-supervisionado ou por observação: o sistema precisa descobrir a classe dos objetos através das propriedades que os mesmos têm em comum. Neste tipo de aprendizado, o indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados, formando clusters. Após isso é necessária análise

para determinar o que cada agrupamento significa no contexto do problema que está sendo analisado.

De acordo com NAVEGA (2002) padrões são unidades de informação que se repetem, ou então são sequências de informações que dispõe de uma *estrutura* que se repete. A tarefa de localizar padrões não é privilégio do Data Mining. Nosso cérebro utiliza-se de processos similares, pois muito do conhecimento que temos em nossas mentes é, de certa forma, um processo que depende da localização de padrões.

O exemplo a seguir sugerido por NAVEGA (2002) mostra os possíveis procedimentos na descoberta de padrões em uma sequência de letras. Utilizando a sequência "ABCXYABCZKABDKCABCTUABEWLABCWO procura-se algum dado relevante. Abaixo seguem algumas possibilidades:

1. A primeira etapa é perceber que existe uma sequência de letras que se repete bastante. Temos as sequências "AB" e "ABC" e concluimos que elas ocorrem com frequência superior à das outras sequências.
2. **A**pós determinarmos as sequências "ABC" e "AB", verificamos que elas *segmentam* o padrão original em diversas unidades independentes:
   3. "ABCXY","ABCZK","ABDKC","ABCTU","ABEWL","ABCWO"
4. Fazem-se agora induções, que geram algumas *representações genéricas* dessas unidades:
5. "ABC??" "ABD??" "ABE??" e "AB???", onde o símbolo (interrogação) '?' representa qualquer letra
6. No final do processo, toda a sequência original é substituída por regras genéricas indutivas que simplificou (reduziu) a informação original a algumas expressões simples.

Este processo é um dos pontos essenciais do Data Mining, ou seja, extrair certos padrões de dados brutos através da indução, reduzindo a informação bruta. Além disso, esse processo nos permite gerar formas de *predizer* futuras ocorrências de padrões.

AVI *apud* BOGORNY (2003) salienta que com relação ao tipo de aprendizado em banco de dados, existe o aprendizado de máquina e o aprendizado em mineração de dados. Um sistema de aprendizado de máquina utiliza informações de um conjunto de treinamento, ou seja, uma amostra de dados selecionada. A amostra é gerada de

acordo com uma técnica : *supervisionada* ou *não-supervisionada*. Na supervisionada o sistema busca as descrições das classes definidas pelo usuário e, se for não-supervisionada, o sistema gera um conjunto de novas classes de acordo com suas descrições.

HAL *apud* BOGORNY (2003) diz que um sistema de aprendizado em mineração de dados busca descrições de dados em uma base inteira e não em uma amostra, resultando em um processo mais complexo e demorado.

DIAS & PACHECO (2005) afirmam que os principais objetivos da mineração são descobrir os relacionamentos entre dados e fornecer subsídios para fazer uma previsão das tendências futuras de acordo com os dados anteriores. Os resultados obtidos da mineração podem ser usados para tomada de decisão, controle de processo e muitas outras aplicações.

2.5.3.6 Ferramentas utilizadas em Data Mining

Existem diversas ferramentas disponíveis no mercado que realizam várias tarefas de mineração. A escolha pode ser feita de acordo com o problema e com o resultado pretendido. São elas:

- WEKA: um software de domínio público, desenvolvido em Java, pela Universidade de Waikato, que implementa uma série de algoritmos de Data mining.
- INTELLIGENT MINER: desenvolvido pela IBM, também é uma ferramenta de mineração, possui interligação com banco de dados DB2 da IBM.
- ORACLE DATA MINER: é uma ferramenta de mineração que possui interligação com o banco de dados Oracle Enterprise.
- ENTERPRISE MINER: nova versão do SAS Enterprise Miner, usado na área de negócios, marketing e inteligência competitiva.
- STATISTICA DATA MINER: este software acrescenta as ferramentas de mineração ao tradicional pacotes de aplicações estatística.

2.5.3- Descoberta de Conhecimento em Bases de Dados de Monitoramento de Qualidade da Água

O processo de monitoramento ambiental gera um grande conjunto de dados que são armazenados em uma base de dados. Normalmente são pouco favoráveis a análise de gestores e pesquisadores ambientais devido a limitação humana em analisá-los e guardá-los na memória. Neste caso a descoberta de conhecimento, através de suas técnicas de mineração de dados poderá , através de algoritmos descobrir padrões de classificação da qualidade da água.

As técnicas de data mining já são utilizadas em diversas áreas e recentemente estão sendo utilizadas para descoberta de conhecimento em bases de dados de monitoramento ambiental através do reconhecimento de padrões inicialmente ocultos, inclusive bases de dados com dados geo-espaciais como na Figura 07.

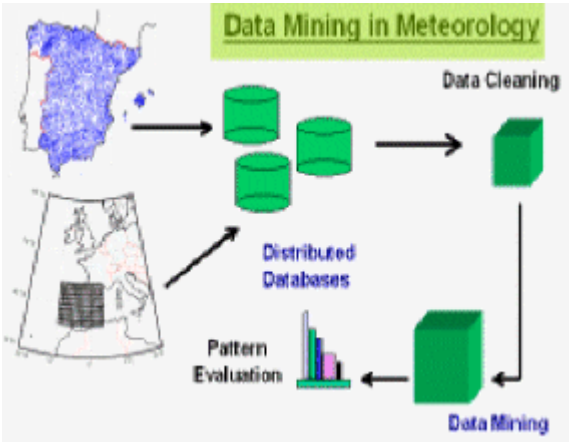

Figura 7 – DM em Meteorologia. (Fonte: www. crossgrid.org/.../meteo_pollution.html)

A Figura 07 mostra a aplicação de data mining na área de meteorologia, primeiramente as informações geo-espaciais são armazenadas em bases de dados distribuídas, em seguida é feita a pré-seleção ou limpeza de dados incorretos para aplicar a técnica de DM, o resultado é a interpretação dos padrões de dados encontrados pela técnica.

DELAVAR (2005) retrata a utilização de data mining em bases de dados geo-espaciais para monitoramento da qualidade da água. Esta aplicação é pela análise da vizinhança que especifica os rios com a mesma poluição. A técnica de data mining

geoespacial pode ser usada para mostrar a quantidade de poluentes e estimá-las para proteção ambiental.

PEREIRA (2005) salienta as comunidades ecológicas representam uma complexa rede de interações entre muitas espécies. Sua análise visa acessar padrões de comportamento trófico do ecossistema, explicitando esses conhecimento através de regras geradas de modelos de classificação e previsão, para investigar padrões da variabilidade temporal de espécies e examinar relações destes padrões `a disponibilidade de alimento e a relação deste alimento com parâmetros físico-químicos de qualidade de água.

BIRANT (2006) apresenta o algoritmo ST-DBSCAN, usado para agrupamento (clustering) de dados espaciais-temporais obtidos por satélite, onde cluster é um dos métodos de KDD largamente utilizados em bases de dados, também GIBERT (2006) apresenta estudos realizados no projeto GESCONDA: um sistema de análise de dados inteligente para descoberta de conhecimento e gerenciamento em bases de dados ambientais.

Vislumbrando o contexto deste trabalho, o objetivo principal é obter conhecimento relevante em uma base de dados de monitoramento de qualidade da água. Neste caso, é importante empregar uma tarefa de classificação, capaz de extrair regras de previsão dos estados de qualidade da água para avaliar os dados atuais da qualidade da água da bacia e possíveis situações futuras.

Como exposto nos estudos citados, a classificação é uma das tarefas de DM mais referenciadas na literatura para KDD. Nela o objetivo é descobrir as relações entre um atributo meta (cujo valor, ou classe será previsto) e um conjunto de atributos previsores. O sistema deve descobrir este relacionamento a partir de exemplos com classe conhecida, ou seja com a classificação de uma amostra *a priori* realizada. O relacionamento descoberto será usado para prever o valor do atributo meta (ou a classe) para exemplos desconhecidos. Neste trabalho pode-se definir classificação da qualidade da água como atributo meta a ser atingida, com os valores: ótimo, bom , regular, ruim e péssimo, e os atributos previsores as variáveis indicadoras da qualidade da água, tais como: pH, OD, DBO, DQO, TURBIDEZ, SÓLIDOS DISSOLVIDOS TOTAIS, NTK, P, COLIFORMES TOTAIS E COLIFORMES

FECAIS. A tarefa de classificação é composta por vários algoritmos que podem ser aplicados resultando em conhecimento obtido sob a forma de regras de classificação. Conforme ROMÃO (2002), as regras do tipo SE...ENTÃO..., também chamadas regras de produção constituem uma forma de representação simbólica e possuem o seguinte formato:

SE <antecedente> ENTÃO <conseqüente>

O antecedente é formado por expressões condicionais envolvendo atributos do domínio da aplicação existentes no banco de dados. O conseqüente é formado por uma expressão que indica a previsão de algum valor para um atributo meta (neste caso qualidade da água), obtido em função dos valores encontrados nos atributos que compõem o antecedente. Assim, a tarefa é descobrir regras de classificação capazes de avaliar a qualidade da água objetivando auxiliar gestores e órgãos governamentais no planejamento de ações futuras.

2.5.4- Medidas de Avaliação de Regras

Independente do tipo de algoritmo usado na tarefa de classificação, as regras descobertas precisam ser avaliadas. Para isto vários autores utilizam técnicas estatísticas para avaliar a capacidade de representação do conhecimento adquirido sob a forma de regras.

Existem poucos estudos analíticos sobre o comportamento de algoritmos de aprendizagem. A análise de classificadores é fundamentalmente experimental. Dimensões da análise: taxa de acerto, complexidade dos modelos, tempo de aprendizagem.

Segundo ROMAO (2002) existem diversas abordagens para avaliar o processo de descoberta de conhecimento, incluindo-se: exatidão dos resultados (alguma medida de taxa de acerto), eficiência, compreensão do conhecimento extraído.

A cobertura (razão entre o número de casos cobertos e o número total de casos) também é uma métrica relevante neste processo de avaliação. É uma medida da

generalidade da regra, quanto maior a cobertura maior o número de exemplos cobertos pelo corpo da regra.

A maior parte da literatura utiliza taxa de acerto como principal meio de avaliação das técnicas de KDD (FREITAS, 1997a), principalmente para tarefa de classificação.

Em algumas situações o tempo de aprendizagem é de extrema importância, especialmente quando o sistema está inserido em um ambiente altamente dinâmico.

Facilidade de compreensão nos resultados da classificação (ex. nas regras) é outra forma de avaliação do processo de descoberta que favorece a credibilidade no sistema por parte do usuário. Em muitas aplicações a compreensão por parte dos humanos é extremamente importante, exigindo algum complemento no pós-processamento. Em geral considera-se que quanto menor o número de regras descobertas e menor o número de condições por regra, maior a compreensibilidade do conjunto de regras. No entanto, esse critério é puramente sintático e objetivo, ignorando aspectos semânticos e subjetivos das regras. (ROMÃO, 2002)

PAZZANI *apud* ROMÃO (2002) salienta que "não há nenhum estudo que mostre que pessoas acham modelos menores mais compreensíveis, ou que o tamanho de um modelo é o único parâmetro que determina sua compreensibilidade".

WITTEN & FRANK (2000) salientam que o conjunto de dados utilizados para extração do conhecimento deve ser dividido em dois grupos, um conjunto de treinamento e um conjunto de teste. O algoritmo de classificação é aplicado ao conjunto de treinamento, as regras são descobertas. Em seguida a performance do algoritmo para estas regras é realizado pela aplicação das mesmas no conjunto de treinamento, caracterizando uma forma de aprendizagem supervisionada.

2.5.4.1- Taxa de Acerto

A exatidão de um resultado de extração de conhecimento, pode ser medido pela taxa de acerto. Conforme ROMAO (2002), a taxa de acerto pode ser definida pela proporção de classificações realizadas corretamente pelas instâncias dos dados aplicáveis. Ela pode ser medida sobre os dados de treinamento e sobre os dados de teste, sendo que a exatidão do conjunto de teste alcançada possui geralmente taxas inferiores, devido ao fato de se utilizar dados que não foram considerados

anteriormente durante o treinamento. Logo, a exatidão perante os dados de teste é considerada mais representativa para caracterização do universo classificado, fornece uma medida mais precisa da confiabilidade das regras (ROMÃO, 2002).

Salienta-se que a Taxa de acerto (relação das instâncias dos dados classificadas corretamente, dividido pelas instâncias dos dados aplicáveis), não se mostrou muito eficaz em situações nas quais a distribuição das classes é desbalanceada, ou seja, as classes mais raras são mais difíceis de serem previstas.

Segundo WITTEN & FRANK (2000), os usuários citam três conjuntos de dados como ideal: treinamento, a validação e ao conjunto de teste. O conjunto de treinamento pode ser usado para esquema de aprendizagem do classificador, a validação para otimizar parâmetros destes classificadores e o conjunto de teste é usado para calcular o erro no final. Segundo os mesmos autores não há problema quando se tem um grande conjunto de dados, usa-se uma boa parte para fase de treinamento para descobrir conhecimento e testa-se independentemente em outro conjunto de teste, aí temos um erro no conjunto de teste representativo, que indica verdadeiramente a performance futura do conhecimento encontrado.

Geralmente quanto maior o conjunto de treinamento melhor o classificador. Quanto maior o conjunto de teste mais exata a estimativa do erro.

2.5.4.2- "Hold Out"

Este método é adequado quando uma grande quantidade de registros está disponível. O conjunto de dados é separado em dois subconjuntos: dados para treinamento e dados para para teste. Os registros do conjunto devem ser separados dos registros de teste de forma aleatória, com uma exceção: manter a mesma proporção de registros de cada classe nos dois conjuntos, o que é chamado de *hold out* estratificado. O conjunto pra treinamento é usado para descobrir regras de classificação. O conjunto para teste deve permanecer separado e ser utilizado apenas para validação dos resultados obtidos pelo algoritmo de treinamento. Esse método é simples e apropriado quando se tem abundância de dados. Segundo WEISS (1991) um conjunto de dados disponíveis para teste precisa de pelo menos 1000 registros, caso contrário deve-se usar outro método de validação, tal como validação cruzada.

2.5.4.3- Validação Cruzada

Este método é utilizado quando se tem poucos dados para treinamento e teste. Considerando que a amostra de dados pode não ser representativa todos os dados são utilizados para treinamento e para teste. Os dados são divididos em n partes e cada parte contendo o mesmo número de registros. Uma parte é omitida, aplica-se o esquema de aprendizagem nas outras nove partes, assim o algoritmo de aprendizagem é executado n vezes nos diferentes conjuntos de treinamento. É realizada a média do erro de todas as aplicações. WITTEN & FRANK (2000) consideram, após extensivos testes em bases de dados que um n de dez é um valor padrão apropriado para conseguir uma boa estimativa de erro.

2.5.4.4- Taxa de erro

WITTEN &s FRANK (2000) salientam que para os problemas de classificação é natural mensurar a performance do classificador em termos da taxa erro (error rate). O classificador prediz a classe de cada instância: se é correta, isso é contado como sucesso, se não for, considera-se um erro. Estado de erro (error rate) é justamente a proporção de erros realizados no conjunto inteiro de instâncias e mostra a performance do classificador. Ele é chamado de erro de re-substituição ou enviesado quando estimado no conjunto de treinamento, e o erro de generalização é estimado no conjunto de testes independente por isso não é um erro enviesado. O *error rate* no conjunto de treinamento não é um bom indicador de performance futura do classificador. Isso por que o classificador aprendeu as regras no mesmo conjunto de treinamento, sendo qualquer estimativa de performance nesse conjunto otimista. Para predizer a performance de um classificador em uma nova informação, precisa-se obter o *error rate* num conjunto independente de dados chamado conjunto de teste.

O erro pode ser decomposto em erros provenientes de ruído no conjunto de dados, erro de viés (estimado no conjunto de treinamento) e de variância. Se o tipo de classificador não é adequado para problema o erro de viés será alto. Fazer uma pré-

seleção dos classificadores elimina o erro de viés. O erro de variância existe na medida em que outro conjunto de dados está sendo coletado e oferece melhor desempenho que a amostra utilizada.

2.5.4.5 Fator de Confiança

A utilização de um fator de confiança é uma forma simples de se avaliar a precisão das regras obtidas apenas nos dados de treinamento. Uma regra é uma expressão na forma *X=> Y,* onde *X* é chamado de antecedente e *Y* denominado conseqüente da regra. Tanto *X* como *Y* pode ser formado por conjuntos de dados. X é formado por expressões condicionais envolvendo atributos previsores do domínio da aplicação presente na base de dados e Y é formado por uma expressão que indica a previsão de algum valor para um atributo meta, obtido em função dos valores previsores encontrados nos atributos que compõem o antecedente. Assim, o fator de confiança é calculado pela razão X/Y, onde X é o número de registro que satisfaz o antecedente e o conseqüente da regra (número de casos que a regra cobre) e Y é o número total de registros que satisfazem o antecedente da regra(total de casos aplicáveis).
Segundo FREITAS (1999), quando uma regra cobre poucos (registros, este método apresenta limitações, resultando em superestimações da confiabilidade das regras. Para minimização destas distorções, QUINLAN (1987) propôs a subtração de uma constante de 0,5 do valor X [1]:

$$\text{Fator de Confiança} = (X - 1/2)/Y \qquad [1]$$

De acordo o fator de confiança é reduzido fortemente no caso de regras com poucos registros, mas não penaliza a avaliação de regras baseadas em muitos registros, permitindo a distinção entre regras especializadas e generalizadas.

2.5.4.6- Matriz de Confusão

Há situações onde os métodos simples de validação das regras são insuficientes. Existem casos onde é necessário descobrir os custos das classificações erradas, os

tipos dos erros. A matriz de confusão foi exemplificada por ROMÃO (2002), para um sistema que auxilia um diagnóstico médico. São diferenciadas decisões negativas falsas (paciente classificado como não doente, sendo ele na verdade doente) e decisões positivas falsas (paciente classificado como doente, sendo ele na verdade não doente). No método é construída uma matriz de confusão NxN (Quadro 4) com domínio composto por duas classes, denominadas aqui de "sim" e "não".

Quadro 3-Matriz de Confusão

	Classe prevista		
Classe atual		**sim**	**Não**
	sim	SC	SE
	não	NE	NC

O algoritmo de classificação deve ser executado sobre os dados de treinamento e o resultado da classificação dividido em quatro categorias:

- Classificações "sim" classificadas corretamente (SC);
- Classificações "sim" classificadas erradamente (SE);
- Classificações "não" classificadas erradamente (NE);
- Classificações "não" classificadas corretamente (NC);

Os contadores internos da matriz de confusão indicarão as quantidades de exemplos sim corretamente e sim erroneamente classificados. Quanto mais coberturas corretas (SC e NC) e menos erradas (SE e NE) maior a precisão da regra.

Para ROMAO (2002) a qualidade da regra pode então ser calculada por uma das seguintes equações (2 a 4):

QUALIDADE=(SC/(SC+NE))*(NC/NC+SE)) [2]

QUALIDADE=(SC/(SC+SE))*(SC/(SC+NE)) [3]

QUALIDADE=(SC-1/2)/(SC+SE) [4]

A equação 4 é uma adaptação da equação 1, conhecida como fator de confiança. A equação 3 é conhecida como Precision*Recall (HAND, 1997). Na equação 2 o primeiro termo é chamado de "sensibilidade" ou "completeza de positivos" ou "taxa

de acerto na classe positiva". O segundo termo é chamado de "especificidade" ou "completeza de negativos" ou "taxa de acerto na classe negativa". Os dois termos da equação são multiplicados para forçar a descoberta de regras que tenham alta sensibilidade e alta especificidade (Carvalho & Freitas, 2000).

Sensibilidade (abrangência) é a habilidade do teste em identificar corretamente os casos que são positivos. Especificidade é a habilidade do teste em identificar corretamente os casos negativos.

3 – MATERIAIS E MÉTODOS

Avaliando o crescente processo de urbanização das cidades banhadas pela bacia do Rio Cuiabá e nas implicações deste processo optamos por aplicar técnicas de Descoberta de Conhecimento (KDD) para descobrir conhecimentos ocultos em um dos módulos do Sibac (na base de dados) para avaliar qualidade da água nos trechos do rio Cuiabá . As amostras foram selecionadas na base de dados do SIBAC através de consultas ao site http://www.geohidro.ufmt.br , especificando o trecho, a data de coleta dos mesmos e as variáveis indicadoras de qualidade sugeridas pelos especialistas. Em seguida, estes registros passaram pela avaliação sistemática de especialistas da área que os classificaram segundo os critérios: ótimo, bom, regular, ruim e péssimo. Posteriormente a amostra foi submetida a ferramenta de mineração de dados Weka, para aplicação das tarefas de mineração e descoberta de padrões.

3.1 ÁREA DE ESTUDO

O Estado de Mato Grosso está localizado na região Centro Oeste do Brasil, com cerca de 900.000 km^2 de extensão territorial e ocupa 11% do território brasileiro. Abriga as principais nascentes de três bacias hidrográficas brasileiras: Bacia Amazônica, Araguaia/Tocantins e Platina. A Bacia Platina, no Estado de Mato Grosso, é chamada de Bacia do “Alto Rio Paraguai”, estende-se até o Estado de Mato Grosso do Sul e pode ser dividida em cinco sub-bacias. São elas: a do rio Cuiabá, Rio Paraguai, Rio São Lourenço/Rio Vermelho, Rio Itiquira/Rio Correntes e o Pantanal (ALVARENGA,1984).

Um dos principais rios desta bacia é o Rio Cuiabá, que drena aproximadamente

28.000 km^2 até a cidade de Barão de Melgaço, abrangendo os seguintes municípios: Acorizal, Campo Verde, Barão de Melgaço, Chapada dos Guimarães, Cuiabá, Jangada, Nobres, Nossa Senhora do Livramento, Nova Brasilândia, Poconé, Rosário Oeste, Santo Antônio do Leverger e Várzea Grande. O rio Cuiabá tem suas nascentes no município de Rosário Oeste e é inicialmente formado por dois pequenos cursos d'água, Cuiabá Bonito e Cuiabá da Larga, que afloram entre as Serras Azuis e Cuiabá. O ponto de união desses dois cursos d'água é chamado Limoeiro, onde o rio passa a chamar-se Cuiabazinho, logo abaixo recebe águas do Manso dobrando de volume, sendo desse ponto em diante chamado de Cuiabá (SONDOTÉCNICA *apud* LIMA,2001).

O alto curso do rio Cuiabá e seus afluentes drenam nos municípios de Nobres, Rosário Oeste, Nova Brasilândia, Campo Verde, Acorizal, Jangada e Chapada dos Guimarães em áreas do planalto. O médio rio Cuiabá atravessa uma extensa superfície aplainada com baixas altitudes entre 200m e 450m, conhecida por Depressão ou Baixada Cuiabana nos municípios de Cuiabá, Várzea Grande, Nossa Senhora do Livramento e Santo Antônio do Leverger, compreende uma área de depressão que fica entre as partes mais altas do planalto e o início da planície inundável do Pantanal Mato-grossense setentrional, nos municípios de Barão de Melgaço e Poconé (SALOMÃO,1999).

LIMA (2001) salienta que a bacia do rio Cuiabá é constituída por três regiões geomorfológicas, com características bióticas e abióticas definidas e próprias, que correspondem `a área de planalto e serras circunvizinhas, `a Baixada Cuiabana e `a planície do pantanal. Assim em função da declividade ele possui comportamento de rio de planalto, controlado pela estrutura geológica, com diversas corredeiras até atingir o nível da base regional, como um rio de planície, representado pelo Pantanal Matogrossense.

Na área de estudo denominada de Baixada Cuiabana, o clima predominante é do tipo quente tropical sub-úmido, com temperatura média anual de 26 graus, ocorrendo as máximas médias diárias em torno de 36 graus, em setembro, e as mínimas de 15 graus, em julho. A precipitação média anual chega a valores de 1342 mm/ano, de acordo com a série temporal medida entre 1989-2000 (INMET, 2000). A bacia apresenta sazonalidade marcada por dois períodos distintos de estiagem, de maio a

outubro ,e cheia de novembro a abril.

A bacia do Rio Cuiabá, uma das mais importantes para a formação bacia do Alto Paraguai, é a principal bacia considerando o ponto de vista econômico e também a que está mais comprometida pelos impactos ambientais, devido a atração de atividades altamente consumidoras e degradadoras de recursos hídricos, tais como: agropecuária, garimpos, construção civil, aglomerado urbano.

3.1.1- Aspectos Demográficos

Segundo dados do IBGE (2000), a população da bacia do rio Cuiabá é predominantemente urbana, com somente 7% na zona rural. A maior concentração populacional encontra-se na parte média da bacia, onde são localizados os municípios de Cuiabá e Várzea Grande que concentram 35% da população de todo Estado de Mato Grosso.

As principais cidades em Mato Grosso, localizadas na Região Hidrográfica, que promovem a economia da região são: Cuiabá, com 483.346 habitantes e Várzea Grande, com 215.298 habitantes. As Sub-bacias mais populosas são as do Cuiabá 02, com 39,80%, devido a localização das cidades de Cuiabá e Várzea Grande respectivamente. Em termos gerais, a densidade habitacional na Bacia é de 5,21hab/km2

O Caderno da Região Hidrográfica do Paraguai/MMA (2006) apresenta que as principais atividades desenvolvidas na região Hidrográfica estão relacionadas historicamente com agropecuária, embora em diversas regiões, tanto no Mato Grosso como no Mato Grosso do Sul, exista o uso localizado mais intenso para mineração, turismo, pesca e industrial. Em termos da ocupação das terras, a atividade mais intensa se refere a pecuária seguida pela atividade agrícola. Parte da pecuária extensiva também se utiliza de parcela significativa dos campos naturais, especialmente na planície pantaneira, que corresponde a 25,7% da área dos estabelecimentos agropecuários, e em pastagens artificiais, 38,8% das terras ocupadas pela atividade. Na mesma bacia, os estudos demonstraram que dos 103.510 km^2, 64% estão em área urbana, que conta com uma população de 1.072.985 hab, dos quais 91,37% residem em área urbana com uma densidade populacional de

14,85 hab/km^2 e 2,47 hab/km^2 na área rural.

TUCCI (2000) salienta que a grande concentração urbana é a principal responsável pelos conflitos e problemas gerados nessa bacia, tais como: degradação e assoreamento de mananciais; aumento de riscos das áreas de abastecimento por poluição orgânica e química; contaminação dos rios por esgoto doméstico e industrial; gerenciamento inadequado da drenagem urbana e falta de coleta e disposição do lixo urbano.

Nas décadas de 1970 e 1980, a maior fonte de impactos ambientais para a região foi a atividade agropecuária no planalto adjacente, em grande parte realizada sem os cuidados previstos na legislação, causando desmatamentos em área de preservação permanente(mata ciliar e nascentes), erosão, perda de solos e assoreamento de rios. (GOMES e PADOVANI-CATELLA, 2004).

A mudança de cenários no setor da pecuária regional, com a mudança da propriedade das terras, passando dos pecuaristas pertencentes a comunidades tradicional- para empresários ou grupos empresariais do centro-sul do País, e a subdivisão das fazendas elevou em muito as taxas de desmatamento e de queimadas, inclusive na áreas de planície. No Mato Grosso 35% da área do pantanal está desmatada (PADOVANI *et al*, 2004).

O Caderno da Região Hidrográfica do Paraguai/MMA (2006) apresentou que, na última década, houve aumento da atividade agrícola na região de planície (aumento da área de cultivo de feijão e de arroz irrigado), da exploração mineral, da navegação fluvial com impactos expressivos na margens do rio Paraguai em especial, e da carga de poluentes domésticos, industriais e agrícolas lançados em muitos dos seus rios. A atividade mineral teve importância no processo de ocupação da Bacia Hidrográfica. A extração de ouro descoberto no rio Coxipó até os tempos atuais tem sido por garimpagem. A extração do ouro é histórica e tem sido combatida em função dos riscos ambientais provocados na região devido ao uso do mercúrio, ao assoreamento provocado pelas deagas e ao desmatamento ciliar. A Universidade Federal de Mato Grosso(UFMT) pesquisa, ao longo dos anos, nos rios próximos a Poconé a contaminação por mercúrio, que já foi identificada na cadeia alimentar.

Segundo LIMA (2001) o rio Cuiabá tem seus usos diretos e indiretos tais como: abastecimento público e rural, produção de energia, irrigação, diluição de esgoto

domiciliar e industrial, pesca e turismo. As alterações na ocupação da bacia, principalmente a intensificação da atividade agrícola nas últimas décadas, vêm causando impactos nos recursos hídricos tais como erosão, assoreamento, eutrofização e contaminação por agrotóxicos.

De acordo com a classificação CONAMA (2005), todos os rios da bacia do rio Cuiabá são considerados de classe 2, áreas destinadas:

a) ao abastecimento doméstico, após tratamento convencional;
b) a proteção das comunidades aquáticas;
c) a recreação de contato primário (esqui aquático, natação e mergulho);
d) a irrigação de hortaliças e plantas frutíferas;
e) a criação natural e/ou intensiva(aqüicultura) de espécies destinadas `a alimentação humana

3.2- SISTEMA INTEGRADO DE MONITORAMENTO AMBIENTAL DA BACIA DO RIO CUIABÁ (SIBAC)

Os dados de qualidade da água analisados neste estudo são provenientes do Sistema Integrado de Monitoramento Ambiental da Bacia do Rio Cuiabá (SIBAC). SIBAC visa subsidiar o controle e manejo dos recursos hídricos desta bacia, na área compreendida entre as nascentes até o inicio da planície pantaneira na região de Barão de Melgaço. Foi projetado e implementado pelo grupo de pesquisa GEOHIDRO da na Universidade Federal de Mato Grosso (UFMT). O componente de Banco de Dados de SIBAC inclui um conjunto de registros multivariados de mais de 150 pontos de monitoramento, provenientes de diversos órgãos tais como UFMT, SEMA, Furnas etc., abrangendo um período de cerca de 30 anos.

O banco de dados possui interface WWW, que pode ser acessado pelo endereço http://www.geohidro.ufmt.br, permitindo ao usuário buscas interativas, escolhendo pontos de monitoramento, período e variáveis de qualidade de água. A Figura 08 mostra a interface de consulta aos dados.

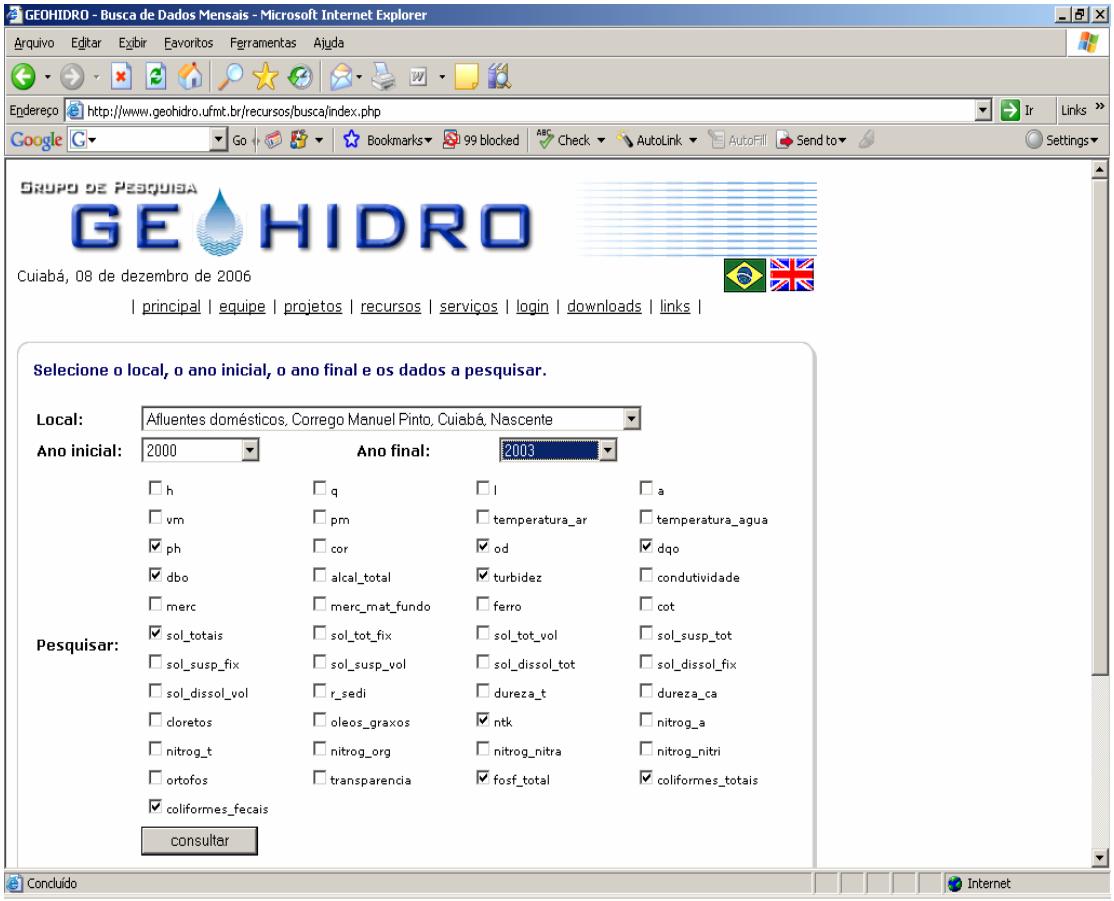

Figura 8 - Interface WWW para consulta de médias mensais de qualidade de água no Banco de Dados SIBAC. (Fonte: www.geohidro.ufmt.br)

Acoplado ao componente de Banco de Dados, o SIBAC possui um Sistema de Informação Geográfica (SIG) implementado em ArcView, para análises e visualizações espaciais com modelos hidrológicos e de qualidade de água integrados, entre eles o NGFlow (SANTOS & ZEILHOFER,), QUAL2E (ZEILHOFER *et al*. 2001) e Outorga (ZEILHOFER et al ,2001).

O modelo Outorga, proposto por HORA (2001), permite a simulação da disponibilidade hídrica na rede hidrográfica em função da quantidade e qualidade da água nos mananciais superficiais. Como resultado tem-se que a disponibilidade hídrica para a concessão da outorga será positiva ou negativa, ou seja, o modelo transforma o aspecto qualitativo da outorga em disponibilidade hídrica (quantitativo).

3.3- CONJUNTO DE DADOS E SUA CLASSIFICAÇÃO POR ESPECIALISTAS

Foram extraídos do banco de dados SIBAC um total de 300 amostras (registros de dados), como mostra tabela exemplar. Devido as classificações realizadas pelos especialistas, alguns usos tem número menor de registros que outros, isto porque os especialistas mostraram-se insatisfeitos com tamanho número de registros, alguns responderam só 50 registros, alegando ser dispendiosa a classificação.

Inicialmente selecionamos a amostra através de diversas consultas ao site. Por meio de entrevistas com 3 (três) engenheiros(as) sanitaristas foram escolhidas dez variáveis que tiveram, de acordo com a opinião dos especialistas, a maior importância na caracterização da qualidade da água para os seguintes usos: **Abastecimento, Irrigação, Balneabilidade** e **Manutenção dos Ciclos Biogeoquímicos**, conforme Tabela1.

O uso "**Abastecimento**" compreende as atividades de consumo humano após tratamento convencional; A "**Balneabilidade**" corresponde ao uso da água para recreação de contato primário (probabilidade alta do usuário ingerir água: natação, esqui aquático,mergulho), na "**Irrigação**" a água é usada para irrigar campos, hortaliças, plantas frutíferas, jardins; e, "**Manutenção dos Ciclos Biogeoquímicos**" deve ser entendido como uma avaliação composta da similaridade ou dissimilaridade das características de qualidade de água sob condições naturais (sem introdução de poluição pontual ou difusa nas mananciais).

As variáveis selecionadas para este estudo de caso tem o limite de suas concentrações e unidades especificados segundo a classe 2 da resolução CONAMA n° 357/2005 classe onde o rio Cuiabá está enquadrado (Quadro 4).

Quadro 4-Variáveis indicadoras de qualidade da água e seus limites segundo o CONAMA 357/2005

Variável	**Descrição**	**Unidade**	**Limite CONAMA, classe 2**
CF[2]	Coliforme Termotolerante	NMP/100ml	Até 1000 por 100ml em 80% de 6 amostras
CT[3]	*Eschrichia Coli*	NMP/100ml	Até 5000 por 100ml em 10% de 5 amostras
NTK	Nitrogênio Total Kjheldal	Mg/L	Até 10 mg/l
P	Fósforo	Mg/L de P	Até 0,025mg/l de fósforo
OD	Oxigênio Dissolvido	Mg/L	Acima de 5mg/l de O_2
DBO	Demanda Bioquímica de Oxigênio	Mg/L	em dias a 20°C até 5 mg/l de O_2
pH	Potencial Hidrogeniônico		6 a 9
ST	Sólidos Dissolvidos Totais	Mg/L	Até 500
DQO	Demanda Química de Oxigênio	Mg/L	sem limite para classe 2
Turb	Turbidez	UNT (unidade nefelométrica de turbidez)	até 100 UNT

Os registros foram repassados para quatro especialistas (três engenheiros(as) sanitaristas, e uma bióloga), para classificação de cada registro de acordo com a sua qualidade em cinco classes (1-péssimo, 2-ruim, 3-regular, 4-bom, 5-ótimo). A classificação foi efetuada separadamente para cada uso. A tabela 2 mostra, de forma

[2] O CONAMA 357 alterou a nomenclatura de Coliformes Fecais para Coliformes Termotolerantes.

[3] Os dados coletados no Sibac ainda estão com a nomenclatura do CONAMA 20, assim justifica-se o valor da variável indicadora CT (antes Coliformes Totais) com a descrição *Eschrichia Coli* e valores conforme CONAMA 20 .

exemplar a classificação de alguns registros para os usos “Abastecimento” (Abast) e “Irrigação” (Irrig).

Para não induzir a classificação dos especialistas foram retirados dos registros as informações sobre o local e data das coletas das tabelas da amostra, conforme Tabela 1.

Tabela 1 - Parte da tabela de registros do SIBAC classificados por um especialista para os usos de abastecimento e irrigação (5: ótimo até 1: péssimo)

pH	OD	DQO	DBO	Turb	ST	NTK	P	CT	CF	Abast	Irrig
6.48	4.64	19	1.46	36.9	116	14	0.04	14670	100	4	5
6.46	7.25	24	5.4	124	238	1.26	0.69	120331	1610	3	3
6.91	1.58	32	7.26	37.8	264	16.35	0.53	24192	9208	1	1
6.7	1.74	6	5.94	100	42	8.4	0.17	20140	840	3	5
6.43	1.76	25	1.64	36.8	147	5.27	0.38	43520	3540	2	2
6.6	3.42	31	3.24	96	174	6.72	0.58	198628	43520	2	1
6.58	3.04	13	1.8	77.2	178	2.96	0.38	61310	6970	2	2
6.3	4.58	33	2.02	128	94	3.68	0.36	86640	2430	3	3
6.26	2.96	21	8.56	173	10	6.19	0.07	241920	1350	2	3
6.13	2.7	36	1.64	187.4	149	0.63	0.03	198628	8160	2	1
6.32	2.02	18	1.82	132.5	307	5.99	0.06	26130	2980	3	3

3.4- MINERAÇÃO DE DADOS

3.4.1 Tarefas de Mineração de Dados implementadas na Weka

O processo de mineração de dados foi realizado a partir da ferramenta de extração de conhecimento denominada *Weka Knowledge* (WITTEN & FRANK, 2000) no intuito de descobrir regras ou padrões de qualidade da água desta bacia. Weka é um software implementado em JAVA, segundo o paradigma de orientação a objetos, e é composto de uma série de algoritmos de aprendizagem para solucionar problemas de

mineração de dados. Os algoritmos podem ser aplicados diretamente a uma série de dados. Os motivos da escolha desta ferramenta foram:

1. É uma ferramenta desenvolvida na linguagem JAVA, que tem como característica principal a portabilidade (facilidade de ser executada em várias plataformas de Sistema Operacional), facilidade para pesquisadores;
2. Tem o código fonte aberto e as vantagens de uma linguagem orientada a objeto (modularidade, polimorfismo, encapsulamento, reutilização de código) para os desenvolvedores, importante para continuidade da pesquisa;
3. É de fácil acesso pela internet.

São distinguidas sete tarefas na mineração de dados, cada uma é adequada para atingir um objetivo, ou seja, possuem melhores soluções para determinados problemas. São elas: **Regressão, Associação, Agrupamento (segmentação ou clusterização), Sumarização, Desvio, Dependência e Classificação**. O software Weka é formado por pacotes que disponibilizam em sua estrutura algumas dessas tarefas.

Uma forma de representar conhecimento é a utilização de regras do tipo “Se <condição> então <faça> senão <atributo meta>”. Os métodos de classificação geram regras apropriadas para este tipo de representação de conhecimento

Como a meta deste trabalho é descobrir um modelo de avaliação para qualidade da água e em seguida aplicá-lo em toda a base de dados do SIBAC, foi utilizada a tarefa de classificação de dados. Esta tarefa possui uma série de algoritmos que podem ser utilizados para gerar um modelo de dados (regras), baseado em uma amostra. Posteriormente esse modelo de dados encontrado (regras) pode ser utilizado para classificar novos conjuntos de dados pertencentes a toda a base de dados e assim predizer classificações sobre a qualidade da água.

3.4.2- Entrada de Dados na Weka

O software Weka lê um formato de arquivo texto padronizado com extensão .arff. Para importar as amostras do SIBAC, inicialmente armazenadas em tabelas, foi

necessário transformá-las neste formato. Este procedimento foi realizado da seguinte maneira e pode ser visualizado na Figura 9:

- Salva-se a planilha com os dados, sem nenhuma formatação, no formato CSV (separado por vírgula);
- Abre-se este arquivo em formato csv em um editor de textos para acréscimo dos parâmetros requeridos do formato de arquivo com extensão .arff ;
- A primeira linha contém o nome do arquivo salvo (@relation irrigação), portanto o nome do arquivo foi irrigação (Figura 09).
- Os itens @atributte são seguidos das variáveis que serão avaliadas no sistema. Contém uma lista de todos os atributos seguidos do tipo (real, inteiro, string) do atributo e dos valores dependentes (categorias da classificação); no caso é a variável classificação com valores de 1 a 5.
- A terceira parte @data corresponde aos dados que a ferramenta avaliará, os registros de qualidade de água e sua respectiva classificação. Esta parte consiste das instâncias a serem mineradas com o valor dos atributos. A última variável (classificação) corresponde a classificação realizada pelos especialistas, onde o número 5 significa ótimo estado de qualidade da água, 4 é bom, 3 é regular, 2 é ruim e 1 é péssimo.
- O arquivo é salvo com o nome e a extensão .arff que pode ser aberto a partir do módulo “Weka Explorer” para visualização e escolha do algoritmo adequado para classificação.

```
@relation irrigacao

@attribute ph real
@attribute od real
@attribute dqo real
@attribute dbo real
@attribute turbidez real
@attribute solTotais real
@attribute ntk real
@attribute fosfTotal real
@attribute coliformesTotais real
@attribute coliformesFecais real
@attribute Classificacao {1,2,3,4,5}

@data
6.48,4.64,19,1.46,36.9,116,14,0.04,14670,100,4
6.17,0.86,64,0.68,100,207,9.07,0.42,73287,23590,3
6.46,7.25,24,5.4,124,238,1.26,0.69,120331,1610,2
6.91,1.58,32,7.26,37.8,264,16.35,0.53,24192,9208,3
6.7,1.74,6,5.94,100,42,8.4,0.17,20140,840,4
6.43,1.76,25,1.64,36.8,147,5.27,0.38,43520,3540,3
7.66,7.12,94,35,57,254,9.83,0.33,3820000,1640000,1
5.28,7.48,7,3.04,1.15,213,14.67,0.08,4130,100,3
4.95,2.8,18,1.42,4,145,0.06,0.08,1730,100,4
4.94,1.48,8,0.29,0.3,148,20.57,0.02,6131,10,4
5.33,5.54,3,1.64,2.3,154,1.34,0.03,7701,275,4
6.92,5.76,132,44,7.3,224,11.55,0.53,262000,63000,2
7.02,0.2,227,62.56,20.4,310,9.86,0.89,677000,213000,1
7.46,6.52,34,2,34,211,11.2,0.35,461100,179300,2
7.41,5.74,6,3.06,13,126,9.63,0.2,65700,39300,3
7.37,7,22,8.5,23,66,0.42,0.12,13000,1000,2
```

Figura 9 - Formato do arquivo arff para entrada na Weka

3.4.3- Ambiente Weka

A Weka possui interface para o usuário. Existem duas possibilidades de utilização : *GUI* interface gráfica e *Simple CLI* que é uma interface simples onde os usuários entram com linhas de comando (Figura 10).

Figura 10 - Janela de entrada do Weka

A janela *Weka Gui Chooser* é usada para introduzir o ambiente de aplicação Weka. Possui seis botões:

1. ***Simple CLI****:* é uma interface que permite a execução direta de linhas de comando do Weka pelo sistema operacional;
2. ***Explorer***: Um ambiente para explorar informação no Weka. *Explorer* contém as funcionalidades de pré-processamento, análise, (classificação, associação e clusterização) e visualização dos resultados.
3. ***Experimenter***: Um ambiente onde o usuário pode conduzir testes estatísticos entre os esquemas de aprendizagem da ferramenta; usuário pode usar diferentes algoritmos simultaneamente e comparar os resultados, e então escolher o melhor algoritmo para o seu conjunto de dados.
4. ***Knowledge Flow***: Ambiente essencialmente com as mesmas funções do Explorer mas com interface *draw-and-drop*, com a vantagem de suportar aprendizagem incremental;
5. ***Arff Viewer***: Um visualizador de arquivos .arff que permite edição, alteração de dados e atributos.

6. ***Log***: É uma área que registra todas as ações efetuadas durante uma sessão com Weka.

Neste trabalho utiliza-se como ambiente de estudo o Explorer e o Arff Viewer para visualização e análise dos arquivos de dados antes da mineração.

3.4.2- Weka Explorer

O botão Explorer (Figura 10) aciona a interface exposta na Figura 11. Podemos verificar as principais tarefas que a ferramenta realiza, ou seja as tarefas de classificação (*Classify*), associação (*Associate*) e agrupamento (*Cluster*). Neste trabalho utilizaremos especificamente a tarefa de classificação (*Classify*).

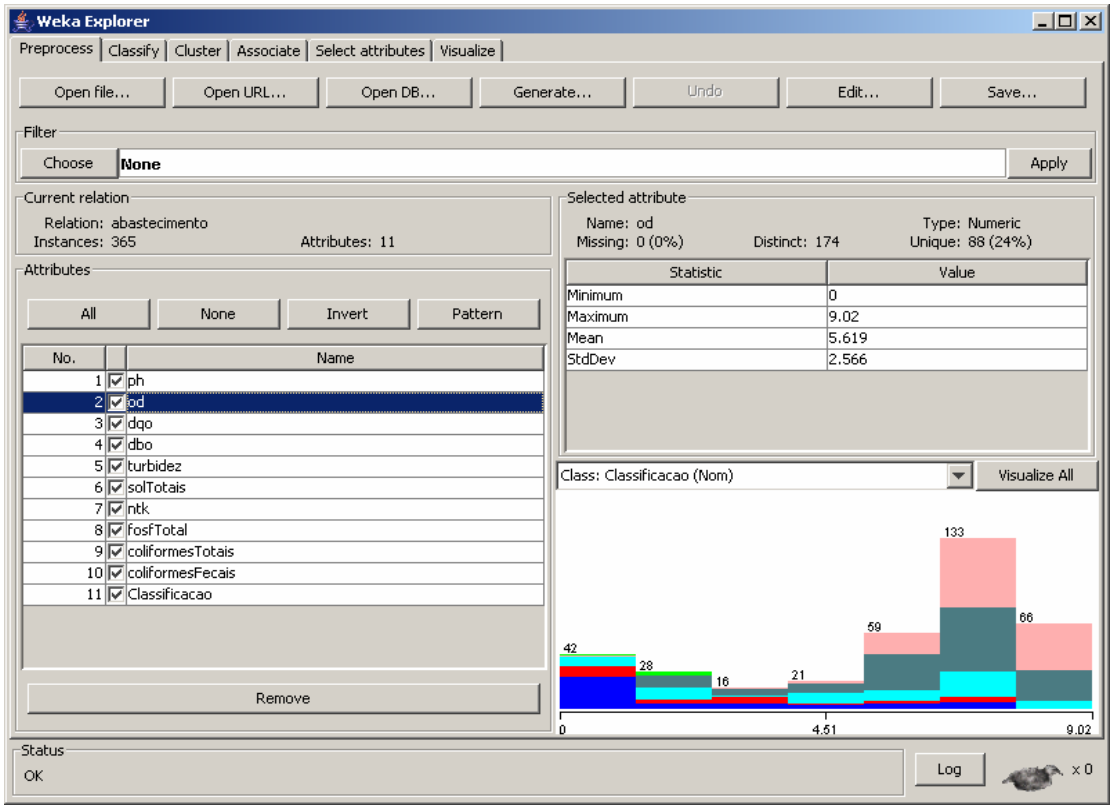

Figura 11 - Tela inicial da Weka com as variáveis e o arquivo abastecimento.arff aberto

A seção “Tabs” (guias) é composta por sete itens:

1. ***Preprocess*:** escolhe e modifica a informação antes da ação de mineração;

2. ***Classify****:* testa ou treina esquemas de aprendizagem em classificação ou regressão;
3. ***Cluster***: executa tarefas de agrupamento (“Cluster”);
4. ***Associate***: aprende regras de associação para a informação;
5. ***Select Attribute***: seleciona os atributos relevantes para a informação;
6. ***Visualize***: visualiza a informação em gráfico 2D.

3.4.2.1- Preprocessing

A seção *Preprocessing* habilita o usuário a introduzir os dados a serem minerados na ferramenta de quatro maneiras:

1. ***Open file***: Abre uma caixa de diálogo permitindo o usuário digitar o nome do arquivo a ser minerado ou o local de origem do mesmo. Utilizamos este botão pois introduzimos um arquivo com extensão .arff.
2. ***Open URL***: Abre uma caixa de diálogo onde o usuário deve digitar o endereço da URL (Uniform Resource Locator) onde a informação está armazenada (acesso em ambientes de rede);
3. ***Open DB****:* Lê a informação de uma base de dados. Para isto deve-se fazer uma edição no arquivo a partir do utilitário weka/experiment/DatabaseUtils.props.
4. ***Generate***: permite o usuário gerar dados através de funções específicas denominadas Datagenerators.

3.4.2.2- Current Relation

O item “**Current relation**”, permite a visualização das características básicas do conjunto de dados sendo elas:

Relation: o nome do arquivo que será minerado;
Instances: número de instâncias que serão analisadas neste arquivo
Attributes: o número de atributos analisados

3.4.2.3- Attributes

O item de menu chamado "**Attributes"** mostra as variáveis (atributos) da amostra, pode-se visualizar três colunas caracterizando as variáveis (Figura 12). São eles:

No.: identifica a ordem do atributo especificado no arquivo da relação;

Selection tick boxes: permite o usuário marcar ou desmarcar o atributo na relação;

Name: nome do atributo como foi declarado no arquivo .arff.

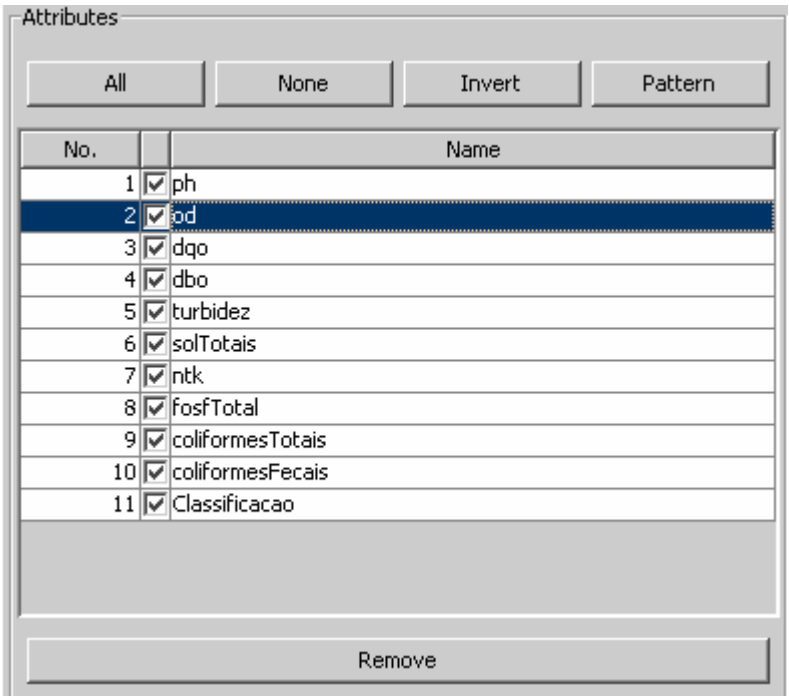

Figura 12 - Seção *Attributes* da Weka

Na lista de atributos especificada pela Figura 12, seguem os botões *all, none, invert e pattern*, usados para alterar a seleção dos atributos. Assim:

***All*:** todos os atributos são selecionados;

None: nenhum atributo selecionado, as caixas de seleção são desmarcadas;

Pattern: habilita o usuário a selecionar atributos baseados em expressões regulares da linguagem Perl5.

O botão ***Remove*** localizado abaixo da lista de variáveis pode ser utilizado para remover atributos indesejados. Esta ação pode ser desfeita acionando o botão ***Undo*** no painel de Preprocess.

3.4.2.4- Selected Attribute

Ao selecionar individualmente um atributo no painel *Attributes*, obtém-se a visualização das características do mesmo na janela chamada *Select Attribute*, como mostra Figura 13. São elas:

***Name*:** nome do atributo ou variável;

Type: tipo do atributo, normalmente nominal ou numérico;

Missing: o número ou percentagem de instâncias para este atributo que não foram especificadas;

Distinct: o número de valores distintos na relação para este atributo.

Unique: especifica o número ou percentagem de instâncias únicas, ou seja, valores que nenhuma outra instância possui.

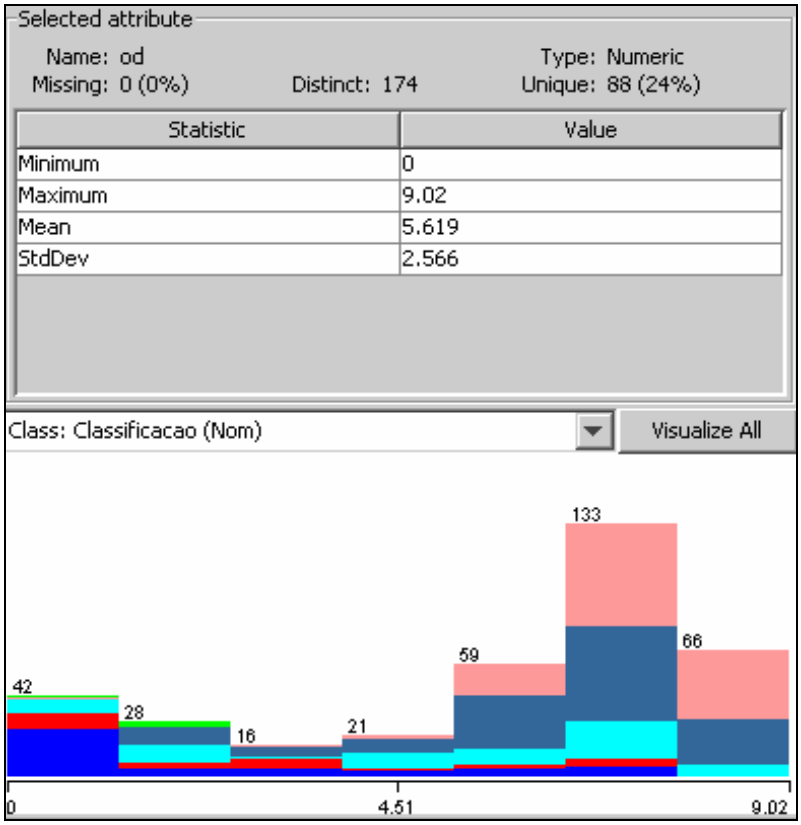

Figura 13 - Seção *"selected attribute"*

Os valores estatísticos presentes nesta janela dependem do tipo do atributo. Se ele for numérico como no exemplo da Figura 13, os valores ***minimum* (mínimo), *maximum* (máximo), *mean* (média) e *stdDev* (desvio padrão)** especificados, se for

nominal a lista conterá a lista dos valores permitidos para o atributo e suas respectivas quantidades. Abaixo dos valores estatísticos é apresentado um histograma com faixas de valores da respectiva variável e o número de casos conforme a classificação adotada pela variável dependente. O histograma e os valores estatísticos da Figura 13 correspondem ao atributo OD (oxigênio dissolvido) e sua avaliação relativo ao uso “Abastecimento”. O botão *Visualize All* mostra o histograma para todos os atributos da relação. O histograma mostra a variabilidade de cada variável no conjunto de dados especificado, neste caso a variabilidade do atributo OD (Figura 13).

3.5- ALGORITMOS DE CLASSIFICAÇÃO E SUA VALIDAÇÃO

A tarefa de classificação é a técnica de mineração de dados mais apropriada para o objetivo deste estudo, que visou, a partir da análise numérica de amostras de variáveis analíticas multivariadas, a descoberta de regras para avaliação da qualidade de água de mananciais superficiais. Segue o Quadro 5 com alguns classificadores e seus métodos de classificação disponíveis na Weka. Os classificadores do grupo *Lasy* (os algoritmos utilizados nestes classificadores são baseados no método do vizinho mais próximo), os *Trees* (os algoritmos destes classificadores utilizam árvores de decisão para criar as regras), *Rules* (criam conjunto de regras capazes de predizer ações futuras), *Bayes* (classificam utilizando redes bayesianas), *Functions* (os algoritmos implementam método de regressão linear).

Dentre os algoritmos disponíveis pelo software, selecionamos o grupo de classificadores chamado **Regras (rules)** por gerar o tipo de conhecimento mais apropriado para este trabalho. Este algoritmo gera um modelo de dados (conjunto de regras) que posteriormente permitirá a classificação dos dados do SIBAC com relação a qualidade da água.

Quadro 5-Classificadores para predições na Weka

Classificador	Grupo	Método
Weka.classifiers.Ibk	Lazy	MBR (K-vizinho mais próximo)
Weka.classifiers.j48.J48	Trees	Árvore de decisão
Weka.classifiers.PART	Rules	Árvore de decisão
Weka.classifiers.NaiveBayes	Bayes	Bayes ingênuo
Weka.classifiers.OneR	Rules	Regras
Weka.classifiers.ZeroR	Rules	Regras
Weka.classifiers.RIDOR	Rules	Regras
Weka.classifiers.JRIP	Rules	Regras
Weka.classifiers.NNGE	Rules	MBR (K-vizinho mais próximo)
Weka.classifiers.ConjuntiveRules	Rules	Regras
Weka.classifiers.Decision Table	Rules	Tabela de decisão
Weka.classifiers.MultilayerPerceptron	Functions	Regressão Linear

Dentre os classificadores disponíveis no Software utilizou-s os seguintes:

OneR

Produz regras simples baseadas em apenas um atributo, usa o atributo do mínimo-erro para a predição(**IREP- Incremental Reduced Erro Prunning,** técnica de simplificação de árvores que melhoram erros em conjuntos de dados com ruídos).

ZeroR

É o mais primitivo dos algoritmos de aprendizagem. Para as classes nominais ele prediz a classe de maior ocorrência nos dados do conjunto de treinamento. Para as classes numéricas este algoritmo prediz a média. Mesmo sendo um algoritmo com baixa performance, o seu esquema de predições pode ser usado para testar outros classificadores e servir com padrão de *benchmark.*

PART (Partial Decision Trees)

É uma variação do J48, que constrói regras a partir da árvore de decisão. O processo de geração de regras para classificação de sistemas normalmente atua em dois estágios: Regras são induzidas inicialmente e posteriormente refinadas . Isto é feito através de dois métodos, através da geração da árvore de decisão e posteriormente o mapeamento da árvore de decisão em regras aplicando processos de refinamento, ou pela utilização do paradigma dividir-para– conquistar.

FRANK E WITTEN (1999) combinam estas duas aproximações em um algoritmo chamado PART (trabalha construindo a regra e estimando sua cobertura como no processo de dividir–para–conquistar repetidamente até que todas as instâncias estejam cobertas). Constrói uma árvore de decisão parcial em cada interação e converte os ramos com a mais alta cobertura em regras.

Conjuntive Rules

Esta classe cria uma única regra conjuntiva (simples) de aprendizagem capaz de predizer para variáveis de classes numéricas e nominais. Se o exemplo do teste não for coberto por esta regra a seguir, prediz usando as distribuições da classe default dos dados não cobertos pela regra nos dados do treinamento.

JRIP (Optimizing IREP-Incremental Reduced Erro Prunning)

O IREP integra a simplificação de árvores pela redução do erro através da técnica dividir-para-conquistar. Este algoritmo possui um conjunto de regras e testa todas as regras, uma por vez. Depois que uma regra é encontrada, todos os exemplos que são cobertos por ela são deletados. Este processo é repetido até que não exista exemplos corretamente classificados, ou até que a regra encontrada pelo IREP (*Incremental* Reduced Erro Prunning) possua um erro inaceitável. Ele é uma variação do REP (Reduced Error Pruning), algoritmo de poda pela redução do erro.

Decision Table

Cria uma tabela de decisão para classificar as condições.

NNGE (Nearest Neighbor With Generalization)

Classifica utilizando o método do vizinho-mais-próximo (árvore de decisão) com o algoritmo usando generalização (poda). É um classificador no qual o aprendizado é baseado em analogia. O conjunto de treinamento é formado por vetores de *n* dimensões e cada elemento deste conjunto representa um ponto no espaço *n*-dimensional. Ele funciona de maneira que é calculado os *k*' vizinhos mais próximos de um determinado conjunto de dados, isso é feito usando métricas específicas, tal como a distância euclidiana. Para determinar a classe de um elemento que não pertença ao conjunto de treinamento, o classificador procura k elementos do conjunto de treinamento que estejam mais próximos deste elemento, que tenham a menor distância. Esses são os chamados vizinhos mais próximos. Verificam -se as classes desses k vizinhos e a classe mais freqüente será a atribuída.

RIDOR (RIpple-DOwn Rule)

Gera uma primeira regra default do erro e então as exceções para esta regra. Daí gera as "melhores" exceções para cada exceção e as itera até criar o modelo. Assim executa a árvore como a expansão das exceções. As exceções são conjuntos de regras que predizem classes à exceção do erro geradas pelo IREP (Incremental Reduced Erro Pruning).

Na Figura 14, é apresentada a tela do *Classify*, na qual o software Weka disponibiliza várias implementações de classificadores (algoritmos). Após a escolha de um deles deve-se executá-lo, clicando no botão C***hoose***, como mostra a Figura 14. Posteriormente, seleciona-se na opção ***Test Options*** as opções de teste do classificador. Existem quatro modos de teste:

> ***Training test***: Faz a predição (regras) e testa com o próprio conjunto de treinamento submetido ao classificador.
>
> ***Supplied test set***: Faz a predição (regras) e testa em outro conjunto de teste inserido pelo botão *set* pelo usuário.
>
> ***Cross-validation***: O classificador é avaliado por validação cruzada. O conjunto de teste é dividido em partes iguais e a predição é aplicada a cada um separadamente.

Percentage split: Faz a predição baseada na porcentagem dos dados que o usuário determina na própria ferramenta. A quantidade de dados capturados para teste pela ferramenta, depende do valor atribuído ao campo porcentagem.

Essas opções de teste permitem a validação da performance do classificador, se há abundância de dados o ideal é testar as regras obtidas pelo classificador em outro conjunto de dados utilizando para isso a opção conjunto de teste (*test set*). Se os dados são poucos deve-se utilizar a validação cruzada (*cross-validation*) ou o porcentagem (*percentage split*) para validar as regras obtidas.

Após a escolha da opção de teste a ser realizada pelo classificador pode-se acionar o botão ***start***. Assim tem-se como resultado no lado direito da Figura 14, a saída do classificador (*Classifier Output)* que mostra os dados estatísticos para avaliação do modelo gerado.

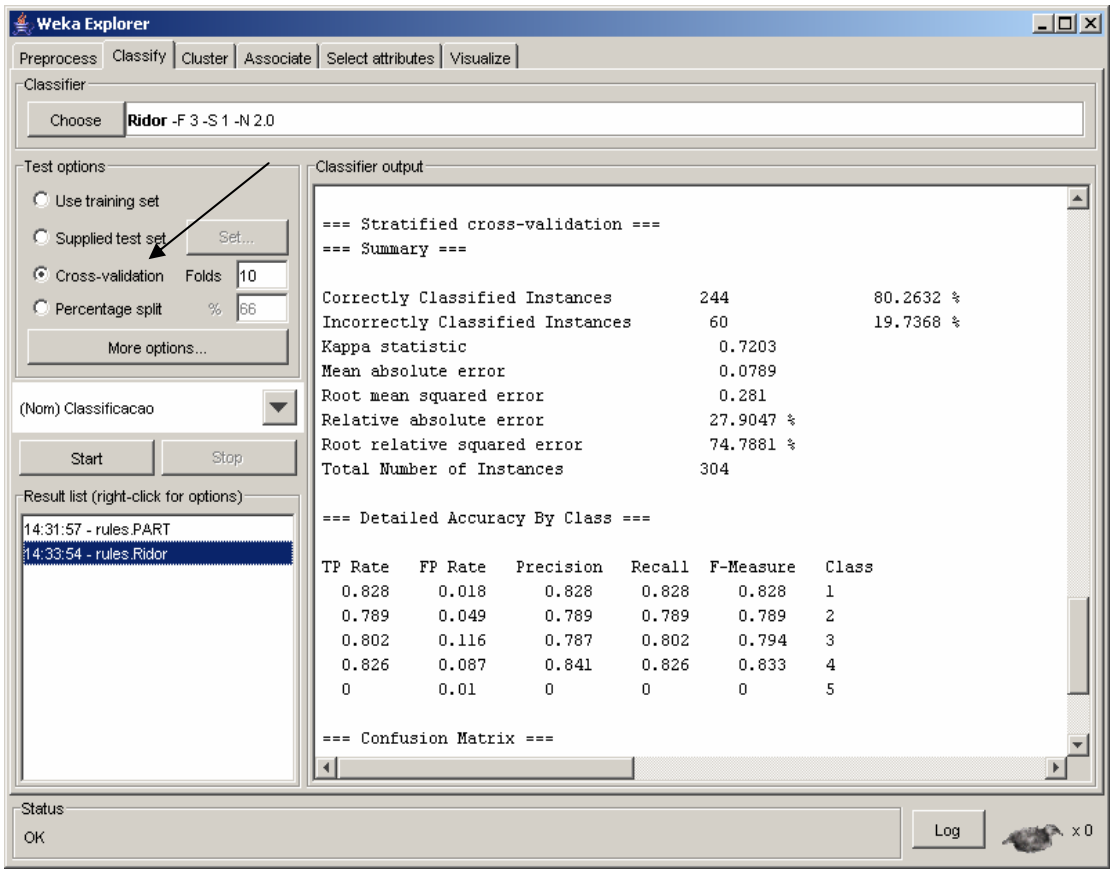

Figura 14 - Tela do Weka na guia Classify (classificadores)

3.5.1- Classifier Output Text

A interface “Classifier Output Text” pode ser visualizada com a utilização da barra de rolagem e possui as seções:

***Run information**:* contém todas as informações do esquema de aprendizagem, nome da relação, instâncias, atributos e modo de teste envolvido no processo (Figura 15).

Classifier model (full training test): representação textual das regras produzidas para o conjunto de treinamento inteiro, ou seja, o modelo de classificação gerado para a amostra (Figura 16).

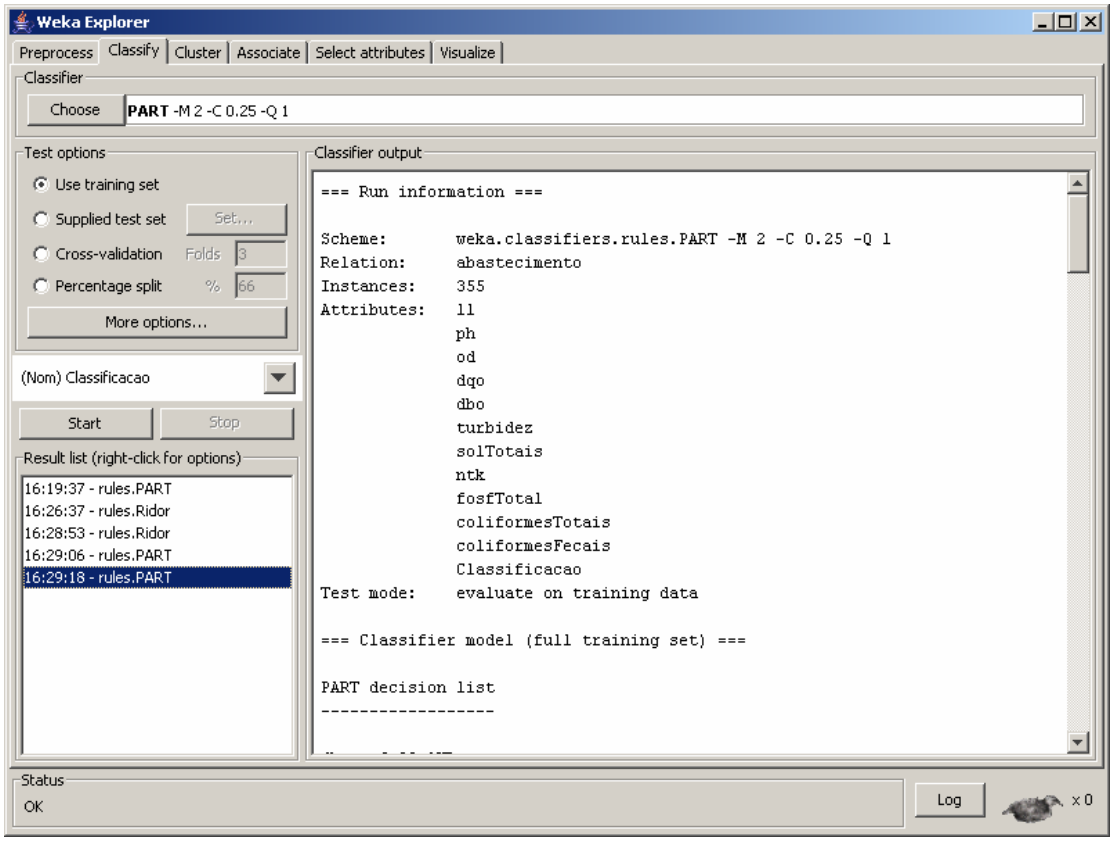

Figura 15 - Seção Run Information do Classifier Output Text

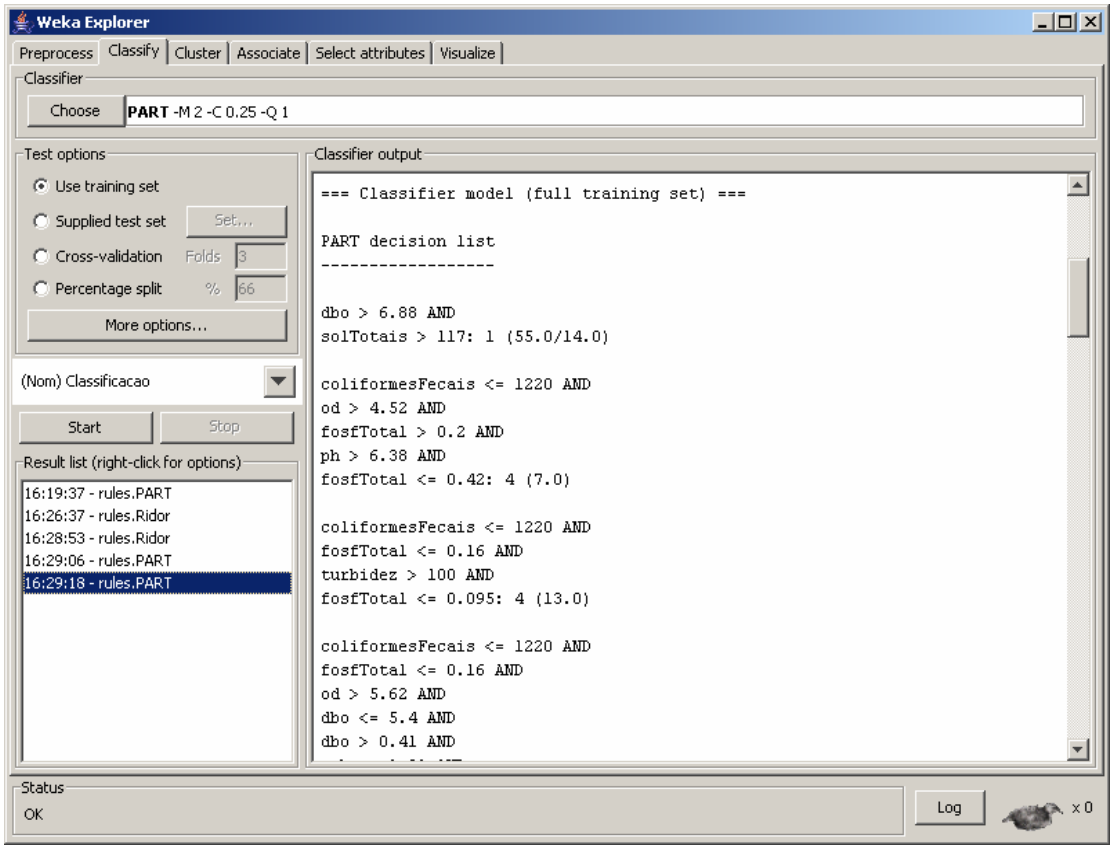

Figura 16 - Seção Classifier Model do Classifier Output Text com as regras

Os resultados do modo de teste podem ser desmembrados em:

Summary*:* lista de resumo dos valores estatísticos que mostram como aquele classificador está habilitado a predizer as classes corretas de acordo com aquele modo de teste (Figura 17).

Detailed accuracy by class*:* acurácia da predição do classificador detalhada por classe, (Figura 18).

Confusion matrix: Visualização da matriz de confusão; mostra como as instâncias foram classificadas em cada classe, (Figura 18).

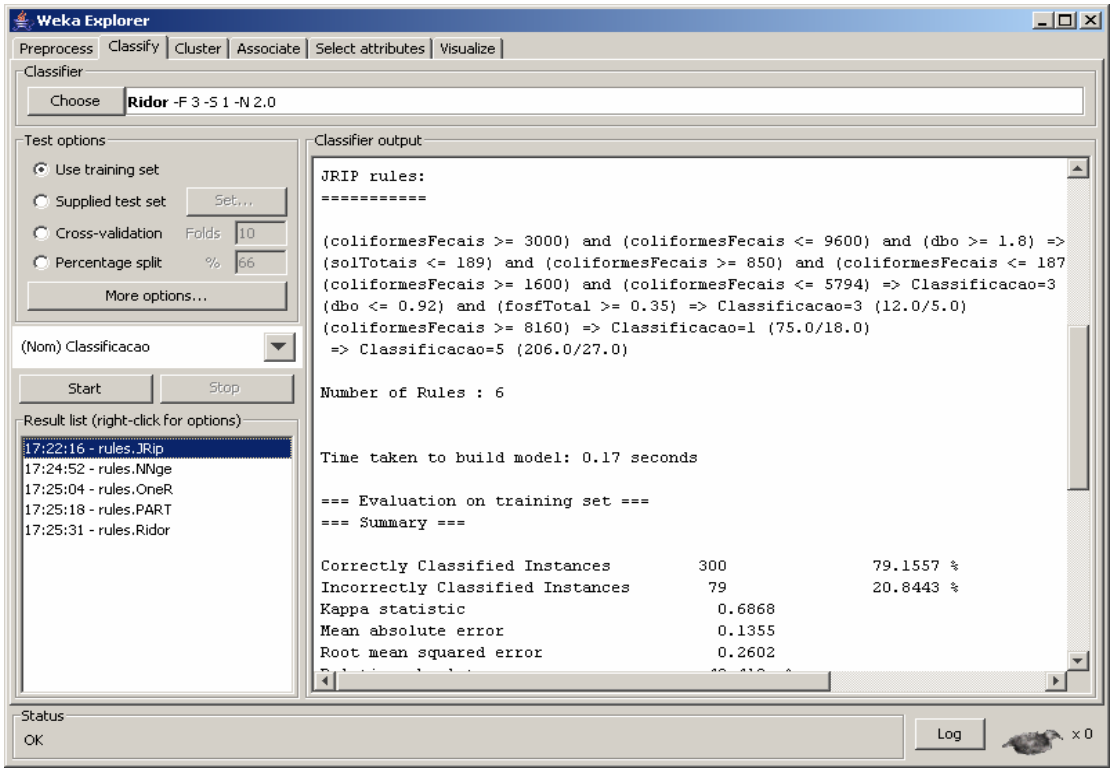

Figura 17 - Regras geradas pelo algoritmo JRIP e porcentagem de instâncias classificadas corretamente.

Alguns pontos de avaliação no resultado:

A primeira linha do *Classifier Output* indica o algoritmo usado e as regras geradas pelo conjunto de treinamento (modelo classificador escolhido no modo de teste), neste exemplo o algoritmo utilizado foi o JRIP, Figura 17. Ainda na Figura 17 tem-se o **número de regras geradas (6 regras)** e o **tempo que a ferramenta levou para construir o modelo** (0.17 segundos), eficiência do classificador.

A seguir tem-se a performance do classificador, no item avaliação pelo conjunto de treinamento (Porcentagem de Instâncias Corretas e Incorretas), primeiro treina-se o algoritmo com os dados (gera as regras) e depois verifica-se a performance com os mesmos dados de treinamento (opção de teste), Figura 17. Existem outras possibilidades de teste dessas regras, tais como conjunto de teste (com outro conjunto de dados), validação cruzada e porcentagem *split* etc. A performance desta opção de teste é mostrada nos itens Classificação de Instâncias Corretas e Incorretas onde está a porcentagem de cada uma. Para validar o

classificador é melhor utilizar conjunto de teste validação cruzada ou porcentagem *split.*

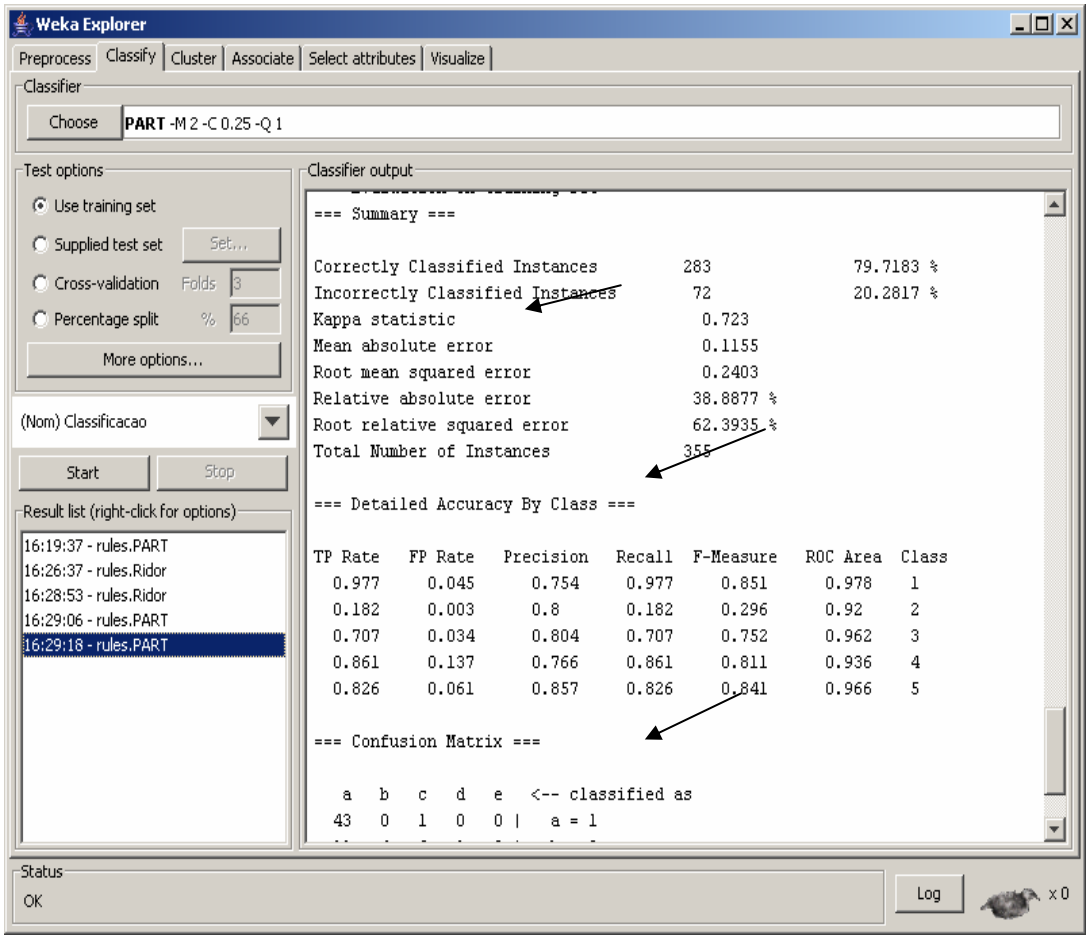

Figura 18 - Seção summary, detailed accuracy by class e confusion matrix do classifier output text

Ainda na tela de saída do resultado é apresentada a **matriz de confusão (Confusion Matrix)** que indica que instâncias foram classificadas de forma correta e incorreta por classe. Importante observar a taxa de falsos positivos na matriz de confusão. Se houve 100% de classificação correta podemos esperar uma matriz de confusão onde todo elemento fora das diagonais é igual a zero.

O Kappa Statistic é um índice que compara o valor encontrado nas observações com aquele que se pode esperar do acaso. É o valor calculado dos resultados encontrados na observações e relatado como um decimal (0 a 1). Quanto menor o valor de kappa menor a confiança da observação, o valor 1 implica a

correlação perfeita, difícil de ser encontrada. Para ser boa uma observação, com 95% confiável, o valor de kappa deve estar no intervalo (0.279, 0.805).

Root Mean-Squared Error é muito usado para medir o sucesso de uma predição numérica.. Este valor é calculado pela média da raiz quadrada da diferença entre o valor calculado e o valor correto. O root mean-squared error é simplesmente a raiz quadrada do mean-squared-error (dá o valor do erro entre os valores atuais e os valores preditos, Figura 18.

Mean Absolute Error: média da diferença entre os valores atuais e os preditos em todos os casos, é a média do erro da predição, Figura 18.

Root Relative Squared Error: reduz o quadrado do erro relativo na mesma dimensão da quantidade sendo predita incluindo raiz quadrada. Assim como a raiz quadrada do erro significativo(root mean-squared error), este exagera nos casos em que o erro da predição foi significativamente maior do que o erro significativo, Figura 18.

Relative Absolute Error: É o erro total absoluto. Em todas as mensurações de erro, valores mais baixos significam maior precisão do modelo, com o valor próximo de zero temos o modelo estatisticamente perfeito, Figura 18.

True Positives (TP): são os valores classificados verdadeiramente positivos.

False Positives (FP): são os falsos positivos, são os dados classificados erroneamente como positivos pelo classificador.

Precisão (precision): É o valor da predição positiva (número de casos positivos por total de casos cobertos), muito influenciada pela especificidade e pouco pela sensitividade. Sensitividade é o número de casos positivos que são verdadeiramente positivos e especificidade é o número de casos negativos que são verdadeiramente negativos.

Recall (Cobertura): É o valor da cobertura de casos muito influenciada pela sensitividade e pouco pela especificidade. É calculada por número de casos cobertos pelo número total de casos aplicáveis.

F-measure: Usada para mensurar a performance pois combina valores de cobertura e precisão de uma regra numa única fórmula. É calculada por 2TP/2TP+FP+FN, onde TP (true positives) são os verdadeiros positivos, FP (False

positives) são os falsos positivos, FN (false negative) são os falsos negativos ou (2*recall*precision/recall+precision)

ROC Curve (Curva ROC): A curva ROC plota os números de positivos no eixo vertical(x), e os expressa como uma porcentagem do total do número de casos positivos, contra o número de casos negativos incluídos no exemplo, expressos como uma porcentagem do total de casos negativos no eixo horizontal. Uma curva ROC perfeita corresponde a uma linha horizontal no topo do gráfico, ponto (100,100) que corresponde a 100% sensitividade contra o ponto (0,0) no eixo horizontal que corresponde a 0% de especificidade. Curvas abaixo desses pontos indicam métodos de decisão menos perfeitos, mas qualquer curva situada acima da reta diagonal que atravessa o gráfico entre os pontos [0,0] e [100,100] pode ser considerada como boa, Figura 19.

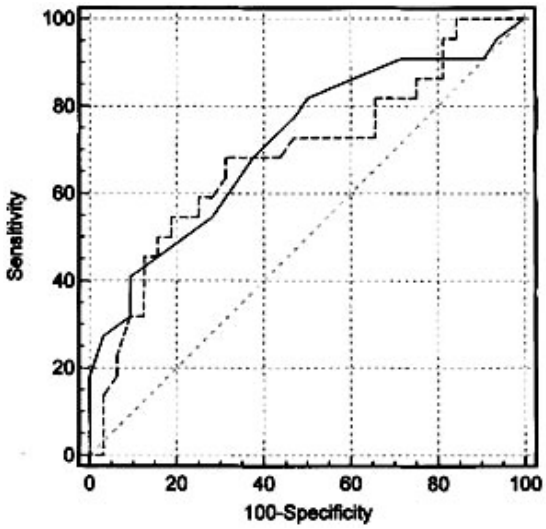

Figura 19 - Curva ROC. (Fonte:www.scielo.cl/fbpe/img/rmc/v134n3/fig06-02.jpg)

4. RESULTADOS

A apresentação dos resultados é dividida em quatro sub-items: Pré-processamento (4.1), Mineração dos dados (4.2), Análise dos Resultados e Validação (4.3). É exposta a performance dos algoritmos para classificação dos registros de qualidade de água proveniente do SIBAC referente a sua aptidão para quatro formas de uso: Abastecimento Doméstico, Irrigação, Manutenção dos Ciclos Biogeoquímicos Naturais e Balneabilidade.

4.1 PRÉ-PROCESSAMENTO

Foram testadas quatro amostras de dados, classificadas pelos especialistas de acordo com a sua aptidão qualitativa para cada um dos quatro tipos de uso de água. Cada amostra contém 11 variáveis (ou atributos), sendo uma delas o atributo “meta” chamado classificação que representa a avaliação das amostras.

Inicialmente os especialistas foram solicitados para classificar um conjunto de 300 registros de acordo com cada tipo de uso. Este volume, entretanto, foi considerado muito extenso pelos especialistas, resultando em um retorno limitado de registros classificados. Foram recebidos os seguintes números de registros classificados (somatório dos registros obtidos dos quatro especialistas):

1. Amostra de Abastecimento com 366 registros de dados;
2. Amostra de Balneabilidade com 351 registros de dados;
3. Amostra de Irrigação com 379 registros de dados;
4. Amostra de Manutenção dos Ciclos Biogeoquímicos com 114 registros de dados.

As Figuras 20 a 23 mostram os histogramas das variáveis de qualidade de água para cada amostra, construídos no pré-processamento na Weka.

Observando os histogramas dos usos de água verifica-se a disposição dos registros classificados pelos especialistas ainda na fase de pré-processamento. As cores vermelho, laranja, amarelo, verde e azul presentes no histograma da variável classificação nas Figuras 20 a 23 representam, respectivamente, as classes péssimo, ruim, regular, bom e ótimo na classificação dos especialistas. Assim é possível observar nas figuras a distribuição das classes nas barras do histograma de cada variável (pH, OD, TURB, P, DQO, ST, NTK, DBO, CF e CT).

Observa-se que existem quantidades de registros diferentes para cada classe. O ideal seria que o número de classificações realizadas pelos especialistas em cada amostra(presente nas Figuras 20 a 23) estivesse uniforme, isto para não direcionar o aprendizado das regras, uma vez que alguns algoritmos tem como critério classificar pelo atributo majoritário da amostra.

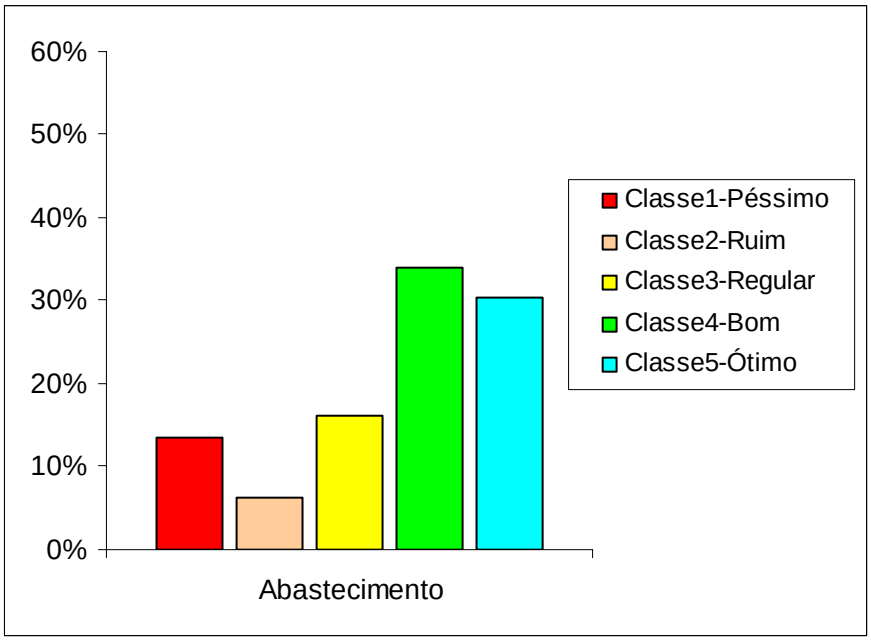

Figura 20-Histograma das variáveis da amostra de abastecimento (n:366)

Verifica-se na amostra de “Abastecimento”, a maior quantidade de registros para a classe bom e ótimo.

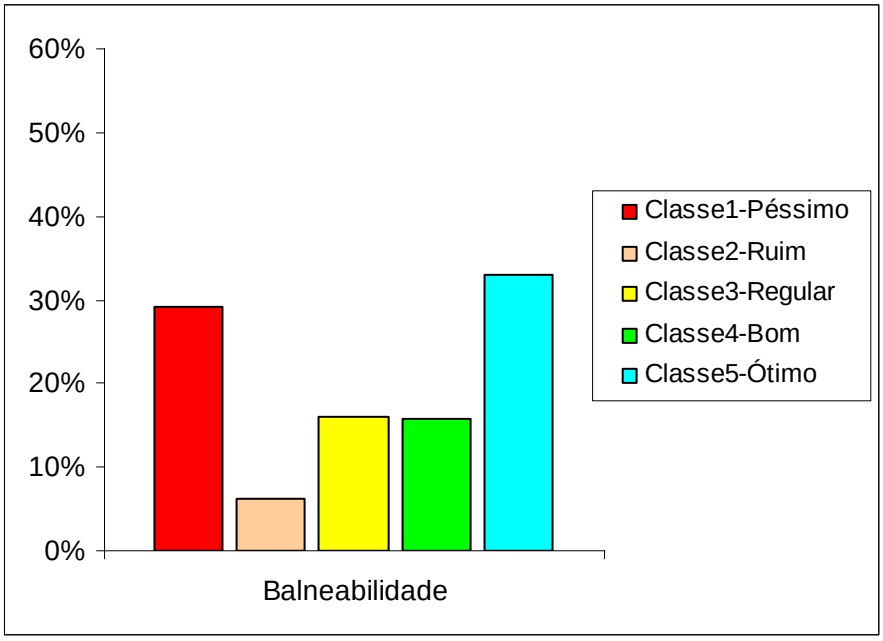

Figura 21-Histograma das variáveis da amostra de Balneabilidade (n:351)

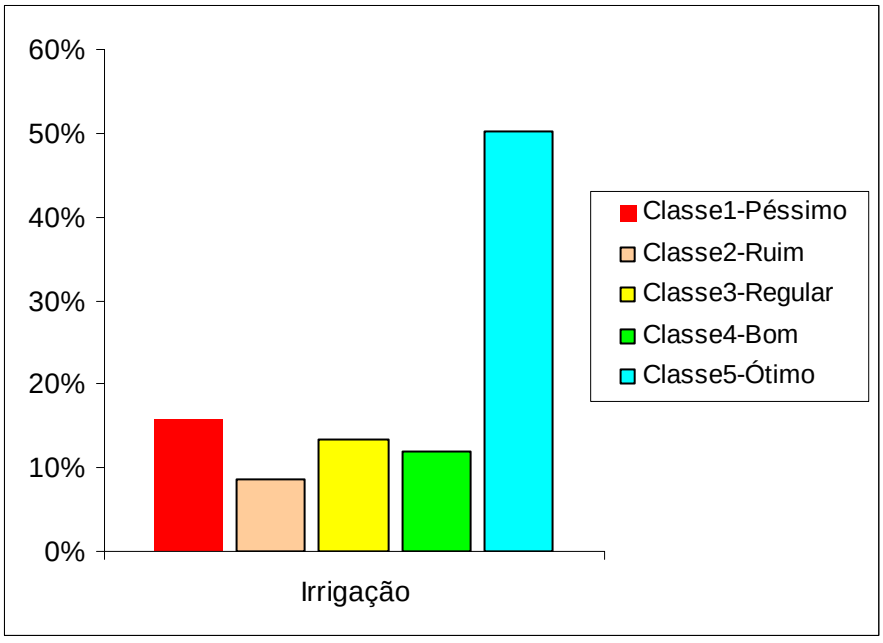

Figura 22-Histograma das variáveis da amostra de irrigação (n:379)

Nas amostras de “Balneabilidade” e “Irrigação” o número maior de registros estão nas classes péssimo e ótimo.

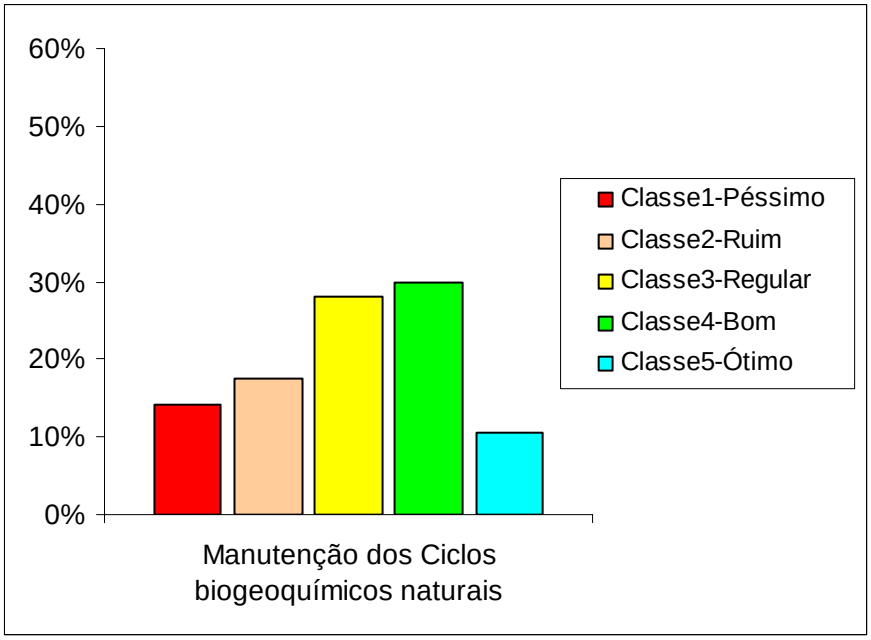

Figura 23-Histograma das variáveis da amostra de Manutenção dos Ciclos Biogeoquímicos Naturais (n:114)

Na amostra de "Manutenção dos Ciclos Biogeoquímicos" tem-se o maior número de registros nas classes regular e bom.

Verifica-se que o número de registros classificados nas classes intermediárias ruim (laranja), regular (amarelo) e bom (verde) tem similaridade na distribuição das quantidades presentes em todas as amostras (Figuras 20 a 23). Isso é relevante para não direcionar o aprendizado do algoritmo e consequentemente o resultado das regras de classificação.

Pode-se acrescentar ainda que a disposição diferenciada dos registros classificados nas quatro amostras (Balneabilidade, Irrigação, Abastecimento e Manutenção dos Ciclos Biogeoquímicos) deve-se aos:

1. limites impostos pelo CONAMA para cada uso, associado a subjetividade da classificação utilizada pelos especialistas; e,
2. aos trechos selecionados no SIBAC que envolvem nascentes(sem poluição), trechos de média poluição e trechos poluídos.

Observa-se na Tabela 2 as freqüências das classes avaliadas para cada amostra (uso), de acordo com a classificação dos especialistas.

Tabela 2-Tabela comparativa da freqüência de classificação das classes

Amostra	Classe1 (Péssimo)	Classe2 (Ruim)	Classe3 (Regular)	Classe4 (Bom)	Classe5 (Ótimo)	Total de instâncias
Abastecimento	49	23	59	**124**	111	366
Balneabilidade	102	22	56	55	**116**	351
Irrigação	60	33	51	45	**190**	379
Manutenção dos Ciclos Biogeoquímicos	16	20	32	**34**	12	114

A seguir são apresentados histogramas das classificações dos três especialistas por tipo de uso (Figuras 24 até 27).

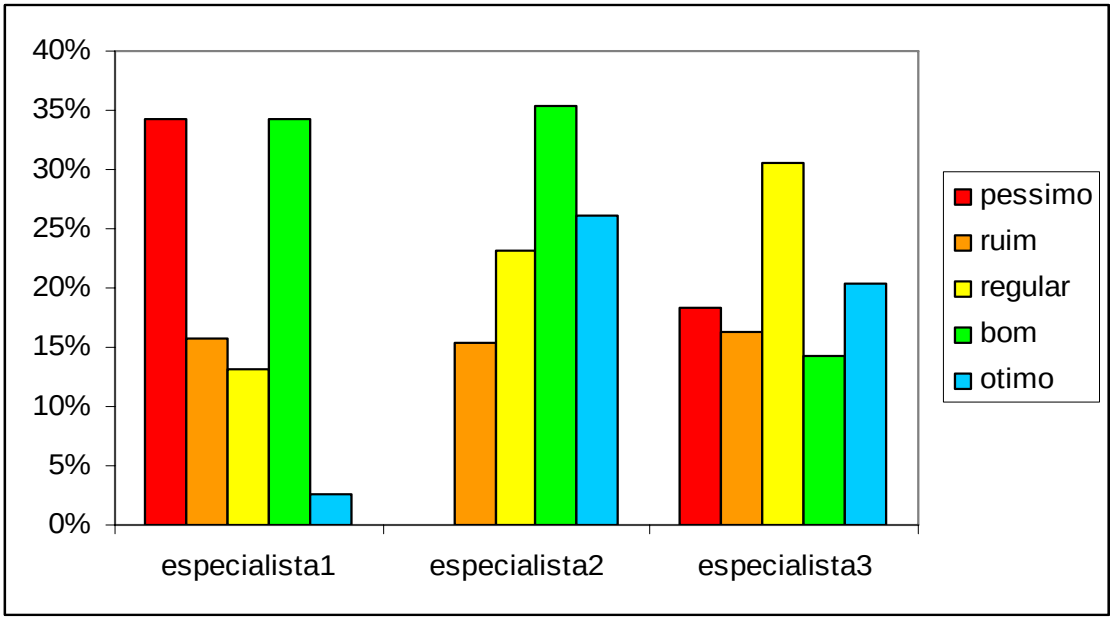

Figura 24-Histograma da amostra abastecimento de acordo com a classificação dos especialistas (n: 50)

Observa-se uma discordância maior na classificação realizada pelos especialistas em relação a classe 1 (péssimo) e classe 5 (ótimo) da amostra “Abastecimento”. Nas demais classes dessa amostra têm-se uma regularidade nos registros.

Na amostra “Balneabilidade” verifica-se discordâncias entre os especialistas nos valores de classificação onde a qualidade é péssima (1), ruim (2) e ótima (5), Figura 25.

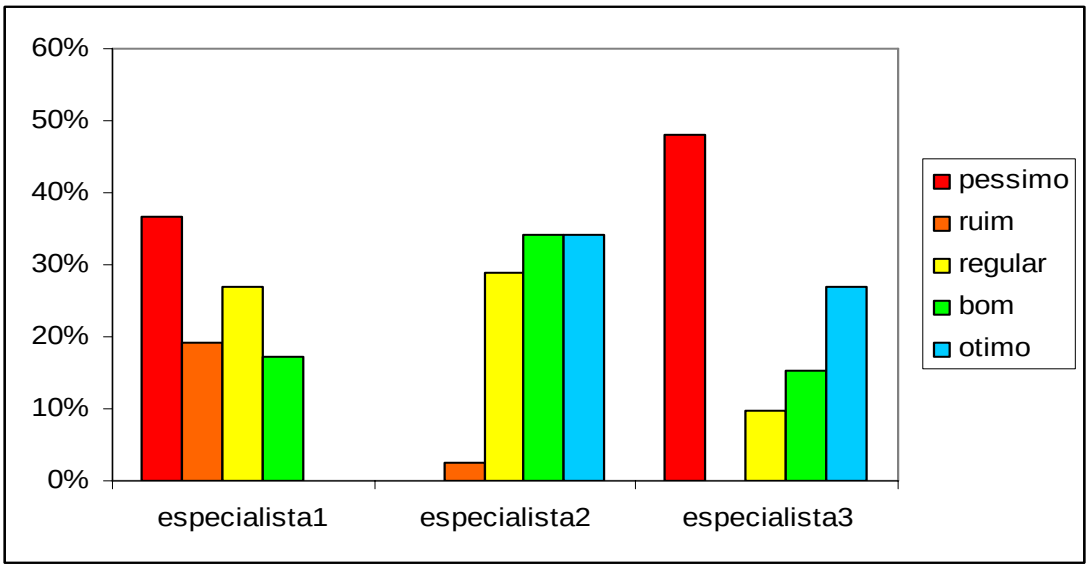

Figura 25-Histograma da amostra balneabilidade de acordo com a classificação dos especialistas (n: 50)

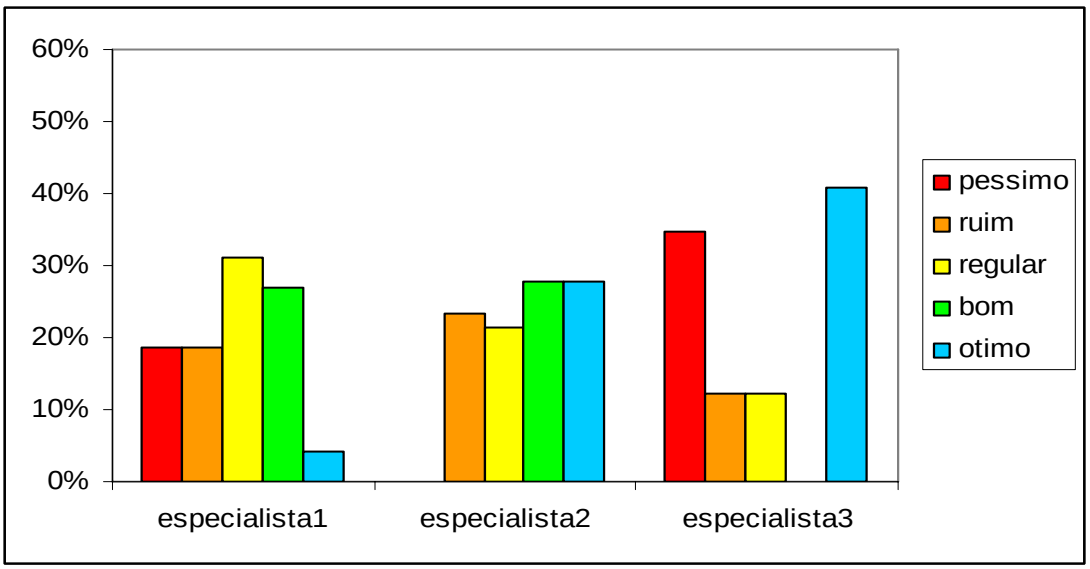

Figura 26-Histograma da amostra irrigação de acordo com a classificação dos especialistas (n: 50)

Na amostra de “Irrigação” tem-se as maiores discordâncias entre os especialistas nas classes 1 (péssimo), 2 (bom) e 5 (ótimo), Figura 26.

Para a amostra de “Manutenção dos Ciclos Biogeoquímicos Naturais” tem-se discordâncias nas classificações realizadas nas classes 1 (péssimo), e 2 (ruim), e 5 (ótimo), conforme Figura 27.

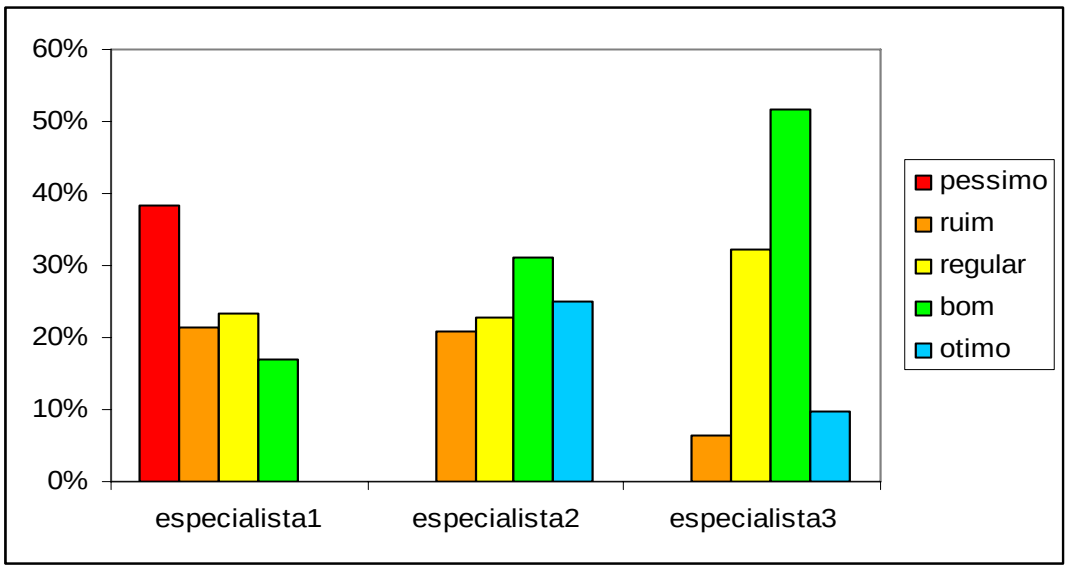

Figura 27-Histograma da amostra Manutenção dos ciclos biogeoquímicos naturais de acordo com a classificação dos especialistas (n: 50)

Analisando as amostras, observa-se que os especialistas se divergiram em relação as classes 1 (péssimo) e 5 (ótimo), isto mostra a incerteza em classificar os extremos.

Por se tratar de um critério subjetivo, onde o especialista é quem decide o que é ótimo, bom, regular, ruim e péssimo baseado em 10 valores de variávei,s é aceitável que haja concordâncias e discordâncias na classificação.

4.2- MINERAÇÃO DE DADOS

4.2.1- Visão geral dos classificadores

Em seguida a esta análise exploratória das variáveis de qualidade de água e sua classificação, foram realizados os procedimentos de mineração dos dados, que se constituiu em um comparativo empírico de oito algoritmos de classificação do tipo “rules”, sendo eles:

- *ZeroR*
- *OneR*
- *Ridor*
- *NNge*
- *JRIP*
- *Decision Table*
- *PART*
- *ConjuntiveRule*

As Figuras 28 até 30 e Tabelas 5 até 12 mostram o desempenho comparativo dos algoritmos de classificação, cada um validado por três opções de teste: *treinamento, validação cruzada* e *porcentagem split.*

No conjunto de treinamento todas as amostras são usadas para testar as regras do classificador, fato este que comprova o bom desempenho de todos os algoritmos para esta opção de teste. As regras são testadas nos mesmos dados que a formaram, não é boa opção para medir a performance,WITTEN & FRANK (2000).

Na k- validação cruzada (*cross-validation-10 folds*) a amostra é dividida em k partes de igual tamanho, de preferência as partes (ou *folds*) devem possuir a mesma quantidade de padrões, garantindo a mesma proporção de classes para cada subconjunto. O algoritmo é executado sob k-1 "folds"(subconjuntos) gerando as regras, posteriormente ele é validado sob o *fold* que sobrou. Assim neste caso onde usamos 10 partes (10 *folds*), o conjunto de treinamento é criado sob 9 (nove) partes e testado calculando a taxa de acerto sob os dados da parte não utilizada, a parte que sobrou. Ao final a taxa de acerto é uma média das taxas de acerto nas k iterações realizadas. É uma boa opção para medir a performance dos algoritmos e dos erros, WITTEN & FRANK (2000).

Na opção porcentagem *split* separa-se uma porcentagem dos dados para testar as regras do classificador, neste caso selecionou-se 66% das amostras.

Neste trabalho ilustra-se nas Figuras 28 a 34 a opção de teste conjunto de treinamento, mas não a utiliza para validar os algoritmos devido a performance otimista que apresenta.

Existem outras medidas de avaliação de resultados disponíveis pela ferramenta, apresentaremos um resumo com algumas métricas para cada uso. Neste trabalho utiliza-se as medidas de predições numéricas somente para os algoritmos que obtiveram a melhor porcentagem de acerto, fez-se uma pré-seleção.

4.2.2- Resultado do uso “Abastecimento”

Para a amostra “Abastecimento” tem-se a porcentagem correta dos classificadores para as três opções de teste na Figura 28.

Ao observar o número de instâncias corretas para o teste de validação cruzada destacou-se o algoritmo PART, Tabela 3.

A taxa de erro ou porcentagem de instâncias incorretas é de muita relevância para avaliar a aproximação do modelo gerado e a função que governa o fenômeno em estudo. Erro de treinamento pode significar muitos ruídos e valores vazios nas amostras. Para minimizar este problema deve-se procurar algoritmos que minimizem o problema.

A melhor porcentagem de erro confirma-se no algoritmo PART, modo de teste validação cruzada. Verifica-se que os algoritmos que possuem altas porcentagens de erro para o conjunto de treinamento podem indicar ruídos nas amostras.

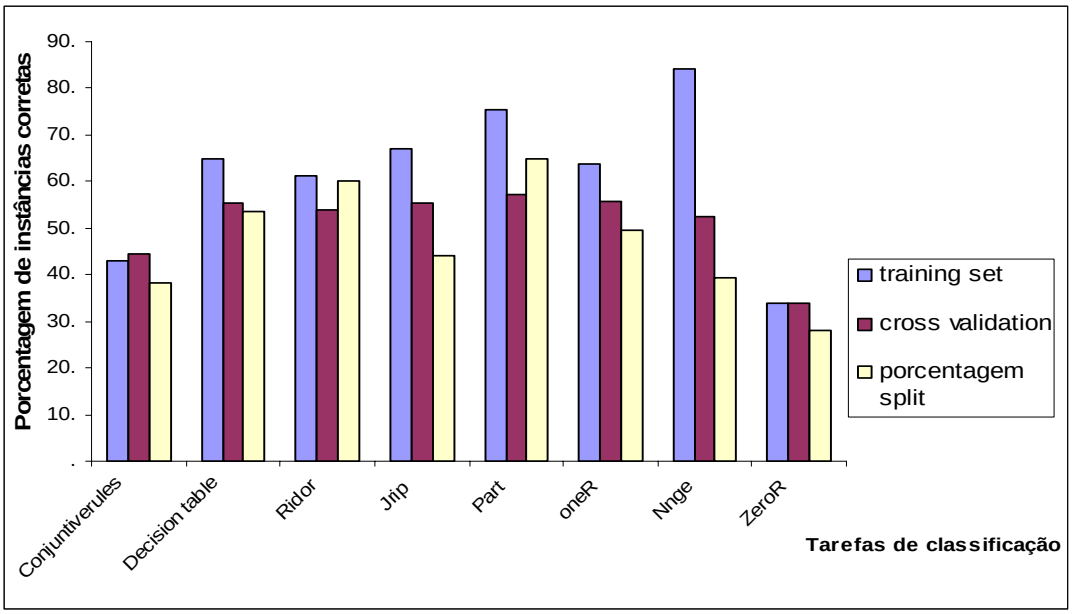

Figura 28-Desempenho dos classificadores para classificação da amostra " Abastecimento"

Tabela 3- Desempenho dos classificadores para a amostra "Abastecimento".

Porcentagem de instâncias corretas - 366 instancias			
	training set	**cross validation**	**porcentagem split**
Conjuntive rules	43.16	44.53	38.4
Decision table	65.02	55.46	53.6
Ridor	61.2	53.82	60
Jrip	66.93	55.46	44
Part	75.4	**57.1**	64.8
OneR	63.66	55.73	49.6
Nnge	84.15	52.45	39.2
ZeroR	33.87	33.87	28

O valor de *Kappa Statistic* é um índice que mede a confiança na observação, valores de *Kappa* acima de 0.6 são considerados significativos, LANDIS & KOCH (1977). Para a amostra de abastecimento observa-se os melhores valores no algoritmo PART com valor 0,42, Figura 29.

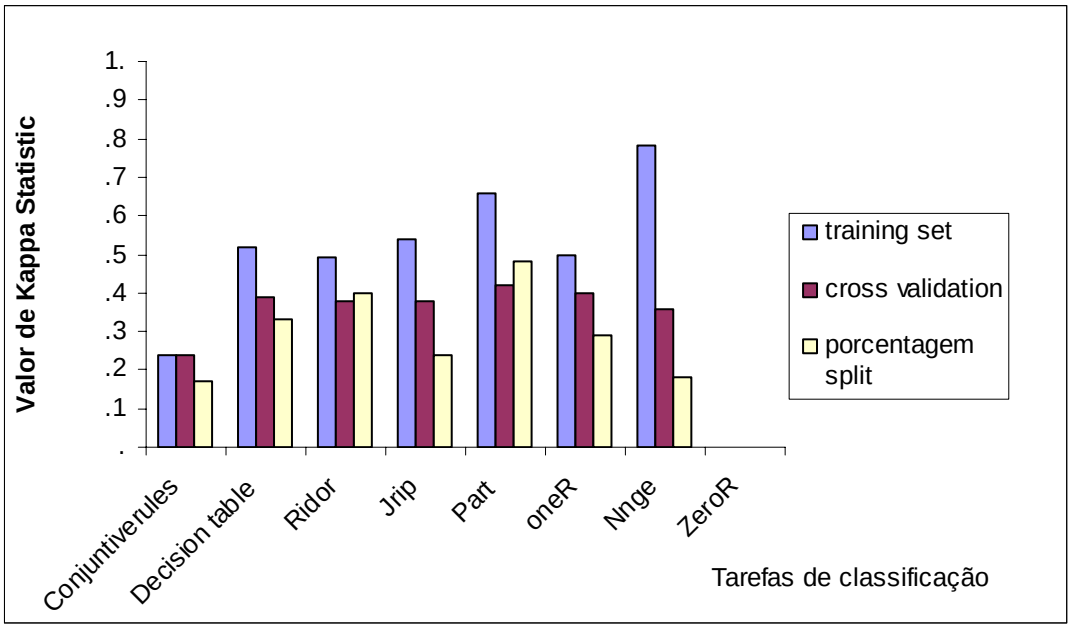

Figura 29-Valores de Kappa Statistic para amostra "Abastecimento"

Pode-se observar além das porcentagens corretas, incorretas e valor de *kappa* destes algoritmos para o uso de abastecimento, os itens: TP (*True Positive*), FP (*False Positive*), Precisão (*Precision*), Cobertura (*Recall*), *F-measure*, Curva *ROC* e Matriz de Confusão, Figura 30.

=== Summary ===		
Correctly Classified Instances	209	57.1038 %
Incorrectly Classified Instances	157	42.8962 %
Kappa statistic	0.42	
Total Number of Instances	**366**	

=== Detailed Accuracy By Class ===						
TP	FP	Precision	Recall	F-Measure	ROC Area	**Class**
0.816	0.057	**0.69**	**0.816**	**0.748**	**0.91**	**1**
0.087	0.038	0.133	0.087	0.105	0.74	**2**
0.373	0.121	0.373	0.373	0.373	0.67	**3**

0.573	0.244	**0.546**	**0.573**	**0.559**	**0.72**	**4**
0.667	0.118	**0.712**	**0.667**	**0.688**	**0.86**	**5**

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
43	1	5	0	0	**a = 1**
9	**5**	8	1	0	**b = 2**
1	0	**45**	12	1	**c = 3**
0	0	5	**110**	9	**d = 4**
0	0	0	38	**73**	**e = 5**

Figura 30 - Métricas de avaliação do algoritmo PART na amostra "Abastecimento" (Validação Cruzada)

True positives **(TP)**: são os valores verdadeiramente positivos classificados para cada classe (Figura 30). A classe onde esta taxa foi mais baixa, ou seja, onde houve menos classificações verdadeiramente positivas foi a classe 2. Na matriz de confusão verifica-se apenas 2 classificações de TP .

False positives **(FP)**: os falsos positivos são os dados classificados erroneamente como positivos pelo classificador para cada classe, neste índice verifica-se o maior problema na classe 4, verificando a matriz de confusão Figura 30.

Precision **(Precisão*)***: é o valor do número de casos positivos por total de casos cobertos, muito influenciada pela especificidade. É calculada por TP/TP+FP, Figura 30. Observa-se a maior precisão na classe 1 onde há maior número de casos TP e a pior precisão na classe 2 onde há muitos casos classificados erroneamente.

Recall **(Cobertura):** É o valor muito influenciado pela sensibilidade e pouco pela especificidade. É calculada por número de casos cobertos pelo número total de casos aplicáveis, Figura 30. Verifica-se a maior cobertura para classe 1.

F-measure**:** Usada para mensurar a performance pois combina valores de cobertura e precisão de uma regra numa única fórmula. É calculada por 2TP/2TP+FP+FN, ou (2*recall*precision/recall+precision), Figura 30. Para classe 3 é a pior e para a classe 1 é melhor.

ROC curve **(Curva ROC):** Qualquer curva situada acima da reta diagonal que atravessa o gráfico entre os pontos [0,0] e [100,100] pode ser considerada boa, assim

observando a Figura 30 verifica-se que todos os valores de ROC de cada classe estão dentro do intervalo mencionado.

As regras geradas pelo algoritmo PART para classificação (em negrito) a amostra de Abastecimento são expostas na Figura 31.

Scheme: **weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1**

Relation: **Abastecimento**

Instances: **366**

Attributes: **11**

Test mode: **10-fold cross-validation**

Classifier model (full training set)➔ **PART decision list**

dbo > 6.88 AND solTotais > 117 AND coliformesFecais > 63000: **1** (53.0/10.0)

coliformesFecais <= 1220 AND fosfTotal > 0.16 AND coliformesFecais > 32.8 AND ntk <= 1.68 AND dbo > 0.87 AND fosfTotal <= 0.25: **4** (9.0)

fosfTotal <= 0.16 AND coliformesFecais <= 1220 AND turbidez > 100 AND fosfTotal <= 0.095: **4** (13.0)

fosfTotal <= 0.16 AND turbidez > 100 AND coliformesTotais <= 73287 AND fosfTotal > 0.07 AND dqo > 19 AND dqo > 24: **4** (5.0/2.0)

coliformesFecais <= 1600 AND turbidez > 96 AND dqo > 18: **3** (7.0/1.0)

coliformesFecais <= 1600 AND fosfTotal <= 0.16 AND od > 5.62 AND dqo > 2.5 AND dbo <= 5.4 AND turbidez > 4.66 AND fosfTotal <= 0.1: **5** (79.0/9.0)

coliformesFecais <= 2950 AND fosfTotal > 0.16 AND coliformesFecais > 32.8 AND coliformesTotais <= 51720: **4** (18.0/7.0)

fosfTotal > 0.2 AND dbo <= 5.4 AND dbo <= 1.32: **3** (8.0/1.0)

fosfTotal <= 0.2 AND coliformesFecais <= 2950 AND od > 4.02 AND dqo > 2.5: **4** (88.0/38.0)

turbidez <= 26 AND dbo > 0.36 AND ph > 3.795 AND fosfTotal <= 0.29 AND ph > 5.13 AND coliformesTotais > 4130 AND solTotais > 77 AND od <= 6.44: **4** (12.0/3.0)

turbidez <= 4.2 AND turbidez > 1.7 AND coliformesFecais <= 110 AND fosfTotal > 0.02: **4** (11.0/1.0)

coliformesFecais <= 3000 AND turbidez > 4.2 AND od > 4.02: **3** (6.0)
turbidez <= 4.8 AND dbo <= 0.36: **3** (4.0)
coliformesTotais <= 9804 AND turbidez <= 4.66: **5** (4.0/1.0)
od <= 4.52 AND coliformesFecais <= 6300: **3** (13.0/6.0)
od > 4.02 AND coliformesFecais > 5470: **3** (25.0/10.0)
coliformesFecais > 5830: 2 (6.0/1.0): **4** (5.0)
Number of Rules : 18

Figura 31-Regras geradas pelo algoritmo PART amostra Abastecimento (classes em negrito).

4.2.3- Resultado do uso "Balneabilidade "

Para a amostra "Balneabilidade" tem-se a porcentagem correta dos classificadores para as três opções de teste na Figura 32.

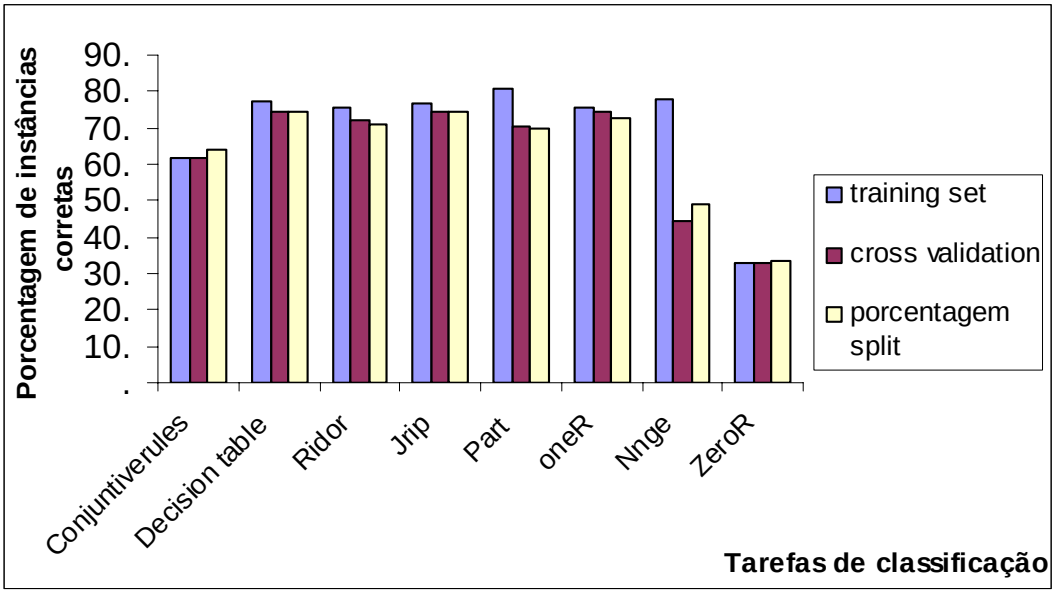

Figura 32- Desempenho dos classificadores para classificação da amostra "Balneabilidade".

Ao observar o número de instâncias corretas para o teste de validação cruzada destacou-se o algoritmo JRIP, Tabela 4.

Tabela 4-Porcentagens corretas dos algoritmos de classificação para Balneabilidade

Porcentagem de instâncias Corretas - 351 instâncias			
	training set	**cross validation**	**porcentagem split**
Conjuntiverules	61.82	61.82	64.17
Decision table	77.2	74.25	74.06
Ridor	75.49	72.36	70.83
Jrip	76.63	**74.35**	74.16
Part	80.6	70.37	70
OneR	75.44	74.35	72.5
Nnge	78.06	44.4	49.16
ZeroR	33.04	33.04	33.3

O valor de *kappa* da amostra de Balneabilidade no algoritmo Jrip é 0.64, valores acima de 0.6 indicam confiabilidade na relação conforme Figura 33.

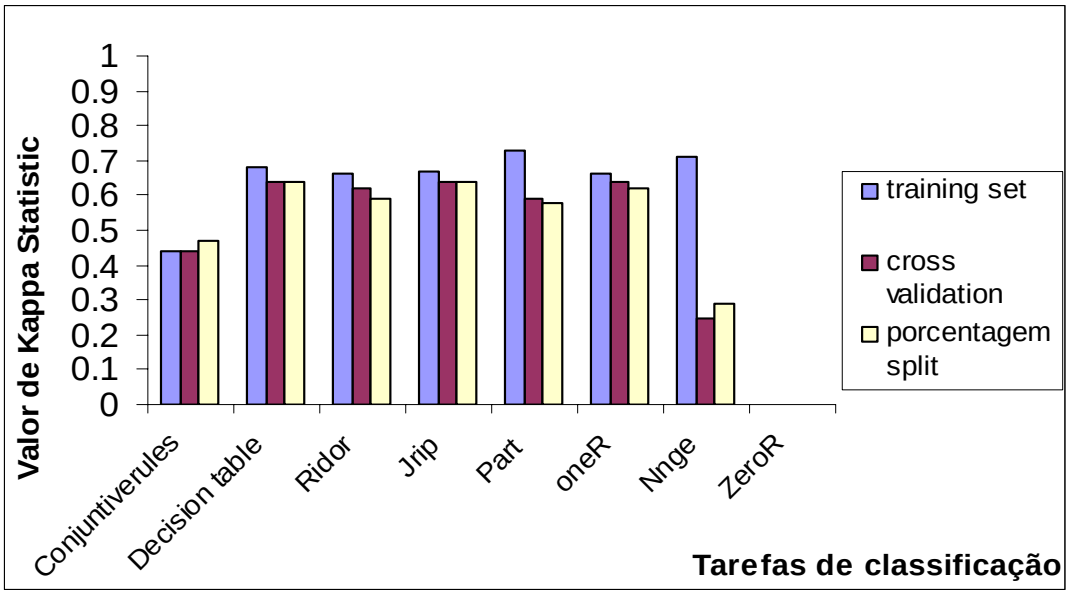

Figura 33- Valor de Kappa Statistic dos algoritmos para amostra de "Balneabilidade"

Com relação a acurácia (soma de positivos verdadeiros e negativos verdadeiros) por classe detectou-se os maiores problemas na classe 2, Figura 34.

=== Summary ===						
Correctly Classified Instances	261	74.359 %				
Incorrectly Classified Instances	90	25.641 %				
Kappa statistic	0.647					
Total Number of Instances	**351**					
=== Detailed Accuracy By Class ===						
TP	FP	Precision	Recall	F-Measure	ROC Area	**Class**
0.931	**0.133**	**0.742**	**0.931**	**0.826**	**0.90**	**1**
0	0	0	0	0	0.77	**2**
0.375	0.037	0.656	0.375	0.477	0.76	**3**
0.618	**0.068**	**0.63**	**0.618**	**0.624**	**0.85**	**4**
0.957	**0.111**	**0.81**	**0.957**	**0.877**	**0.93**	**5**
=== Confusion Matrix ===						
a b c d e <-- classified as						
95 0 5 1 1 | a = 1						
18 **0** 0 0 4 | b = 2						
11 0 **21** 15 9 | c = 3						
4 0 5 **34** 12 | d = 4						
0 0 1 4 **111** | e = 5						

Figura 34- Métrica de avaliação do algoritmo Jrip para a amostra "Balneabilidade" (Validação Cruzada)

As regras geradas pelo algoritmo JRIP são, Figura 35:

Scheme: weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1

Relation: **Balneabilidade**

Instances: **351**

Attributes: **11**

Test mode: **10-fold cross-validation**

Classifier model (full training set)➔ **JRIP rules**:

(coliformesFecais <= 1000) and (coliformesFecais >= 259) => Classificacao= **4** (64.0/23.0)

(coliformesTotais >= 10170) and (coliformesFecais <= 2920) and (coliformesFecais >= 1010) => Classificacao= **3** (24.0/4.0)

(coliformesFecais >= 1236) => Classificacao= **1** (126.0/30.0)

=> Classificacao= **5** (137.0/25.0)

Number of Rules : 4

Figura 35- Regras geradas pelo algoritmo JRIP amostra Balneabilidade (classes em negrito)

4.2.4- Resultado do uso "Irrigação"

Para a amostra "Irrigação" tem-se a porcentagem correta dos classificadores para as três opções de teste, Figura 36.

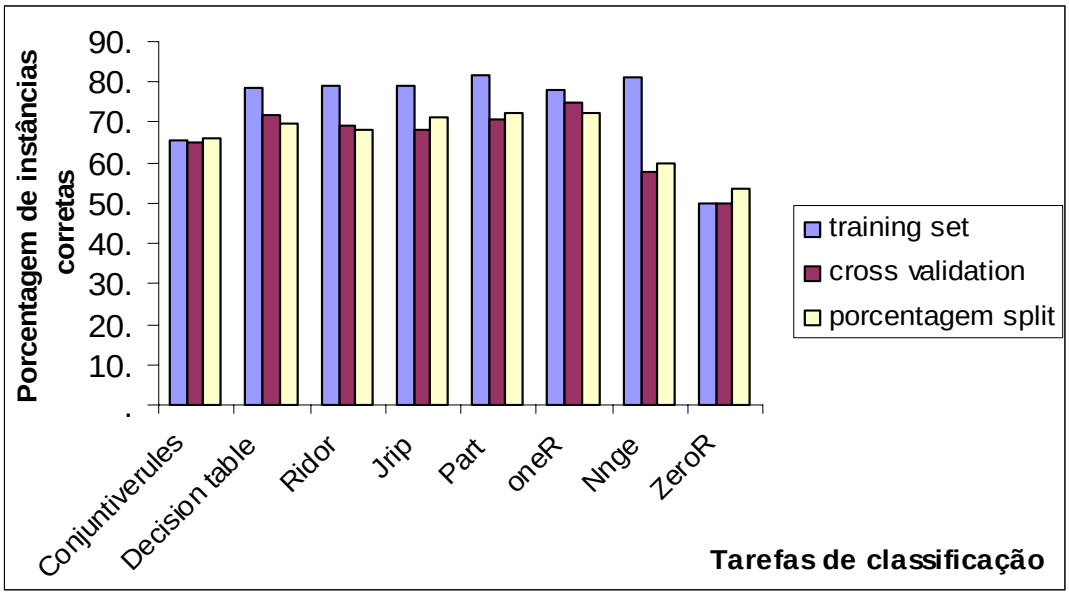

Figura 36- Desempenho dos classificadores para classificação da amostra "Irrigação".

Observa-se que para o uso de irrigação a melhor porcentagem de acerto, para o modo de teste *cross-validation* (validação cruzada) no algoritmo *OneR*, Tabela 5.

Tabela 5-Porcentagem de Instâncias corretas dos algoritmos de classificação para irrigação

Porcentagem de instâncias Corretas - 379 instâncias			
	Training set	**Cross validation**	**Porcentagem split**
Conjuntiverules	65.43	65.17	65.89
Decision table	78.62	71.65	69.76
Ridor	78.89	69.12	68.21
Jrip	79.15	68.07	71.31
Part	81.79	70.71	72.09
OneR	78.1	**74.67**	72.09
NNge	81.26	57.7	59.6
ZeroR	50.13	50.13	53.48

O valor de *kappa* da amostra de Irrigação no algoritmo OneR é 0.61, valores acima de 0.6 indicam confiabilidade na relação conforme Figura 37.

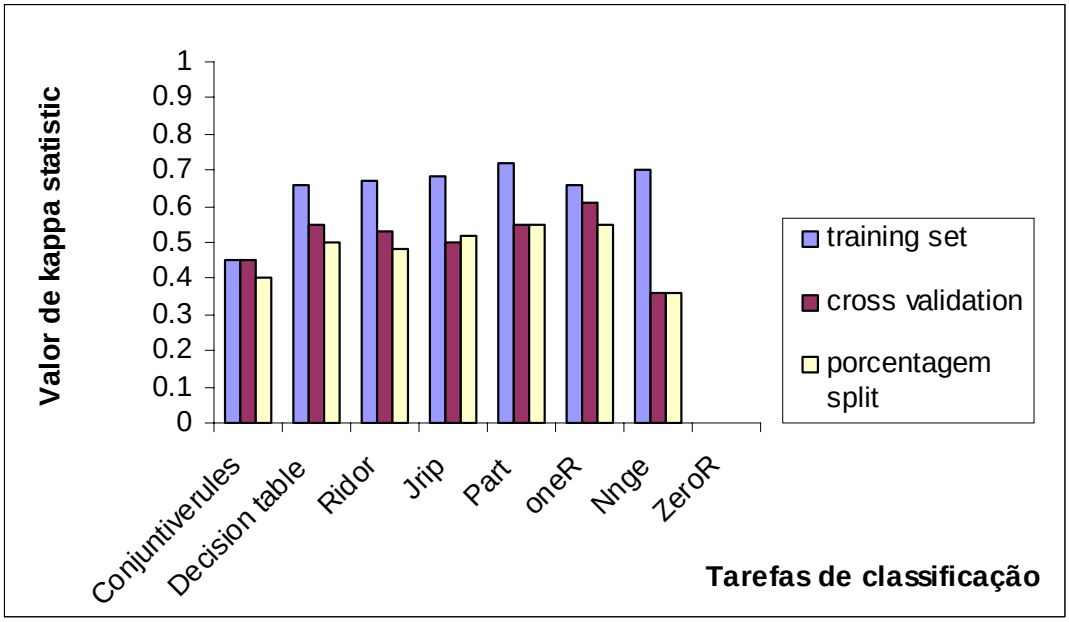

Figura 37- Valor de kappa statistic para os algoritmos da amostra de "Irrigação"

Para a amostra de irrigação obtivemos um bom resultado, basta verificar a porcentagem de acerto elevada, a baixa porcentagem de erro (instâncias incorretas) e o valor de *Kappa* acima de 0.6.

Com relação a acurácia (soma de positivos verdadeiros e negativos verdadeiros) por classe tem-se que os maiores problemas de classificação ocorreram nas classes 2 e 4 onde as taxas de TP *(True Positives)* são mais baixas, indicando erro nessas classificações, Figura 38.

=== Summary ===		
Correctly Classified Instances	283	74.6702 %
Incorrectly Classified Instances	96	25.3298 %
Kappa statistic	0.6119	
Total Number of Instances	**379**	

=== Detailed Accuracy By Class ===						
TP	FP	Precision	Recall	F-Measure	ROC Area	**Class**
0.967	**0.078**	**0.699**	**0.967**	**0.811**	**0.944**	**1**
0.273	0.026	0.273	0.353	0.623	0.5	**2**

0.353	0.046	0.545	0.353	0.429	0.654	**3**
0.311	0.039	0.519	0.311	0.389	0.636	**4**
0.968	**0.18**	**0.844**	**0.968**	**0.902**	**0.894**	**5**

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
58	1	0	0	1	a = 1
15	**9**	8	1	0	b = 2
8	7	**18**	6	12	c = 3
2	1	7	**14**	21	d = 4
0	0	0	6	**184**	e = 5

Figura 38- Métrica de avaliação do algoritmo OneR para amostra "Irrigação" (Validação Cruzada)

As regras geradas pelo algoritmo OneR foram (Figura 39):

Scheme: weka.classifiers.rules.OneR -B 6

Relation: **Irrigacao**

Instances: **379**

Attributes: **11**

Test mode: **10-fold cross-validation**

=== Classifier model (full training set) ===

coliformesFecais:

< 1095.0 -> **5**

< 1605.0 -> **4**

< 5560.0 -> **3**

< 10031 -.> **2**

>= 1003 1-> **1**

(296/379 instances correct)

Figura 39-Regras geradas pelo algoritmo OneR amostra Irrigação

4.2.5- Resultado do uso "Manutenção dos Ciclos Biogeoquímicos Naturais"

Para a amostra "Manutenção dos Ciclos Biogeoquímicos Naturais" tem-se a porcentagem correta dos classificadores para as três opções de teste, Figura 40. Observa-se que para este uso a melhor porcentagem de acerto, para o modo de teste *cross-validation* (validação cruzada) no algoritmo *OneR*, Tabela 6.

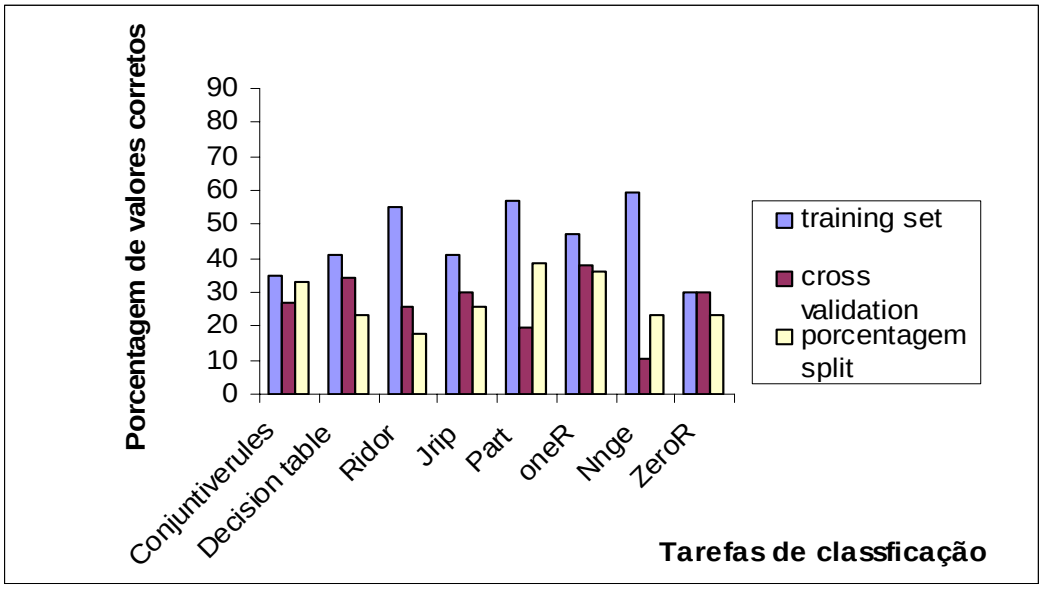

Figura 40- Desempenho dos classificadores para classificação da amostra "Manutenção dos Ciclos Biogeoquímicos" (validação cruzada).

Tabela 6 - Porcentagem de dados corretos para amostra "Manutenção dos Ciclos Biogeoquímicos Naturais"

Porcentagem de instâncias corretas			
	training set	**cross validation**	**porcentagem split**
Conjuntiverules	35.08	27.15	33.33
Decision table	41.22	34.21	23.07
Ridor	55.26	25.43	17.94
Jrip	41.22	29.82	25.64
Part	57.01	19.29	38.46
OneR	47.36	**37.7**	35.89
Nnge	59.64	10.52	23.07
ZeroR	29.82	29.82	23.07

O valor de *kappa* desta amostra foi de 0.61 para o algoritmo OneR. Como valores acima de 0.6 indicam confiabilidade na relação percebe-se que o resultado não foi bom, Figura 41.

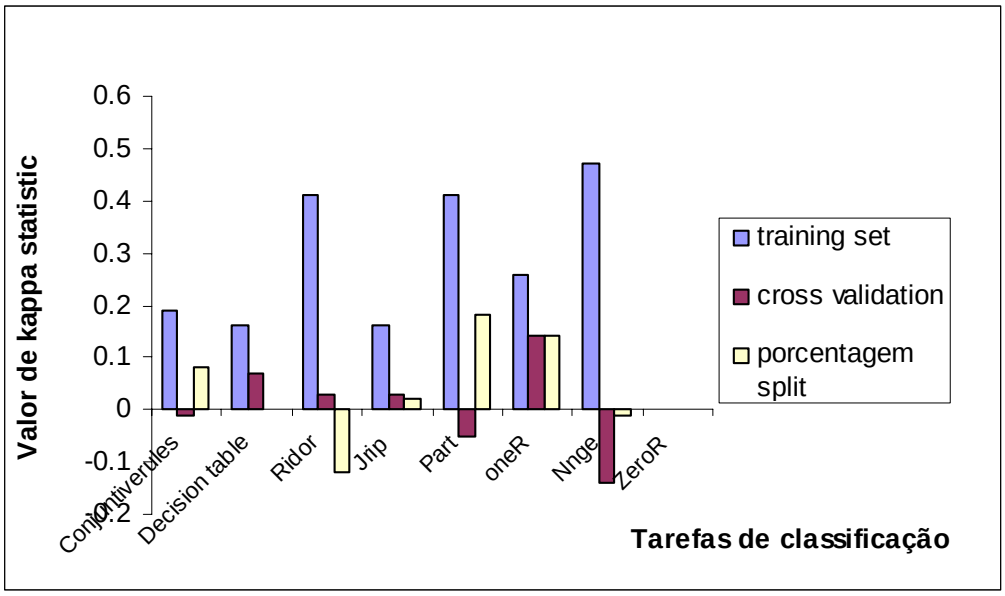

Figura 41- Valor de kappa para ciclos biogeoquímicos naturais

Observando a Figura 42, tem-se o desempenho do algoritmo OneR selecionado, visto que este obteve melhor performance de acertos perante os outros analisados. Verificamos na Figura 42 baixos índices de acerto deste algoritmo para a amostra (Manutenção dos Ciclos Biogeoquímicos Naturais) comprovados pela alta taxa de erro (porcentagem incorreta), baixo valor de kappa e baixos índices de TP (verdadeiros positivos). Ao olhar a matriz de confusão observa-se os maiores problemas de classificação nas classes 1, 2, 3 e 5, apenas a classe 4 obteve boa classificação.

=== Summary ===		
Correctly Classified Instances	43	37.7193 %
Incorrectly Classified Instances	71	62.2807 %
Kappa statistic		0.1451
Total Number of Instances	**114**	

=== Detailed Accuracy By Class ===						
TP	FP	Precision	Recall	F-Measure	ROC Area	**Class**
0	0.061	0	0	0	0.469	**1**
0.25	0.085	0.385	0.25	0.303	0.582	**2**
0.375	0.207	0.414	0.375	0.393	0.584	**3**
0.765	**0.5**	**0.394**	**0.765**	**0.52**	**0.632**	**4**
0	0	0	0	0	0.5	**5**

=== Confusion Matrix ===					
a	b	c	d	e	<-- classified as
0	5	4	7	0	a = 1
4	**5**	3	8	0	b = 2
2	3	**12**	15	0	c = 3
0	0	8	**26**	0	d = 4
0	0	2	10	**0**	e = 5

Figura 42- Métrica de validação do algoritmo OneR para amostra "Manutenção dos Ciclos Biogeoquímicos Naturais" (Validação Cruzada)

As regras geradas pelo algoritmo OneR foram (Figura 43):

Scheme: weka.classifiers.rules.OneR -B 6

Relation: **Ciclosbiogeoquimicos**

Instances: **114**

Attributes: **11**

Test mode: **10-fold cross-validation**

=== Classifier model (full training set) ===

dbo:

< 3.05 -> **4**

< 72.2 -> **3**

>= 72.2-> **2**

(54/114 instances correct)

Figura 43 - Regras geradas pelo algoritmo OneR amostra Manutenção dos Ciclos Biogeoquímicos Naturais

5- DISCUSSÃO

5.1- ENQUETE COM ESPECIALISTAS

Para a realização deste estudo realizou-se uma classificação prévia de um conjunto de registros (selecionados no SIBAC) por especialistas em avaliação de qualidade da água, por se tratar de um aprendizado supervisionado (onde há necessidade de um guia para o algoritmo). Foram entrevistados três Engenheiros (as) Sanitaristas e uma Bióloga. No decorrer deste procedimento foram encontradas várias dificuldades, dentre elas:

- Desinteresse pela pesquisa;
- Limitações do tempo disponível dos profissionais para classificar a enquete de 300 registros (alegaram não ter tempo mesmo sabendo da importância do trabalho para a área);
- Desistência: um quinto especialista previsto para a pesquisa não preencheu a enquete alegando sobrecarga de trabalho;
- Dificuldade para recolher as classificações: parcialmente a tabela para classificação foi recebida com aceitação, mas teve demora demasiada para responder, causando atraso no andamento da pesquisa.
- Preenchimento da pesquisa com número menor de registros: dois dos especialistas só concordaram em responder 50 registros.

Na mineração de dados é aconselhável que se tenha um número elevado de registros, principalmente para o conjunto de treinamento (que gera as regras). Neste trabalho optou-se pelo número de registros obtidos (classificados) dos especialistas. Certamente após a realização desta enquête com os especialistas concordamos que a

mesma deva ser realizada de outra maneira e com menos variáveis para a avaliação, determinando as variáveis que são mais problemáticas para cada uso.

A escolha dos algoritmos foi baseada no tipo de conhecimento a ser encontrado, visando a obtenção de regras de classificação para avaliar qualitativamente o Banco de Dados de Monitoramento Ambiental (SIBAC). Supõe-se que os melhores desempenhos de alguns algoritmos em detrimento de outros se deve ao método de busca de conhecimento que ele utiliza.

Para o uso de abastecimento por exemplo (Figuras 23, 24 e 25), observa-se que as maiores dúvidas na classificação dos especialistas estão nos valores extremos, onde a qualidade da água deve ser considerada de péssima qualidade (1) ou ótima qualidade (5). Ao observar as Figuras 26, 27 e 28 verificamos a mesma situação.

O resultado de balneabilidade e irrigação foram melhores, demonstrando através do valor de *kappa statistic* confiabilidade na amostra. Com relação as amostras de abastecimento e ciclos biogeoquímicos tem-se baixos índices de kappa, altos índices para as porcentagens de erro e baixos índices de porcentagens corretas, isto no melhor algoritmo selecionado para a amostra. Podemos enumerar possíveis causas para este resultado:

- Valores altos de porcentagens incorretas no conjunto de treinamento (training test) pode significar ruído nas duas amostras. Elas possuem esse índice elevado;
- Como o algoritmo ZeroR (considerado um algoritmo primitivo, que classifica as amostras pelo atributo majoritário) apresentou elevada porcentagem de erro em todas as amostras a que foi submetido, se ocorrer outro algoritmo com pior eficiência tem-se *overfitting* (regras muito especializadas) do modelo gerado. Assim observamos que o desempenho do OneR foi pior do que ele, isso demonstra *overfitting* das regras(regras muito especializadas). Poucos registros nesta amostra (Manutenção dos Ciclos Biogeoquímicos) pode ter contribuído para esta especialização;
- Baixos valores de *Kappa Statistic* também demonstram pouca confiabilidade nas classificações da amostra, as duas possuem valores de *kappa* abaixo de 0.6, valores acima deste demonstram confiabilidade na amostra.

Observando a acurácia dos resultados por classe, verifica-se que os índices de classificação correta são mais baixos para as classes intermediárias, ou seja, classe 2 (ruim), classe 3 (regular), classe 4 (bom). Provavelmente isso se deve ao caráter subjetivo da enquete submetida aos especialistas. Há mais chances de ocorrerem classificações errôneas nas classes intermediárias pela dificuldade em mensurar o que é ruim (classe 2), regular (classe3) e não é péssimo (1), por exemplo, o extremo da relação.

Verifica-se que o conjunto de regras não foi muito complexo, apenas o algoritmo **NNGE (Nearest Neighbor With Generalization)**, que classifica utilizando o método do vizinho-mais-próximo (árvore de decisão), gerou elevado número de regras (90) com muitos testes lógicos (em todos os quatro tipos deste estudo de caso), o que poderia causar pouca compreensão dos fatos por especialistas no assunto.

Com relação as regras geradas pelo algoritmo de melhor performance nas amostras obteve-se regras com número de variáveis diferenciadas, ou seja, o conhecimento dos níveis de qualidade da água encontrados na saída dos algoritmos (expressos sob a forma de regras) não incluíram todas as variáveis iniciais do arquivo de entrada. (Quadro 6).

Quadro 6- Quadro demonstrativo das variáveis de saída nas regras

Uso	**Algoritmo de melhor desempenho**	**Variáveis consideradas nas regras**	**Variáveis importantes segundo especialistas**
Abastecimento	**(Part)**	OD, DBO, DQO, CF, CT, NTK, P, TURB, ST	pH, DBO, DQO, CT, CF, NTK, P, TURB, ST
Balneabilidade	**(JRIP)**	CF e CT	CF e CT
Irrigação	**(OneR)**	CF	pH, DQO, DBO, NTK, CT, CF, P
Manutenção dos Ciclos Biogeoquímicos Naturais	**(OneR)**	DBO	PH, OD, DQO, DBO, NTK, P

Pode-se observar no Quadro 6 que as variáveis de saída estão presentes na avaliação inicial realizada pelos especialistas como uma das mais importantes para cada uso. Observa-se, que as variáveis consideradas para classificação dos usos "Abastecimento" e "Balneabilidade correspondem basicamente com as variáveis consideradas pelos especialistas, enquanto o algoritmo OneR, por exemplo, o de melhor performance nos usos irrigação e Manutenção dos Ciclos Biogeoquímicos Naturais é um algoritmo que escolhe o atributo mais relevante para construir a sua regra baseada em um único atributo. Para o uso de irrigação percebe-se que o mesmo foi rígido escolhendo a variável CF, isto porque para classificar irrigação de hortaliças esta variável é relevante, porém, para irrigação de campos não seria necessário. Para Manutenção dos Ciclos Biogeoquímicos Naturais tem-se a variável DBO escolhida pelo algoritmo para compor as regras. Embora seja uma variável importante para este uso, tem-se P, NTK, DQO, OD e pH como variáveis relevantes para esta classificação, embora desconsideradas pelo algoritmo OneR.

Para o uso de Balneabilidade, contato primário (natação, esqui-aquático) tem-se como variáveis mais relevantes para avaliar este uso CF (Coliformes Termotolerantes) e CT*(Eschrichia Coli)*, segundo os especialistas (embora existam outras também importantes). O algoritmo JRIP também as utilizou para representar o modelo de classificação (regras) criado para este para avaliar a qualidade da água.

Para o uso de abastecimento, o algoritmo PART considerou todas as variáveis relevantes para construir o modelo de classificação de regras, descrevendo todas as variáveis nas regras, só não utilizou a variável pH.

Observou-se através de comparação, que as regras produzidas neste estudo tem analogia com a classificação realizada pela Agência Estadual de Meio Ambiente e Recursos Hídricos do Estado de Pernambuco- CPRH, na determinação da qualidade dos corpos dágua. Pode-se observar nos Quadros 7 e 8, um comparativo entre as Regras de classificação descobertas e a classificação adotada pela CPRH.

Quadro 7-Comparação dos limites de Classificação da CPRH e regras geradas pelos algoritmos

Classificação CPRH	**Descrição dos limites**	**Estudo do Sibac**	**Usos com Regras semelhantes a CPRH**
Não Comprometida	Limites obedecem Decreto Estadual N° 7.269/81 para classe 1	5(ótima)	Abastecimento Balneabilidade Irrigação
Pouco Comprometida	Limites obedecem Decreto Estadual N° 7.269/81 para classe 2	4(Boa)	Abastecimento Balneabilidade Irrigação Manutenção dos Ciclos Biogeoquímicos Naturais
Moderadamente Comprometida	Limites obedecem Decreto Estadual N° 7.269/81 para classe 3	3(Regular)	Abastecimento Balneabilidade Irrigação Manutenção dos Ciclos Biogeoquímicos Naturais
Poluída	Limites obedecem Decreto Estadual N° 7.269/81 para classe 4	2(Ruim)	Manutenção dos Ciclos Biogeoquímicos Naturais
Muito Poluída	Limites não obedecem Decreto Estadual N° 7.269/81	1(Péssima)	Abastecimento Balneabilidade Irrigação

Observa-se que as regras tem uma certa equivalência , Quadro 8:

Quadro 8- Quadro comparativo das regras classificadas pelo algoritmo por uso e CPRH

Uso	Regras	Comparação com a CPRH
Irrigação	CF < 1095 ➔5	Intervalo equivalente a classe 2 CONAMA 357 e pouco comprometida (CPRH)
	CF < 1605 ➔4	Equivalente a classe 3 CONAMA 357 e moderadamente comprometida (CPRH)
	CF < 5560 ➔3	Classe 3 e 4 CONAMA 357 e poluída (CPRH)
	CF < 10.031 ➔2	Entre a classe 3 e 4 CONAMA 357, poluída (CPRH)
	CF>=10.031 ➔1	Sem limites e muito poluída CPHH
Manutenção dos Ciclos Biogeoquímicos Naturais	DBO < 3.05➔ 4	Classe 1 CONAMA357 e Não comprometida CPRH
	DBO<72.2➔3	Valores na Classe 2 , 3 e 4 (sem limite), moderadamente poluída a poluída na CPRH
	DBO>=72.2➔2	Sem limites no CONAMA 357 e muito poluída CPRH
Balneabilidade	CF<1000 and CF>=259➔4	Classe 2 CONAMA, Pouco comprometida CPRH
	CF>=10.170 and CF<=2920 and CF>=1010➔3	Entre Classe 3 e 4 CONAMA, Moderadamente Comprometida e Poluída CPRH
	CF>=1236➔1	Classe 3 e 4, de Moderadamente Comprometida a muito poluída
	CF➔5	Classe 1 CONAMA e Não Comprometida pela CPRH

Dentre os melhores algoritmos testados nestes conjunto de dados (Abastecimento, Balneabilidade, Irrigação e Manutenção dos Ciclos Biogeoquímicos Naturais), observou-se que possivelmente os métodos de classificação utilizados por eles sejam mais indicados para pequenos conjuntos de dados em processos de mineração. Assim, tem-se a descrição dos algoritmos e dos métodos que eles utilizam:

- **PART** (Partial decision trees) : Constrói regras a partir de árvores de decisão parciais usando o J4.8. Constrói uma árvore de decisão parcial em cada iteração e converte a melhor folha (nó da árvore) em regra. O processo de geração de regras tem dois estágios: Regras são induzidas inicialmente e posteriormente refinadas (dividir-para-conquistar). Uma árvore de decisão (*Decision Trees*) (Quinhan,1993) utiliza a estratégia de dividir-para-conquistar, onde um problema complexo é decomposto em subproblemas mais simples e recursivamente, a mesma estratégia é aplicada a cada subproblema. Algoritmo de uma árvore decisão:
    - Apresenta-se um conjunto de dados ao nó inicial (ou nó raiz) da árvore;
    - Dependendo do resultado do teste lógico aplicado ao nó raiz, a árvore ramifica-se para um dos nós filhos (ou uma sub-árvore);
    - Este procedimento é repetido até que um nó terminal seja alcançado.
- **JRIP (Optimizing IREP-Incremental Reduced Erro Prunning**): Existem muitas técnicas de aprendizado que vêm sendo adaptadas para o aprendizado e árvores de decisão. A maioria das árvores de decisão utiliza a estratégia de sobrecarregar para posteriormente simplificar, para tratar dados com ruído. Assim, nessa estratégia, uma hipótese é formada gerando inicialmente uma árvore complexa que super-utiliza os dados, e depois simplifica esta árvore utilizando técnicas de poda. As técnicas de poda melhoram taxas de erros de dados não vistos quando o conjunto de dados possui ruído. Existem inúmeros métodos propostos para poda de árvores e uma técnica eficiente é a poda do mínimo erro *REP* (*Reduced Erro Pruning*) (Cohen, 1995). Dentre as variações deste método está o *IREP* (*Incremental Reduced Erro Pruning*) (Cohen, 1995), utiliza árvore de decisão e as simplifica pela redução do erro, com um algoritmo que trabalha a técnica dividir-para-conquistar. Depois que

uma regra é encontrada, todos os exemplos que são cobertos por ela são deletados. Um caminho para melhorar a abordagem incremental do IREP é adiar o processo de produção de regras deste método, assim esse método se aproxima do método de poda pelo erro, uma otimização conhecida como JRIP(Cohen, 1995).

- **OneR:** É uma das formas mais elementares de encontrar regras muito simples a partir de um conjunto de instâncias, é o método 1R(1-*rule*). O método 1R gera uma árvore de decisão de apenas um nível, que é expressa através de um conjunto de regras que testa apenas um atributo em particular. É um método simples, econômico e frequentemente obtém boas regras para caracterizar a estrutura de dados. Muitas vezes a estrutura dos registros são simples, tornando um único atributo capaz de classificar uma instância com um bom nível de precisão. Mesmo que isto não aconteça, é sempre bom começar os testes pelos métodos mais simples. Usa o atributo do mínimo-erro para a predição (**REP- Reduced Erro Prunning** técnica de simplificação de árvores que melhoram erros em conjuntos de dados com ruídos). Exemplo do algoritmo 1R:
  - **Para cada valor do atributo, criar uma regra:**
  - **Contar o numero de vezes que a classe aparece**
  - **Determinar a classe mais freqüente**
  - **Criar a regra associando a classe a este valor do atributo**
  - **Calcular as taxas de erro das regras**
  - **Escolher as regras com menor taxa de erro**
- O pior comportamento observado foi do algoritmo **NNGE (Nearest Neighbor With Generalization),** elevado número de regras criando uma certa complexidade em todos os usos. O método de classificação que ele utiliza é o método do vizinho-mais-próximo (árvore de decisão). É um classificador onde o aprendizado é baseado em analogia. O conjunto de treinamento é formado por vetores de *n* dimensões e cada elemento deste conjunto representa um ponto no espaço *n*-dimensional. Ele calcula os *k*' vizinhos mais próximos (distância euclidiana) de um determinado conjunto

de dados e atribui como da mesma classe. Representa um dos paradigmas da aprendizagem indutiva, ou seja, objetos semelhantes pertencem ao mesmo grupo. Vantagens: a aprendizagem consiste em memorizar exemplos, indicada para problemas complexos. Desvantagem: não obtém uma representação compacta dos exemplos, tempo de aplicação é lento (para cada exemplo de teste calcula a distância a cada exemplo de treino). É altamente afetado por atributos redundantes (exemplos onde os vizinhos são da mesma classe ou com ruído) e irrelevantes.

Outra observação que se pode fazer é com relação as regras geradas por cada conjunto de dados (usos). Para as amostras:

- Abastecimento: o maior caso de ocorrências (classificações) na classe 4 e 5. Percebe-se nas regras geradas pelo algoritmo PART (algoritmo de melhor performance) que existem 10 regras que contemplam os casos de classificação 4 (Boa) e 5 (Ótima), de um total de 18 (dezoito) regras geradas pelo mesmo algoritmo.
- Balneabilidade: o resultado ignora a Classe 2 (classe de menor ocorrência de casos), elabora regras para as Classes 4 , 3 e 1. O que não se enquadra nestes casos é considerado Classe 5 (maior número de ocorrências).
- Irrigação: este uso possui uma equivalência entre a ocorrência das classes, elaborou portanto regras para todos os casos.
- Manutenção dos Ciclos Biogeoquímicos Naturais: criou regras para as maiores ocorrências, Classes 2, 3 e 4.

A técnica de descoberta de conhecimento tem a vantagem de aprender num conjunto de dados para posteriormente ser aplicada a outros, podendo ser utilizada para predições futuras até em outros conjuntos de dados. Esse é o motivo que nos levou a escolher esta técnica em detrimento de outras para avaliar a qualidade da água em Bases de Dados (SIBAC)

As regras resultantes deste trabalho podem futuramente compor o módulo baseado em conhecimento (Sistema Especialista) de um Sistema Inteligente para Monitoramento Ambiental e assim avaliar outras bacias hidrográficas.

A automatização de regras pode ser realizada com o desenvolvimento de aplicações em uma linguagem de programação (Java é mais indicada para a Weka), que

especifique o Banco de dados de monitoramento ambiental como entrada e utilize as regras geradas em um sistema inteligente para avaliar qualitativamente outras bacias hidrográficas.

5 - CONCLUSÕES

Consta-se uma disponibilidade crescente de dados multi-variados de qualidade da água mantidos em base de dados ambientais, cuja avaliação conclusiva por avaliação qualitativa, se torna cada vez mais complexa. Neste contexto, a presente dissertação teve como objetivo principal uma discussão do seguinte questionamento: Como extrair conhecimento implícito em bases de dados de monitoramento ambiental e com isto melhorar o processo de tomada de decisão para avaliação qualitativa da gestão de recursos hídricos?

Foi identificada, a partir da revisão bibliográfica, a área de KDD (Knowledge Discovery Databases) como propícia para subsidiar a extração de conhecimento implícito em bases de dados para melhorar a tomada de decisão no contexto de monitoramento ambiental.

Em um estudo de caso utilizando a Base de Dados do SIBAC, observou-se a viabilidade de se obter conhecimento através da investigação de técnicas de mineração, onde testou-se o desempenho de sete algoritmos de classificação do tipo "*rules*" visando a extração de regas para enquadrar registros de qualidade de água em cinco classes qualitativas (ótimo, bom, regular, ruim, péssimo) para os usos: "Abastecimento", "Balneabilidade", "Irrigação" e "Manutenção dos Ciclos Biogeoquímicos Naturais". Para fins de validação do desempenho dos algoritmos de classificação foram utilizadas métricas como: índice de Kappa, Matriz de Confusão, Valores de TP (*True Positives*), Valores de FP (*False Positives*), Precisão (*Precision*), Cobertura (*Recall*), F-Measure, Curva ROC (*ROC Curve*). Baseado na classificação de um conjunto de treinamento de em média 300 registros, realizada por três especialistas pôde-se observar que:

1. O processo de KDD (descoberta de conhecimento em bases de dados) inclui um volume extenso de técnicas aptas para subsidiar a tomada de decisão no monitoramento ambiental, sendo entretanto, o grupo *rules* da técnica de classificação (os que representam o conhecimento através de regras de decisão), o mais indicado para a tarefa em questão. Essas regras são condições do tipo *if-then* que são generalizadas de forma que resumem o conteúdo da base de dados.
2. Dentre as vantagens do uso das regras está a facilidade da incorporação do conhecimento (ele fica explícito), facilidade na interpretação do resultado e a facilidade de futuramente armazenar essas regras numa base de conhecimento.
3. Existe uma variação expressiva no desempenho dos algoritmos testados para o domínio em questão, sendo o algoritmo de melhor desempenho o “PART” para classificação do uso “Abastecimento”, o algoritmo Jrip para o uso “Balneabilidade”, o algoritmo OneR para Irrigação e para “Manutenção dos Ciclos Biogeoquímicos Naturais”.
4. Os algoritmos aplicados na classificação tem métodos de classificação distintos, sendo uns bastante simples (ONeR) e outros consideravelmente mais complexos, como o Nnge. Neste estudo a simplicidade provou ter bons resultados, uma vez que o algoritmo OneR obteve bom desempenho para todos os usos, embora não tenha apresentado o melhor desempenho em “Abastecimento” (está em segundo lugar na lista dos melhores classificadores) e em “Balneabilidade” teve o mesmo desempenho do JRip. Optou-se por considerar o Jrip porque em avaliação de qualidade da água faz-se necessário o uso de vários atributos (ele utiliza vários), também houve um aumento de performance deste algoritmo no modo de teste porcentagem split.
5. O método utilizado pelo algoritmo NNge (método do vizinho mais próximo), tornou a gama de resultados muito extensa, ou seja, o conjunto de regras muito grande, difícil de ser interpretado e armazenado numa base de conhecimento. Estas características estiveram presentes nos conjuntos de regras obtidas por este algoritmo para todos os usos.

6. As regras geradas pelo algoritmo OneR são baseadas em um único atributo. O resultado mostrou que o algoritmo escolheu a variável mais relevante para cada uso: em “Balneabilidade”, “Abastecimento” e “Irrigação” escolheu CF, somente em Ciclos Biogeoquímicos Naturais escolheu DBO. Provavelmente os classificadores acharam a estrutura desta amostra de dados simples, por isto este algoritmo obteve um bom desempenho.
7. Observou-se a importância da Cobertura (Recall) das regras geradas para cada classe em seus usos distintos. Faz-se necessário que o modelo cubra o maior número de casos e que além disso seja preciso na sua cobertura. A cobertura minimizará a ocorrência de problemas matemáticos que nunca sejam classificados. Se algum conjunto de dados não for coberto, o sistema pode não ser capaz de acompanhar o aprendizado que ocorre na resolução desse problema. Por outro lado tem-se a exatidão (casos cobertos corretamente sob o número de casos cobertos), esta minimizará os problemas relativos a predições incorretas, que dependendo da classe pode ter um alto custo (por exemplo, água com estado de qualidade péssimo (1) avaliada como ótima (5)). Como em alguns casos é interessante utilizar uma única medida para avaliar o algoritmo, e visto que cobertura e precisão cobrem diferentes aspectos de um algoritmo, medidas alternativas foram criadas, tal como *F-Measure,* que combina em sua fórmula precisão e cobertura. Assim, além de verificar as porcentagens corretas, valor de kappa, deve-se prestar atenção aos FP(falsos positivos), Cobertura(Recall), Precisão (precision), Matriz de Confusão (erros e acertos de cada classe), principalmente erros nas classes 1(péssimo), 2 (Ruim) e 3 (Regular) . Erros de classificação nestas três classes podem significar riscos para a população e investimentos desnecessários na recuperação de determinado trecho da bacia hidrográfica.
8. Entende-se que as técnicas avaliadas consistem em uma complementação de outras ferramentas em uso operacional como o sistema CONAMA. A avaliação automatizada pode subsidiar um diagnóstico periódico de monitoramento, indicando o nível de um eventual desvio da qualidade de água pré-estabelecida pelas classes estáticas da CONAMA, melhorando o processo de tomada de decisão.

9. As técnicas de mineração de dados (etapa do processo de KDD) podem auxiliar na descoberta automática ou semi-automática de conhecimento pois permitem a previsão de valores de atributos com maior probalidade de acerto. São previsões seguras, baseadas em informações reais disponíveis nas bases de dados ambientais.
10. Em um sistema operacionalizado acoplado a uma base de conhecimento, as regras geradas podem ser aperfeiçoadas, dependendo da alimentação de novos registros classificados por especialistas do domínio. Da mesma forma permitiria a classificação periódica de toda a base de dados, através de uma análise automatizada, classificando toda a bacia em níveis de qualidade propostos neste trabalho.
11. A desvantagem frente aos índices IQA (Índice de Qualidade da Água) é que nesta técnica as regras mudam em função do conjunto de treinamento, há produção de regras a cada conjunto de treinamento da base submetido a mineração.
12. Constatam-se que os algoritmos não somente geram as regras, mas incluem, a partir das saídas geradas pelas técnicas de validação, uma avaliação da confiabilidade do conhecimento gerado, indicando desta forma a necessidade específica de aprimorar os conhecimentos sobre a qualidade de água no contexto de um determinado uso e sugerir o consenso dos especialistas neste processo.
13. Entende-se que a aplicação das técnicas possui, sobretudo, utilidade na avaliação constante de grandes conjuntos de dados dinâmicos. Permitem uma visão sinóptica em áreas abrangentes, ex. bacia hidrográficas inteiras, tarefa complexa e dispendiosa se realizada por avaliação interativa por um técnico/especialista.

6 – BIBLIOGRAFIAS CITADAS

ALVARES, Lillian. **Aplicação de data mining em bases de dados especializadas em ciência da informação para obtenção de informações sobre a estrutura de pesquisa e desenvolvimento em ciência da informação no Brasil**. Brasília, 2000. Monografia (Especialização) UFRJ/ECO, MCT/INT/IBICT.

ALVARENGA, S.M.; BRASIL,A.E.; PINHEIRO,R. *et al.* **Estudo Geomorfológico Aplicado `a Bacia do Alto Rio Paraguai e Pantanais Matogrossenses.** Projeto RADAMBrasil, série Geomorfologia, Salvador-BA,1984, pp. 89-183.

AZEVEDO, L.G.T.; Rego, M.F.;Baltar, A.M.; Porto,R. **Sistemas de suporte a decisão para a outorga de direitos de uso da água no Brasil: uma análise da situação brasileira em alguns estados**. Revista Bahia e Análise de Dados, Salvador, v.13, n.ESPECIAL,P.481-296,2003.

BARRETO,J.M. **Inteligência Artificial no Limiar do Século XXI**. 2 Ed. Florianópolis,SC, 1999.

BIRANT, D. ; KUT, A. **ST-DBSCAN: An Algorithm for clustering spatial-temporal data**. In: Data & Knowledge Engineering 60. 208-221. Elsevier, 2007. Disponível em: www.sciencedirect.com. Acesso em : 11/11/2006.

BOGORNY, V. **Algoritmos e Ferramentas de Descoberta de Conhecimento em Bancos de Dados Geográficos**. Porto Alegre: PPGC de UFRGS, 2003. Disponível em: http://www.inf.ufrgs.br/%7Evbogorny/publication.html. Acesso em 07/09/2006.

BRANCO, S. M. (1986). **Hidrobiologia aplicada à engenharia sanitária**, São Paulo, 3 ed., CETESB/ASCETESB, 616p.

BRASIL. Ministério do Meio Ambiente. Conselho Nacional de Recursos Hídricos. Resolução n.12, de 19 de junho de 2000. Brasília, 2000.

_____.**Caderno da Região Hidrográfica do Paraguai/MMA**, Secretaria de Recursos Hídricos. Brasília: MMA, 2006. 140p.:il. Color; 27cm.

_____.Ministério do Meio Ambiente. Agência Nacional de Águas(ANA). **Disponibilidades e Demandas de Recursos Hídricos Brasil.** Brasília**: Superintendência de Planejamento de Recursos Hídricos**, Superintendência de Conservação de Água e o Solo, Superintendência de Usos Múltiplos, Agência Nacional de Águas(ANA), 2005. 134p.(arquivo em pdf).

CANO, J. R.; HERRERA, F.; LOZANO, M. Evolutionary Stratified Training Set Selection for Extracting Classification Rules with Trade Off precision-interpretability. In: Data & Knowledge Engineering 60. 90-108. Elsevier, 2007. Disponível em: www.sciencedirect.com. Acesso em : 11/11/2006.

CANHAMERO, M. **Apostila do curso e capacitação Técnica em Recursos Hídricos com ênfase na sub-bacia hidrográfica Billings Tamanduaté-" O saneamento Ambiental através do Conhecimento de Novas Técnicas de Análise para os Recursos Hídricos**. Disponível em: www.agds.org.br/cursos/pdf/apostilaprofmagali.pdf. Acesso em: 15/01/2007.

CONAMA - Conselho Nacional do Meio Ambiente. Resolução n° 357, de 17 de Março de 2005.

CARVALHO, D. R. & FREITAS, A. A. **A hybrid decision tree/genetic algorithm for coping with the problem of small disjunts in Data Mining**. In: Porc. Generic Evolucionary Computation(GECCO-2000), las Vegas, NV, USA. July,2000.

_____. Ministério do Meio Ambiente. Secretaria de Recursos Hídricos. Política Nacional de Recursos Hídricos. Lei n. 9.433, de 8 de Janeiro de 1997. Brasília, 1997.

_____. Ministério do Meio Ambiente. Conselho Nacional do Meio Ambiente. Resolução CONAMA n.357, de 25 de março de 2005. Brasília, 2005.

DEBERDT, André Jean. Projeto PROCIENCIAS -Qualidade de Água. Disponível em: http://www.educar.sc.usp.br/biologia/prociencias/qagua.htm. Acesso em 10/10/2006.

DELAVAR, M.R.; KARIMIPOUR, F. ; REZAYAN, H. **Neighborhood Analysis in Water Pollution Estimation**. In: Environmental Informatics Archives, volume 3(2005), 232-238. ISEIS Publication Series number p002.

DIAS, M.M.; PACHECO,R.C.S**. Uma metodologia para o desenvolvimento de sistemas de descoberta de conhecimento**. Publicado em Acta Sci. Technol.. Maringá, v.27, n.1,p.61-72. Jan./june,2005. Disponível em: http://www.ppg.uem.br.Acesso em 10/10/2006

ESTEVES, F. A. **Fundamentos de Limnologia**. Rio de Janeiro: Editora Interciência/Financiadora de Estudos e Projetos. 602p. 1998.

FAYYAD, U.M. *et al.* **Advances in knowledge discovery and datamining**. Cambridge, Ma : AAAI Press, 1996.

FAYYAD, U. *et al.* **From data mining to knowledge discovery in databases.** *In: AI Magazine.* Cambridge, Ma : AAAI Press, 1996.

FAYYAD, U.; PIATETSKI-SHAPIRO,G.;SMYTH,P. **The KDD Process for Extracting Useful Knowledge from volumes of Data**. In: Communications of the ACM, p.27-34, 1996.

FEITOSA, F. A. C.; FILHO, J. M**. Hidrogeologia: conceitos e aplicações**, LABHID-UFPE. Fortaleza-CE. Ed. CPRM, 1997.

FELDENS, M. A. **Descoberta de Conhecimento em bases de dados e Mineração de dados**. Pelotas, jun. 1997. Disponível em: <http://gpia.ucpel.tche.br/voia/ioia/public.htm> Acesso em 20/03/2004.

FERRAZ A.R.G.; BRAGA JR. **Modelo Decisório para a Outorga de direito ao uso da água no Estado de São Paulo**. RBRH-Revista Brasileira de Recursos Hídricos, vol 3 n.1 jan/mar 1998, 5-19.

FREITAS, A.A. **On objetive measures of rule surprisingness**. In: Proceedings of the 2nd European Symposium Principles of Data Mining and Knowledge Discovery,1997. Disponível em: http://dainf.cefetpr/~alex/thesis.html.

GIBERT, K.; SANCHES-MARRE, M.;RODA-RODRIGUES, I. **GESCONDA: An Intelligent data Analysis system for Knowledge discovery and management**.In: Environmental Modelling & Software, volume 21, Issue 1, 2006, p.115-120.

GOMES, A. K. **Análise do Conhecimento Extraído de Classificadores Simbólicos utilizando medidas de avaliação e de interessabilidade.**(Dissertação) Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo. USP-São Carlos,2002.

GOMES, M.M.; PADOVANI ,C. R. **A agonia da pecuária no baixo Rio Taquari (MS)**. *In* Simpósio sobre Recursos Naturais e Sócio-econômicos do Pantanal-Sustentabilidade Regional 4. Corumbá(MS): Embrapa Pantanal/ UFMS/UCDB/ SEBRAE, 2004. Resumo. Cd-Rom.

HAND, D.J. **Construction and Assessment of classification rules**. John Wiley & Sons, 1997.

HARRISON, T.H. **Intranet Data Warehouse**. São Paulo:Editora Berkeley, 1998.

IBGE, 2000, Instituto Brasileiro de Geografia e Estatística: Censo 2000. Disponível em: www.ibge.gov.br. Acesso em: 07/08/2005.

LACERDA, M.P.; SOUZA, R.C.F. **Aplicação da Mineração de Dados em Sistema de Avaliação de professor e aluno**. Monografia(conclusão de curso). Universidade Federal do Pará. Belém, 2004.

LANNA, A.E.L;1999. **Aspectos Institucionais da Gestão de Recursos Hídricos, Capitulo 5. Publicação do Instituto de Pesquisas Hidráulicas da Universidade Federal do Rio Grande do Sul-** home-page: atlantico.iph.ufrs.br/portaliph/ppg/disciplinas/hip78/2.pdf. Rio Grande do Sul.

LAVRAC, N; FLACH, P.; ZUPAN,B.**Rule evaluation measures: a unifying view**. In: Proceeding of the Ninth International Workshop on Inductive Logic Porgramming. LNAI. Volume 1634.(1999)74-85.

LIMA, E. B. N. R. **Modelagem Integrada para Gestão da Qualidade da água na Bacia do Rio Cuiabá.** Tese(doutorado). Programa de Pós-Graduação da Engenharia Civil da Universidade Federal do Rio de Janeiro-COPPE. UFRJ.2001.

MANCUSO, P. C. S.; SANTOS, H. F. S. **Reuso de Água.** 1.ed. Barueri, SP: Ed. Manole, 2003. 579p.

MEIRELLES, M.; BUENO, M.C.D.; DIAS, T. C.S.; COUTINHO, H.L.C. **Sistema de Suporte a Decisão para avaliação de impactos ambientais em bacias hidrográficas por redes de dependência e lógica fuzzy. In: XII Simposio Brasileiro de Sensoriamento Remoto, 2005, Goiânia, Brasil. Anais.** XII Simposio Brasileiro de Sensoriamento Remoto,Goiânia, Brasil, 16-21 abril 2005, INPE, P.2259-2266. Disponível em: http://marte.dpi.inpe.br/col/ltid.inpe.br/sbsr/2004/11.21.21.19/doc/2259.pdf. Acesso em 02 maio 2006.

MOTA, S. **Preservação de recursos hídricos**. Rio de janeiro, ABES, 1995.

MOTTA, C.G. L., **Sistema Inteligente para avaliação de Riscos em Vias de Transporte Terrestre.** Tese de Mestrado, COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2004. Disponível em:www.coc.ufrj.br. Acesso em: 20/07/2006.

MOTTA, C.G. L. **Introdução a Técnicas de Data Mining- Classificação de Dados**,2005. Disponível em: www.arquivosevt.Incc.br/pdfs/Intoducao%20data%Mining%2002.pdf. Acesso em: 20/07/2006.

NALINI, R**. Ética Ambiental**. 2ª Ed. Campinas,SP: Ed. Millenium,2003.424p.

NASCIMENTO, L. V.; VON SPERLING, M. **Os padrões brasileiros de qualidade das águas e os critérios para proteção da vida aquática, saúde humana e animal**. In: Congresso Interamericano de Ingeniera Sanitária y Ambiental, AIDIS,26, 1998, Lima. Anais Lima: 1998, p. 1-6.

NAVEGA, S. **Princípios Essenciais do Data Mining**. In: Infoimagem, Cenadem, novembro,2002. Anais: Infoimagem, Cenadem, 2002. Disponível em: http://www.intelliwise.com/reports/i2002.pdf Acesso em 11/10/2006.

NORTON, M. J**. Knowledge Discovery in databases**. In: Library Trends, v.48, n.1, p.9-21, 1999.

PADOVANI, C. R. 2004. **Fire Monitoring and analysis for the Brazilian Pantanal**. *In* Simpósio Internacional de Projetos Ecológicos de Longa Duração, 1.CNPq: Manaus: Resumos. p.47.

PELLEGRINI, G.F.; COLLAZOS,K. **Extração de Conhecimento a partir dos Sistemas de Informação.** Disponível em: http://www.inf.ufsc/~I3C/artigos/Pellegrini00.pdf Acesso em 08/09/2006.

PEREIRA, G. C. **Mineração de dados para análise e diagnóstico Ambiental**. Tese(doutorado). Programa de Pós-Graduação de Engenharia da Universidade Federal do Rio de Janeiro-COPPE/UFRJ,2005.

QUINLAN, J. R. **Generating productives rules from decision trees.** In: Proceesings of the tenth International Joint Conference on Artificial Intelligence, pages 304-307, Italy, 1987.

QUONIAM, L. ; TARAPANOFF,K.; JUNIOR,R.H.A.;ALVARES, L**. Inteligência obtida pela aplicação de Data Mining em Bases de teses Francesas sobre o Brasil.** In: Ciência da Informação, Brasília, v.30, n.2, p.20-28, 2001.

REBOUÇAS, A.C.; BRAGA, B; TUNDISI, J. G.. **Águas Doces do Brasil: Capital Ecológico, uso e conservação.** São Paulo: Escritura Editora, 1999.717p.

REZENDE, S.O. (Coord.), **Sistemas Inteligentes: Fundamentos e Aplicações.** Barueri, SP, Brasil, Rezende, S.O., 2003.Editora Manole.

ROBERTO, A. N. ; PORTO, R. L. ; ZAHED, K. **Sistema de Suporte a Decisões para Análise de Cheias em Bacias Complexas**. In: XII Simpósio Brasileiro de Recursos Hídricos, 1997, Vitória. Anais do XII Simpósio Brasileiro de Recursos Hídricos, 1997.

ROMÃO, W. **Descoberta de Conhecimento relevante em Bancos de Dados sobre Ciência e Tecnologia.** Tese(Doutorado)-Universidade Federal de Santa Catarina.. Florianópolis,2002. Disponível em: http://teses.eps.ufsc.br/resumo.asp?3150. Acesso em 09/10/2006.

SALOMÃO, F. X. T. **Rio Cuiabá: A geologia e a problemática da erosão e do assoreamento.** In: Ferreira, M. S. F. D.(org.), O rio Cuiabá como subsídio para a educação Ambiental, Ed. UFMT,1999.

SIBAC-Sistema de Monitoramento da Bacia do Rio Cuiabá, 2000, Relatório de Modelagem Integrada da Bacia do Rio Cuiabá, PROPEP.

SPRAGUE,R. H. , WATSON, H. J. **Sistemas de apoio a decisão: colocando a teoria em prática.** Trad. Ana Beatriz G. R. Silva. Rio de Janeiro,Campus, 1989. p.43-54.

SILVEIRA, Rosemari de F. **Mineração de Dados aplicado a Definição de Índices em Sistemas de Raciocínio Baseado em Casos. Porto Alegre: CPGCC da UFRGS, 2003.** Monografia(especialização). Universidade Federal do Rio Grande do Sul . Porto Alegre, BR-RS, 2003.

SPERLING, M.V. **Introdução `a Qualidade das Águas e ao tratamento de esgotos**, ed 2, Belo Horizonte-MG, Ed. SEGRAC,1996.

TUCCI, C. E. **Modelos Hidrológicos**, ed. 1. Porto Alegre, 1998. Ed. da Universidade Federal do Rio Grande do Sul

WAIKATO, U.D. **Weka Knowledge Explorer(Waikato Enviroment for Knowledge Analys).** Nova Zelândia, 2000.

WATERMAN,D.A. **A Guide to expert systems**. Addison-Wesley Publishing Company, 1986.

WEISS,S.I.;KULIKOWSKI,C. **Computer Systems that learn: classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems**. San Francisco, California,1991. Morgan Kauffmann.

WITTEN, I.H.; FRANK, E. **Data Mining: Practical Machine Learning Tools and a Techniques with Java Implementations**. Morgan Kaufmann Publishers. Waikato, 2000,USA.

7 – BIBLIOGRAFIAS CONSULTADAS

BERRY, M.J. ; LINOFF, G . **Data Mining Techniques: For Marketing, Sales and Customer Suport**. New York, NY: John Wiley & Sons, 1997. 444p.

BRATKO, I: Prolog Programming for Artificial Intelligence. Addison-Wesley(1990).

GOEBEL, M. et al. **A Survey of Data Mining and knowledge Discovery Software Tools.** 1999. Disponível em: http://citeseer.nj.nec.com/goebel99survey.html> Acesso em 28/03/2004.

LUCENA, C.J.P. **Inteligência Artificial e Engenharia de Software**. Publicações acadêmico-científicas Puc-RJ/IBM. Brasil. Jorge zahar Editor. Rio de Janeiro,1987.

GANASCIA, J. G. **Inteligência Artificial**. Ed. Ática, Sao Paulo,1997.

DATE,C. J. **Introdução a Sistemas de Bancos de Dados, tradução da 4 Ed Americana**,p.674, Editora Campus,1996.

DESENVOLVIMENTO E MEIO AMBIENTE: Riscos Coletivos- Ambiente e Saúde- CO-EDIÇÃO: REVISTA NATURES SCIENCES E SOCIÉTÉS. Curitiba,PR: Ed. da UFPR, n.5,2002. Semestral. ISSN 1518-952X.

LANDIS J.R., KOCH G.G. **The Measurement of Observer Agreement for Categorical Data**, *Biometrics*, 1977a, **33**, 159-174.

LEVINE, R. J.; DRANG,D. E.; EDELSON, B. **Inteligência Artificial e sistemas especialistas.** Trad. RATTO, M.C.S.R. São Paulo, Mcgrawhill, 1988. 264p.

NAVEGA, S. C. (2000) Inteligência Artificial, Educação de Crianças e o Cérebro Humano. Publicado em Leopoldianum, Revista de Estudos de Comunicações da Universidade de Santos (Ano 25, No. 72, Fev. 2000, pp 87-102). Disponível em http://www.intelliwise.com/reports/p4port.htm

NAVEGA, S. C. (2002) Projeto CYC: Confundindo Inteligência com Conhecimento. In: KMBrasil 2002, 3° Workshop Brasileiro de Inteligência Competitiva. Disponível em http://www.intelliwise.com/reports/kmbscn.htm

NAVEGA, S. C. (in press) Pensamento Crítico e Argumentação Sólida. Intelliwise Publicações. Trechos em http://www.intelliwise.com/books

PORTO, M. **Sistema de Gestão da qualidade das águas: uma proposta para o caso brasileiro.** Tese(livre Docência)- Escola Politécnica de Universidade de São Paulo. Departamento de Engenharia Hidráulica e Sanitária . São Paulo, 2002.

PROENÇA, C. N. O.; MEDEIROS, Y.D.P.; CAMPOS,V.P**. Metodologia para Definição de Parâmetros de Qualidade d a Água visando o enquadramento de corpos d'água em região semi-árida**. Disponível em: www.grh.ufba.br/Publicacoes/Artigos/Artigos%202004/Em%20andamento/Clélia/artigo%20clelia%20abrh%2030-08.pdf . Acesso em: 13/11/2006.

UNIVERSIDADE FEDERAL DO CEARÁ(UFC). Laboratório de Inteligência Artificial. Disponível em http://www.lia.ufc.br/~bezerra/exsinta/exsintashell.htm Acesso em: 20 março 2006.

RICH, Elaine; KNIGHT, Kevin. **Artificial Intelligence**. New York: Mc Graw Hill Book, 1983.

SANTOS, I. M. **Data Warehouse como ferramenta de auxílio em sistemas de monitoramento Ambiental**. Monografia(graduação). Departamento de Ciências da Computação da Universidade Federal de Mato Grosso-UFMT, Cuiabá-MT, 2003.

NILSSON, Nils J. **Artificial Intelligence: a new synthesis**. San Francisco, CA: Morgan Kaufmann, 1998. 513p.

NILSON, Neils S. Principles of Artificial Intelligence, Spring Verlag, Berlin, 1982.

BRANCO, S. & ROCHA, A. A. **Elementos de ciência do ambiente**. 2a. ED. SÃO PAULO, CETESB/ASCETESB, 1987. 206 P.

BUSS, D. F.; BAPTISTA, D. F. & NESSIMIAN, J. L. *Bases conceituais para a aplicação de biomonitoramento em programas de avaliação da qualidade da água de rios.* Cad. Saúde Pública, Rio de Janeiro, 19(2): p. 465-473, mar-abr, 2003.

MMA - Ministério do Meio Ambiente, Programa Monitore: Diretório das Instituições que Realizam Monitoramento Ambiental. Brasília: MMA. 1998.

NAVAS-PEREIRA, D. & HENRIQUES, R. M. *Aplicação de índices biológicos numéricos na avaliação da qualidade ambiental.* Revista Brasileira de Biologia, 56: p441-450. 1995.

CARMOUZE, J. P. (1994). **O Metabolismo dos ecossistemas aquáticos: fundamentos teóricos, métodos de estudo e análises químicas.** São Paulo - Editora Edgard Blücher – FAPESP. 253p.

ESTEVES, F. ª (1988). **Fundamentos de limnologia**, Rio de Janeiro, - Editora Interciência Ltda – FINEP. 574p

CARVALHO, B. A. Ecologia aplicada ao saneamento ambiental. Rio de janeiro, abes/bnh/feema, 1980.

FELLENBERG, G.. Introdução aos problemas de poluição ambiental. São paulo, epv/springer/edusp, 1980. 196 p.

ACM special interest group on knowledge discovery in data and data mining. Disponível em: <http://www.acm.org/sigkdd/>

CABENA, Peter *et al. Discovering data mining:* from concept to implementation. New Jersey : Prentice Hall, 1997.

DATAMATION magazine. Disponível em: <http://www.datamation.com/>

INFORMATION discovery. Disponível em: <http://www.datamining.com/>

INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING. Disponível em: <http://www.digimine.com/usama/datamine/kdd99/>