# O PROCESSO *KDD* APLICADO NA EXTRAÇÃO DE REGRAS: UM ESTUDO DE CASO DA ÁREA MÉDICA

**Anderson Roges Teixeira Góes**
UFPR: Programa de Pós-Graduação em Métodos Numéricos em Engenharia (PPGMNE); e
Departamento de Expressão Gráfica; CP: 19081; CEP: 81531-990; Curitiba, PR
artgoes@ufpr.br

**Maria Teresinha Arns Steiner**
UFPR: Programas de Pós-Graduação em Métodos Numéricos em Engenharia (PPGMNE) e de Engenharia de Produção (PPGEP); e
PUCPR: Programa de Pós-Graduação em Engenharia de Produção e Sistemas (PPGEPS); CP: 19081; CEP: 81531-990; Curitiba, PR
tere@ufpr.br; maria.steiner@pucpr.br

**RESUMO**
Neste trabalho é apresentado o processo KDD no qual se insere um algoritmo para realizar a extração de regras para classificação de padrões e a seleção de atributos simultaneamente, por meio de Algoritmos Genéticos, aplicado a um problema da área médica. Tal aplicação apresentou resultados satisfatórios para os conjuntos de treinamento, mas não tão satisfatórios para os conjuntos de teste. Desta forma, tem-se que o algoritmo proposto deverá ainda ser melhorado com o intuito de se obter uma melhor generalização para os resultados.
**Palavras-chave:** Extração de Regras de Classificação; Seleção de Atributos; Algoritmo Genético.

**ABSTRACT**

This paper presents the KDD process in which it was showed an algorithm to extract rules for pattern classification and selection of attributes simultaneously, through Genetic Algorithms, applied to a problem of a medical area. Such an application presented good results for the training sets, but not so satisfactory for the testing sets. In this way, the proposed algorithm must be still improved in order to get a better generalization to the results..
**Key words:** Classification Rules Extraction; Attributes Selection; Genetic Algorithm.

**Área principal (AS-PO na Área de Saúde)**

## 1. Introdução

A geração de bancos de dados ocorre de forma natural nos dias atuais, pois os meios computacionais são práticos para seu armazenamento. Os dados são originados de indústrias dos mais diversos ramos de produção, empresas de telecomunicações, instituições educacionais, hospitais, instituições financeiras, de saúde, dentre tantas outras. No entanto, a tarefa de simplesmente armazenar tais dados não é suficiente; é necessário, também, verificar se os dados coletados possuem informações relevantes, se há algum conhecimento a ser descoberto.

Com isso, neste trabalho é proposto um algoritmo que realiza, simultaneamente, a extração de regras para a classificação de padrões e a seleção de atributos por meio de Algoritmos Genéticos (AG). O algoritmo proposto foi aplicado a uma base de dados da área médica, com a finalidade de obter regras para classificar pacientes com cálculo e pacientes com câncer no duto biliar. A metodologia aqui aplicada pode ser enquadrada no processo KDD (*Knowledge Discovery in Database* ou Descoberta de Conhecimento em Bases de Dados).

Este trabalho está dividido em seis seções, sendo que a segunda realiza estudo referente ao processo KDD. A terceira seção descreve o sistema de extração de regras por meio de Algoritmo Genético de uma forma geral; na quarta seção é apresentado o algoritmo proposto. A aplicação na área médica e seus respectivos resultados são apresentados na quinta seção e, finalmente, a sexta e última seção contém as considerações finais e indicação de próximas etapas de trabalho.

## 2. KDD (*Knowledge Discovery in Database*)

A "Descoberta de Conhecimento em Bases de Dados" (*Knowledge Discovery in Database*, ou simplesmente, *KDD*) é um processo que visa encontrar informações em bancos de dados de uma maneira automatizada, criando relações de interesse que muitas vezes não são observadas por especialistas no assunto. O termo "*KDD*" surgiu no final da década de 80 e se mantém fortemente nos dias atuais, o que se pode verificar pelos trabalhos de Brachman e Anand (1994), Fayyad et al. (1995), Weiss e Indurkhya (1998), Han e Kamber (2006), Steiner et al. (2006), dentre outros.

Ao realizar busca pela definição de *KDD*, tem-se que a sua primeira menção foi no trabalho de Frawley, Piatetsky-Shapiro e Matheus (1991) como sendo "o processo, não trivial, de extração de informação, implícitas, previamente desconhecidas e úteis, a partir dos dados armazenados em um banco de dados".

Três anos depois Braschman e Anand (1994) definiram *KDD* como "uma tarefa de descoberta de conhecimento intensivo, consistindo de interações complexas, feitas ao longo do tempo entre o homem e uma grande base de dados possivelmente suportada por um conjunto heterogêneo de ferramentas".

No entanto, a definição mais comum na literatura é de Fayyad *et al.* (1995) onde se tem que *KDD* é "o processo não-trivial de identificação válida, em dados, novos, potencialmente úteis e finalmente com padrões compreensíveis".

Anterior a sua denominação, o processo *KDD* era, muitas vezes, confundido por muitos com o conceito de *Data Mining* (Mineração de Dados ou, simplesmente, *DM*). Mas, *DM* é a principal das cinco etapas do processo *KDD*. Estas etapas, ilustradas na figura 01 a seguir, são as seguintes: seleção dos dados; limpeza dos dados ou pré-processamento; transformação dos dados; *DM* e, finalmente, interpretação do conhecimento gerado (FAYYAD *et al*., 1995).

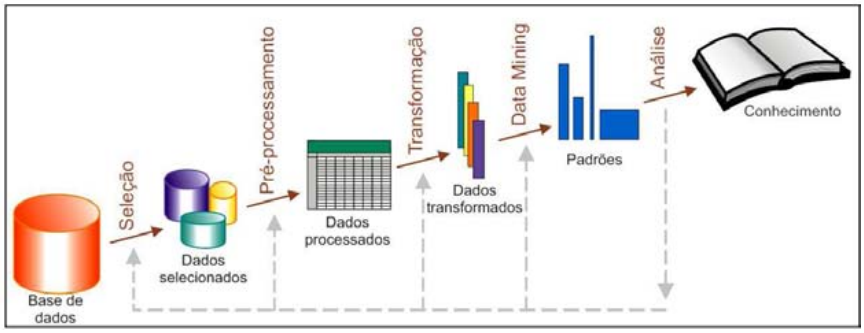


**Figura 01** – Etapas do processo KDD
Fonte: Fayyad, Piatetsky-shapiro e Smyth (1995)

No decorrer deste processo pode-se encontrar conhecimentos explícitos ou não, isto é, as informações podem ser as que já se tinha conhecimento ou informações inesperadas que ao analisar a base de dados não se verificava nenhuma relação óbvia. Podem ainda ocorrer informações sem nenhuma relação significativa, pela falta de atributos ou por não haver conhecimento novo a ser descoberto.

Geralmente, têm-se como respostas, informações que não se podem detectar quando se aplica métodos tradicionais na análise de dados para posterior tomada de decisão, pois, em sua grande maioria, os métodos tradicionais são capazes de verificar apenas as relações explicitas nos bancos de dados.

## 2.1. Etapas do Processo *KDD*

No desenvolvimento deste trabalho é utilizado o processo *KDD* definido por Fayyad *et al.* (1995) e, assim sendo, as etapas são explicitadas a seguir:

**Seleção dos dados:** nesta fase é escolhido o conjunto de dados que se pretende analisar, definindo assim os atributos e os eventos (registros). Em sua grande maioria, esta seleção é realizada por um especialista da área proveniente dos dados, pois possui papel fundamental no resultado final;

**Limpeza dos dados ou pré-processamento:** é a fase que determina a qualidade dos dados, onde são eliminados dados redundantes, ruídos possíveis de serem detectados e discrepância nos dados. Além disto, é verificada a possibilidade de diminuir o número de variáveis. Para isto, podem ser aplicados métodos estatísticos, a fim de melhorar a eficácia dos algoritmos de classificação, como apresentado por Steiner *et al.* (2006);

**Transformação dos dados:** após o pré-processamento dos dados, estes precisam ser armazenados e formatados de forma adequada à aplicação do algoritmo na próxima fase. Também é nesta fase que são determinados atributos faltantes que podem ser obtidos de outros atributos como, por exemplo, a duração de certo evento por meio do horário inicial e horário final da ocorrência do mesmo;

***Data Mining*:** esta é a etapa mais importante de todo o processo *KDD*, uma vez que é neste momento que se aplicam técnicas para análise dos dados por meio de algoritmos, heurísticas ou metaheurísticas para a descoberta de padrões. O tempo de execução desta fase deve ser compatível com o tempo disponível na espera da solução. Muitos são os métodos, sendo que alguns dos mais conhecidos são Árvores de Decisão, Redes Neurais e Algoritmos Genéticos;

**Interpretação do conhecimento gerado:** após da fase de *DM*, deve-se interpretar o conhecimento apresentado, verificando a relevância (ou não) na obtenção dos padrões e com isso, analisar a eficácia do método aplicado na etapa de *DM*. Caso o analista julgue que o conhecimento não é válido, o processo deverá ser reiniciado, analisando todas as etapas em busca de melhorar e/ou refazer o que for necessário, até que o conhecimento obtido seja julgado como verdadeiro por quem o analisa.

Sendo a etapa de *DM* a mais importante do processo *KDD*, é evidente que a busca por métodos eficientes são exigidos à medida que os bancos de dados se mostram cada vez maiores em relação número de informações que podem armazenar, sem muitas vezes explicitar as relações existentes entre os vários atributos.

## 3. Extração de regras para classificação de padrões por meio de Algoritmos Genéticos

Para Carvalho (2005) existem pelos menos duas abordagens por AG para a descoberta de regras para classificação de padrões: *Abordagem de Michigan,* onde cada indivíduo representa uma única regra. Assim, a população é o conjunto de regras para o problema; e *Abordagem de Pittsburgh,* onde cada indivíduo da população representa um conjunto de regras para o problema.

Na primeira abordagem pode-se destacar os sistemas classificadores, cujo problema principal é a solução para o funcionamento do AG, pois a avaliação da aptidão dos indivíduos é, na verdade, a avaliação de toda a população.

A segunda abordagem é baseada no AG clássico, uma vez que a população consiste em diferentes soluções possíveis que competem entre si. Conforme Navas e Rouco (2009) esta abordagem admite dois tipos de representação dos indivíduos que influenciam na definição dos operadores:

**Clássica:** os indivíduos são representados pela forma binária podendo-se, assim, aplicar os operadores clássicos. No entanto, nesta representação os tamanhos dos indivíduos na população não precisam ser fixos.

**Ad-hoc:** a codificação é mais próxima ao problema real precisando, no entanto, de operadores especiais. Ao desenvolver os operadores para este tipo de codificação, estes podem ser mais complexos, mas são mais eficientes por serem operadores específicos para o problema em questão. Outro fator importante para a utilização desta codificação é a fácil hibridação do algoritmo, que pode ocorrer no desenvolvimento dos operadores genéticos.

O trabalho de Navas e Rouco (2009) apresenta um sistema denominado GABIL (*Genetic Algorithm-Based Inductive Learning* ou Algoritmo Genético Baseado em Aprendizagem Indutiva) que utiliza a abordagem Pittsburgh. Neste sistema, os indivíduos são conjuntos de regras com codificação binária, mas com tamanho variável. Os operadores genéticos utilizados são os clássicos, mas com pequenas adaptações devido à variação do tamanho de cada indivíduo.

A representação comumente utilizada consiste na disjunção de conjuntos de regras. Cada regra que compõe o indivíduo consiste, na parte esquerda, da conjunção de "validações" e, na parte direita, a classe (JONG *et al.*, 1993).

O conceito de "validações" refere-se ao teste realizado para cada atributo: se o valor do atributo e do exemplo está no conjunto dado de valores, então é verdadeiro (1), senão falso (0). Não se perde a generalidade ao impor uma validação para cada atributo de cada classe, uma vez que as regras são compostas de formas conjuntivas com disjunção interna (pode-se ter mais um valor para cada atributo).

Para melhor entendimento do sistema, Navas e Rouco (2009) ilustram o procedimento com o seguinte exemplo: seja o conjunto de valores permitidos para cada atributo Cor {vermelho, amarelo, verde} e para a Forma {esfera, cubo, pirâmide, cone, cilindro}.

A representação do indivíduo na figura 02, a seguir, representa o conjunto de regras da figura 03.

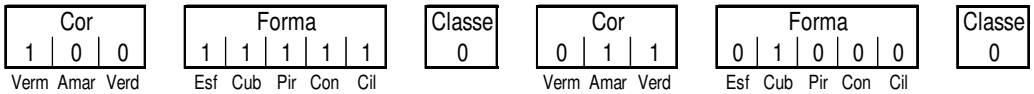


**Figura 02** - Representação do indivíduo (Navas e Rouco, 2009)

**Se** (cor= vermelho) **Então** classe=0
**Se** (cor=amarelo **ou** verde) **E** (forma=cubo) **Então** classe=0

**Figura 03** - Conjunto de regras (Navas e Rouco, 2009)

A função *fitness* evolui para conjuntos de regras que são completos e coerentes: F(indivíduo *i*) = (porcentagem de acertos)$^2$ . Com isso, podemos verificar que o sistema GABIL utiliza o formato básico do AG para resolver problemas complexos, como a extração de regras para classificação de padrões.

Na próxima seção é apresentado o algoritmo proposto que tem sua essência no sistema GABIL.

## 4. Algoritmo proposto para seleção de atributos e extração de regras para classificação de padrões

A metodologia utilizada inicialmente neste trabalho, a fim de se extrair regras para classificação de padrões e, simultaneamente, selecionar atributos por meio de AG está baseada no Sistema GABIL, diferindo apenas pelo fato dos indivíduos possuírem tamanhos fixos, pois assim o algoritmo proposto se torna simples para implementação ou pode ser resolvido por diversos *softwares*.

No entanto, por estarmos selecionando atributos, além da extração de regras, ao codificamos a solução em linguagem usual, e não a binária, observa-se que os indivíduos não possuem tamanhos iguais.

O algoritmo proposto consiste basicamente dos passos contidos no pseudocódigo da figura 04. As variáveis apresentadas no algoritmo indicam: q(i) = quantidade de registros na classe i; cl = classe que possui maior número de registros.

*Enquanto for possível extrair regras*
*Atualizar: q(i) e cl*
*For i=1 to 4*
*Aplicar Algoritmo Genético*
*Guardar a solução se for melhor, maior fitness, que anterior ou se i=1*
*Critério: 1º. A solução classificou corretamente algum registro e não classificou incorretamente nenhum registro;*
*2º. A solução classificou corretamente algum registro, mas há registro classificado incorretamente;*
*3º. Qualquer solução que não satisfaça os critérios acima é desprezada.*
*Next*
*Se o algoritmo genético apresentou solução, então é possível extrair a regra, continue. Caso contrário pare.*
*Guarde a regra obtida (solução do AG)*
*Extrair do banco de dados todos os registros classificados (corretamente ou incorretamente) pela regra gerada, permanecendo apenas os registros que a regra não conseguiu classificar.*
*Volte*

**Figura 04** - Pseudocódigo do algoritmo proposto

Quanto ao pseudocódigo apresentado acima, o AG é aplicado quatro vezes na extração de cada regra com a finalidade de buscar a melhor solução conforme os critérios descritos na sequencia do algoritmo. Isto é necessário, uma vez que ao aplicar o AG, por ser uma

metaheurística, não se pode garantir que a solução encontrar é a solução ótima, ou seja, a melhor solução entre todas. Além disto, com os critérios descritos, as regras sempre classificaram ao menos um elemento da classe em questão.

### 4.1. A modelagem do AG no algoritmo proposto

A representação do indivíduo é composta de dois grupos, o primeiro representa os atributos (um bit para cada atributo) e o total de bits do segundo grupos é igual ao total de categorias dos atributos.

Como forma de ilustração, temos na figura 05 três atributos (Cor, Material, Forma), sendo que o primeiro possui três categorias (Azul, Verde, Amarelo), o segundo atributo com duas categorias (Aço e Madeira) e o terceiro atributo com cinco categorias (Círculo, Quadrado, Retângulo, Triângulo e Trapézio).

Supondo que estamos extraindo regra para classificação de padrões e selecionando atributos para uma classe X, o indivíduo (figura 05) seleciona apenas os atributos Material e Forma, definindo a seguinte regra:

**Se** *Material* (aço) **e** *Forma* (retângulo **ou** triângulo) **então** *Classe X*

| Atributos | | | Categorias dos Atributos | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *Cor* | | | *Material* | | *Forma* | | | | |
| Cor | Material | Forma | Azul | Verde | Amarelo | Aço | Madeira | Círculo | Quadrado | Retângulo | Triângulo | Trapézio |
| 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |

**Figura 05** – Exemplo de indivíduo na modelagem proposta

Para verificar a aptidão do indivíduo da figura 05, não são levados em consideração os bits das categorias do atributo Cor no grupo 2, uma vez que este atributo não está ativado no grupo 1.

Esta codificação permite a aplicação dos operadores tradicionais de mutação e de *crossover*, sendo utilizado nestes experimentos o *crossover* de "1 ponto". Já a função *fitness* ficou definida por:

$$ECC_k/TEC_k - ECC_y/TEC_y$$

onde:

*K* é a classe que possui maior quantidade de registros em cada iteração.
$ECC_k$ = quantidade de exemplos classificados corretamente da classe *k.*
$TEC_k$ = quantidade de exemplos da classe *k.*
$ECC_y$ = quantidade de exemplos classificados incorretamente.
$TEC_y$ = quantidade de exemplos, excluindo os exemplos da classe k.

O *software* utilizado para os testes foi o MatLab versão 7, do qual se fez uso do *toolbox* para AG que, segundo o tutorial, utiliza todas as técnicas de seleção, operadores e outros elementos do AG foram implementados conforme Goldberg (1989).

## 5. Aplicações do algoritmo proposto na extração de regras para distinção de cálculo e câncer no duto biliar

A fim de situar o leitor quanto às etapas do processo *KDD*, estas estarão indicadas no decorrer desta seção com caracteres em negrito

A **seleção dos dados** para este trabalho foi realizada através de pesquisa em históricos médicos apresentados por Steiner *et al.* (2006). Este banco de dados possui informações de 118

pacientes, sendo que comprovadamente 35 possuíam câncer no duto biliar e 83, cálculo. Alguns destes registros são apresentados no quadro 01 a seguir.

QUADRO 01 – **ALGUNS REGISTROS DO PROBLEMA EM QUESTÃO**

| ID | SE | BT | BD | BI | FA | SGOT | SGPT | TAP | AL | AMI | CR | LE | VG | DI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 46 | F | 41,8 | 21,10 | 20,60 | 234 | 178 | 646,25 | 92 | 14 | 3,3 | 0,80 | 9,0 | 36,8 | CF |
| 52 | M | 21,4 | 12,95 | 8,45 | 55 | 80 | 229,57 | 92 | 12 | 3,5 | 0,55 | 7,8 | 40,6 | CF |
| 73 | M | 26,2 | 13,60 | 12,60 | 90 | 97 | 116,38 | 104 | 14 | 2,7 | 0,80 | 12,6 | 32,3 | CF |
| 47 | M | 31,6 | 16,50 | 15,40 | 31 | 59 | 174,46 | 92 | 12 | 3,0 | 0,70 | 11,40 | 39,0 | CF |
| 39 | F | 7,9 | 5,09 | 2,86 | 20 | 27 | 158,62 | 219 | 12 | 3,00 | 3,30 | 10,3 | 33,6 | CA |
| 66 | F | 4,0 | 2,00 | 2,00 | 20 | 53 | 285,00 | 76 | 18 | 2,80 | 1,20 | 10,7 | 44,0 | CA |
| 29 | F | 1,0 | 0,50 | 0,50 | 104 | 57 | 370,00 | 189 | 15 | 3,50 | 0,80 | 8M3 | 44,2 | CA |

*ID: Idade; SE: Sexo; BT: Bilirrubina total; BD: Bilirrubina direta; BI: Bilirrubina Indireta; FA: Fosfatases alcalinas; TAP: Tempo de atividade da protrombina; ALB: Albumina; AMI: Amilase; CR: Creatinina; LE: Leucócitos; DI: Diagnóstico (CF: Câncer no duto biliar; CA: Cálculo)*

Para a aplicação do algoritmo, foram necessárias algumas **transformações nos dados**, como por exemplo, a categorização dos atributos. Esta categorização é apresentada no quadro 02, num total de 40 categorias distribuídas nos 14 atributos, com os respectivo número de registros em cada categoria.

QUADRO 02 – **ATRIBUTOS CATEGORIZADOS**

**(CONTINUA)**

| Atributo | Categorias | Qtd. |
|---|---|---|
| Idade | Criança/ Adolescente – 0 aos 19 anos | 1 |
| | Adulto Jovem - 20 aos 40 anos. | 35 |
| | Adulto – 41 aos 65 anos. | 54 |
| | Idoso - a partir dos 66 anos. | 28 |
| Sexo | Masculino | 71 |
| | Feminino | 47 |
| Bilirrubina Total (bt) (1) | Até 1,2 mg/100 ml | 2 |
| | Superior a 1,2mg/100ml | 116 |
| Bilirrubina Direta (bd) (1) | Até 0,4 mg/100 ml | 2 |
| | Superior a 0,4mg/1ooml | 116 |
| Bilirrubina Indireta (bi) (3) | Abaixo de 0,20 mg/dl | 0 |
| | 0,20 a 0,80 mg/dL | 7 |
| | Superior a 0,8 mg/dl | 111 |
| Fosfatases alcalinas (fa) (1) | Até 50 U/l | 54 |
| | 50 a 136 U/l | 39 |
| | Superior a 136 U/l | 25 |
| SGOT (sgot) (1) | Até 15 U/l | 1 |
| | 15 a 37 U/l | 20 |
| | Superior a 37 U/l | 97 |
| SGPT (sgpt) (1) | Até 30 U/l | 2 |
| | 30 a 65 U/l | 5 |
| | Superior a 65 U/l | 111 |
| Tempo de atividade da protrombina (tap) (1) | Até 75% de atividade | 23 |
| | 75 a 100% de atividade | 39 |
| | Superior a 100% de atividade | 56 |
| Albumina (alb) (1) | Até 3,5 g/100ml | 0 |
| | 3,5 a 5,5 g/100 ml | 0 |
| | Superior a 5,5 g/100ml | 118 |
| Amilase (ami) (1) | Até 25 U/100ml | 118 |
| | 25 a 115 U/100 ml | 0 |
| | Superior a 11 U/100ml | 0 |

QUADRO 02 – **ATRIBUTOS CATEGORIZADOS (CONCLUSÃO)**

| Atributo | Categorias | Qtd. |
|---|---|---|
| Creatinina (cr) (1) | Até 0,2 mg/100 ml | 0 |
| | 0,2 a 0,7 mg/100 ml | 38 |
| | Superior a 0,7 mg/100 ml | 80 |
| Leucócitos (le) (2) | Até 5.000 leucócitos/ µL de sangue | 3 |
| | 5.000 a 10.000 leucócitos/ µL de | 78 |
| | Superior a 10.000leucócitos/ µL de | 37 |
| Vg (vg) (3) | Até 36% de Ht | 26 |
| | 36% a 54% de Ht | 92 |
| | Superior a 54% de Ht | 0 |

*FONTE: (1) Labes (2011); (2) Bonelli (2011); (3) Shinohara (2011) .*

Uma breve análise dos dados nos levou a desconsiderar dois dos atributos (Albumina e Amilase), **limpeza dos dados**, pois todos os registros pertencem a uma única categoria de cada atributo. Desta forma, são utilizados na execução do algoritmo 12 atributos e 34 categorias.

Os indivíduos foram codificados conforme apresentado na seção 4.1. Assim, cada indivíduo possui 46 bits, dos quais os 12 primeiros referem-se aos atributos e os demais 34, referem-se às categorizações realizadas nos atributos. A figura 06 apresenta o primeiro registro do quadro 01, já codificado em forma binária para aplicação do AG.

**0010010101001001001001010001100001010010101**

FIGURA 6 – EXEMPLO DE INDIVÍDUO CODIFICADO PARA APLICAÇÃO DO AG

Na figura acima os quatro primeiro bits, em vermelho referem-se à idade, como o terceiro bit deste conjunto está ativo, indica pelo quadro 02 que é “Adulto – 41 aos 65 anos”. Os próximos dois bits, em azul, referem-se ao sexo, o segundo bit está ativo indicando ser do sexo “feminino”. Esta análise pode ser realizada para os próximos grupos considerando o quadro 01 e quadro 02. Já o último bit indica a classe pertencente, neste caso é um paciente com câncer no duto biliar.

Depois de codificados ocorreu a aplicação da técnica de ***Data Mining***, neste caso o AG, o quadro 03 apresenta as regras obtidas no “treinamento” com 94 registros, sendo 28 referentes a pacientes que tiveram câncer e 66 referentes a pacientes que tiveram cálculo. Os demais registros foram utilizados para testar as regras: sete referentes a câncer e 17 referentes a cálculo.

QUADRO 03 – **REGRAS OBTIDAS COM O CONJUNTO DE TREINAMENTO (CONTINUA)**

| Ordem aplica. | Regra | Então | Classe | Classificação Certa | Classificação Errada | Sem Classif. |
|---|---|---|---|---|---|---|
| 1 | Se Bi (superior a 0,20 mg/dl) e se Fa (até 50 U/l) e se SGOT (até 37 U/l) e se Le (superior a 5.000 leucócitos/µL de sangue) | **então** | *Cálculo* | 16 | 0 | 78 |
| 2 | Se Idade (de 20 a 65 anos) e se Sexo (masculino) e se Fa (50 a 136 U/l) e se SGOT (superior a 37 U/l) e se SGPT (superior a 65 U/l) e se TAP (até 75% de atividade ou superior a 100% de atividade) e se CR (superior a 0,2 mg/100ml) | **então** | *Cálculo* | 16 | 0 | 62 |
| 3 | Se Idade (superior a 20 anos) e se Sexo (masculino) e se Bi (superior a 0,8 mg/dl) e se Fa (até 50 Ul) e se Le (5.000 a 10.000 leucócitos/ µL de sangue) e se Vg (36% a 54% de Ht) | **então** | *Cálculo* | 8 | 0 | 54 |

QUADRO 03 – **REGRAS OBTIDAS COM O CONJUNTO DE TREINAMENTO (CONCLUSÃO)**

| Ordem aplica. | Regra | Então | Classe | Classificação Certa | Classificação Errada | Sem Classif. |
|---|---|---|---|---|---|---|
| 4 | Se Idade (superior a 40 anos) e se Bi (superior a 0,8 mg/dl) e se SGOT (superior a 37 U/l) e se SGPT (superior a 65 U/l) e se Cr (superior a 0,2 mg/100ml) e se Vg (até 36% de Ht) | **então** | *Câncer no fígado* | 8 | 0 | 46 |
| 5 | Se Idade (20 a 40 anos) e se Bi (superior a 0,20 mg/dl) e se SGOT (superior a 37 U/l) e se SGPT (superior a 65 U/l) e se Cr (superior a 0,7 mg/100ml) e se Vg (36% a 54% de Ht) | **então** | *Cálculo* | 7 | 0 | 39 |
| 6 | Se Fa (50 a 136 U/l) e se SGPT (superior a 65 U/l) e se TAP (superior a 75% de atividade) e se Cr (superior a 0,7 mg/100ml) e se Le (superiro a 10.000 leucócitos/ µL de sangue) | **então** | *Câncer no duto biliar* | 9 | 0 | 30 |
| 7 | Se Sexo (feminino) e se Bt (superior a 12 mg/100ml) e se Fa (50 a 136 U/l) e se Le (superior a 10.000 leucócitos/ µL de sangue) e se Vg (36% a 54% de Ht) | **então** | *Cálculo* | 7 | 0 | 23 |
| 8 | Se sexo (masculino) e se TAP (até 75% de atividade ou superior a 100% de atividade) e se Le (superior a 10.000 leucócitos/ µL de sangue) | **então** | *Câncer no duto biliar* | 3 | 0 | 20 |
| 9 | Se sexo (masculino) e se Bd (superior a 0,4mg/100ml) e se Bi (superior a 0,8 mg/dl) e se SGOT (superior a 37 U/l) e se Cr (superior a 0,7 mg/100ml) | **então** | *Cálculo* | 4 | 0 | 16 |
| 10 | Se Cr ( até 0,2 mg/100ml) e se Le (superior a 10.000 leucócitos/ µL de sangue) e se Vg (36% a 54% de Ht) | **então** | *Câncer no duto biliar* | 3 | 0 | 13 |
| 11 | Se Le (inferior a 5.000leucócitos/ µL de sangue ou superior a 10.000 leucócitos/ µL de sangue) | **então** | *Cálculo* | 3 | 0 | 10 |
| 12 | Se Idade (inferior a 41 anos) e se TAP (superior a 100% de atividade) | **então** | *Cálculo* | 2 | 0 | 8 |
| 13 | Se idade (20 a 40 anos ou superior a 65 anos) e se SGPT (superior a 65 U/l) e se Le (de 5.000 a 10.000 leucócitos/ µL de sangue) | **então** | *Câncer* | 2 | 0 | 6 |
| 14 | Se Sexo (masculino) | **então** | *Câncer no duto biliar* | 1 | 0 | 5 |
| 15 | Se Idade (41 a 65 anos) e se Bt (superior a 1,2 mg/100ml) e se Bd (superior a 0,4 mg/100ml) e se TAP (de 75% a 100 de atividade) | **então** | *Cálculo* | 1 | 0 | 4 |
| 16 | Caso contrário | Não há conclusão | | | 4 | |

A primeira regra obtida com o algoritmo proposto (quadro 03), quando apresentados os 94 registros, classificou corretamente 16 registros e nenhum errado - neste momento o algoritmo procurou regra para classificar a classe "cálculo" visto que do total de registros esta classe possuía 66 e a outra (câncer no duto biliar) 28. A segunda e terceira regras também obtiveram classificação correta de registros e ambas são regras para a classe "Cálculo", uma vez que para a

obtenção da segunda regra são apresentados os 50 registros de cálculo (não classificados na primeira regra) e 28 de câncer no duto biliar, e na terceira tem-se 34 registros (não classificados na segunda regra) de pacientes com cálculo e 28 com câncer fígado.

Para obter a quarta regra, são apresentados ao algoritmo 26 registros de pacientes com cálculo (os demais – 40 – foram classificados pelas regras 1 a 3) e 28 de pacientes com câncer no duto biliar. Como esta última classe é a que possui maior quantidade, a regra quatro classifica registros deste tipo (Câncer no duto biliar). Assim, esta análise pode ser realizada para as demais regras até a de número 15.

Ao aplicar o algoritmo para a obtenção da regra de número 16, foram apresentados quatro registros, dois referentes a câncer no duto biliar e dois de cálculo. Nesta etapa o algoritmo obteve regras que classificavam os quatro registros como cálculo ou como câncer no duto biliar, assim não foi possível chegar a uma conclusão e definição para tal regra.

**Analisando o resultado** obtido com o algoritmo no "treinamento" tem-se que as regras classificam corretamente 90 registros, não classifica erroneamente nenhum registro e não consegue classificar quatro registros. Assim, tem-se 95,75% de eficácia.

Utilizando os 20% de registros restantes para testar as regras (quadro 04), afirmamos que a eficácia do algoritmo é de 70,83%, o que representa um desempenho relativamente bom, devido ao número de registros presente no banco de dados (118). Esse total corresponde a apenas $6{,}9 \times 10^{-9}\%$ de todas as possíveis combinações conforme a modelagem do AG utilizada, onde todos os indivíduos possuem 34 bits (não considerando os bits referentes aos atributos), ou seja, $2^{34}$ possíveis combinações.

QUADRO 04 – **APLICAÇÃO DAS REGRAS DO CONJUNTO DE TESTE**

| **Regra** | **Classe correta** | **Classificou** | | **Não Classifica** |
|---|---|---|---|---|
| | | **Certo** | **Errado** | |
| 1 | Cálculo | 4 | 0 | 20 |
| 2 | Cálculo | 5 | 0 | 15 |
| 3 | Cálculo | 1 | 1 | 13 |
| 4 | Câncer | 1 | 0 | 12 |
| 5 | Cálculo | 0 | 0 | 12 |
| 6 | Câncer | 0 | 0 | 12 |
| 7 | Cálculo | 0 | 0 | 12 |
| 8 | Câncer | 0 | 0 | 12 |
| 9 | Cálculo | 1 | 0 | 11 |
| 10 | Câncer | 1 | 2 | 8 |
| 11 | Cálculo | 0 | 0 | 8 |
| 12 | Câncer | 0 | 0 | 8 |
| 13 | Câncer | 2 | 2 | 4 |
| 14 | Câncer | 0 | 1 | 3 |
| 15 | Cálculo | 0 | 1 | 2 |
| 16 | Não classif. | | 2 | |
| **Total** | | **15** | **7** | **2** |

## 6. Considerações Finais

Ao realizarmos os experimentos, utilizamos o processo *KDD*: seleção dos dados; pré-processamento e transformação dos dados (dois atributos descartados no problema médico e codificação dos atributos em indivíduos para o AG); utilização de *Data Mining*, através da aplicação de AG e, finalmente, interpretação dos resultados, obtendo assim as regras de classificação que podem auxiliar o profissional na tomada de decisão no tratamento do pacientes, por como exemplo, solicitação de exames específicos para cada doença.

Quanto ao algoritmo proposto, este se difere dos demais apresentados na literatura pelo fato de extrair as regras de classificação e, simultaneamente, selecionar os atributos utilizando AG, tornando as regras mais simples.

Em trabalhos futuros pretendemos utilizar banco de dados com maior número de registros, pois, como já comentado, o total de registros utilizados corresponde a apenas 6,9x10-9% de todas as possíveis combinações, conforme a modelagem do AG aqui proposta.

Tal algoritmo também foi aplicado à outra base dados (área elétrica) com classificação correta de 98,61% dos registros no treinamento. No entanto, no teste o algoritmo classificou 57% dos registros apresentados, mas mesmo assim, é um bom resultado, uma vez que neste problema também havia poucos registros considerando o total de possíveis soluções.

Como próxima etapa para este trabalho pretende-se hibridizar o AG com outras metaheurísticas como, por exemplo, com Redes Neurais Artificiais, Busca Tabu, dentre outras, realizando comparações dos resultados com outras técnicas de classificação, verificando a eficácia e eficiência do algoritmo aqui proposto.

## Referências

**Bonelli, (2011)** Tipos de Exames. Disponível em <http://www.bonelli.com.br/paginas/asp/relativo/pgnExames_tipos_b.asp> Acessado em 18 de fev. de 2011.

**Brachman, R. J.; Anand, T.** (1994), The Process of Knowledge Discovery in Databases: A First Sketch. KDD Workshop: 1-12.

**Carvalho, D. R.** (2005) *Árvore de Decisão / Algoritmo Genético para tratar o Problema de Pequenos Disjuntos em Classificação de Dados.* Rio de Janeiro. Tese (Doutorado em Ciências em Engenharia Civil – Universidade Federal do Rio de Janeiro).

**Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R.** (1995) *Advances in Knowledge Discovery & Data Mining.* 1 ed. American Association for Artificial Intelligence, Menlo Park, Califórnia, 1995.

**Frawley, W. J.; Piatetsky-Shapiro, G.; Matheus, C. J.** (1991) *Knowledge Discovery in Databases - An Overview.* In: Knowledge Discovery in Databases, p. 1--30.

**Goldberg, D. E.** (1989) *Genetic algorithms in search, optmization, and machine learning.* Addison-Wesley Publishing Company, Inc. Massachusetts.

**Han, J.; Kamber, M.** (2006) *Data Mining: Concepts and Techniques.* 2a ed. Morgan Kauffmann Publishers.

**Jong, K. A., Spears. W. M., Gordon, D. F.** (1993) *Using Genetic Algorithms for Concept Learning. Computer Science - Machine Learning.* Volume 13, Numbers 2-3, p. 161-188. November.

**Labes, (2011)** Hemograma completo. Disponivel em <http://www.labes.com.br/hemograma_completo.htm> Acessado em 18 de fev. de 2011.

**Navas, R. P.; Rouco, F. J. M.** *Algoritmos Genéticos. Facultad de Informática de Sevilla.* España. Disponível em <http://www.cs.us.es/~delia/sia/html98-99/pag-alumnos/web11/indice.html> Acessado em 09 de mai. 2009.

**Shinohara, E. M. G. (2011)** Células Sanguíneas. Material de aula da disciplina de Hematologia Clínica – USP. Disponível em < http://www.fcf.usp.br/Ensino/Graduacao/Disciplinas/Exclusivo/Inserir/Anexos/LinkAnexos/C%E9lulas%20sangu%EDneas.pdf> Acessado em 18 de fev. de 2011.

**Steiner, M. T. A.; Soma, N. Y.; Shimizu, T.; Nievola, J. C.; Steiner Neto, P. J.** (2006) *Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados.* Revista Gestão & Produção, v.13, n.2, p.325-337, mai.-ago. 2006

**Weiss, S.; Indurkhya, N.** (1998) *Predictive Data Mining: a pratical guide.* Morgan Kauffmann Publishers, Inc.