ALGORITMOS GENÉTICOS APLICADOS EM *DATA MINING* PARA OBTENÇÃO DE REGRAS SIMPLES E PRECISAS

Clay Rulliam dos Santos Miranda, Gina Maira Barbosa de Oliveira, Jamilson Bispo dos Santos

Universidade Prebiteriana Mackenzie

Rua da Consolação, 930 - CEP 01302-907 - São Paulo - SP - Brasil

Resumo No presente trabalho, desenvolveu-se um ambiente computacional que utiliza Algoritmos Genéticos como ferramenta de *Data Mining*, com o objetivo de obter regras de classificação precisas e compreensíveis. Baseado no modelo proposto por Fidelis, Lopes e Freitas (2000), realizamos experimentos para a descoberta de regras de diagnóstico em uma base de dados de doenças dermatológicas, objetivando mensurar a qualidade e a compreensibilidade das regras obtidas. Para isso, a avaliação das regras foi composta por duas parcelas, sendo que a primeira representa o grau de precisão das regras e a segunda o grau de simplicidade das mesmas. Nos experimentos, a importância relativa de cada parcela na avaliação total da regra foi alterado e os resultados analisados.

Palavras Chaves: Algoritmos Genéticos, *Data Mining*, classificação, predição.

Abstract: In the present work, we developed a computational environment that uses Genetic Algorithms as Data Mining tools aiming to obtain accurate and comprehensible classification rules. Based on a model proposed by Fidelis, Lopes and Freitas (2000), some experiments have been performed to discover diagnostic rules in a dermatological diseases database. The purpose was to obtain rules for quality and comprehensibility measurements. The rule evaluation was composed by two components. The first one, representing the accuracy level of the rules and the second one their comprehensibility level. In our experiments, the relative importance of each component in the total rule evaluation was modified and the results have been analyzed.

Keywords: Genetics Algorithms, Data Mining, classification, prediction.

1 INTRODUÇÃO

Com a evolução da tecnologia da informação, cada vez mais dados são armazenados pelas organizações. No entanto, é preciso encontrar formas eficientes de extrair os conhecimentos estratégicos implícitos nestes dados. *Data Mining* (DM) é a etapa principal de um processo maior chamado de "descoberta de conhecimento em bancos de dados" (Knowledge Discovery in Databases) [1]. DM consiste num processo de busca de conhecimentos novos, úteis e interessantes em base de dados, que possam ser utilizados em áreas como teoria da decisão, estimação, predição e previsão.

Neste trabalho, foi desenvolvido um ambiente computacional que utiliza Algoritmos Genéticos (AGs) aplicados em *Data Mining*, para obtenção de regras de classificação simples e precisas.

O modelo que serviu de base para nossa investigação foi proposto em [2], no qual um AG foi utilizado como ferramenta de DM na descoberta de regras de diagnóstico de doenças dermatológicas. O ambiente que desenvolvemos faz uma investigação mais específica no método de avaliação, para mensurar o quanto as regras obtidas são simples, além de precisas. A base de dados utilizada para validar os experimentos contém o histórico de diagnósticos para identificar a doença do grupo Erythemato-Squamous [12], que possui 6 classes, e é um problema real da área dermatológica.

Data Mining tem como objetivo a busca de conhecimento em bancos de dados sendo que diferentes métodos podem ser empregados nesta busca. A utilização de Algoritmos Genéticos é justificável por estes apresentarem resultados eficientes quando temos um problema em que o espaço de busca é grande, como nas tarefas de classificação. A tarefa de classificação, mais precisamente, a obtenção de regras de predição do tipo SE-ENTÃO, foi pesquisada neste trabalho. E dentre as propriedades que podemos utilizar para avaliar as regras obtidas, que são grau de precisão, de compreensão e de interesse para o usuário, consideramos a primeira e segunda.

Na próxima seção, os principais conceitos sobre *Data Mining* são apresentados. Na seção 3, são apresentados os AGs e como eles podem ser utilizados na tarefa de descoberta de regras. O modelo original no qual este trabalho foi baseado é apresentado na seção 4. A secão 5 discute as propriedades desejáveis nas regras a serem obtidas e como mensurar a compreensibilidade da regra. Os resultados dos experimentos e suas análises são demonstrados na seção 6. Enfim, na última seção são apresentadas as conclusões e propostas para trabalhos futuros.

2 *DATA MINING*

O processo de identificação e extração das informações em bases de dados que possam ser úteis é conhecido como

mineração de dados ou *Data Mining* (DM) [3]. O objetivo do *Data Mining* é automatizar a descoberta dos conhecimentos que estão nas bases de dados e não são perceptíveis.

O uso do DM em uma base de dados é considerado o passo central de um processo maior chamado Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases), que inclui vários outros processos, que podem ser divididos em pré-processamento e pós-processamento [1].

O pré-processamento possui algumas etapas, tais como, integração, limpeza e discretização dos dados e a seleção dos atributos relevantes para a tarefa de *Data Mining*. O pós-processamento tem como objetivo aprimorar a compreensão dos resultados obtidos com o DM. No caso deste trabalho, no qual visamos a descoberta de regras do tipo SE-ENTÃO precisas e compreensíveis, poderíamos utilizar um tipo de pós-processamento para filtrar as regras mais interessantes.

Existem diferentes tarefas de *Data Mining*, sendo que cada uma tem como objetivo buscar um determinado tipo de conhecimento que está associado a um problema específico. As principais tarefas executadas pelos algoritmos de DM são: classificação, *clustering*, associação e modelos de dependência.

A tarefa estudada neste trabalho é a classificação e consiste na tarefa de descobrir um relacionamento entre os atributos previsores e o atributo meta, usando registros históricos nos quais as classes são conhecidas. Nesta tarefa, o objetivo é predizer o valor, ou seja a classe, de um atributo meta especificado pelo usuário, baseado nos valores dos atributos previsores [3]. Esta tarefa pode ser considerada um tipo particular de predição de regras. Tal regra é constituída de uma "parte SE", o antecedente da regra, que constitui um conjuntos de condições dos atributos de predição e de uma "parte ENTÃO", o consequente da regra, que possui o valor de predição para o atributo meta.

Para alcançarmos o objetivo do DM, é necessária a aplicação de técnicas ou algoritmos que possam realizar a tarefa de extração de informações das bases de dados. Estas técnicas estão diretamente ligadas ao tipo de tarefa onde se deseja aplicar o DM, ou seja, ao tipo de problema em questão. Podemos destacar como algumas das principais técnicas utilizadas em *Data Mining*, as Redes Neurais, as Árvores de Decisão e os Algoritmos Genéticos, que utilizamos neste trabalho e que são apresentados na próxima seção.

3 ALGORITMOS GENÉTICOS

O Algoritmo Genético (AG) é uma metáfora biológica idealizada por John Holland na década de 60 [13]. Baseado nas teorias evolutivas de Charles Darwin e R. A. Fischer, o AG utiliza a evolução como meio poderoso de executar funções de otimização em um computador. Holland procurou especificar um modelo no qual os mecanismos de adaptação natural pudessem ser transportados para um sistema computacional.

Como a base para a teoria dos Algoritmos Genéticos está diretamente associada aos conceitos biológicos, adotou-se os mesmos termos da biologia na área computacional. O cromossomo, ou indivíduo, é formado por uma cadeia de símbolos que representa uma solução possível para o problema em questão. O cromossomo pode ser representado de forma binária, inteira ou real, de acordo com o tipo de problema que se deseja resolver. A essa representação denominamos alfabeto do AG. A representação de cada parâmetro de acordo com o alfabeto adotado é chamado de gene. A codificação do cromossomo representa o genótipo do indivíduo. Podemos definir a população como um conjunto de pontos no espaço de busca utilizado, representada por um conjunto de indivíduos.

Uma geração representa uma iteração completa do Algoritmo Genético, que resulta em uma nova população. A aptidão é o valor da função objetivo para um indivíduo, que representa sua adaptação como solução ótima para o problema estudado.

Um AG se inicia com a geração aleatória de uma população com n indivíduos, que representam possíveis soluções para o problema, ou seja, alguns pontos no espaço de busca. Para cada indivíduo, calcula-se a aptidão através da função objetivo, que mensura a adaptação do indivíduo como solução do problema. Verifica-se então se os critérios de término do problema já foram atingidos, sendo que estes critérios geralmente são a aptidão atingida pelo melhor indivíduo, em conjunto com o número de gerações.

O processo de seleção é composto basicamente de dois métodos: crossover e mutação. No primeiro, escolhem-se x pares de indivíduos, baseado na taxa de crossover, e, para cada par, sorteiam-se pontos de crossover que permitirão que novos indivíduos sejam gerados. A estes novos indivíduos gerados no processo anterior é aplicada uma mutação, também especificada por uma taxa probabilística.

Após os métodos de crossover e mutação, calcula-se a função objetivo desses novos indivíduos e selecionam-se aqueles com melhor adaptação entre os indivíduos da população inicial, e os novos indivíduos gerados. Os selecionados constituirão a população na próxima geração do AG.

Estes passos se repetem até que a condição de término seja satisfeita. O resultado apresentado é o melhor indivíduo, ou seja, o que possui maior aptidão para a solução do problema proposto.

Para a tarefa de descoberta de regras alguns aspectos dos AGs são relevantes, tais como, a representação do indivíduo, os operadores genéticos e a função de avaliação.

Para representarmos as regras do tipo SE - ENTÃO, tem-se duas abordagens: Michigan e Pittsburgh [1]. Na abordagem Michigan, cada indivíduo representa uma única regra e a população representa um conjunto de regras. Na abordagem Pittsburgh, cada indivíduo representa um conjunto de regras.

No modelo proposto por Fidelis, Lopes e Freitas [2], que serviu de base para esta pesquisa, os autores adotaram a abordagem Michigan.

4 MODELO DE FIDELIS, LOPES E FREITAS

O modelo de mineração de regras investigado neste trabalho é baseado em [2], no qual um AG foi aplicado em uma base de dados médica, com o objetivo de obter regras capazes de diagnosticar uma doença de um paciente, baseado em seus sintomas. A base de dados utilizada em [2], e também no presente trabalho, contém o histórico de identificação do grupo de doença Erythema-Squamous e possui 366 registros [12]. Cada registro é composto por 34 campos de atributos, sendo estes atributos avaliados numa escala de 0 a 3, onde 0 indica que o aspecto não está presente, 3 grande possibilidade de presença e os valores 1 e 2 indicam valores intermediários. Existem duas exceções: o atributo 'family history' pode ter

dois valores, 0 para nenhum caso na família e 1 se já houve algum caso, e o atributo 'age' apresenta a idade do paciente que varia entre 0 e 79. O último campo do registro refere-se à classe da doença diagnosticada pelo médico no caso clínico. Na base existem 6 classes de doença pertencentes ao grupo Erythema-Squamous: psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis e pityriasis rubra pilares, as quais no referiremos de forma simplificada como classes 1, 2, 3, 4, 5 e 6.

A representação do indivíduo e os operadores genéticos utilizados no modelo original foram mantidas neste trabalho. O cromossomo é representado seguindo a abordagem Michigan, sendo composto por 34 genes correspondentes aos atributos especificados na base de dados.

Cada gene do cromossomo corresponde à sua condição na parte SE da regra e o cromossomo todo corresponde ao antecedente da regra. A parte ENTÃO da regra não é representada no cromossomo e corresponde à classe preditiva, que é especificada na execução do AG. Ou seja, a cada execução do AG, uma única classe é utilizada como meta e é previamente especificada. Assim, para obtermos as regras para as 6 classes, são necessárias pelo menos 6 execuções do AG, uma para cada classe.

Um gene é constituído de três segmentos: peso, operador e valor, conforme a Figura 1.

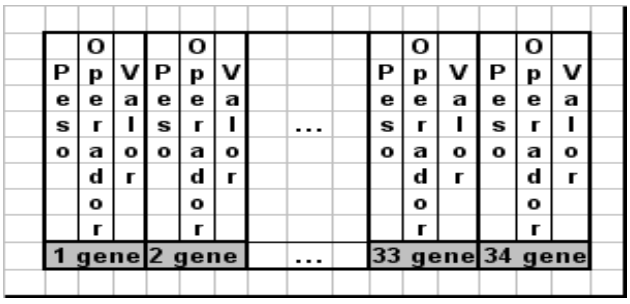

Figura 1 - Codificação do cromossomo

O segmento *peso* é definido como um valor compreendido entre 0 e 1 e é utilizado para definir se o atributo fará parte da regra ou não. Esta definição é dada por um limiar pré-definido: se o peso é menor que o limite especificado, o gene em questão não fará parte da condição da regra, e caso seja maior ou igual, a característica entra na condição da regra. O limiar estipulado no modelo original foi de 0,3, valor este que mantivemos em nossa investigação. O segmento *operador* representa o operador relacional aplicado na n-ésima condição. Os operadores possíveis são: =, ≠ , ≥ e <. O segmento *valor* representa um dos valores que os atributos podem assumir, de acordo com o seu domínio.

Os operadores genéticos utilizados foram os seguintes: torneio estocástico, de tamanho 3, como método de seleção; crossover de dois pontos, com probabilidade de 100% e elitismo como estratégica de reprodução, preservando apenas o melhor indivíduo.

Para a mutação, utilizou-se três operadores de mutação, sendo cada um deles associado aos segmentos existentes em cada gene do cromossomo. Assim, aplicou-se uma taxa de mutação de 30%, para cada tipo de mutação separadamente.

A função de avaliação utilizada em [2] busca mensurar a precisão da regra. Para isso, foram utilizados dois indicadores comumente usados em medicina: a especificidade e a sensitividade.

Considerando-se uma regra do tipo SE *A* ENTÃO *C*, onde *A* é o antecedente e *C* o conseqüente, uma forma de medir a precisão da regra é calcular o fator de sensitividade da regra. Primeiramente, deve-se calcular o número de falsos positivos, falsos negativos, verdadeiros negativos e verdadeiros positivos, considerando-se a aplicação da regra em todos os registros de uma base, conforme a Tabela 1. A classe preditiva do exemplo é C se o exemplo satisfaz A.

		Classe presente	
		C	não C
Classe	C	*Tp*	*Fp*
preditiva	não C	*Fn*	*Tn*

Tabela 1 -Falsos positivos (Fp), falsos negativos (Fn), verdadeiros negativos (Tn) e verdadeiros positivos (Tp)

Na tabela, Tp representa o número de registros que satisfazem A e C. Fp representa o número de registros que satisfazem A mas não satisfazem C. Fn representa o número de registros que não satisfazem A mas satisfazem C. Tn representa o número de registros que não satisfazem A nem C. Assim, a sensitividade preditiva pode ser calculada através da equação:

$$Sp = \frac{Tp}{Tp + Fp} \qquad (1)$$

A especificidade da regra pode ser medida determinando-se o quanto a regra é completa, que é o número de exemplos que possuem a classe preditiva C e que são de fato conferidos pela regra antecedente. Assim, a especificidade pode ser definida como:

$$Se = \frac{Tn}{Tn + Fn} \qquad (2)$$

Combinando-se Sp e Se, pode-se definir a função de avaliação para medir o quão a regra é precisa:

$$Avaliação = Sp * Se \qquad (3)$$

Para realizar os experimentos, a base de dados foi dividida, aleatoriamente, em duas: base de treinamento e base de teste. A primeira é composta por 2/3 dos registros da base original e é utilizada durante a execução do AG no cálculo da função de avaliação. A segunda é utlizada para mensurar a qualidade das regras descobertas [1] [2].

Os resultados publicados em [2] são reproduzidos na Tabela 2.

Classe da Doença	Rules	Fitness Treinamento	Fitness Teste
1	IF (clubbing of the rete ridges >= 1 AND	0,973	0,973
	perifollicular parakeratosis = 0)		
2	IF (koebner phenomenon = 0 AND	0,855	0,855
	vacuolisation and damage of basal layer < = 1 AND		
	spongiosis >= 2 AND)		
3	IF (band-like infiltrate >= 2)	1	0,979
4	IF (knee and elbow involvement = 0 AND	0,86	0,783
	family history = 0 AND		
	acanthosis < 3 AND		
	focal hypergranulosis < 2 AND		
	spongiosis >= 1 AND		
	inflammatory monoluclear infiltrate >= 1)		
5	IF (melanin incontinence < 1 AND	1	1
	fibrosis of the papillary dermis ≠ 0 AND		
	munro microabcess = 0)		
6	IF (follicular papules >= 1 AND	1	1
	perifollicular parakeratosis >= 1)		

Tabela 2 - Resultados publicados no experimento original [2].

5 PRECISÃO VERSUS SIMPLICIDADE

Na tarefa de DM, três aspectos com relação às regras a serem obtidas são importantes: a precisão, a compreensão e o interesse das mesmas [1].

Quanto à medida do quão a regra é interessante, pode-se dizer que existem duas abordagens básicas: subjetiva, controlada pelo usuário e objetiva, controlada pelo próprio sistema [1]. Este aspecto não foi considerado neste trabalho.

Como apresentado na seção 4, a função de avaliação utilizada em [2], visava exclusivamente a definição de regras precisas. Neste trabalho, é proposta uma análise do uso de AGs em DM com o objetivo de obter regras simples, ou seja, compreensíveis, e ao mesmo tempo precisas.

Existem vários métodos para medir o quanto a regra pode ser considerada compreensível e a seguir apresentamos dois. Em ambos, a simplicidade da regra é dada por um valor normalizado entre 0 e 1, sendo que 1 corresponde ao máximo de simplicidade possível.

Uma primeira forma de se mensurar a simplicidade das regras é dada pela proporção inversa do número de condições (*N*) que constituem a parte antecedente da regra, ou seja:

$$Si = \frac{1}{N} \qquad (4)$$

Uma segunda maneira de se mensurar a simplicidade de uma regra é dada pela equação (5), sendo *Nm* o número máximo de condições que podem constituir a parte antecedente da regra e *N* o número de condições que efetivamente constituem a parte antecedente da regra:

$$Si = \frac{Nm - N + 1}{Nm} \qquad (5)$$

No presente trabalho, a simplicidade foi mensurada pela equação (5), por apresentar uma avaliação que decai de forma menos abrupta com o aumento do número de condições.

6 EXPERIMENTOS

6.1 Avaliação simples: reprodução do experimento original

Primeiramente, foi implementado um programa na Linguagem Java, com o objetivo de reproduzir o experimento descrito na seção 4 [2]. Foram criadas as duas bases de dados em Access, treinamento e teste, a partir da base de dados dermatológica original [12].

Na reprodução do experimento de Fidelis, Lopes e Freitas [2], a única mudança efetuada no modelo do AG refere-se à estratégia de reinserção da população. No modelo original, apenas o melhor pai é preservado de uma geração para a outra (elitismo) e os demais são substituídos pelos filhos. Em nosso modelo, os pais competem com os filhos no final de cada geração e aqueles com maior aptidão sobrevivem, quer sejam eles pais ou filhos na geração corrente.

Os parâmetros do AG utilizados tanto na reprodução do experimento original quanto nos experimentos relatados na seção 6.2, foram os seguintes: população de 50 indivíduos, crossover de 100%, mutação de 30% (para cada segmento do gene), 50 gerações por execução, método de seleção torneio estocástico de tamanho 3, crossover de dois pontos e três execuções do AG para cada classe de doença.

Os resultados obtidos na reprodução do experimento original são apresentados na Tabela 3.

Classe da Doença	Rules	Fitness Treinamento	Fitness Teste
1	IF (melanin incontinence = 0 AND	0,948	1
	clubbing of the rete ridges >= 1)		
2	IF (fibrosis of the papillary dermis = 0 AND	0,758	0,847
	spongiosis != 0 AND		
	saw-tooth appearance of retes = 0 AND		
	perifollicular parakeratosis = 0)		
3	IF (spongiform pustule = 0 AND	0,989	1
	band-like infiltrate >= 2)		
4	IF (oral mucosal involvement = 0 AND	0,817	0,788
	knee and elbow involvement = 0 AND		
	melanin incontinence >= 0 AND		
	elongation of the rete ridges = 0)		
5	IF (oral mucosal involvement = 0 AND	1	1
	fibrosis of the papillary dermis != 0)		
6	IF (hyperkeratosis >= 0 AND	0,995	1
	perifollicular parakeratosis != 0)		

Tabela 3 - Resultados da reprodução do experimento original.

Comparando-se as tabelas 2 e 3 conclui-se que a qualidade das regras obtidas está bem próxima da publicada em [2]. Assim, a reprodução pode ser vista como bem sucedida. Os resultados obtidos com a reprodução serão utilizados como base para as análises em relação à incorporação de uma nova componente na avaliação, que mensura a simplicidade da regra.

A seguir, são apresentados os resultados dos experimentos, incluindo o critério de simplicidade da regra na função de avaliação.

6.2 Avaliação composta: simplicidade e precisão

A alteração principal que efetuamos em nosso modelo, refere-se à função de avaliação, que passou a ter duas componentes. A primeira, denominada *fA1*, visa medir a precisão da regra e é idêntica à utilizada em [2] e é dada pela equação (3). A segunda, denominada *fA2*, visa medir a simplicidade e é dada pela equação (5). A função de avaliação é dada por uma soma ponderada das duas componentes, ou seja:

$$fA = \rho * fA1 + (1 - \rho) * fA2 \qquad (6)$$

ρ é o peso que deve ser previamente especificado pelo usuário.

Realizaram-se experimentos para cada classe de doença variando-se os pesos aplicados em cada fração da função de avaliação (ρ). Ou seja, avaliamos o efeito final provocado sobre as regras obtidas ao modificarmos a importância relativa entre a precisão e a simplicidade. Devido à natureza estocástica da busca evolutiva, foram realizadas 3 execuções do AG, por experimento.

As tabelas 4 a 9, apresentadas a seguir, contêm as regras descobertas nas execuções do AG, a função de avaliação (Fitness média ou *fA)*, encontrada durante o processo de mineração, a parte da função que representa a acuidade da regra (Precisão ou *fA1*), a parte que representa a compreensibilidade da regra (Simplicidade ou *fA2*), o peso ρ aplicado na equação (6) (0.1, 0.5 ou 0.9) e a capacidade de generalização da regra, denominada Precisão no teste.

O valor da variável Precisão no teste é obtido calculando-se a função de avaliação na base de teste, considerando-se apenas a precisão da regra, ou seja, utilizando-se apenas a função *fA1*.

Para destacarmos o efeito da introdução da componente simplicidade na avaliação das regras, são apresentados nas duas últimas colunas a variação da precisão e a variação do número de condições, ambas medidas em relação às regras obtidas na reprodução do experimento original (Tabela 3).

Os resultados da classe 1 são apresentados Tabela 4. Pode-se verificar que no experimento onde a simplicidade possui menor peso (ρ = 0,9), a regra possui um número de condições maior que o experimento base, mas com uma precisão um pouco maior (1%). Tal constatação, além de inesesperada, mostrou-se única na análise dos resultados. Verificamos que o peso 0,5 foi suficiente para encontrar uma regra com o número mínimo de condições, ou seja, 1, com a mesma eficácia da regra do experimento base, equivalente a um pequeno decréscimo na eficácia (1%) em relação à regra obtida com peso 0,9.

Regras classe 1	Precisão fA1	ρ	Simplicidade fA2	Fitness fA	Precisão teste fA1	Variação precisão	Variação condições
IF (clubbing of the rete ridges != 0)	0,948	0,1	1	0,994	1	0%	-1
IF (clubbing of the rete ridges != 0)	0,948	0,5	1	0,974	1	0%	-1
IF (fibrosis of the papillary dermis = 0 AND	0,958	0,9	0,941	0,956	1	1%	1
elongation of the rete ridges != 0 AND							
perifollicular parakeratosis = 0)							

Tabela 4 - Resultados da classe 1.

Os resultados da classe 2 são apresentados Tabela 5. Neste caso, no experimento com menor peso para a simplicidade, houve a diminuição de uma condição em relação à regra encontrada no experimento base, sem alterar a precisão. Para o peso 0,5, diminiuiram duas condições com um pequeno decréscimo na precisão (5,3%). A regra com simplicidade máxima, foi encontrada com ρ=0,1, mas com uma queda considerável na precisão (21,1%).

Regras classe 2	Precisão fA1	ρ	Simplicidade fA2	Fitness fA	Precisão teste fA1	Variação precisão	Variação condições
IF (spongiosis >= 1)	0,598	0,1	1	0,959	0,643	-21,1%	-3
IF (spongiosis != 0 AND	0,718	0,5	0,97	0,844	0,755	-5,3%	-2
saw-tooth appearance of retes = 0)							
IF (polygonal papules = 0 AND	0,758	0,9	0,941	0,776	0,783	0%	-1
oral mucosal involvement = 0 AND							
spongiosis >= 2)							

Tabela 5 – Resultados da classe 2.

Como podemos observar pelos resultados da classe 3 apresentados na Tabela 6, utilizando-se $\rho=0,9$ obteve-se uma regra com o mesmo número de condições e mesma precisão que no experimento base. A utilização do ρ=0,5 foi suficiente para obter a regra com simplicidade máxima, com um decréscimo pouco relevante na precisão (1,5%).

Regras classe 3	Precisão fA1	ρ	Simplicidade fA2	Fitness fA	Precisão teste fA1	Variação precisão	Variação condições
IF (band-like infiltrate != 0)	0,974	0,1	1	0,997	1	-1,5%	-1
IF (band-like infiltrate != 0)	0,974	0,5	1	0,987	1	-1,5	-1
IF (elongation of the rete ridges >= 0 AND	0,989	0,9	0,97	0,987	1	0%	0
band-like infiltrate >= 2)							

Tabela 6 - Resultados da classe 3.

A Tabela 7 apresenta os resultados da classe 4, que mostram que obteve-se uma condição a menos que na regra do experimento base, com uma pequena diminuição na precisão (1,3 %), utilizando-se ρ=0,9. O mesmo foi observado com ρ=0,5. Já com o maior grau de importância para a simplicidade (ρ=0,1), obteve-se uma regra com o mínimo de condições, mas com uma grande queda na precisão (25,3%).

Regras classe 4	Precisão fA1	ρ	Simplicidade fA2	Fitness fA	Precisão teste fA1	Variação precisão	Variação condições
IF (spongiosis >= 1)	0,61	0,1	1	0,961	0,617	-25,3%	-3
IF (koebner phenomenon != 0 AND	0,806	0,5	0,941	0,874	0,800	-1,3%	-1
melanin incontinence = 0 AND							
clubbing of the rete ridges = 0)							
IF (koebner phenomenon != 0 AND	0,806	0,9	0,941	0,819	0,800	-1,3%	-1
melanin incontinence = 0 AND							
clubbing of the rete ridges = 0)							

Tabela 7 - Resultados da classe 4.

Os resultados da classe 5 são mostrados na Tabela 8, onde nota-se uma igualdade com o experimento base para ρ = 0,9. Para ρ = 0,5, obteve-se uma regra com uma condição a menos, com máxima simplicidade, e com pequeno decréscimo na precisão (1%).

Regras classe 5	Precisão fA1	ρ	Simplicidade fA2	Fitness fA	Precisão teste fA1	Variação precisão	Variação condições
IF (fibrosis of the papillary dermis != 0)	0,99	0,1	1	0,999	1	-1,6%	-1
IF (fibrosis of the papillary dermis != 0)	0,99	0,5	1	0,995	1	-1%	-1
IF (oral mucosal involvement = 0 AND	1	0,9	0,97	0,997	1	0%	0
fibrosis of the papillary dermis != 0)							

Tabela 8 - Resultados da classe 5.

Por último, a Tabela 9 demonstra os resultados da classe 6, na qual se observa que para todas as variações no peso ρ, obteve-se uma regra com uma condição a menos que no experimento base, atingindo-se o máximo de simplicidade e sem descréscimo na precisão.

Regras classe 6	Precisão fA1	ρ	Simplicidade fA2	Fitness fA	Precisão teste fA1	Variação precisão	Variação condições
IF (perifollicular parakeratosis >= 1)	0,995	0,1	1	0,999	1	0%	-1
IF (perifollicular parakeratosis >= 1)	0,995	0,5	1	0,998	1	0%	-1
IF (perifollicular parakeratosis >= 1)	0,995	0,9	1	0,996	1	0%	-1

Tabela 9 - Resultados da classe 6.

Analisando-se os resultados das 6 classes, de uma forma geral, podemos concluir que a incorporação do fator referente à simplicidade na avaliação foi relevante, pois foi possível obter nas classes 1, 3, 5 e 6 regras com o número mínimo de condições, ou seja, uma única condição mais simples que as regras do experimento base e com uma perda pequena na precisão, variando de 0% a 1,5%. Nas classes 2 e 4, embora não se tenha obtido regras com simplicidade máxima, em ambas foi possível reduzir o número de condições. A perda na precisão atingiu a maior variação para a classe 2 (5,3%) e para a classe 4 manteve-se em um valor baixo (1,3%).

Podemos concluir que, dentre os três valores de peso testados, o que retornou um melhor desempenho foi ρ=0,5. Em todas as classes onde este peso foi aplicado, foi possível simplificar a regra em pelo menos uma condição, sem penalizar excessivamente a precisão da regra (apenas a classe 2 excedeu 1,5%).

Por outro lado, o peso ρ=0,9 nunca excedeu uma queda de 1,5% na precisão, mas em alguns casos não foi suficiente para simplificar a regra (classes 1, 3, 5). Poderia ser uma opção interessante, caso a diretriz do usuário fosse só simplificar as regras se o prejuízo na precisão for insignificante.

O peso ρ=0,1 apresentou o pior desempenho pois ele praticamente anula a parcela de precisão na avaliação da regra, fazendo com que busca evolutiva seja focada quase exclusivamente na simplicidade. Em todas as classes, ele levou o AG a encontrar regras com a simplicidade máxima (uma única condição), mas em alguns casos com um prejuízo enorme à precisão, como na classe 4 onde a queda foi acima de 25%.

7 CONCLUSÃO

Os experimentos descritos na seção anterior mostram que, com a introdução de uma componente na avaliação, foi possível aprimorar a busca evolutiva resultando em regras mais simples. Entretanto, verificamos que dependendo da importância relativa das parcelas precisão e simplicidade, o comportamento da busca pode ser fortemente alterado. Para os valores de peso avaliados, o que retornou um melhor desempenho foi aquele que equilibrou as duas medidas, realizando uma média simples entre o grau de precisão e o grau de simplicidade.

A determinação deste peso adequado, entretanto, varia de uma aplicação para outra ou até mesmo de acordo com as medidas adotas como grau de precisão e simplicidade. Por exemplo, em outros experimentos que realizamos, não descritos neste trabalho, foi utilizada a equação (4), no lugar da (5), para mensurar a simplicidade. Neste caso, os melhores resultados foram obtidos para valores de peso entre 0,7 e 0,9, pois a equação (4) é mais severa em relação ao aumento das condições em uma regra.

Uma abordagem que pretendemos investigar em um futuro próximo refere-se à aplicação de AGs Multi-objetivos [14] na mineração de regras simples e precisas. Assim, não seria necessária uma definição *a priori* da importância relativa entre as medidas, além de ser possível obter na população final regras com características mais diversificadas. Seria possível, por exemplo, apresentar ao usuário, desde regras com precisão máxima, sem necessariamente serem simples, até regras simples em detrimento da precisão, além de regras com as duas medidas equilibradas.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Freitas, A. A. (2002). A survey of evolutionary algorithms for data mining and knowledge discovery. In: A. Ghosh and S. Tsutsui. (Eds.) *Advances in Evolutionary Computation*. Springer-Verlag.

[2] Fidelis,M.V., Lopes, H. S. and Freitas, A. A. (2000). Discovering comprehensible classification rules with a genetic algorithm. *Proc. Congress on Evolutionary Computation - 2000 (CEC-2000)*, La Jolla, CA, USA, pp. 805-810.

[3] Freitas, A. A. (2000). *Notas de aula da disciplina Data Mining*, Programa Pós-Graduação em Informática Aplicada, PUC-PR, Brasil.

[4] Santos, J. B. (2002). *Algoritmos Genéticos como Ferramenta para Data Mining*. Dissertação de Mestrado, Universidade Presbiteriana Mackenzie, São Paulo, SP, Brasil.

[5] Freitas, A. A. (2000). Understanding the crucial differences between classification and discovery of association rules: a position paper. *ACM SIGKDD Explorations*, 2(1), pp. 65-69.

[6] Whitley, D. (1994). A genetic algorithm tutorial. The GENITOR Research Group in Genetic Algorithms and Evolutionary Computation at Colorado State University, *Statistics and Computing* , 4, pp. 65-85.

[7] Oliveira, G. M. B. (2001). *Computação Evolutiva*. Apostila de uso interno do curso de Pós-Graduação em Engenharia Elétrica. Universidade Presbiteriana Mackenzie, São Paulo, SP, Brasil.

[8] Prugel-Bennett, A. (2000). *Genetic Algorithms. Image, Speech, and Intelligent Systems*. Research Group Department of Electronics and Computer Science. University of Southampton.

[9] Beasley D., Bull, D.R. and Martin R.R. (1993). An overview of genetic algorithms: Part 1, Fundamentals. *University Computing*, 15(2), pp.58-69.

[10] Beasley D., Bull, D.R. and Martin R.R. (1993). An overview of genetic algorithms: Part 2, Research Topics. University Computing, 15(4), pp. 170-181.

[11] Mitchell, M., Forrest, S. (1994). Genetic Algorithms and Artificial Life. Artificial Life 1 (3), pp. 267-289.

[12] Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases Irvine, CA: University of California, Department of Information and Computer Science. Disponível em: [http://www.ics.uci.edu/ ~mlearn/MLRepository.html].

[13] Goldberg, D. E. (1989). *Genetic Algorithms in search, optimization, and machine learning*. Reading, MA: Addison Wesley.

[14] Coello, C. A. C. (1999). A comprehensive survey of evolutionary-based multi-objective optimization techniques. *Knowledge and Information Systems*, 1(3):269-308.