

2º Trabalho de Inteligência Computacional - Data Mining

Autor: Gustavo Rezende Silva,
Orientador: Gina Maira Barbosa de Oliveira

¹Faculdade de Computação
Universidade Federal do Uberlândia (UFU)
Uberlândia – MG – Brasil

`gustavorezendesilva@hotmail.com, gina@ufu.br`

Resumo. *Este trabalho tem como intuito desenvolver um algoritmo genético para realizar uma mineração de dados em uma base dermatológica, de maneira a gerar regras de classificação para as diversas classes da doença erythemato-squamous presentes neste banco de dados.*

Palavras-Chave. *algoritmos genéticos, data mining, mineração de dados, inteligência computacional*

1. Introdução

Atualmente a quantidade de banco de dados armazenados por supermercados, hospitais, empresas de marketing e muitos outros tipos de empresas vem aumentando significativamente. Com isso, técnicas para extrair e analisar informações de grandes bases têm sido desenvolvidas e este conjunto de ferramentas e técnicas é definida como mineração de dados (data mining).

Na área médica existe um grande interesse em desenvolver ferramentas para extrair informações de bancos de dados de pacientes com o intuito de auxiliar os médicos a diagnosticar os mesmos, devido ao grande potencial do data mining este vem sendo utilizado. Neste trabalho aplicou-se algumas técnicas de mineração de dados para extrair informações de um banco de dados dermatológico disponível no repositório *Machine Learning repository* da UCI.

Esta base contém informações sobre pacientes diagnosticados com alguma das variações da doença chamada *erythemato-squamous*. O diagnóstico dos diferentes tipos desta enfermidade é complicada, pois a variação entre elas é pequena.

O objetivo final da análise deste banco de dados é conseguir extrair uma regra para cada uma das variações da doença *erythemato-squamous* presentes na base. De forma que um paciente informando apenas os seus sintomas possa ser classificado com alguma das enfermidades, e este resultado pode ser utilizado pelo médico como referência para o diagnóstico.

2. Desenvolvimento

Com o intuito de gerar regras para a base dermatológica citada na sec. 1 foi desenvolvido um algoritmo genético seguindo o artigo de Fidelis *et al.*(2000) e Miranda *et al.*(2003). Nestes trabalhos o indivíduo é uma abstração da regra de classificação de uma classe específica da doença *erythemato-squamous*, ou seja, o que define se um paciente é pré-diagnosticado com uma das enfermidades classificadas.

No banco de dados utilizado cada paciente é representado por um número e 34 atributos que representam os sintomas das doenças, por isso um indivíduo é composto por 34 atributos, 34 operadores matemáticos(=, !=, < ou >=) e 34 pesos (fig. 1). De forma que um atributo relacionado com um operador define uma condição para um paciente ser classificado naquela

Classe das Doenças	Regras	Fitness Treinamento	Fitness Teste
1	thinning of the suprapapillary epidermis ≥ 1 focal hypergranulosis < 3	0.929820	0.996290
2	follicular papules = 0 fibrosis of the papillary dermis < 1 exocytosis $\neq 0$ clubbing of the rete ridges < 1 vacuolisation and damage of basal layer $\neq 3$ band-like infiltrate < 1	0.829799	0.765162
3	elongation of the rete ridges = 0 band-like infiltrate ≥ 2	0.992163	0.994803
4	family history $\neq 1$ eosinophils in the infiltrate < 1 elongation of the rete ridges $= 0$ saw-tooth appearance of retes < 2 follicular horn plug $\neq 2$	0.811603	0.814022
5	koebner phenomenon $= 0$ oral mucosal involvement $\neq 1$ knee and elbow involvement < 3 PNL infiltrate $\neq 2$ fibrosis of the papillary dermis $\neq 0$ band-like infiltrate $\neq 3$	0.996662	0.991431
6	follicular papules ≥ 1 perifollicular parakeratosis $\neq 0$	0.991299	0.986765

Tabela 1. Resultados do data mining

Para escolher o melhor resultado foi levado em consideração a avaliação de treinamento e teste, além do número de condições resultantes. Uma vez que uma regra com muitas condições não representa o problema de forma genérica, apenas aquele conjunto de dados específicos.

Com a intenção de se ter uma ideia geral do resultado das regras obtidas, as médias dos melhores resultados de cada execução para a mesma regra foram calculadas e se encontram na tabela 2.

Classe	Média Fitness Treinamento	Média Fitness Teste
1	0.932275	0.981523
2	0.825245	0.850473
3	0.976214	0.969933
4	0.850391	0.725871
5	0.896647	0.831133
6	0.939587	0.809525

Tabela 2. Média das avaliações de cada classe

Ao comparar os resultados obtidos neste trabalho com os de Fidelis *et al.*(2000) e Miranda *et al.*(2003), percebe-se que as avaliações das regras encontradas são bem próximas. Entretanto, as condições associadas a cada regra variam em quase todas elas.

Em seguida, alguns parâmetros do algoritmo genético foram alterados com o intuito de obter melhores resultados.

Alterando apenas o peso máximo permitido os resultados encontrados foram diferentes (tabela 3). Para algumas classes como a 1 e 2 as avaliações foram melhores, já na classe 5 o fitness foi igual porém a quantidade de condições foi menor, a classe 3 permaneceu com as mesmas avaliações e a 4 e 6 pioraram.

Classe das Doenças	Regras	Fitness Treinamento	Fitness Teste
1	clubbing of the rete ridges ≥ 1 perifollicular parakeratosis < 1	0.955212	0.996290
2	koebner phenomenon < 1 polygonal papules $== 0$ fibrosis of the papillary dermis < 1 elongation of the rete ridges $\neq 3$ thinning of the suprapapillary epidermis $== 0$ perifollicular parakeratosis $== 0$	0.910755	0.923471
3	PNL infiltrate $== 0$ band-like infiltrate ≥ 2	0.992163	0.994803
4	koebner phenomenon ≥ 1 scalp involvement $\neq 2$ melanin incontinence < 1 exocytosis ≥ 1 clubbing of the rete ridges < 1 thinning of the suprapapillary epidermis $\neq 2$	0.892055	0.727187
5	polygonal papules < 2 fibrosis of the papillary dermis $\neq 0$ hyperkeratosis < 3 focal hypergranulosis < 1	0.996662	0.991431
6	koebner phenomenon $\neq 2$ perifollicular parakeratosis ≥ 1	0.986932	0.986765

Tabela 3. Peso máximo = 0.2

As maiores dificuldades na reprodução dos modelos foi conseguir analisar os resultados e compara-los, e ainda entender que um número grande de condições não satisfaz o propósito da mineração de dados.

Referências

- Clay R. S. Miranda ; OLIVEIRA, G. M. B. ; SANTOS, J. B. (2003). Algoritmos genéticos aplicados em data mining para obtenção de regras simples e precisas. *SBAI2003: Simpósio Brasileiro de Automação Inteligente*.
- Fidelis M. V., L. H. S. and Freitas, A. A. (2000). Discovering comprehensible classification rules with a genetic algorithm. *Proc. Congress on Evolutionary Computation*, pages 805–810.
- OLIVEIRA, G. M. B. (2017). Apresentação em sala de aula - algoritmos genéticos aplicados em data mining para obtenção de regras de classificação.