Análise Comparativa entre Algoritmos de Aprendizado de Máquina sob a Ferramenta Weka

Marino Souza dos Santos[1]

[1]Departamento de Ciência da Computação – Instituto de Matemática
Universidade Federal Bahia (UFBA)
40.170-110 – Salvador – BA – Brazil

marino@dcc.ufba.br

Resumo. *Este trabalho aprensenta uma análise de desempenho entre algoritmos de Aprendizado de Máquina, de diferentes paradígmas, sobre uma base de diagnósticos de pacientes submetidos à exames de câncer de mama registrados por [Wolberg 1991]. Os algoritmos e métodos usados nos experimentos fora executados pela ferramenta Weka.*

1. Introdução

Nos dias atuais o uso da informática para otimizar processos e tornar dados mais acessíveis é bastante comum, e em prontuários médicos estes dados, que já são digitais, podem se tornar aliados na investigação e otimização de detecção de doenças. Este trabalho faz uso de uma base de diagnósticos de exames oncológicos, mais precisamente de pacientes sob suspeita de câncer de mama. Vale resaltar que estes dados ja foram usados em outros experimentos, citados na sessão 5, contudo os experimentos reaizados neste trabalho são não para confrontá-los, nem apoiá-los, a finalidade dos experimentos conduzidos aqui são apenas para comparar alguns algoritmos de Aprendizado de Máquina da ferramenta [Hall et al. 2009].

O uso de algoritmos de Aprendizado de Máquina neste contexto auxilia agiliza o processo diagnosticação de doenças. Os algoritmos ainda não são perfeitos, e podem dar falsos negativos, como veremos a seguir, mas reduz o esforço do profissinal resposável pelos diagnósticos consideravelmente.

2. Pré-processamento

Esta etapa não foi trabalhosa, pois os dados disponibilizados por [Mangasarian 1992] estavam num modelo muito semelhante ao usado pela ferramenta *Weka*, foi somente necessário remover os atributos que informavam números de identificação dos pacientes, pois não contribuiam para o resultado. Dentre as 699 instâncias, apenas 16 continha dados incompletos, que foram devidamente reportados na tabela de *tokens* da *Weka*.

Domínio	Instâncias	Atributos	Tipo de Classe
Breast Cancer Winscosin	699	9	Discreta

Table 1. Dados do conjunto de dados Breast Cancer Winscosin[Mangasarian 1992]

O domínio do *Breast Cancer Winscosin dataset* disponibilizado por [Mangasarian 1992] possuia 10 atributos mais a classe. Para este experimento o

Base do Hospital de Winscosin	
Atributo	**Intervalo inteiro**
Clump Thickness	Entre 1 e 10
Uniformity of Cell Size	Entre 1 e 10
Uniformity of Cell Shape	Entre 1 e 10
Marginal Adhesion	Entre 1 e 10
Single Epithelial Cell Size	Entre 1 e 10
Bare Nuclei	Entre 1 e 10
Bland Chromatin	Entre 1 e 10
Normal Nucleoli	Entre 1 e 10
Mitoses	Entre 1 e 10

Table 2. As classes são escritas como 2 para beníngo e 4 para malígno

H

Algoritmos de Classificação	
Método	**Tipo**
NaiveBayesSimple	Probabilístico
RBFNetwork	Conexionista
LADTree	Simbólico
LBR	Lazy (extra)
NaiveBayes	Probabilístico
NaiveBayesUpdateable	Probabilístico
SimpleLogistic	Conexionista
J48	Simbólico
IBK	Lazy
MultilayerPerceptron	Conexionista

Table 3. Os 10 algoritmos dispostos foram testados e os 4 primeiros foram os selecionados.

primeiro atributo (`ID de Pacientes`) foi omitido por não ser significante para a análise ficando com a seguinte configuração.

3. Extração de Padrões

Nesta etapa, o método *Cross-Validation 10-fold* foi usado a fim de validar cada algoritmo testado. O método consiste em particionar o dataset em 10 partes onde o treinamento é feito em cima de 9 partições e os testes na partição restante. A ferramenta permite que o usuário informe a quantidade de folds, e para o caso de 10, ela executa 10 vezes fazendo diferentes permitações de partição destinada para teste contra as partições de treinamento.

Dentre os algoritmos dispostos em 3 Foi escolhido o NaiveBayesSimple pois teve o mesmo desempenho que os outros probabilísticos, porém nenhum aluno ainda havia escolhido. Entre os Conexionistas o MultilayerPerceptron teve que ser abortado após levar mais de 60 segundos executando, o RBFNetwork ainda não estava no conjunto de algoritmos de ninguém além de ter tido o melhor resultado entre os do seu paradígma na questão de precisão, com a marca de apenas 20 erros, pois o tempo para ele, o SimpleLogistic (33 erros) e o SMO (29 erros) foram o mesmo, zero segundos. Para os algoritmos

Comparações entre os 4 algoritmos escolhidos	
Algoritmo	Percentual de acerto
(1) NaiveBayesSimple	97.2818
(2) RBFNetwork	97.1388
(3) LADTree	94.8498
(4) LBR	97.2818

Table 4. Percentual de acerto dos algoritmos escolhidos

do paradígma simbólico foram testados o J48, que foi bem usado pela turma, porém o LADTreeteve melhor desempenho em precisão com apenas 36 falsos positivos contra 39 do seu concorrente, entretanto, no quesito tempo, demorou 2,4 segundos contra um tempo muito próximo de zero. Entre os algoritmos extra foram escolhidos do paradígma chamando de LAZY pela ferramenta, LBR com 19 erros contra IBK com 39 erros.

4. Pós-processamento

A tabela 4 mostra o percentual de acerto dos algoritmos escolhidos, onde o NaiveBayesSimple e o LBR ficam empatados com os melhores resultados. Os dados completos podem ser acessados em [Souza 2016].

5. Agradecimentos

Na base de dados é pedido explicitamente que citem os seguintes colaboradores:

O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 and 18.[Wolberg 1991]

William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

K. P. Bennett and O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon and Breach Science Publishers).

References

Hall, M., Frank, E., Holmes, G., Pfahringer, B., and Reutemann, P. (2009). *The WEKA data Mining Software: An update*. SIGKDD Explorations, 11th edition.

Mangasarian, O. (1992). Uci machine learning repository. `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29`.

Souza, M. (2016). Codigo fonte com os experimentos e tex. `http://marinofull.github.io/trabalho2-ia`.

Wolberg, D. W. H. (1991). Wisconsin Breast Cancer Database.