Avaliação de Algoritmos de Classificação
Mineração de Dados

Pedro Batista - pedro@ufpa.br

October 12, 2010

1 Introdução

Este trabalho tem como objetivo, avaliar três algoritmos de classificação. Para isso serão empregadas técnicas estudadas em sala de aula, como a *Student Paired test*, que é uma medida estatística, esta nos ajuda a decidir, se, duas técnicas são significativamente diferentes.

2 A base de dados

A base de dados utilizada é a mesma citada em [1]. Esta tem como classe o atributo diabete, que pode ser positivo ou negativo. Para esta predição, se faz uso de atributos como idade, número de gravidez, pressão sanguínea diastólica, dentre outras. A base é constituída de 768 amostras e nenhum atributo está faltando.

Para este trabalho, a base foi dividia três vezes, cada uma em dois conjuntos, disjuntos. Isto é, primeiramente, embaralhamos toda a base. Então os últimos 68 atributos foram usados para teste, e o resto para treino, esta foi a base T1. Então pegamos novamente a base total e separamos os elementos de 300, a 368 para teste e o resto para treino, que formou a base T2. A base T3 foi então criada utilizando os 68 primeiros elementos para teste, e o restante para treino.

As características das bases são mostradas na tabela 2.

base/diabetes	T1_treino	T1_test	T2_treino	T2_test	T3_treino	T3_test
positivo	255	16	245	26	246	25
negativo	450	52	460	42	459	43

Table 1: Características das bases de treino e teste utilizadas.

3 O experimento

Os algoritmos escolhidos para classificação foram: rede neural (RN) multi-camada, arvore de decisão com J48, e a Naive Bayes.

Para a rede neural, a melhor configuração encontrada foi: 80 camadas, no máximo 500 épocas, e uma taxa de apredizagem de 0.3.

Os resultados para todos os algoritmos são mostrados na Tabela 3.

		T1		T2		T3		Erro Total
		Classificado						
		positivo	negativo	positivo	negativo	positivo	negativo	
RN Multi-camada	positivo	8	8	15	11	12	13	24,51%
	negativo	7	45	9	33	2	41	
	Erro	22,06%		29,41%		22,06%		
J48	positivo	9	7	18	8	17	8	23,04%
	negativo	10	42	11	31	3	40	
	Erro	25,00%		27,94%		16,18%		
Naive bayes	positivo	8	8	16	10	16	9	25,49%
	negativo	12	40	9	33	4	39	
	Erro	29,41%		27,94%		19,12%		

Table 2: Resultados por vários algoritmos.

References

[1] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Symposium on Computer Applications and Medical Care*, pages 261 – 265, 1988.