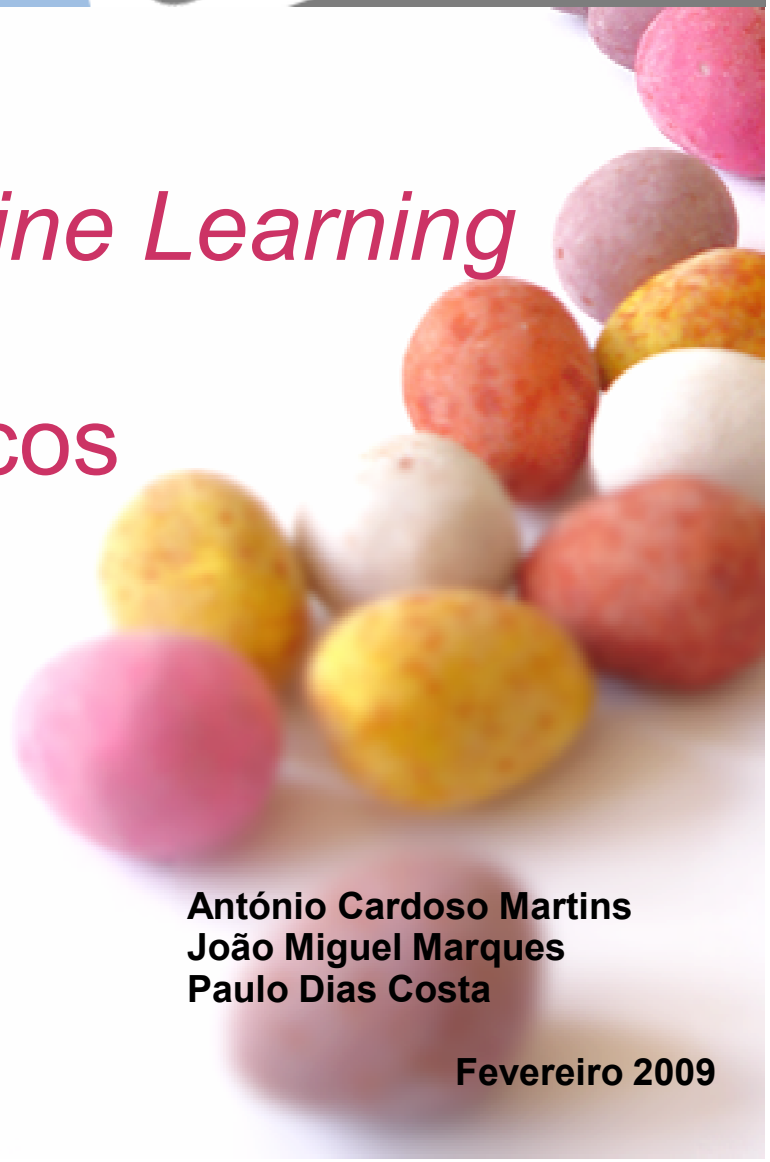
MESTRADO EM
INFORMÁTICA MÉDICA

U.PORTO

Estudo Comparativo de Três Algoritmos de *Machine Learning* na Classificação de Dados Electrocardiográficos



Mestrado em Informática Médica

António Cardoso Martins
João Miguel Marques
Paulo Dias Costa

Fevereiro 2009

U.PORTO
FMUP FACULDADE DE MEDICINA
UNIVERSIDADE DO PORTO

Objectivos

Conhecer principais técnicas *data mining*

Conhecer principais algoritmos de *machine learning*

Análise dados e parâmetros

WEKA

Estudo Experimental
Dataset

Dataset Arrhythmias

Repositório *Machine Learning - University of California*

16 Classes / 3 grupos:

1→ Normal 2 a 15 → Anormal 16 → Não Classificados

452 instâncias 279 atributos

Valores omissos

Estudo Experimental
Tratamento do dataset

Ficheiro disponibilizado em .data → .csv

Ficheiro adicional (*arrhythmia.names*)

- descrição atributos
- proprietários da base de dados

Coligir informação num único ficheiro:

- .arff
- etiquetar cada atributo
- classificar quanto ao tipo: numérico e nominal

Estudo Experimental
Algoritmos Utilizados

FMUP

OneR

J48

Naïve Bayes

Estudo Experimental
Metodologia

Substituição valores omissos por valores probabilísticos

3 x 2 x 3 : três algoritmos, dois testes, três configurações

3: → OneR, J48, Naïve Bayes

2: → *Cross-validation* *Percentage Split*

3: → 50% treino-50% teste

→ 70% treino-30% teste

→ 80% treino-20% teste

Estudo Experimental
Metodologia

Dados a analisar

Número de instâncias correctamente classificadas
(percentagem de acerto)
Tempo para construção do modelo
(tempo de aprendizagem)
Erro médio
Sensibilidade
Especificidade
Área ROC

Estudo Experimental
Metodologia

Sensibilidade

Se = Vp / (Vp + Fn)

TVp = Vp / P = Vp / (Vp + Fn) = Se

⇔ TVp = Se

Especificidade

Sp = Vn / (Fp + Vn)

TFp = Fp / N = Fp / (Fp + Vn)

⇔ Sp = 1 - TFp

WEKA fornece

TVp TFp Área ROC: relação TVp / 1-TFp = Se/Sp

Estudo Experimental
Resultados - Split percentage

Split	J48			OneR			Naïve Bayes		
	Acerto (%)	*Tempo (s)*	*Erro (médio)*	*Acerto (%)*	*Tempo (s)*	*Erro (médio)*	*Acerto (%)*	*Tempo (s)*	*Erro (médio)*
50%/50%	**65,49**	**1,84**	**0,0474**	**58,41**	**0,10**	**0,0520**	**64,16**	**0,15**	**0,0454**
70%/30%	**72,06**	**1,50**	**0,0401**	**58,09**	**0,16**	**0,0524**	**69,85**	**0,14**	**0,0377**
80%/20%	**70,00**	**1,75**	**0,0433**	**55,56**	**0,15**	**0,0556**	**74,44**	**0,14**	**0,0324**

Estudo Experimental
Resultados - Cross validation

Cross Validation	**J48**			**OneR**			**Naïve Bayes**		
	Acerto (%)	*Tempo (s)*	*Erro (médio)*	*Acerto (%)*	*Tempo (s)*	*Erro (médio)*	*Acerto (%)*	*Tempo (s)*	*Erro (médio)*
Folds 10	**63,27**	**1,56**	**0,0500**	**57,08**	**0,12**	**0,0537**	**61,50**	**0,10**	**0,0477**

Estudo Experimental
Resultados - Split percentage

Split 50%/50%

	OneR	J48	Naïve Bayes
Vp	0,584	0,655	0,642
Fp	0,407	0,186	0,135
Se	58,40%	65,50%	64,20%
Sp	59,30%	81,40%	86,50%
Área ROC	0,588	0,728	0,811

Split 70%/30%

	OneR	J48	Naïve Bayes
Vp	0,581	0,721	0,699
Fp	0,467	0,187	0,117
Se	58,10%	72,10%	69,90%
Sp	53,30%	81,30%	88,30%
Área ROC	0,557	0,772	0,847

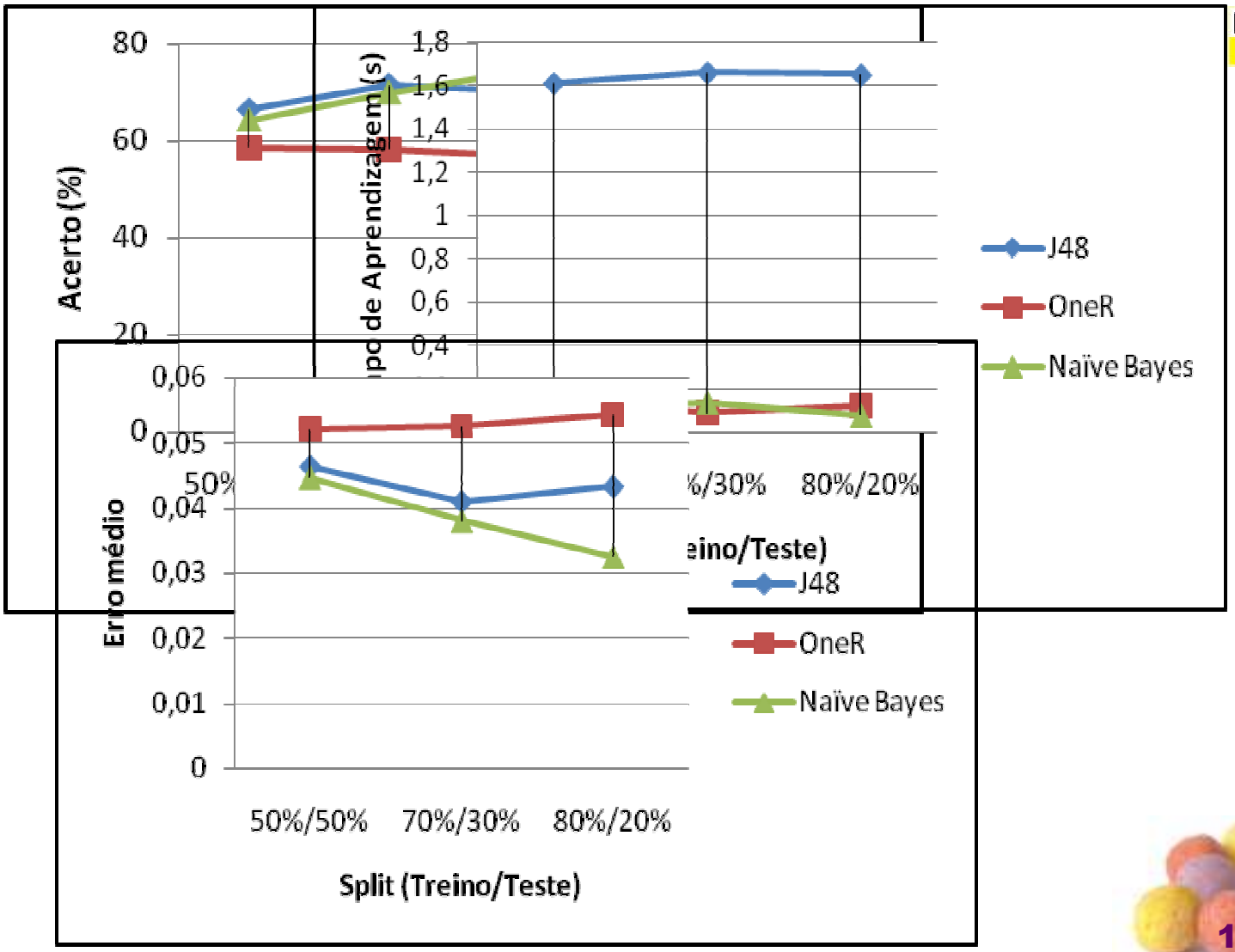
Split 80%/20%

	OneR	J48	Naïve Bayes
Vp	0,556	0,7	0,744
Fp	0,459	0,194	0,097
Se	55,60%	70,00%	74,40%
Sp	54,10%	80,60%	90,30%
Área ROC	0,548	0,795	0,848

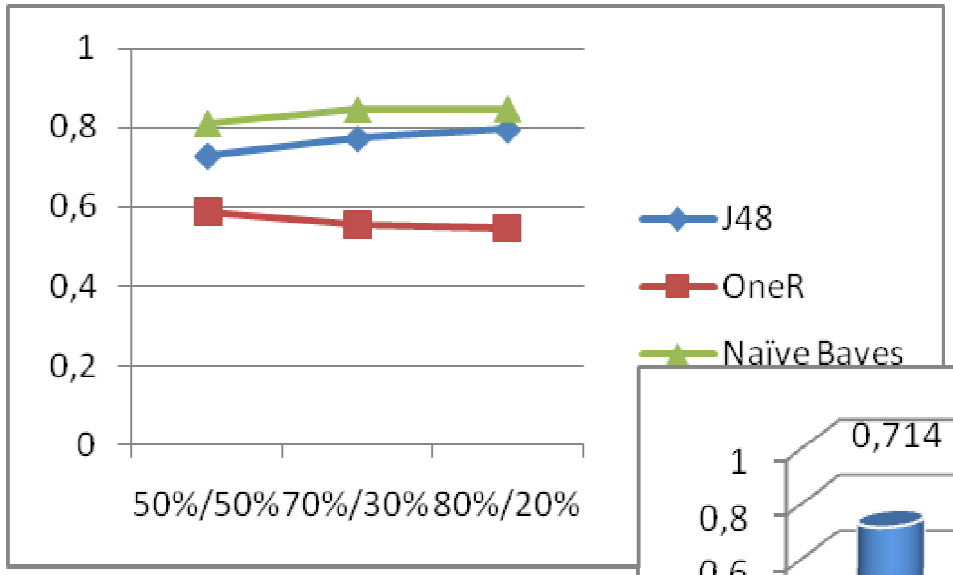
Estudo Experimental
Resultados - Cross validation

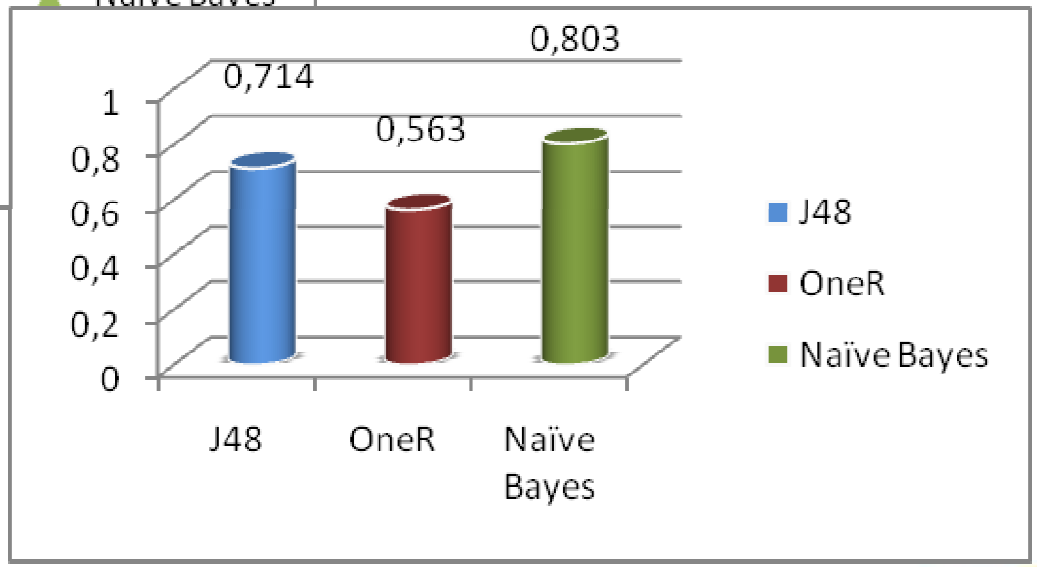
	Cross-validation		
	OneR	**J48**	**Naïve Bayes**
Vp	**0,571**	**0,633**	**0,615**
Fp	**0,444**	**0,176**	**0,164**
Se	**57,10%**	**63,30%**	**61,50%**
Sp	**55,60%**	**82,40%**	**83,60%**
Área ROC	**0,563**	**0,714**	**0,803**

Estudo Experimental
Análise Resultados - Split percentage



Estudo Experimental
Análise Resultados - ROC





Conclusão

FMUP

OneR e Naïve Bayes: aprendizagem mais rápida

J48 e Naïve Bayes: maior acuidade

Dependência forte da *percentage split*

Limitações:

- valores omissos *dataset*; alternativa: descartar dados
- significância valores Se, Sp e ROC

Futuro:

- *dataset* nacional e alterações classes (?)
- teste com outros algoritmos
- significância valores

Conclusão

FMUP

OneR e Naïve Bayes: aprendizagem mais rápida

J48 e Naïve Bayes: maior acuidade

Dependência: forte com *percentage split*

Limitações:

valores omissos *dataset*; alternativa: descartar dados

significância valores Se, Sp e ROC

Futuro:

dataset nacional e alterações classes (?)

teste com outros algoritmos

significância valores

Naïve Bayes:

melhor desempenho

Questões?