Precisamos centos de clasificadores para resolver os problemas de clasificación do mundo real?

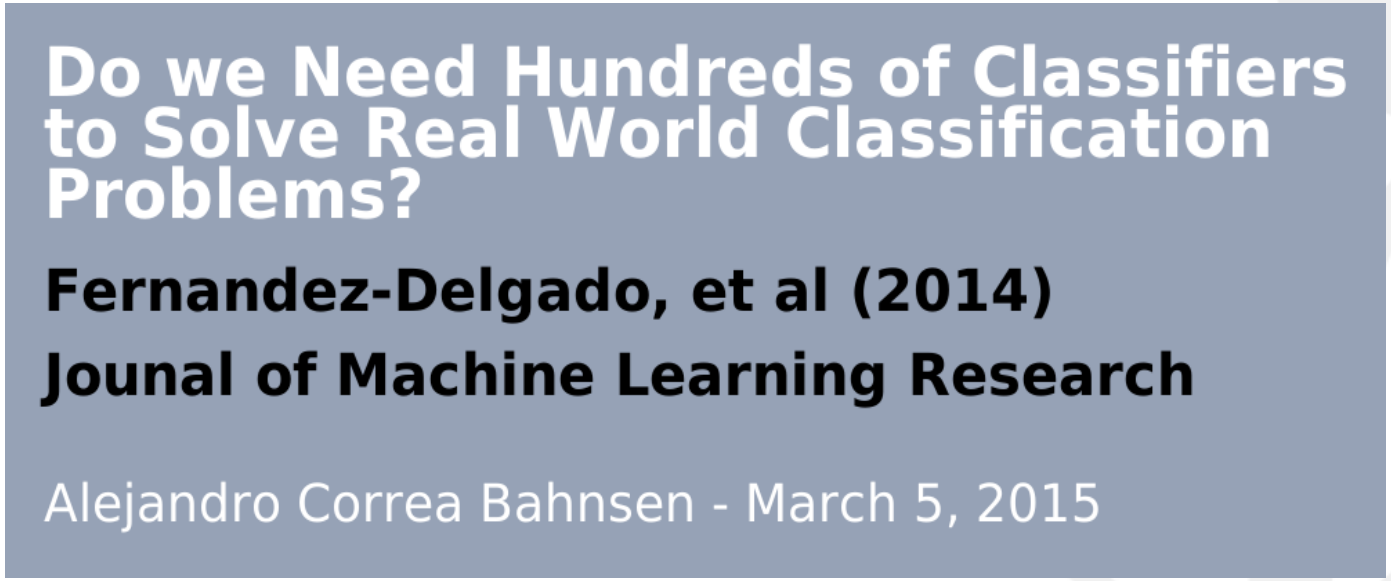
M. Fernández Delgado
E. Cernadas
D. Amorim
S. Barro

Centro Singular de Investigación en Tecnoloxías da Información
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

citius.usc.es

Introducción

- Orixe desta presentación: artigo publicado en Journal of Machine Learning Research (novembro 2014,vol.15,pp. 3133-3181): http://jmlr.org/papers/volume15/delgado14a/delgado14a.pdf



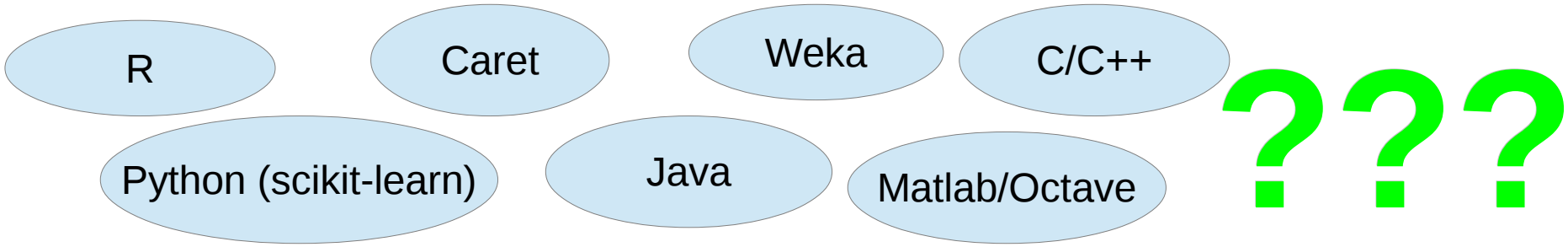
- Presentación de Alejandro Correa Bahnsen sobre este artigo, incluindo máis clasificadores

Para que se fai este traballo?

- Obxectivo persoal
- Clasificadores que usaba: redes neuronais (MLP), Support Vector Machine (SVM)
- Descubrimento de clasificadores en R: principalmente estatísticos
  - Coñeces moitos clasificadores **exóticos**
- Weka: moitos clasificadores, execución manual, interface gráfica
  - Máis clasificadores que non coñecías

CiTIUS

Moitas outras implementacións



- Paquete **caret** de R: http://caret.r-forge.r-project.org
  - Execución de 180 clasificadores e regresores (usando paquetes R)
  - Unha única interface (función **train**) chama a todos
  - Sintonización do clasificador (parámetros e valores)
- Weka: máis alternativas
  - Execución de 60 clasificadores dende liña de comandos
  - Programa en Java que execute clasificadores (permite sintonización de parámetros)
  - Permite combinar ensembles e clasificadores base: explosión de combinacións!

CiTIUS

Coñecer novos clasificadores ...

- Proveñen de campos moi distintos:
  - R: estatísticos: lda, qda, mars, plsr, ...
  - Caret: bayesianos,redes neuronais, simbólicos, ...
  - Weka: Data Mining, IA simbólica (p.ex. árbores decisión), bayesianos, clustering, regresión, ensembles ...
  - C/C++ e Matlab/Octave: redes neuronais
  - De cada clasificador hai variantes ...
  - Non sabes nin que existían ...
  - Quén os emprega?
  - Que tal funcionan na práctica?

… pagará a pena?

- Os clasificadores das familias exóticas, ¿serán mellores que os que eu usaba?
- Mellor dito, se temos un problema de clasificación novo, que clasificador é esperábel que sexa o mellor?
- Isto depende do conxunto de datos:
  - Non hai ningún clasificador que sexa o mellor para tódolos datos.
- Se temos unha **colección de datos moi ampla** podemos:
  - Ter unha idea realista do bo que é cada clasificador.
- Se temos **moitos clasificadores** podemos:
  - Supor que ningún clasificador non incluido nela vai funcionar mellor: medida da dificultade dos datos

Para usar clasificadores novos ...

- Ante clasificadores novos:
  - Preguiza, medo, descoñecemento de como usalos
  - Tendemos a non usalos
- Mantémonos cos clasificadores cos que estamos familiarizados (sintonización de parámetros, pre-procesamento dos datos, ...): somos cómodos
- E se os outros fosen mellores?
- Non por ser novos, senón por proceder doutros campos
- Para usar un clasificador novo, necesitamos:
  - Implementación (a nosa pode non ser correcta)
  - Documentación mínima (valores dos parámetros sintonizábeis)

Beneficios ...

- Resulta útil facer unha taxonomía de clasificadores: mapa de estradas
- Comprobar por min mesmo como son de bos os clasificadores exóticos
- Non me fío dos artigos: seleccionan os datos para que os seus clasificadores se comparen ben
- Agrupación por familias de clasificadores similares (p.ex: lda,qda,rrlda,slda,ldaBag,fda,mda,hdda,...)
- Implementacións alternativas dos mesmos clasificadores (os máis populares, p.ex. redes neuronais en C,Matlab,R,caret,Weka, Python, ...)

Imos comparalos todos

- Non son usuais as comparacións **neutrais** de clasificadores
- Proponse un clasificador e compárase cos máis populares: polarizado para beneficiar ao clasificador proposto
- Case sempre o proposto funciona mellor: que casualidade!
- Queriamos facer unha comparación na que nós non vendéramos nada
- Selección do conxunto de datos: polarizada para favorecer un certo clasificador

Conxuntos de datos

- Seleccionamos todos os problemas de clasificación de la UCI database: 121 conxuntos de datos
- Descartamos os máis grandes (implementacións non large-scale)
- Entradas: 4-262; Patróns: 16-130,064; Clases: 2-100
- Como cambian os resultados co nº de entradas, patróns, clases e dificultade do problema?
- Con moitos clasificadores, podemos asumir que:
  - En cada problema, o mellor clasificador acada o mellor resultado posíbel
  - Comparamos clasificadores a respeito do mellor resultado para cada conxunto de datos.
- Crítica: as entradas categóricas con *n* valores posíbelmente sería mellor codificalas como *n* entradas binarias

Familias de clasificadores (I)

Colección de **179 clasificadores** pertencentes a **17 familias**

LDA (20)

LDA-R	LDA2-T	RRLDA	SDA	SLDA	StepLDA	sddaLDA	PLDA	spLDA	QDA
QDACov	sddaQDA	stepQDA	FDA	FDA2	MDA-R	MDA-T	PDA	RDA	HDDA

Bayesianos (6)

naiveBayes	vbmpRadial	NB	NBU	BayesNet	NBSimple

Redes neuronais (21)

RBF-M	RBF-T	RBFN-W	rbfDDA	MLP-M	MLP-FANN	MLP-T	avNNet	MLPWD	NNET
PCANNET	MLP-W	PNN	ELM	KELM	LVQ-R	LVQ-T	BDK	DKP	DPP

SVM (10)

LibSVM	SVMLIGHT	LIBSVM-W	LIBLINEAR	SVMRADIAL	SVMRC	SVM-LINEAR	SVM-POLY	LSSVM-RADIAL	SMO

Árbores de decisión (14)

RPART-R	RPART-T	RPART2-T	OBLIQUE TREE	C5.0 TREE	CTREE-T	CTREE2-T	J48-W	J48-T	RANDOM SUBSPACE-W
NBTREE	RANDOM TREE	REPTREE	DECISION STUMP						

CiTiUS

Varios clasificadores **caret** deron erros e foron executados en **R**

Familias de clasificadores (II)

Clasificadores baseados en regras (12)

PART-W	PART-T	C5.0RULES	JRIP-T	JRIP-W	ONER-T	ONER-W	DTNB	RIDOR	ZEROR
DECISIONTABLE		CONJUNCTIVERULE							

Boosting (20)

ADABOOST	LOGITBOOST-R	LOGITBOOST-W	RILB	ABM1-DS	ABM1-J48	C5.0	MBAB-DS	MBAB-DT	MBAB-IBK
ADABOOST	MBAB-LGST	MBAB-MLP	MBAB-NB	MBAB-ONER	MBAB-PART	MBAB-RF	MBAB-RT	MBAB-REPT	

Bagging (24)

BG	TREEBG	LDABG	PSLBG	NBBG	CTREEBG	SVMBG	NNETBG	MCST	BG-DS	BG-DT
BG-HP	BG-IBK	BG-J48	BG-LIBSVM	BG-LGST	BG-LWL	BG-MLP	BG-ONER	BG-PART	BG-RF	BG-RT
BG-RT	BG-RT									

Stacking (2)

STACKING	STACKINGC

Random Forests (8)

RF-R	RF-T	RRF	CFOREST	PARRF	RRF-GLOBAL	RF-W	ROTATION-FOREST

Familias de clasificadores (III)

Outros ensembles (11)

RC	OCC	MS	MCC	CSS	GRD	END	DEC	VOTE	DG	LWL

Modelos Lineares Xeralizados (GLM, 5)

GLM	GLMNET	MLM	BAYESGLM	GLM-STEPAIC
MBAB-LGST	MBAB-MLP	MBAB-NB	MBAB-ONER	MBAB-PART

K veciños máis cercanos (KNN,5)

KNN-R	KNN-T	NNGE	IBK	IB1

Partial Least Squares and Principal Component Regression (PLSR,6)

PLS	GPLS	SPLS	SIMPLS	KERNELPLS	WIDEKERNELPLS

Logistic and Multinomial Regression (3)

SIMPLELOGISTIC	LOGISTIC	MULTINOM

Multivariate Adaptive Regression Splines (MARS,2)

MARS	GCV-EARTH

Outros métodos (10)

PAM	VFI	HYP	FC	CVPS	CVC	ASC	CVR	KSTAR	GAUSS-PRRADIAL

Metodoloxía experimental

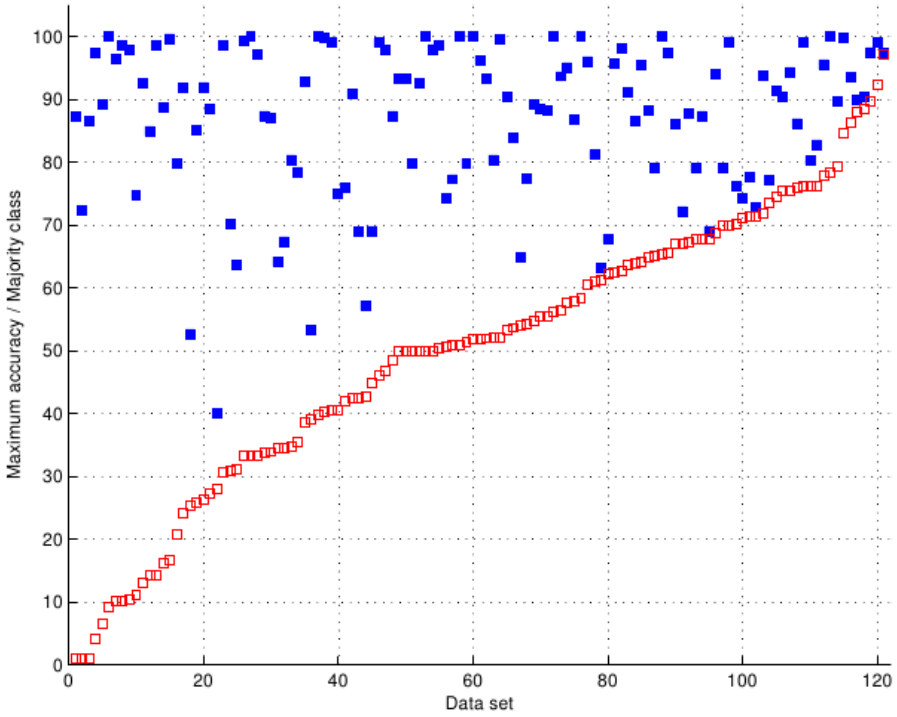
- Conxuntos de entrenamento e validación (50% cada un) para a sintonización de parámetros.
  - Cada clase ten as mesmas proporcións nos conxuntos de entrenamento e teste
- Teste: validación cruzada 4-fold (para evitar custe computacional elevado) usando valores seleccionados de parámetros sintonizábeis
  - Particións de entrenamento/teste distintas das usadas na sintonización de parámetros
- Conxuntos con datos de teste separados:
  - Sintonización con datos de entrenamento (separados en entrenamento+validación)
  - Teste con datos de teste
- Particións e resultados dispoñíbeis públicamente

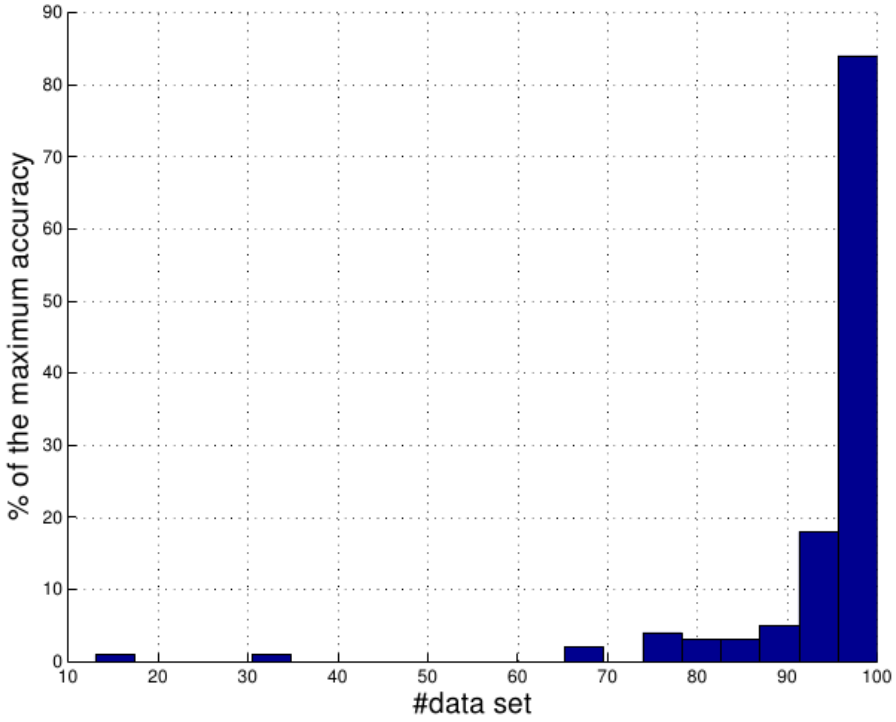
Resultados: mellor clasificador e familia

- Acerto, ranking de Friedman e Kappa de Cohen
- Mellor clasificador: Parallel Random Forest e Random Forest, caret. Acerto medio 82.0%, kappa 63.6%
- As diferencias a respeito dos clasificadores seguintes non son estatísticamente significativas (Test T) até o 9º clasificador (C5.0)
- Mellor familia: RF. Non é casualidade: entre os 10 mellores hai 3 RF, e os 8 clasificadores RF están entre os 25 primeiros
- Sorprendente: esta versión de RF é moi antiga (1996), e RF non concentra moitos esforzos de investigación
- Segunda mellor familia: SVM (acerto 81.8%, kappa 62.2%), 5 clasificadores entre os 20 primeiros.

Mellores resultados

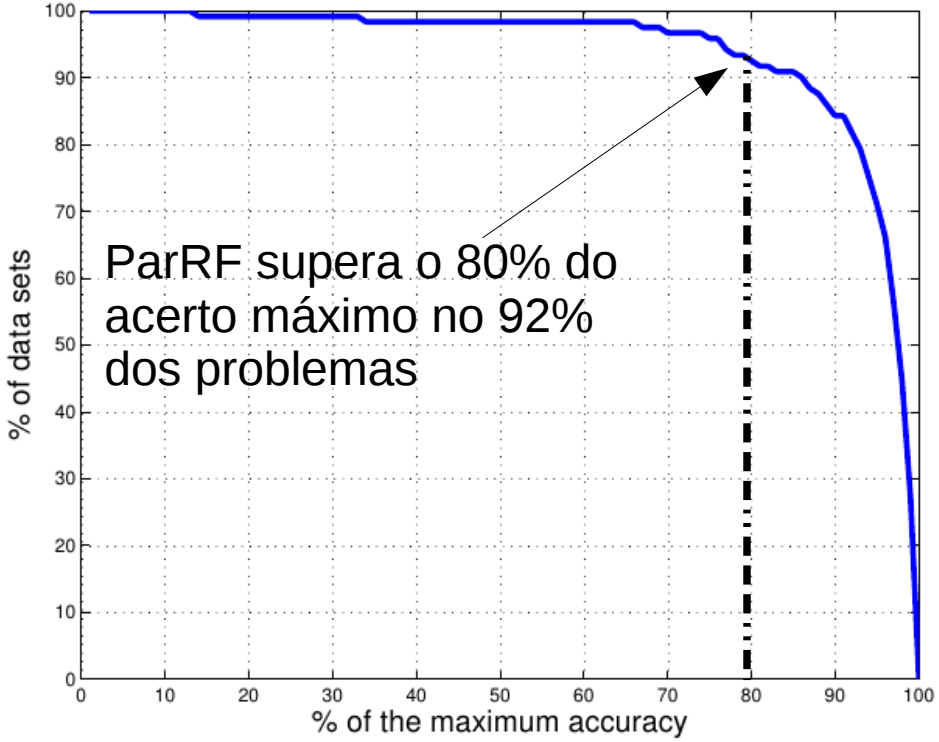
- Esquerda: acerto máximo (azul), %clase minoritaria (vermello)
- Dereita: histograma do % do acerto máximo para o mellor clasificador (Random Forest)



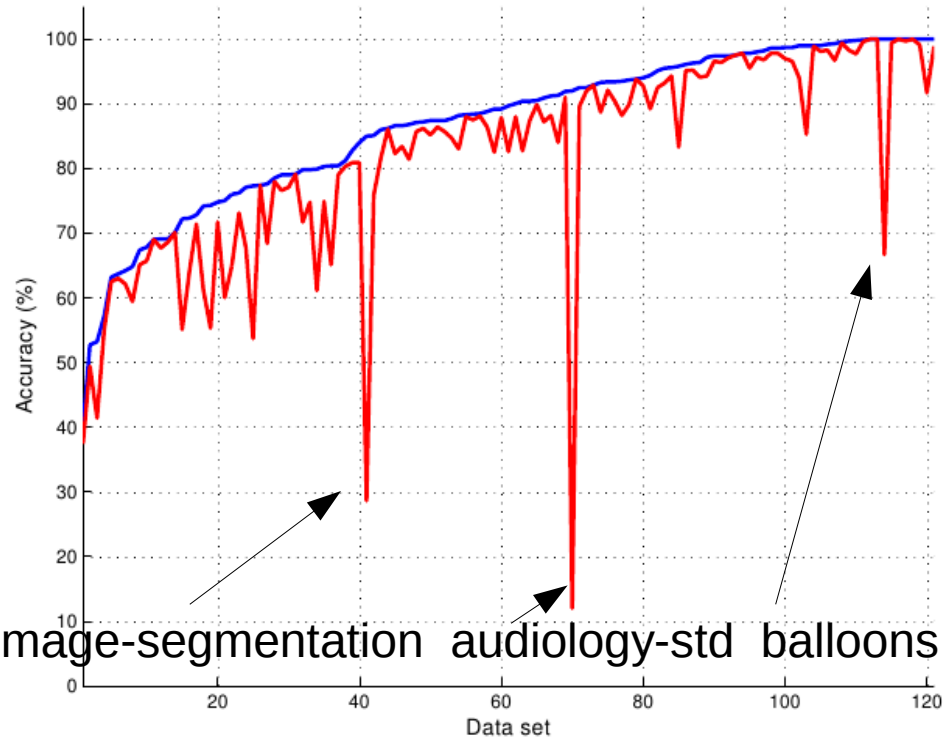


Resultados de parRF en detalle

- Esquerda: % de problemas (Y) para os que parRF supera cada nivel de acerto (X)

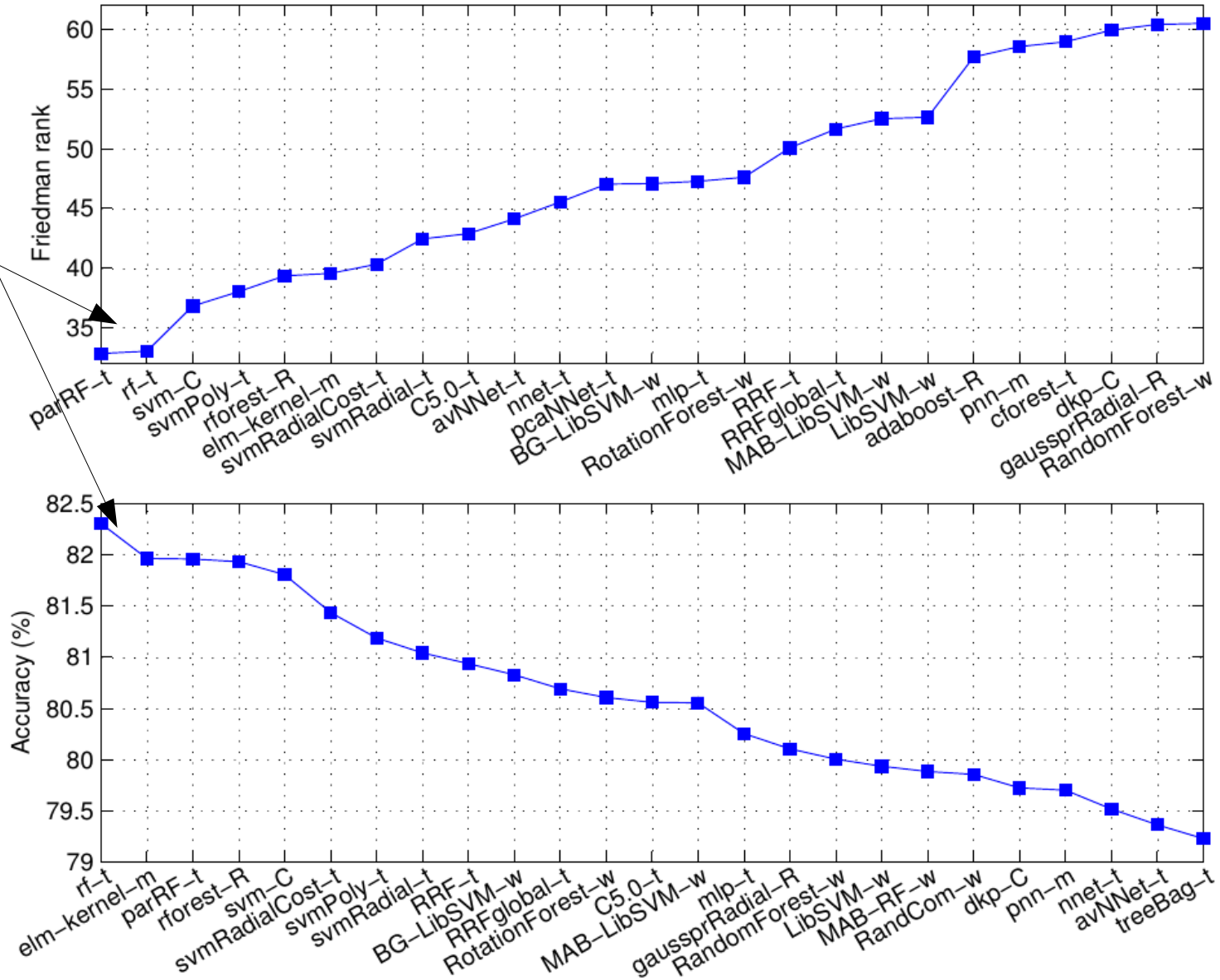


- Dereita: acerto máximo (azul) e de parRF (vermello) ordeado por acertos crecentes



Ranking de Friedman e acerto
25 mellores clasificadores

Salto entre primeiro e segundo



Outros clasificadores entre os 25 primeiros

10ª-12ª posición: avNNet, nnet e pcaNNet

6ª : Extreme Learning Machine (ELM) con núcleo gausiano:

9º: C5.0: Boosting de árbores de decisión

13ª: Bagging de LIBSVM

20ª: Adaboost de árbores de clasificación

21ª: Probabilistic Neural Networks (PNN, é de 1990!)

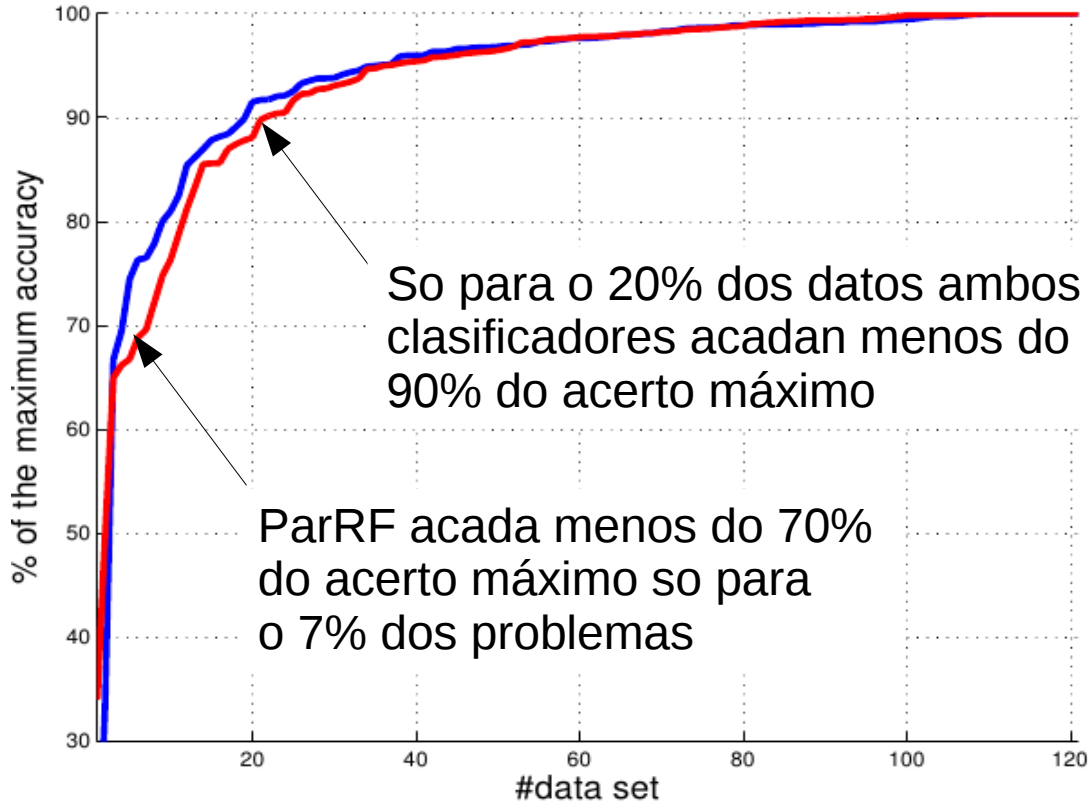
23ª: Direct Kernel Perceptron (o noso)

24ª: GaussPrRadial, clasificador baseado en procesos gausianos

O resto nos 25 mellores son Random Forests ou SVMs!

Comparación entre parRF (1º) e LIBSVM (2º)

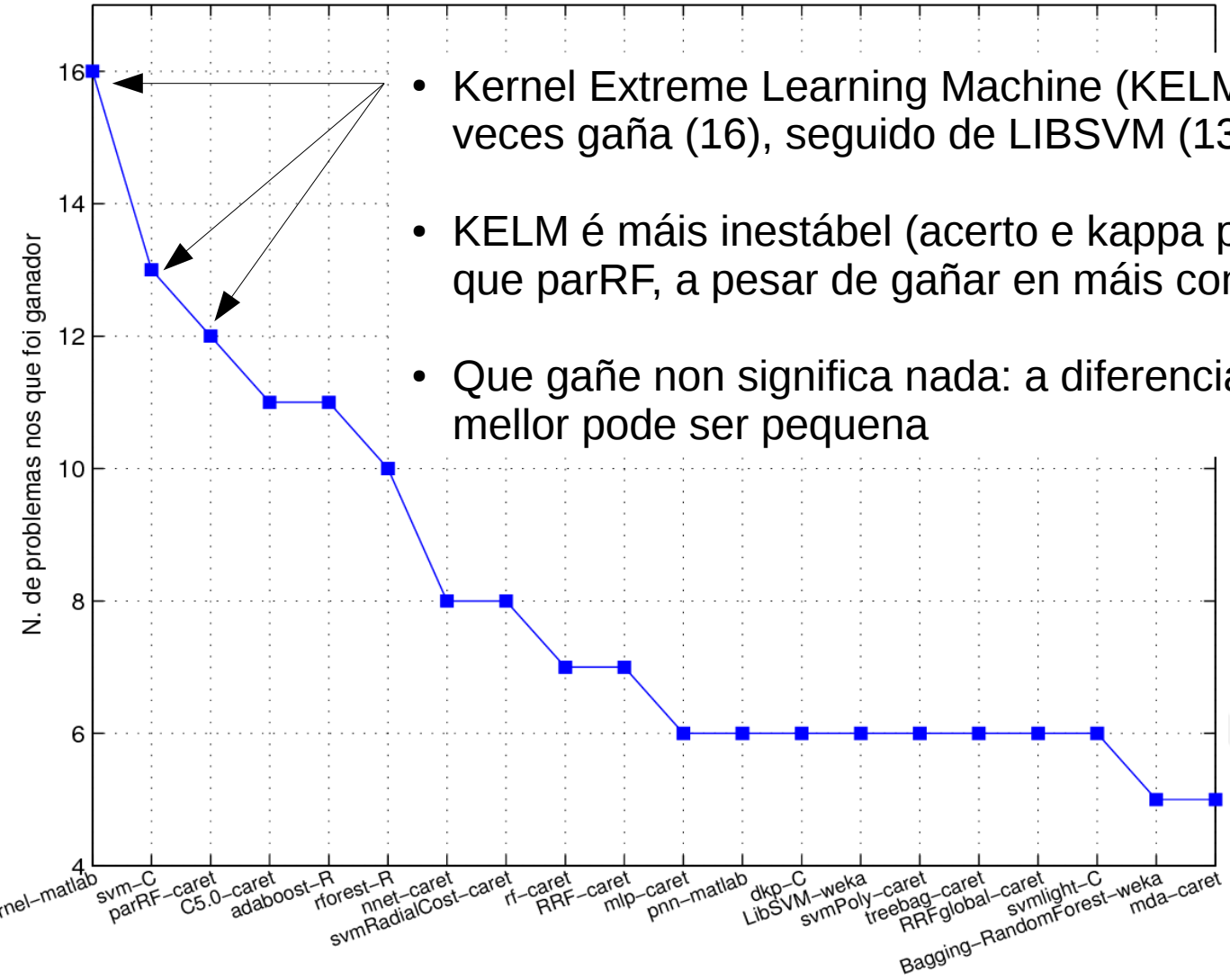
- Porcentaxe do acerto máximo para parRF (azul) e LIBSVM (vermello) ordeado por porcentaxes crecentes



- Conclusión: sería raro que parRF ou LIBSVM acaden nun problema novo menos do 90% do acerto máximo posíbel

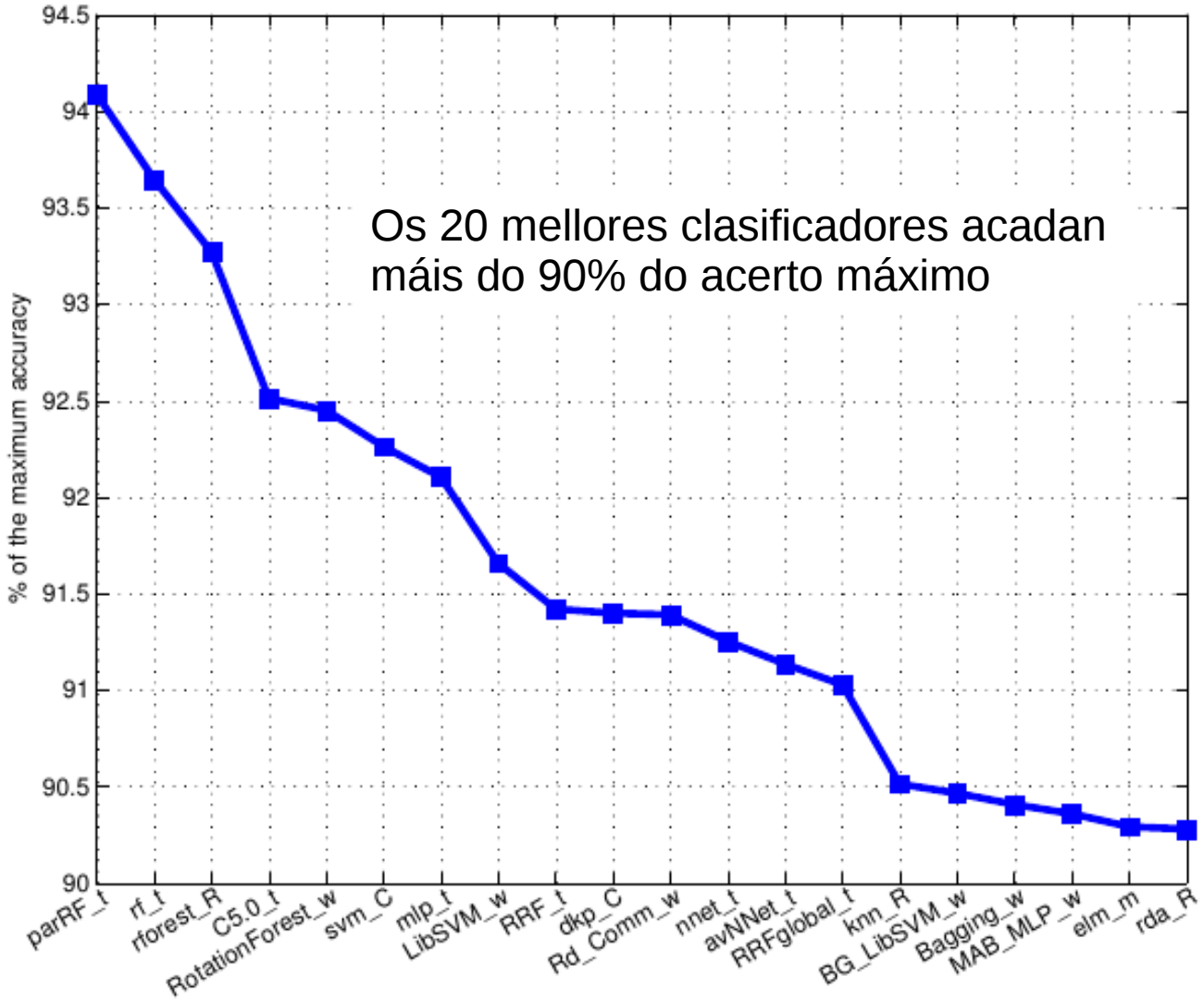
CiTIUS

En cantos problemas un clasificador é o mellor?



- Kernel Extreme Learning Machine (KELM) é a que máis veces gaña (16), seguido de LIBSVM (13) e parRF 12)
- KELM é máis inestábel (acerto e kappa promedio menores) que parRF, a pesar de gañar en máis conxuntos de datos
- Que gañe non significa nada: a diferencia co segundo mellor pode ser pequena

20 clasificadores coa mellor porcentaxe do acerto máximo



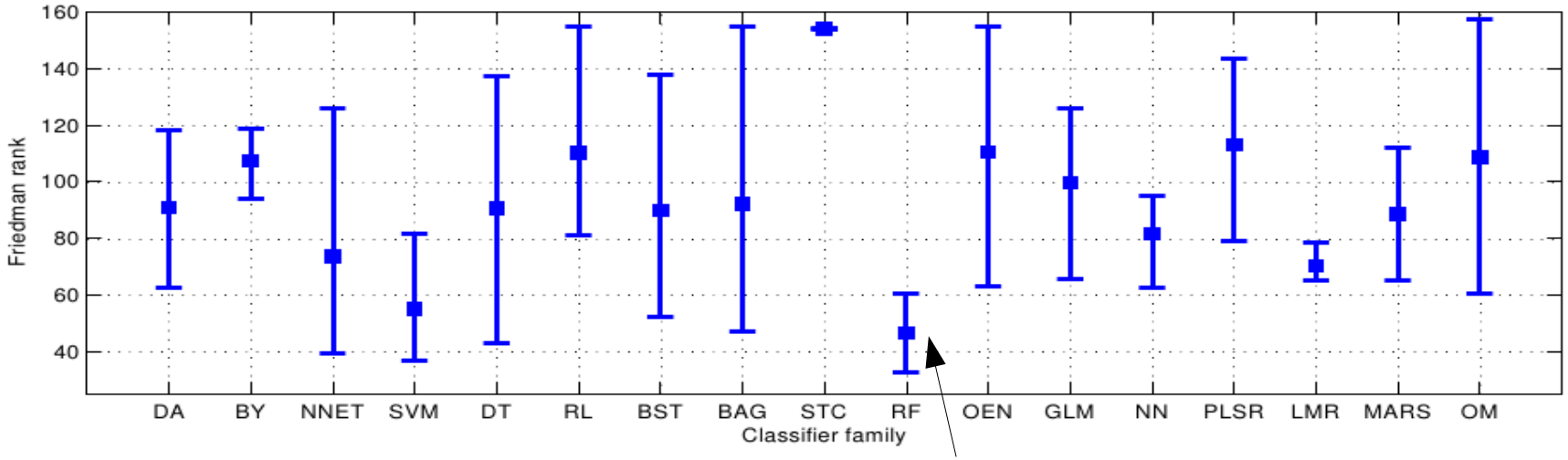
Si pero, con que probabilidade acadan perto do acerto máximo?

- Probabilidades de que os mellores clasificadores superen o 95% do acerto máximo:

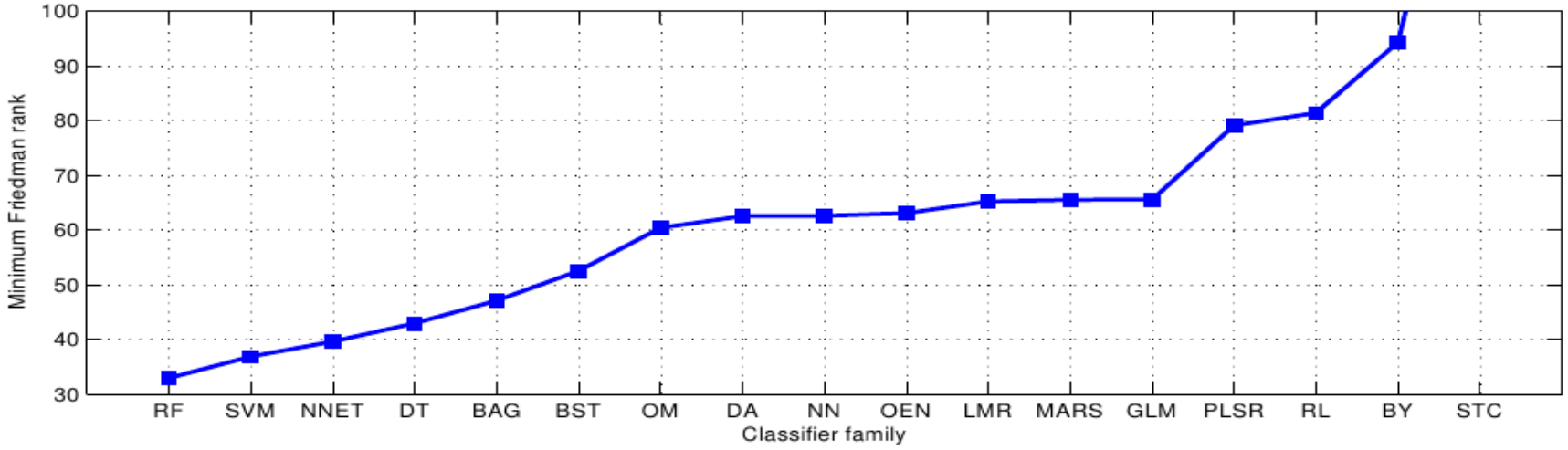
No.	Classifier	P95	No.	Classifier	P95
1	parRF_t	71.1	11	elm_kernel_m	60.3
2	svm_C	70.2	12	MAB-LibSVM_w	60.3
3	rf_t	68.6	13	RandomForest_w	57.0
4	rforest_R	65.3	14	RRF_t	56.2
5	Bagging-LibSVM_w	63.6	15	pcaNNet_t	55.4
6	svmRadialCost_t	63.6	16	RotationForest_w	54.5
7	svmRadial_t	62.8	17	avNNet_t	53.7
8	svmPoly_t	62.8	18	nnet_t	53.7
9	LibSVM_w	62.0	19	RRFglobal_t	53.7
10	C5.0_t	61.2	20	mlp_t	52.1

CiTIUS

Resultados por familias



Random Forest exhibe o mellor ranking de Friedman e o intervalo máis estreito



Mellores clasificadores de cada familia

Familia	Mellor clasificador	Acerto/ Posición	Familia	Mellor clasificador	Acerto/ Posición
DA	FDA-T	78.4/26ª	Stacking	Stacking	49.3/170ª
Bayesianos	BayesNet-W	75.1/64ª	Outros ensembles	Random Committees	79.9/30ª
Redes neuronais	KELM	82.0/6ª	GLM	GLMNET-R	77.8/39ª
SVM	LIBSVM	81.8/3ª	Veciños máis cercanos	KNN-T	78.6/27ª
DT	Random SubSpace-W	77.4/64ª	PLSR	PLS-T	74.4/77ª
RL	C5.0Rules-T	76.7/82ª	LMR	MULTINOM-T	77.9/37ª
Boosting	C5.0-T	80.6/9ª	MARS	GCV-EARTH	77.4/38ª
Bagging	BG-LIBSVM	80.8/13ª	Outros	GAUSS-PR-RADIAL	80.1/24ª

Problemas de dúas clases

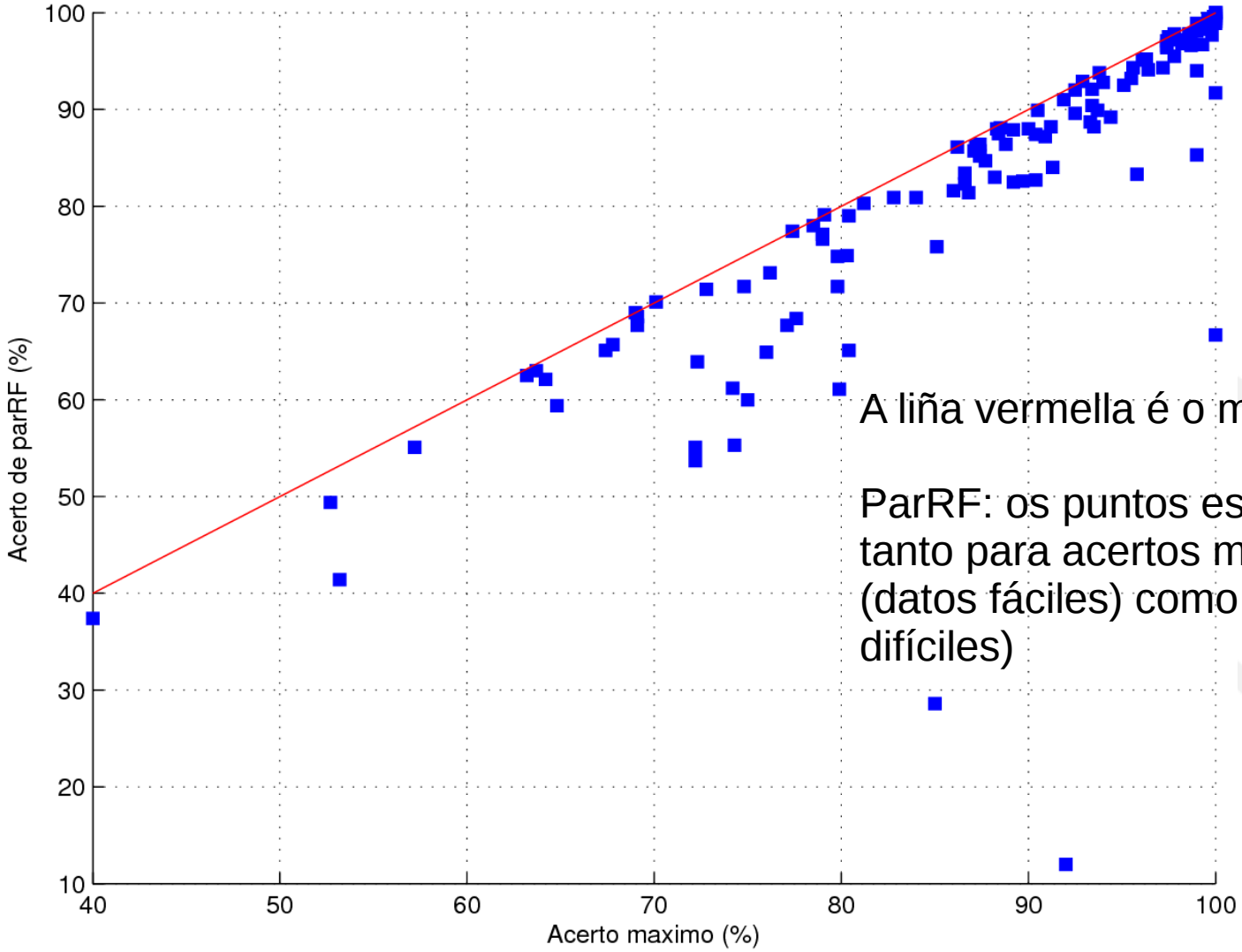
- AvNNet (comité de 5 MLPs) e SVM-POLY son os mellores
- ParRF e RF están na 5ª e 6ª posicións, LIBSVM na 8ª
- O mellor (avNNet) acada 95% do acerto máximo

Rank	Classifier	Acc. (%)	Rank	Classifier	Acc (%)
36.2	avNNet_t	83.0	50.0	mlp_t	82.2
39.9	svmPoly_t	79.9	51.4	elm_kernel_m	77.5
41.0	pcaNNet_t	82.9	54.1	RotationForest_w	82.0
42.2	svmRadialCost_t	80.0	54.9	rforest_R	80.9
44.2	parRF_t	82.6	57.6	mlpWeightDecay_t	79.7
44.7	rf_t	81.2	57.7	svmBag_R	78.8
47.1	C5.0_t	82.0	59.7	fda_t	81.0
47.2	svm_C	79.0	60.8	cforest_t	74.7
47.5	nnet_t	82.1	61.5	Bagging_LibSVM_w	77.9
48.0	svmRadial_t	79.4	62.9	knn_t	80.4

CiTIUS

Discusión segundo as propiedades dos datos

- Dificultade dos datos: caso de parRF



A liña vermella é o mellor clasificador

ParRF: os puntos están xunto á diagonal tanto para acertos máximos grandes (datos fáciles) como pequenos (datos difíciles)

Calculamos acertos ponderados en función de ...

- Acerto máximo, para a dificultade do problema
- Nº de patróns (crecente e decrecente)
- Nº de clases
- Nº de entradas

Acerto máximo

No.	Classifier	μ^C	No.	Classifier	μ^C
1	parRF_t	69.9	11	nnet_t	67.7
2	rf_t	69.6	12	dkp_C	67.6
3	rforest_R	69.3	13	RRFglobal_t	67.4
4	C5.0_t	69.0	14	Bagging_LibSVM_w	67.3
5	RotationForest_w	68.6	15	Decorate_w	67.1
6	svm_C	68.4	16	knn_t	67.1
7	mlp_t	68.4	17	Bagging_REPTree_w	67.0
8	RRF_t	68.1	18	elm_m	67.0
9	LibSVM_w	67.8	19	pda_t	67.0
10	avNNet_t	67.8	20	RandomCommittee_w	66.9

Aumentando e reducindo o nº de patróns

Aumentando nº patróns

No.	Classifier	μ^P	No.	Classifier	μ^P
1	rf_t	91.1	11	Bagging_LibSVM_w	89.9
2	parRF_t	91.1	12	RandomCommittee_w	89.9
3	svm_C	90.7	13	Bagging_RandomTree_w	89.8
4	RRF_t	90.6	14	MultiBoostAB_RandomTree_w	89.8
5	RRFglobal_t	90.6	15	MultiBoostAB_LibSVM_w	89.8
6	LibSVM_w	90.6	16	MultiBoostAB_PART_w	89.7
7	RotationForest_w	90.5	17	Bagging_PART_w	89.7
8	C5.0_t	90.5	18	AdaBoostM1_J48_w	89.5
9	rforest_R	90.3	19	Bagging_REPTree_w	89.5
10	treebag_t	90.2	20	MultiBoostAB_J48_w	89.4

Reducindo nº patróns

No.	Classifier	μ^D	No.	Classifier	μ^D
1	rf_t	82.1	11	MultiBoostAB_LibSVM_w	79.7
2	rforest_R	81.8	12	LibSVM_w	79.6
3	svm_C	81.6	13	RandomCommittee_w	79.5
4	parRF_t	81.6	14	dkp_C	79.5
5	RRF_t	80.8	15	nnet_t	79.3
6	RotationForest_w	80.3	16	elm_kernel_m	79.2
7	C5.0_t	80.2	17	avNNet_t	79.2
8	mlp_t	80.0	18	treebag_t	79.0
9	Bagging_LibSVM_w	80.0	19	MAB_MLP_w	78.8
10	RRFglobal_t	79.8	20	knn_R	78.7

CiTIUS

Segundo o número de clases e entradas

Número de clases

No.	Classifier	μ^L	No.	Classifier	μ^L
1	svm_C	80.5	11	RotationForest_w	76.6
2	rf_t	80.5	12	RRFglobal_t	76.1
3	rforest_R	79.8	13	MultilayerPerceptron_w	76.1
4	Bagging_LibSVM_w	79.7	14	rda_R	76.0
5	parRF_t	79.5	15	knn_R	75.9
6	MultiBoostAB_LibSVM_w	79.5	16	SMO_w	75.6
7	LibSVM_w	79.5	17	hdda_R	75.4
8	RRF_t	77.9	18	KStar_w	75.3
9	dkp_C	77.7	19	elm_m	75.1
10	MAB_MLP_w	76.9	20	RandomCommittee_w	75.1

Número de entradas

No.	Classifier	μ^I	No.	Classifier	μ^I
1	parRF_t	84.0	11	mlp_t	81.5
2	rf_t	83.3	12	SMO_w	81.3
3	rforest_R	82.9	13	Bagging_RandomTree_w	81.3
4	RotationForest_w	82.8	14	elm_kernel_m	81.1
5	MAB_MLP_w	82.5	15	mlp_C	81.0
6	LibSVM_w	82.4	16	dkp_C	80.8
7	MultilayerPerceptron_w	82.0	17	fda_t	80.8
8	svm_C	82.0	18	rda_R	80.8
9	RandomCommittee_w	81.8	19	SimpleLogistic_w	80.7
10	C5.0_t	81.6	20	RRF_t	80.4

CiTIUS

Conclusións (I)

- O Random Forest é o mellor e a mellor familia de clasificadores (94% do acerto máximo; supera o 95% do acerto máximo no 71% dos casos): 7 RFs entre os 20 mellores
- Sorprendente: clasificador vello (20 anos), con pouco esforzo en investigación no seu campo
- Non hai ningún clasificador que sexa o mellor en todo … pero case.
- A SVM é a segunda familia (92% do acerto máximo, 5 SVMs entre os 20 mellores)
- Resultados similares ao traballo de *Meyer, Leisch, Korner (2003): The Support Vector Machine under Test, Neurocomputing, 55(1-2),pp. 169-186*, pero incluindo moitos máis clasificadores e problemas

Conclusións (II)

- Parece que o progreso é unha ilusión (artigo de 2006)
- Outros clasificadores bos: KELM, C5.0-T, avNNet-T, nnet-T, RotationForest, PNN
- As mellores implementacións parecen ser as de Caret porque permiten a sintonización de parámetros
- Recibín comentarios sobre a ausencia de *Gradient Boosted Machines* (GBM)
- Nun traballo posterior usamos 428 clasificadores, combinando en Weka ensembles e clasificadores base
- Ao principio pensaba: ¿será de interese unha comparativa tan grande?
- Si, porque así non está condicionada nin polos conxunto de datos elexidos, nin pola colección de clasificadores, que se pode considerar completa

Grazas pola vosa atención

Manuel Fernández Delgado: manuel.fernandez.delgado@usc.es

citius.usc.es