Proyecto final
Métodos avanzados en aprendizaje automático

Alejandro Bellogín Kouki
Universidad Autónoma de Madrid
alejandro.bellogin@uam.es

19 de mayo de 2008

Outline

Preprocesamiento I

1. Selección de atributos:
   - Análisis de los atributos junto con su descripción
   - Estudio de matriz de covarianza y correlaciones
   - Árboles de decisión y algoritmos de búsqueda de atributos (greedy, algoritmos genéticos). Los atributos que se encontraron más relevantes fueron los siguientes: ANT_MUT, ANT_POL, FP, TIPO, BONF, recibosdev.
   - Se probó PCA (subconjunto de 30 atributos)pero no se obtuvieron buenos resultados

Preprocesamiento II

2. Reemplazar los *missing values* por la media del atributo.
3. Centrar.
4. Normalizar.
5. Eliminar *outliers*. Detectamos *outliers* utilizando la siguiente fórmula basada en el identificador de Hampel [3]: si el valor de un elemento x para el atributo i cumple

$$x_i - \mu_i > 3{,}8 \cdot \sigma_i$$

lo contabilizamos como *outlier*

Preprocesamiento III

- Antes:

Núm. ejemplos	Núm. atributos
79999	70

Cuadro: Estadísticas del conjunto inicial

- Después:

Núm. ejemplos	Núm. atributos
72702	42

Cuadro: Características del conjunto final de datos

Outline

Conjuntos de datos

- Entrenamiento: clases balanceadas (submuestreo).
- Test: misma proporción que en construcción.
- No validación! (problemas: distribución a seguir, medida de error, menos casos de entrenamiento)

Outline

Métodos de clasificación I

Clasificador	Sensibilidad media	Sensibilidad máxima con entrenamiento	Sensibilidad máxima con test
Bagging con regresión	0,2783	0,7676	0,2742
Regresión lineal	0,2773	0,7594	0,2766
Least Med Sq	0,2772	0,7660	0,0 [error]
$M5'$	0,2756	0,7684	0,2763
Stacking	0,2756	0,7684	0,2763
Bagging con $M5'$	0,2750	0,9296	0,5723
NaiveBayes	0,2418	0,7234	0,2434
NBTree	0,2418	0,7234	0,1563
Comité aleatorio	0,2408	0,9182	0,0
J48	0,2247	0,6997	0,0
J48graft	0,2236	0,6997	0,0
REPTree	0,2124	0,7701	0,2812
Perceptrón multicapa	0,1905	0,8437	0,2561
Bosque aleatorio	0,1820	1,0	1,0
Reglas Part	0,1596	1,0	0,5219
Reglas conjuntivas	0,1554	0,6645	0,1309

Cuadro: Sensibilidad de los clasificadores estudiados para un conjunto de entrenamiento formado por el 60 % de elementos de fuga del conjunto original

Métodos de clasificación II

Proporción del conjunto de fuga en entrenamiento	Bagging con regresión	Regresión lineal	Least Med Sq	$M5'$	Stacking	Bagging con $M5'$
20 %	0,2773	0,2756	0,2744	0,2761	0,2761	0,2829
30 %	0,2784	0,2786	0,2781	0,2768	0,2769	0,2639
40 %	0,2811	0,2827	0,2793	0,2819	0,2819	0,2771
50 %	0,2787	0,2780	0,2809	0,2802	0,2802	0,2770
60 %	0,2775	0,2751	0,2766	0,2721	0,2721	0,2754
70 %	0,2847	0,2839	0,2855	0,2887	0,2887	0,2799
80 %	0,2722	0,2745	0,2758	0,2734	0,2733	0,2843

Cuadro: Sensibilidad sobre distintos conjuntos de entrenamiento y test para los seis mejores algoritmos

Métodos de clasificación III

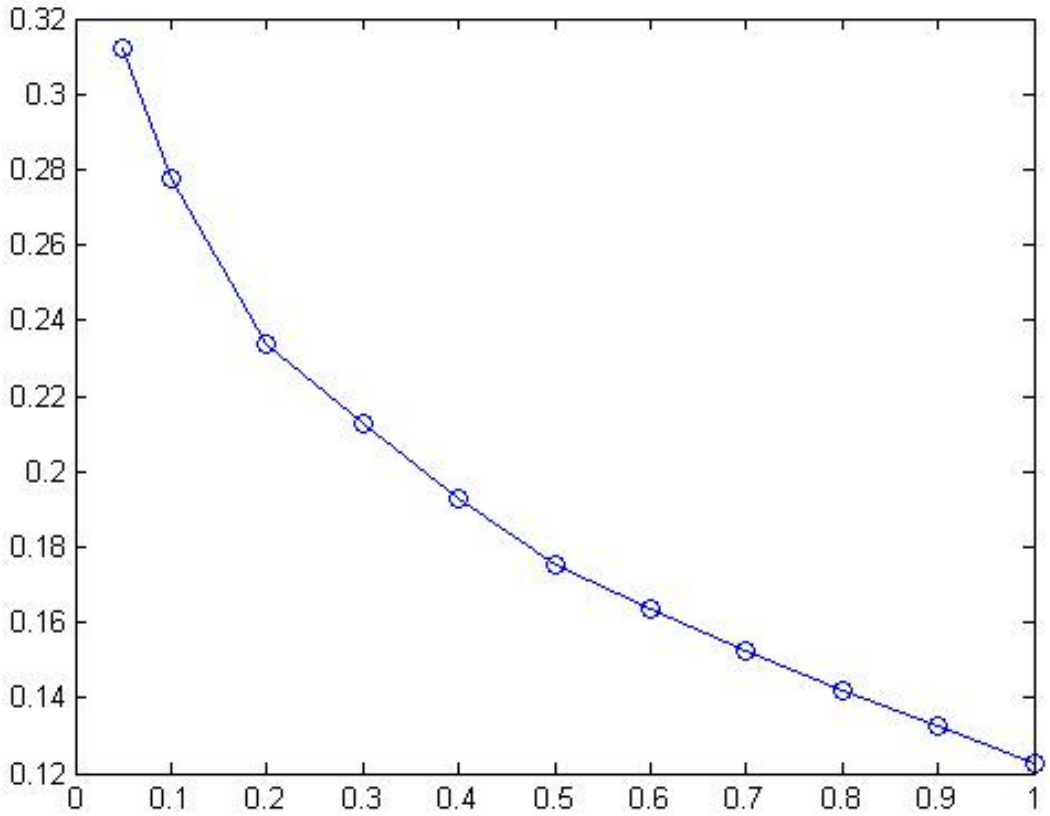

Figura: Curva Hit rate frente a tamaño de campaña para bagging con regresión lineal

Métodos de clasificación IV

- Se eligió *bagging con M5′*:
  - En media no ha sido el mejor que se ha comportado
  - Pero ha demostrado ser robusto frente a cambios de proporción del conjunto de fuga en entrenamiento (tabla 4)
  - Se comporta bien frente a conjuntos de datos no balanceados (cuarta columna de la tabla 3).

Métodos de clasificación: Weka

Clasificador	Necesita datos discretos	Parámetros
Bagging con $M5'$	No	`-P 90 -S 1 -I 100 -W weka.classifiers.trees.M5P -- -M 50`
Bagging con regresión	No	`-P 90 -S 1 -I 50 -W weka.classifiers.functions. LinearRegression -- -S 0 -R 0.1`
Bosque aleatorio	Sí	`-I 10 -K 1 -S 1`
Comité aleatorio	Sí	`-S 1 -I 10 -W weka.classifiers.trees. RandomTree -- -K 5 -M 50.0 -S 1`
J48	Sí	`-C 0.0657 -M 50`
J48graft	Sí	`-C 0.15 -M 50`
Least Med Sq	No	`-S 4 -G 0`
$M5'$	No	`-M 50`
NaiveBayes	Sí	(ninguno)
NBTree	Sí	(ninguno)
Perceptrón multicapa	No	`-L 0.3 -M 0.2 -N 150 -V 0 -S 0 -E 20 -H a`
Regla conjuntiva	No	`-N 3 -M 2.0 -P -1 -S 1`
Reglas Part	Sí	`-M 2 -C 0.25 -Q 1`
Regresión lineal	No	`-S 0 -R 0.1`
REPTree	No	`-M 2 -V 0.0010 -N 3 -S 1 -L -1`
Stacking	No	`-X 10 -M "weka.classifiers.functions. LinearRegression -S 0 -R 0.1S 1 -B "weka.classifiers.trees.M5P -M 50.0"`

Cuadro: Algunas características de los clasificadores utilizados

Outline


Christopher J. C. Burges.
A tutorial on support vector machines for pattern recognition.
Data Mining and Knowledge Discovery, 2(2):121–167, 1998.


Iván Cantador.
Aplicación de perceptrones paralelos y adaboost a problemas de clasificación desequilibrados.
Master's thesis, Departamento de Ingeniería Informática de la Universidad Autónoma de Madrid (UAM). Madrid, España, Junio 2005.


Laurie Davies and Ursula Gather.
The identification of multiple outliers.
Journal of the American Statistical Association, 88(423):782–792, 1993.


Richard O. Duda, Peter E. Hart, and David G. Stork.
Pattern Classification (2nd Edition).
Wiley-Interscience, Noviembre 2000.


Gonzalo Martínez-Muñoz and Alberto Suárez.
Aggregation ordering in bagging.
In *Proc. of the IASTED International Conference on Artificial Intelligence and Applications*, pages 258–263. Acta Press, 2004.


G. Weiss and F. Provost.
The effect of class distribution on classifier learning.
Technical report, 2001.


Ian H. Witten and Eibe Frank.
Data Mining: Practical Machine Learning Tools and Techniques.
Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, Junio 2005.