MINERÍA DE ASPECTOS AVANZADOS

Adrián Calzadilla González

19/4/2017

Contents

List of Figures

List of Tables

Información general

- Nombre y apellido: Adrián Calzadilla González
- Usuario kaggle: Adrián
- Código: https://github.com/AdCalzadilla/md_aspectosavanzados_nb

Introducción

Práctica donde se ejecutan los conocimientos y métodos de aprendizaje vistos durante el desarrollo de la asignatura *Minería de datos: Aspectos Avanzados*. Para ello se hará uso de la plataforma *Kaggle*, que permite establecer una competición de clasificación avanzada entre todos los alumnos.

Clasificación no balanceada

Problema de carácter matemático con un total de 6.400 instancias, que se han dividido al 50% entre entrenamiento y test. Este conjunto de datos está representado por un total de 22 atributos con valores numéricos y dos clases, con un ratio de desbalanceo (IR) de aproximadamente 2.

El objetivo será alcanzar la máxima precisión en términos de la medida AUC(2008).

Bitácora

Viernes 24 de Marzo

Primer contacto con el problema. En ese momento sólo se aceptaban dos subidas a kaggle y el fichero *csv* debía tener valores de 0 y 1.

En la primera entrada, simplemente se le aplicó el algortimo *Random Forest*(2006) al conjunto de datos y se subió a la plataforma *kaggle* para ver el funcionamiento de la plataforma, el comportamiento del dataset sin preprocesar y un resultado que diera un punto de partida.

En la segunda entrada del día se realizó preprocesamiento que consistió:

- Normalización: mediante el paquete *Caret* interpolar las variables continuas en el intervalo 0, 1.
- $SMOTE$(2006): a partir de la librería *unbalanced* se ejecuta este algortimo para balancear las clases.

Finalmente se le volvió a aplicar el algoritmo *Random Forest*.

Lunes 27 de Marzo

Siguen habiendo dos subidas permitidas a *kaggle*. Aunque los ficheros *csv* siguen teniendo que tener los valores 0 o 1, un grupo de alumnos nos damos cuenta que subiendo probabilidades los resultados mejoran considerablemente.

En la primera entrada de ese día, se le aplica al dataset *gbm* (*Gradient Boosting Machine*)(2001) para ver como responde ante este algoritmo.

En la segunda, se vuelve a realizar normalización y *SMOTE*. Posteriormente se le vuelve a aplicar *gbm*.

Viernes 31 de Marzo

Las entradas permitidas a *kaggle* suben a 6 y el formato del fichero *csv* tiene que tener probabilidades entre 0 y 1.

En este día se realizan dos subidas: en la primera se realiza detección de *outliers*, se aplica normalización y *SMOTE* al dataset. Posteriormente se ejecuta *gbm*. El resultado es peor que los obtenidos el Lunes.

En la segunda subida, se vuelve a aplicar normalización y *SMOTE*, pero no se detectan los *outliers*. A continuación, se aplica el algoritmo *gbm* y se le cambia la distribución por defecto, *bernoulli*, por *multinomial*. El resultado es peor que el realizado anteriormente.

Martes 4 de Abril

En el preprocesamiento se utiliza detección de *outliers*, se normalizan los datos, se realiza selección de variables mediante el algoritmo *mrMR* y se utiliza *SMOTE* para el balanceo de datos.

Posteriormete se le aplica el algorimt *svm* (*Support Vector Machines*)(1998).

Se realizan dos entradas a *kaggle* una con cada columna de probabilidad devuelta por el algoritmo. Dan el mismo resultado.

Miércoles 5 de Abril

Se realizan los siguientes pasos en el preprocesamiento:

- Detección de *outliers*.
- Normalización de los datos.
- Selección de variables mediante *mrMr*(*max-relevance min-redundancy*)(2005).

Una vez preprocesado el dataset, el siguiente paso es utilizar las distintas técnicas de balanceo sobre él:

- Sin balanceo.
- *Undersampling*.
- *Oversampling*.
- Variables sintéticas.

Y se generan modelos con *Random Forest* y *svm*, los algoritmos que mejor resultado habían ofrecido, para cada uno de los dataset obtenidos con las diferentes técnicas de balanceo.

Una vez ejecutado los algoritmos se realiza la entrada a *kaggle* de los dos mejores resultados obtenidos en local, *undersampling* con *svm* y *undersampling* con *Random Forest*.

Seguramente debería haber subido todos los modelos al *kaggle*.

Viernes 7 de Abril

Se realizan los mismos pasos anteriores en el preprocesamiento, aunque intenta afinar más en la detección de *outliers* y en la selección de variables. También se realizan las distintas técnicas de balanceo realizadas el día anterior y se vuelven a aplicar los algoritmos *Random Forest* y *svm*.

En este caso, se decide realizar la entrada a *kaggle* de cuatro modelos. Los 2 generados con *SMOTE*, uno con *undersampling* y otro sin balancear.

El modelo realizado con *SMOTE* y *Random Forest* obtuvo el mejor resultado en la clasificación final.

Tabla de entradas

En la *Tabla 1* se puede observar todas las entradas a *kaggle* realizadas con sus puntuaciones, algoritmos, si se ha realizado balanceo y de que tipo, y la fecha en que se realizó la subida.

Table 1: Tabla de resultados

Fecha	Nombre	Balanceo	Algoritmo	Punt. Pública	Punt. Privada
2017-03-24	result24M-imb.csv	Sin	Random Forest	0.70224	0.67466
2017-03-24	result24M-SMOTE-imb.csv	SMOTE	Random Forest	0.70587	0.69894
2017-03-27	result27.imb.csv	Sin	gbm	0.78431	0.79264
2017-03-27	result27_2.imb.csv	SMOTE	gbm	0.79248	0.80209
2017-03-31	result_31.imb.csv	SMOTE	gbm	0.76040	0.77283
2017-03-31	result_31_SinOutlier.imb.csv	SMOTE	gbm	0.75029	0.76631
2017-04-04	result_4A.svm.csv	SMOTE	svm	0.82074	0.80672
2017-04-04	result_5A.svm.csv (1)	SMOTE	svm	0.82073	0.80674
2017-04-04	result_5A.svm.csv (2)	SMOTE	svm	0.82073	0.80674
2017-04-05	result_5A.un.svm.csv	Undersampling	svm	0.76716	0.78729
2017-04-05	result_5A.un.rf.csv	Undersampling	Random Forest	0.79487	0.80600
2017-04-07	result_7A.un.svm.csv	Sin	svm	0.80385	0.80230
2017-04-07	result_7A.un.svm.csv	Undersampling	svm	0.80385	0.80230
2017-04-07	result_7A.smote.rf.csv	SMOTE	Random Forest	0.81198	0.81229
2017-04-07	result_7A.smote.svm.csv	SMOTE	svm	0.76714	0.77013

Gráfico de resultados

En la *Figura 1* se puede observar las puntuaciones obtenidas tanto en la clasificación privada como en la pública respecto a la fecha de entrada en *kaggle*.

Conclusiones

En esta práctica he aprendido diferentes formas de aproximación a un problema desbalanceado. Así como las librerías que permiten realizar el balanceo de datos en *r*. Además, he aprendido el funcionamiento de la herramienta *kaggle* y su importancia dentro de la ciencia de datos. También, he visto nuevas formas de preprocesamiento y he aplicado diferentes algoritmos para la generación de modelos y he analizado su comportamiento frente al dataset.

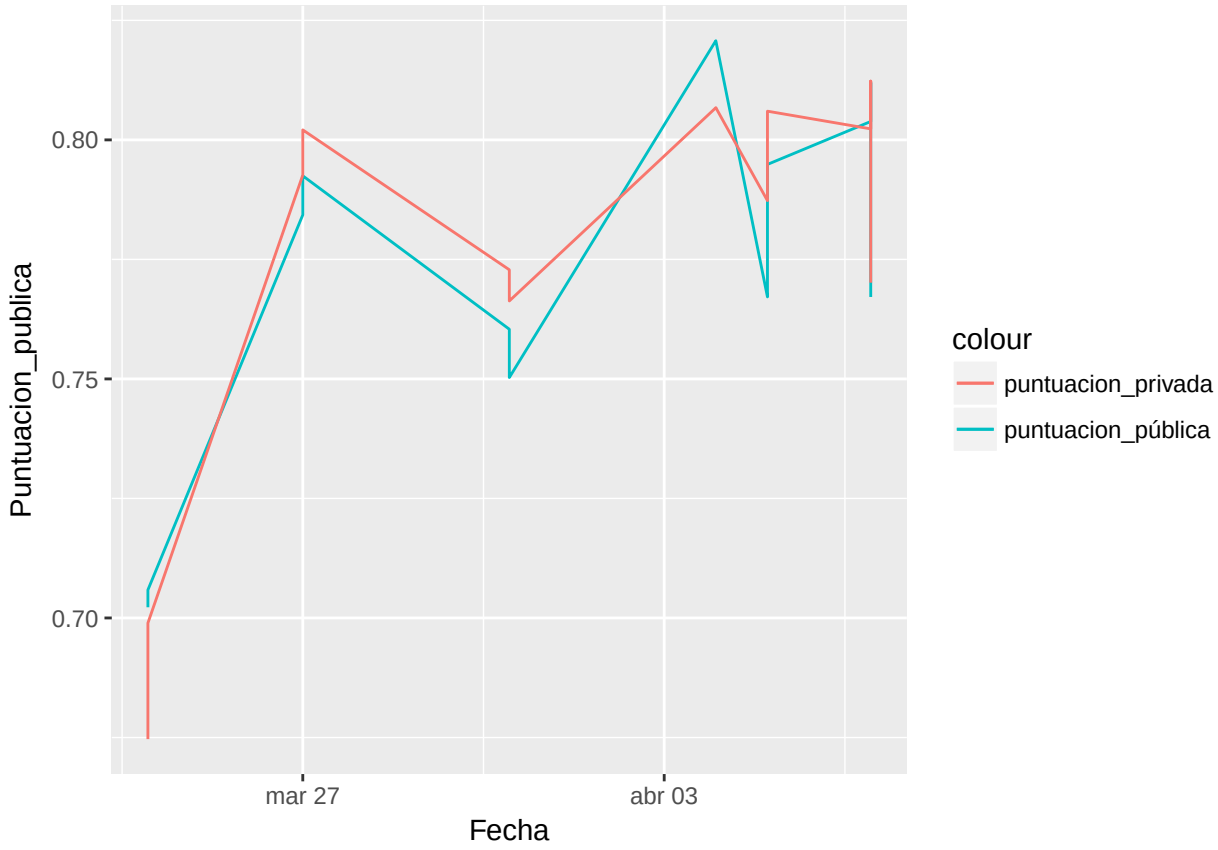

Figure 1: Puntuaciones obtenidas

Bibliografía

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*. JSTOR, 1189–1232.

Joachims, Thorsten. 1998. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." *Machine Learning: ECML-98*. Springer, 137–42.

Lobo, Jorge M, Alberto Jiménez-Valverde, and Raimundo Real. 2008. "AUC: A Misleading Measure of the Performance of Predictive Distribution Models." *Global Ecology and Biogeography* 17 (2). Wiley Online Library: 145–51.

Peng, Hanchuan, Fuhui Long, and Chris Ding. 2005. "Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (8). IEEE: 1226–38.

Rodriguez, Juan José, Ludmila I Kuncheva, and Carlos J Alonso. 2006. "Rotation Forest: A New Classifier Ensemble Method." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (10). IEEE: 1619–30.

Wang, Juanjuan, Mantao Xu, Hui Wang, and Jiwu Zhang. 2006. "Classification of Imbalanced Data by Using the Smote Algorithm and Locally Linear Embedding." In *Signal Processing, 2006 8th International Conference on*. Vol. 3. IEEE.