

Balanceado de bases de datos

Pablo.Bermejo@uclm.es

El Enron Corpus es un conjunto de correos electrónicos pertenecientes a antiguos trabajadores de la empresa Enron. Tras un escándalo financiero, los correos electrónicos de todos los empleados de la empresa fueron hechos públicos para su estudio por parte de las autoridades.

Para cada empleado, disponemos de todos sus correos almacenados en carpetas tal y como el mismo empleado las definió en su día. Así, la carpeta a la que un correo pertenece puede verse como un posible valor de la variable clase a predecir a partir del contenido de los correos electrónicos.

En el enlace http://www.cs.umass.edu/~ronb/datasets/enron_flat.tar.gz puede descargarse la colección preprocesada de correos de 7 de estos usuarios. El preprocesamiento realizado a estas bases de datos consiste en:

- No existen jerarquías de carpetas. Los correos pertenecientes a subcarpetas han sido copiados al primer nivel.
- Las carpetas con menos de 3 e-mails han sido eliminadas.
- El campo *X-folder* de las cabeceras ha sido eliminado ya que contenía el nombre de la carpeta.
- Se han eliminado las carpetas *all_documents*, *calendar*, *contacts*, *deleted_items*, *discussion_threads*, *inbox*, *notes_inbox*, *sent*, *sent_items* and *_sent_mail*.

En clase se os facilitará con las bases de datos de 7 usuarios convertidas en formato .Arff, donde cada instancia representará un correo electrónico y se encuentran ordenadas en orden temporal de llegada a la bandeja de entrada del usuario correspondiente.

El caso de la clasificación de correo electrónico en carpetas es especial debido a su carácter temporal. Así, modelos típicos de evaluación como Cross-Validation no son válidos pues selecciona los conjuntos de entrenamiento y validación aleatoriamente. En este problema utilizaremos el modelo de evaluación: time-based splits evaluation. Como Weka no ofrece este modelo, hay que implementarlo en weka. Esta evaluación consiste simplemente en entrenar con los primeros N documentos y validar con los siguientes 100, entonces entrenar con los primeros $2N$ documentos y validar con los siguientes 100; luego entrenar con los primeros $3N$ y validar con los siguientes 100,... así hasta validar con los últimos 100 o menos documentos.

Como las instancias en las bases de datos que se os entregan ya están ordenadas temporalmente, la evaluación consistirá en crear los conjuntos de entrenamiento desde la instancia 0 hasta la $N-1$, $2N-1$,..., correspondientemente, los conjuntos de validación con las instancias N hasta $N+99$, $2N$ hasta $2N + 99$,... No tenéis que implementar esta evaluación, se os dará una clase `newEvaluation` que extiende la clase `weka.classifiers.Evaluation` conteniendo el método `public void timeBasedSplitEvaluation(Classifier c, Instances data, int N)`.

En esta base de datos nos encontraremos con el problema de que el número de documentos por clase (carpeta) no está uniformemente distribuido, y por lo

tanto nuestro clasificador aprenderá modelos sesgados a partir de los conjuntos de entrenamiento. Así que compararemos los resultados antes y después de aplicar un filtro de balanceado previo a la clasificación.

Las tareas a realizar en esta práctica son:

1. Extraer las características de cada usuario:
 - a. Número de carpetas (cardinalidad de la clase).
 - b. Número de instancias
 - c. Número de atributos
 - d. (base,Peak): número menor y mayor de documentos encontrados a lo largo de todas las carpetas.
 - e. (μ : σ): media y desviación típica del número de documentos por clase.
2. Realizar clasificación supervisada utilizando: Naive Bayes Multinomial, Support Vector Machine e ibK (k=1). Calcular las siguientes métricas: Accuracy general, Precisión para cada clase, Recall para cada clase, ROC, AUC. Fijar el parámetro del time-based splits evaluation N=100.
3. Igual que el anterior pero preprocesando los conjuntos de entrenamiento con el filtro `weka.filters.supervised.instance.SMOTE`.
4. Conclusiones:
 - a. Comenta y compara los resultados antes y después de balancear.
 - b. Compara el beneficio de balancear en los distintos clasificadores.
 - c. Comenta qué métrica/s te ha parecido más útil.