DATA MINING AZKEN PRAKTIKA

ÍNDICE:

PLANTEAMIENTO DEL PROBLEMA

Primeramente, para afrontar esta práctica de trabajo conjunto de la clase, nos dividimos en varios subgrupos (el nuestro de 6 personas), donde cada uno abordaría un apartado del problema.

Así pues, en nuestro grupo decidimos encargarnos de implementar algún filtro (como el “Resample” y/o el “Smote”) para poder hacer frente a los ficheros, inicialmente muy cargados de datos y con un desequilibrio entre los valores positivo y negativo de la clase, y varios sistemas de clasificación (como el “RandomForest”, “Support Vector Machine” o “SVM”, “OneClass”, “NaiveBayes” y “BayesNet”).

Como se puede ver, son muchos los clasificadores que decidimos implementar, por lo que entrenar con ellos para saber cuáles son los valores de los parámetros con los cuales se conseguía un *fmeasure* mejor con respecto al valor de la clase positiva resulta muy trabajoso.

De ahí que decidiésemos implementar, primeramente, un filtro como el “Resample”, que se encarga de reducir el tamaño del dataset inicial (22000 instancias, más o menos; con una diferencia entre clase positiva y negativa de 21980 vs 80), equilibrándolo, a su vez, y consiguiendo que la diferencia entre instancias con estimación de clase positiva y negativa fuese mínima.

En este punto, teníamos la duda de si intentar implementar el “Smote” nos iba a merecer la pena o no, ya que con el “Resample” se podía conseguir un resultado muy similar. Empezamos a implementar ambos módulos, individuales, pero al ver que el filtro del “Smote” nos daba problemas dependiendo de las librerías y las versiones de Java, decidimos centrarnos en el “Resample” y en la implementación y tratamiento de los clasificadores.

Una vez conseguimos el módulo ejecutable del “Resample”, sólo teníamos que integrarlo en el código de los clasificadores (también módulos independientes) para que el entrenamiento con el dataset de entrenamiento original no fuese tan largo, además de ese desequilibrio ya comentado entre los valores de la clase, el cual producía que al entrenar el modelo de los clasificadores el *fmeasure* de la clase positiva fuese muy baja.

Nos llevó trabajo implementar los clasificadores, ya que teníamos que decidir, primero de todo, qué parámetros queríamos tratar para buscar su valor óptimo (de cara a una clasificación cuyo modelo, después del entrenamiento, nos diese un *fmeasure* alto con respecto a la clase positiva), y después, ejecutar las pruebas `para comprobar su validez. Decidimos, pues, sacar los valores óptimos impresos por pantalla, y/o guardarlos en un fichero de texto plano, y extraer, también, el modelo para poder cargarlo, posteriormente en un módulo que se encargaría de hacer las estimaciones para el fichero de test.

Finalmente, para el paso final, decidimos generar un Script que ejecutara los módulos individuales tanto del filtro como de los distintos clasificadores, para que se crearan los modelos de clasificación. Otro script se encargaba de, al igual que el script anterior que cogía los módulos y los ejecutaba, coger los modelos generados recientemente y aplicarlos sobre el fichero de test, para sacar las predicciones respectivas de cada modelo de clasificación. Para terminar, nos valimos de la ayuda de la implementación de un módulo que se encargaba de ver qué predicciones se conseguían con la ejecución del segundo script mencionado, y, elegir, por votación, cuál era el valor de clase más votado en la estimación final para cada instancia del dataset del test, y así, decidir cuál era la clase estimada final.

REPARTO DEL TRABAJO:

El reparto de trabajo ha ido variando a medida que se iba avanzando en la práctica. En un principio, estos eran las tareas repartidas a cada miembro:

Javier Fernández -> Implementación de filtro “Smote” y de clasificador “SVM”

Unai Urkidi -> Implementación de filtro “Resample”

Alex Valero -> Implementación de clasificador “BayesNet”

Ander Orbegozo -> Implementación de clasificador “OneClassClassifier”

Mikel Arrospide -> Implementación de clasificador “BayesNet”

Xabier Apraiz -> Implementación de clasificador “RandomForest”

Pero, a la vista de que algunos compañeros implementaron algunos módulos más rápido que otros el reparto fue variando, de manera que unos se ayudaban a otros en la consecución de los objetivos. Ander y Alex consiguieron implementar sus módulos individuales inicialmente propuestos, mientras que Unai y Javi, entre ambos, consiguieron implementar el “Resample” y el “SVM”, dejando el filtro “Smote” como objetivo secundario por lo mencionado en el apartado anterior. Mikel y Xabi también consiguieron cumplir sus respectivos objetivos.

Aparte de eso, también generamos un ejecutable en java que servía para contar el número de instancias que había por cada valor de clase en el dataset, con el objetivo de que nos sirviera de ayuda para saber si el “Resample” y/o el “Smote” nos funcionaban como esperábamos. Así, con la ayuda de ese módulo, pudimos conseguir probar diferentes reducciones del tamaño del dataset por medio del “Resample” (distintas especificaciones del porcentaje a reducir), y decidir quedarnos con la reducción de instancias que menos diferencia entre valores de clase generase.

Los scripts corrieron cargo de todo el grupo, aunque fuese Ander el que más contribuyó en el proceso, al igual que la implementación del módulo de decisión de estimación final lo hizo Alex.

PROBLEMAS OBTENIDOS Y SOLUCIONES:

Muchos de los problemas que hemos tenido han sido más relativos a cuestiones internas del código Java que al entendimiento de uso de los clasificadores y/o filtros.

Como problema destacado podemos mencionar el uso del módulo “Smote” en Java. La implementación inicial que había conseguido hacer Javi había resultado exitoso; tanto que el jar ejecutable funcionaba correcta y debidamente, habiendo tratado los parámetros de porcentaje y *nearestNeighbors* para buscar una generación de instancias nuevas óptima en base al filtro. Pero, parece ser que el problema era que la versión con la que él había trabajado era de un JDK más antiguo que el del resto, y había errores de compilación. Además, a pesar de que ese módulo funcionaba para un JDK antiguo, el resto de los módulos cono podían ser el “OneClass” o el “SVM” no conseguían funcionar en esa versión. De modo que, tras laboriosas jornadas de trabajo sobre el “Smote”, se decidió dejarlo de lado con el objetivo de terminar el resto de la práctica, y a la vista de que el “Resample” podía servir, perfectamente, también para cumplir con la idea de reducción y equilibrio del dataset de entrenamiento.

//TODO

EXPERIMENTOS Y PRUEBAS:

Lara probar los distintos módulos hemos utilizado varias estrategias.

Para la prueba de los módulos encargados de ejecutar los filtros, hemos utilizado no sólo los ficheros de train y dev referentes a la práctica, sino también otros muchos arff-s, para ver que funcionaba bien.

Para probar los clasificadores y decidir qué valores de parámetros escoger sí que hemos utilizado el fichero de train y el de dev, ya que necesitábamos resultados que nos fuesen representativos para los datos de la práctica.

Para el módulo “Smote” hemos tratado de ver qué valor de *nearestNeighbors”* es la óptima y qué porcentaje de reducción es la más adecuada para aplicar.

Dicho esto, para el módulo de clasificación “SVM”, hemos tratado de encontrar valores óptimos a los parámetros *SVMType, KernelType* y *degree*. El primer parámetro se encarga de decidir si efectuamos clasificación o regresión. Para esta práctica nuestro objetivo es clasificar las instancias, no efectuar una regresión sobre las mismas, de modo que sólo dos valores de este parámetro han sido tratados. El segundo parámetro mencionado hace referencia al modo de calcular el límite que se encarga de separar espacios para la clasificación de instancias. (Es sabido que el “SVM” tiene parámetros como *Gamma, Un,* etc. Que son aplicados en las fórmulas que guarda el parámetro *KernelType*.). Y el tercer parámetro, es necesario para especificar qué grado de polinomios se tratarán para generar las funciones de separación de espacios que utiliza “SVM”.

El resto de parámetros los hemos dejado con sus valores por defecto, ya que hemos considerado que lo importante es saber qué grado de polinomios tratar para poder crear límites en forma de curva o recta, por ejemplo; qué método de clasificación de entre los dos que puede utilizar “SVM” utilizar; y qué fórmula de generación de límites es la más idónea para esa clasificación.

//TODO

CONCLUSIONES:

Se pueden sacar muchas conclusiones de esta práctica. Entre ellas, la más general e importante es la de que hemos trabajado conjuntamente en un trabajo que podría describirse como real. Hemos tenido tres datasets: uno de entrenamiento, otro de evaluación, y finalmente, uno de test, los cuales hemos tenido que usar debidamente (no pasar el test hasta tener los modelos bien entrenados con el train y el dev). Además, los datos estaban ocultos, siendo un claro problema de datos privados o confidenciales, de manera que los nombres de los atributos, así como sus posibles valores, se mostraban de manera que se observaba que habían sido sustituidos por nombres que no eran representativos. Esto nos ha evitado tener la tentación de mirar en exceso los dataset, para saber si los resultados obtenidos tenían sentido o no, y nos ha obligado a tratar de sacar conclusiones de los resultados en base a reglas estadísticas o matemáticas (como pueden ser las probabilidades de estimación de un valor de clase u otro, la frecuencia con que se estimaba una clase u otra, etc.).

Otra de las conclusiones que podemos sacar es que se necesita una idea inicial fija y bien pensada o amueblada para poder trabajar en grupo de manera eficaz y lógica: pensar cómo y cuándo se van a hacer los entregables o los objetivos parciales, pensar cómo podemos unir los distintos módulos generados, cómo representar los resultados, qué parámetros utilizar para no haber incongruencias entre módulos, etc.

Además, nos ha aportado mucho conocimiento el hecho de tener que abordar problemas con distintos clasificadores y sistemas de filtración de instancias, pues al tener que ayudarnos entre todos mutuamente, hemos tenido que preocuparnos de saber y entender las funcionalidades de los trabajos de cada fracción de grupo.

//TODO

## Como iniciar el programa

URL github: <https://github.com/aorbegozo005/Data_Mining_Practica_5>

Para facilitar usar el programa hemos creado la carpeta “Jar finalak”. En esta carpeta se encuentran todos los jar necesarios y los scripts sh.

El script está dividido en dos partes:

### Modeloak.sh

Con este script entrenaremos los modelos y se sacaran los mdl con los parámetros optimizados.

### Sailkatu.sh

Con este script se utilizaran los modelos creados por el anterior y se harán las predicciones. Por un lado se creara un archivo arff que será la copia del test pero con atributos añadidos de cada clasificador indicando su voto en la instancia.

Después se creara utilizando ese arff un archivo de texto plano indicando la predicción de cada instancia. El método utilizado ha sido el de votación por lo que la clase será el valor más votado entre todos los clasificadores.