Notas Reunión 30/03/2022

Presentes en la reunión: Prof. Benjamín Bustos, Matías Vergara

Trabajo realizado entre 23/03/2022 y 30/03/2022

- Lectura: Towards Class-Imbalance Aware Multi-Label Learning

→ La lectura sirve para validar ciertos conceptos aplicados para describir el dataset y el enfoque de los experimentos. En particular, se menciona Label Density y Label Cardinality, además de presentarse los enfoques lineales, no lineales y de ensemble mediante los mismos algoritmos utilizados en el benchmarking inicial.

→ Se presentan experimentos sobre 18 datasets distintos, de los cuales el label density más bajo es de 0,036. Recordar que en nuestro caso, dicha medida toma un valor de 0,015 en el caso de etiquetas normalizadas y 0,005 en el caso original.

→ Se mencionan los árboles de decisión como un buen clasificador base para el caso de datasets desbalanceados. Dado que no los había incluido en el benchmarking inicial, repetí experimentos agregando este nuevo clasificador. No se observaron resultados destacables.

→ Con respecto a la principal innovación del paper (COCOA), se trata de una manera de lidiar con el desbalance de correlaciones, y no de casos positivos/negativos como buscábamos. En particular, COCOA busca contribuir al aprendizaje de relaciones entre etiquetas (digamos, entre dos etiquetas A y B) cuando hay pocos casos de ambas (A, B) y muchos casos de una sola (A,_) mediante la interpretación de (A,_) como (A,B).

→ ... En consecuencia, no creo que nos sirva para nuestro problema, o al menos no por ahora.

- Lectura: Integration of deep learning model and feature selection for multi-label classification

→ Enfoque basado en deep learning. Se realiza una selección de features mediante grafos para reducir la dimensión de entrada y así reducir también la cantidad de datos necesarios para entrenar.

→ No profundicé demasiado en este paper pues se trata de un enfoque para optimizar redes profundas, que por ahora no es lo que buscamos.

- Validación: Al variar el threshold, ¿se equivocan los algoritmos siempre con los mismos ejemplos?:

→ La respuesta resultó ser sí. Al pasar de un threshold a otro más alto, las métricas para las etiquetas que permanecen en estudio son prácticamente idénticas.

→ Lo anterior valida la hipótesis sobre el por qué al podar el problema nos encontrábamos con una Hamming Loss creciente, al igual que el Hamming Score: nuestro porcentaje de error no varía demasiado, pero si lo hace el número de etiquetas en estudio, que actúa dividiendo el error. Y como consecuencia, la fracción da un valor más alto.

- Data Augmentation:

→ Se realiza data augmentation mediante rotaciones y reflexiones, respetando aquellos patrones con etiquetas como vertical u horizontal.

→ Pasamos de 775 patrones a 3285.

→ Se vuelven a calcular descriptores, esta vez incluyendo las entradas sintéticas.

→ La frecuencia de cada etiqueta aumenta alrededor de 4 veces. Ahora con t=15 tenemos 144 labels, mientras que con t=60 tenemos 57.

→ Para el caso desbalanceado, los resultados se mantienen.

→ Para el caso balanceado, la precision cae pero el recall se mantiene. Por ejemplo, en t=15 antes teniamos 53 labels con una precision de alrededor del 0.4 y un recall de 0.8, ahora tenemos una precision de 0.1 con un recall de 0.7 pero con 144 labels.
→ Lo anterior se ve prometedor: no mejoramos en eficacia, pero incorporamos muchas más etiquetas. Este aspecto puede estudiarse más a fondo en las próximas semanas.

Trabajo a realizar entre 30/03/2022 y 06/03/2022

- **Buscar nuevos orígenes para los descriptores**

→ Probar con arquitecturas pensadas para el problema multilabel.
→ Estudiar tanto los resultados que estas redes obtengan por sí solas como el qué tan bien se comportan los descriptores que podamos extraer de ellas como entrada para nuestros algoritmos.

Trabajo futuro/tareas postergadas

→ Aplicar Data Augmentation nuevamente, pero ahora incorporando el enfoque de crops que recomendó el prof. Iván en la reunión del 23/03.
→ Para que realmente nos sirva el data augmentation, necesitamos aplicarlo específicamente a las clases minoritarias (de lo contrario, el desbalance se mantiene). Sin embargo, como las etiquetas están unidas por el patrón, esto es difícil.