Otros paradigmas de clasificación no convencional

UCO ONLINE

Otros paradigmas de clasificación no convencional

UNIVERSIDAD Ð CÓRDOBA

1. Clasificación no convencional

A lo largo de las 5 primeras lecciones del curso hemos estudiado distintos problemas de aprendizaje convencional, tanto regresión (en las dos primeras lecciones), como clasificación (lecciones 3-5). En esta última lección del curso, hemos estudiado dos paradigmas de clasificación no convencionales, como la clasificación multi-etiqueta y clasificación multi-instancia.

Existen otros muchos paradigmas de aprendizaje en general, y de clasificación en particular, que no se encuadran dentro de la clasificación tradicional o convencional por sus características propias. Para estudiar todos estos paradigmas de clasificación serían necesarias semanas o cursos completos para estudiarlos y analizarlos en mayor detalle. Sin embargo, en esta lección se introducen algunos de los otros muchos paradigmas de clasificación existentes.

1.1. Clasificación de datos desbalanceados

Aunque no sea un escenario muy distinto al de clasificación convencional, es necesario al menos mencionar la clasificación donde los datos de entrada se encuentran muy desbalanceados respecto a la clase, ya que es un escenario muy típico en problemas reales. Por ejemplo, supongamos un conjunto de datos médicos; si analizamos la población, por lo general existirán más pacientes que no tengan cierta enfermedad, que que sí la tengan, haciendo que el escenario esté muy desbalanceado hacia la clase negativa.

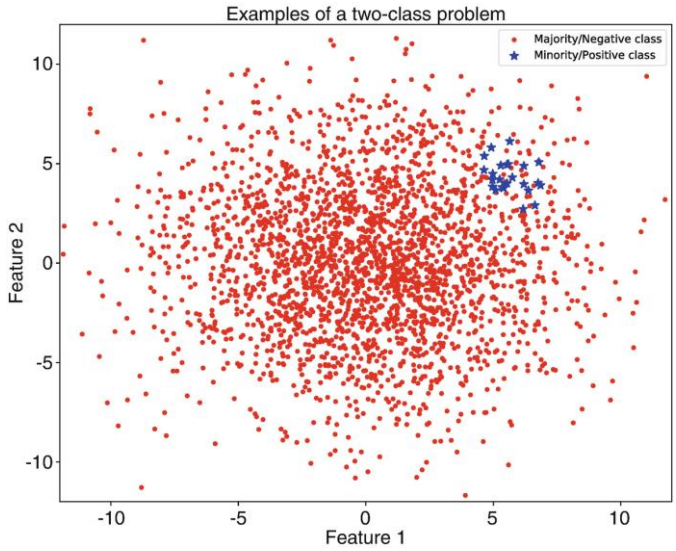

Figura 1. Ejemplo de escenario de clasificación desbalanceada[1].

[1] Fuente: Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Foundations on imbalanced classification. In Learning from Imbalanced Data Sets (pp. 19-46). Springer, Cham.

UCO ONLINE

En estos casos, es bastante probable que los métodos para clasificación tradicional puedan obviar las clases minoritarias, y se necesita de métodos o técnicas específicas que permitan clasificar correctamente aquellos patrones de la clase mayoritaria.

Uno de los principales enfoques para la resolución de estos problemas es realizar un muestreo de los datos para producir un problema menos desbalanceado que el original, o generar sintéticamente patrones de la clase minoritaria, de modo que permita clasificar los patrones de todas las clases forma más igualitaria. Otra opción, es directamente modificar los algoritmos clásicos para trabajar con los datos desbalanceados sin dar de lado dichas clases minoritarias. Por último, el aprendizaje sensible a costes permite asignar distintos costes a los patrones, de tal forma que clasificar incorrectamente un patrón concreto (por ejemplo de la clase negativa) sea más costoso que clasificar incorrectamente otros patrones (por ejemplo, los de la clase positiva). Para más información, consultar [Lop13].

1.2. Clasificación basada en data streams

Hasta el momento, para resolver el problema de clasificación se suponía un conjunto de datos estático, del cual se aprendía o inducía un modelo para utilizar en el futuro sobre nuevos datos desconocidos. Sin embargo, en muchos escenarios reales, los datos no se encuentran estáticos sino que existe un flujo de datos que va llegando y que habrá que ir considerando dinámicamente. Este flujo suele hacer que la cantidad de datos final sea mucho mayor que los escenarios tradicionales, por lo que también es posible que tras utilizar o procesar unos datos, estos sean desechados por completo para no ocupar ingentes cantidades de memoria.

Además, el flujo continuo de datos puede significar que la distribución de los datos varía con el tiempo: la que en un momento específico pudiera ser la clase mayoritaria, podría dejar de serlo; el fragmento del espacio donde soliesen estar distribuidos los datos, podría expandirse o moverse, recibiendo datos en puntos que antes no se estaban considerando.

Uno de los conceptos más resaltados de este tipo de aprendizaje es el de *concept drift*. Esto significa que las propiedades estadísticas de la variable objetivo que el modelo intenta predecir cambian con el tiempo de forma imprevista. Esto, causa problemas porque las predicciones se vuelven menos precisas a medida que pasa el tiempo. Para más información, consultar [Wan20].

1.3. *One-class classification*

La clasificación de única clase, o *one-class classification* podría considerarse un caso especial de clasificación tradicional o multi-clase, donde los patrones vistos durante el proceso de entrenamiento son de una única clase. El objetivo es aprender una representación o un clasificador que permita reconocer los patrones de dicha clase durante un proceso futuro de test. El proceso de aprendizaje, como puede apreciarse, se espera más complejo que el de clasificación tradicional, ya que no se observan en entrenamiento patrones de ambas (o más de dos) clases que hay que diferenciar, sino de únicamente una. En la Figura 2 se puede observar la diferencia principal entre clasificación multi-clase, y clasificación *one-class*. Para más información acerca de este paradigma, consultar [Per21].

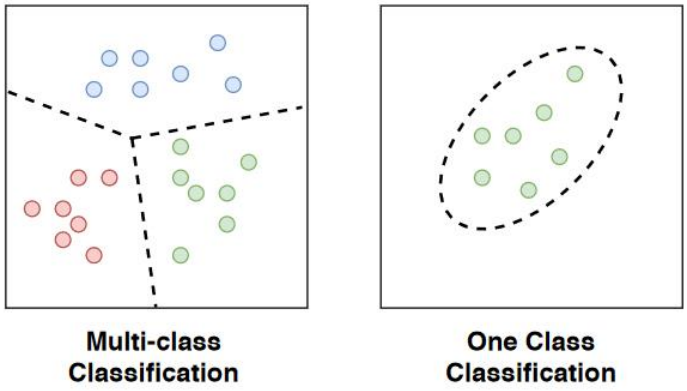

Figura 2. Diferencia entre clasificación multi-clase (izquierda) y one-class (derecha). Fuente [Per21].

1.4. Clasificación difusa (*fuzzy*)

Por lo general, hasta el momento hemos tratado con información precisa para los distintos atributos de entrada (o de salida) en los datos utilizados para entrenar clasificadores. Sin embargo, puede haber escenarios donde la información de entrada se base en conceptos más "difusos", es decir, información incierta o imprecisa. Por ejemplo, supongamos que, para un atributo de entrada, un atributo "altura" puede tomar los valores "bajo", "medio", o "alto"; o que un atributo edad se divide en "joven", "adulto", y "viejo". Dependiendo de la persona que analice estos conceptos, el valor puede ser diferente, ya que se trata de información imprecisa. Para ello, a cada atributo se le asignan unas funciones que indican el grado de pertenencia a cada uno de los posibles conceptos, devolviendo valores entre 0 y 1.

En la Figura 3 se muestra un ejemplo de conceptos difusos para el término edad; cabe destacar que la definición de estos conceptos, sus funciones, y sus valores de pertenencia, dependen por lo general del conocimiento que proporcione el experto. Dicha figura indica que una persona que se encuentre entre los 0 y 25 años aproximadamente, se le considera joven con un 100% de confianza, mientras que esa confianza va disminuyendo gradualmente hasta los 45 años. Similar, se comenzaría a considerar a una persona como madura a partir de los 25 años (con un nivel bajo de confianza); se la considera adulta con total confianza entre los 45 y 55 años, y de nuevo dicha confianza comienza a disminuir hasta los 75 años.

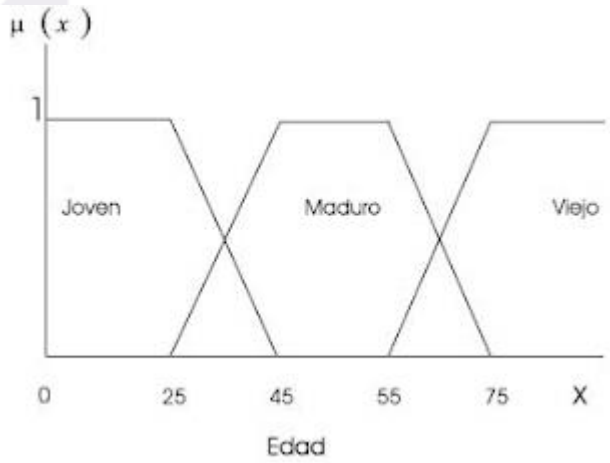

Figura 3. Ejemplo de concepto difuso de edad.

El objetivo de la clasificación difusa es, aprender un clasificador a partir de atributos siguiendo una estructura similar a la presentada anteriormente, de tal modo que se modele una variable objetivo, a la que también se le asignará por lo general un nivel de confianza o membersía a distintos términos.

1.5. Clasificación ordinal

La clasificación ordinal es un tipo especial de clasificación multi-clase para la que existe un orden inherente entre las clases, pero no una diferencia numérica significativa o determinada entre ellas. En clasificación tradicional, podíamos encontrarnos por ejemplo las clases Iris "Versicolor", "Virginica", y "Setosa"; entre estas clases no se puede definir ningún orden entre ellas. Sin embargo, en problemas donde las clases sean por ejemplo "Bajo", "Medio" y "Alto"; o "Bueno", "Regular" y "Malo", sí que se puede definir un orden entre ellas.

Mientras que en el escenario tradicional, el error de clasificar una flor Iris Setosa como Versicolor o como Virginica es el mismo, en el caso de clasificación ordinal, el error de clasificar un patrón de la clase Bajo como Medio, sería menor que clasificarlo como Alto.

Dada la naturaleza de estos problemas, se han definido distintos métodos, y principalmente métricas de evaluación, para tratar estos problemas. Para más información, consultar [Car11].

1.6. Clasificación multi-etiqueta con etiquetas perdidas (*missing labels*)

En la clasificación multi-etiqueta estudiada durante esta semana, se considera que, la información provista de clases/etiquetas en el conjunto de entrenamiento es 100% verídica y confiable. Es decir, supongamos un problema con 10 etiquetas donde un patrón concreto tiene asociadas 3 de ellas; eso significa que no está asociado con ninguna del as 7 etiquetas restantes.

Sin embargo, en problemas reales, principalmente en aquellos que necesiten expertos para la anotación o etiquetado de los patrones, es muy probable que un patrón esté asociado con más etiquetas de las que aparecen como relevantes en el conjunto de datos. Esto puede ser debido a varios aspectos: el anotador etiqueta un patrón con las etiquetas principales, pero puede dejarse algunas secundarias; si existen etiquetas similares o redundantes es posible que no incluya alguna de ellas; etc.

En este escenario, habrá problemas (llamados problemas con etiquetas perdidas) en los que podemos considerar que las etiquetas consideradas relevantes son realmente relevantes, pero además, en los datos de entrenamiento podría haber etiquetas marcadas como no relevantes que realmente sí lo son. De este modo, el proceso de entrenamiento debería en primer lugar, de algún modo, reconocer todas aquellas etiquetas que son relevantes aunque no se indiquen explícitamente, y posteriormente entrenar el modelo correspondiente. Para más información, consultar [Buc11].

Referencias

[Buc11] Bucak, S. S., Jin, R., & Jain, A. K. (2011, June). Multi-label learning with incomplete class assignments. In CVPR 2011 (pp. 2801-2808). IEEE.

[Car11] Cardoso, J. S., & Sousa, R. (2011). Measuring the performance of ordinal classification. International Journal of Pattern Recognition and Artificial Intelligence, 25(08), 1173-1195.

[Lop13] López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information sciences, 250, 113-141.

[Per21] Perera, P., Oza, P., & Patel, V. M. (2021). One-class classification: A survey. arXiv preprint arXiv:2101.03064.

[Wan20] Wankhade, K. K., Dongre, S. S., & Jondhale, K. C. (2020). Data stream classification: a review. Iran Journal of Computer Science, 3, 239-260.