Importancia de los Datos: Afecta de gran manera el número de instancias o cuanto ruido tengamos en nuestros datos?

David Elorza Gabilondo

Universidad del País Vasco

Objetivos

- Objetivo: Evaluar el impacto del ruido y cantidad de instancias en el modelo de clasificación..
- Preguntas de investigación:
  RQ1 ¿Varía la calidad del modelo según la cantidad de instancias?
  1. Borrado de instancias (de 2000 en 2000).
  2. Conjunto original (9000 instancias).
  3. Inserción de instancias desde otro conjunto (citado en Tarea y Datos)

  RQ2 ¿Cómo afecta la introducción gradual de ruido al rendimiento del modelo?
  1. Conjunto Original (0% de ruído artificial).
  2. Conjuntos alterados con porcentajes de ruído

Tarea y Datos

Se busca probar analíticamente como varía el resultado en función de la cantidad de instancias (RQ1) y el ruído (RQ2):

- **Tarea:** En la teoría, tener un gran conjunto de datos y un buen preproceso es diferencial a la hora de obtener unos buenos resultados, como de cierto es?. Fuentes: 1) Suicidal Data = Conjunto Original, Suicide Detection = Ampliacion del Conjunto
- Input: El conjunto de datos original se compone de 9206 instancias, 0 missing values y 300 instancias repetidas.
- Output: Class = label. 0 y 1 (Clasificación binaria).

Representación del texto

- **Pre-proceso** Se ha convertido a minúsculas, quitado stop words, estematizado y tokenizado
- **Vectorización** Se ha utilizado la vectorización w2v.

Clasificador 1: base

El primer clasificador hacia uso de 9000 instancias sin ruido y nos daba un fscore de 0.78.

1. Preproceso: Conversión a minúculas, borrado de puntos de exclamación..., estematización y tokenizacion..
2. Vectorización: Word 2 Vector:
3. ALgoritmo: KMeans.

Resultados Experimentales para RQ1

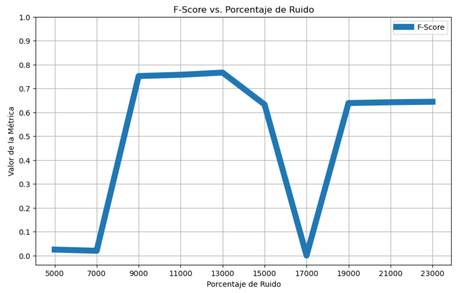

Figure 1:Fscore por cantidad de datos

- **Configuraciones: 1.** Borrado de 2000 y 4000 instancias; **2.** Conjunto original(9000 instancias); **3.** Inserción de 2000 a 14000 instancias en saltos de 2000.
- **Apunte:** Se ha utilizado KMenas como algoritmo de clasificación no supervisada (clustering), por lo que no hay división del conjunto en train y test.
- El mejor modelo esta entre los conjuntos de 9000 instancias (original) y el de 13.000 instancias. a partir de ahí los resultados son algo peores (exceptuando el conjunto con 16.000 instancias), pero se mantienen estables.

Clasificación no supervisada

Al usar Clustering con el algoritmo KMeans estamos usando clasificación no supervisada.

- Si observamos la matriz de confusión para ver como predice los datos:

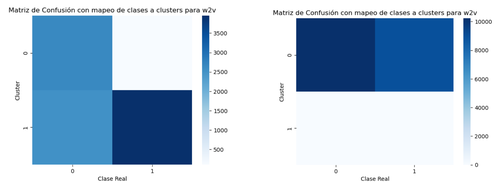

Figure 2:Mejor modelo (9000) VS medio con más instancias (19000)

1. Los resultados son consistentes.
2. Los resultados no son los esperados, se esperaba que que la gráfica tuviese forma de una ecuación logarítmica. Tal vez sea porque las nuevas instancias añadidas necesiten un preproceso más complejo?
3. Se acepta que la calidad es más significante a la cantidad, pero una gran cantidad de datos no siempre garantiza mejores resultados.

Resultados Experimentales para RQ2

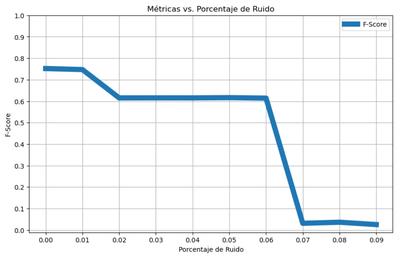

Figure 3:Fscore por porcentaje de ruído

Configuraciones: Se hace un proceso de borrado, inserción y sustitución de caracteres aleatoria y simultáneamente. El porcentaje índica el nº de caracteres afectados.

```
0    [0life, meaningles, 4rjust, wantXBrto, end, li...
1    [muttering, ibwnna, die, mysrelfo, daXy, Tfr, ...
2    [wr5s, slave, eall4, liYykeb, only, PuposO, li...
3    [pdid, simethin5, the, ofoctoeri, overdosd, ij...
4    [ieUl, lVk9Xnoone, cares, j0us, want4, die, ma...
```

Discussion

- Como se puede apreciar con la gráfica a medida que el ruido aumenta, el F-Score desciende, como era de esperar.
- Al aleatorio los resultados no son estables ya que cada vez se modifican diferentes cosas, se ha hecho la media.

Conclusiones

- Resumen: Se ha probado diferentes configuraciones para el clasificador con porcentaje de ruido y número de instancias.
- RQ1: Se concluye que un gran nº de instancias no siempre asegura un mejor resultado.
- RQ2: Se concluye que un modelo es capaz de aguantar cierto ruido hasta un punto.
- Fortalezas: Buenos resultados de F-Score.
- Debilidades: Se podría mejorar con un mejor preproceso.