UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales

Estimación por intervalos de las proporciones de clases en muestras sin etiquetar utilizando la distribución Poisson-Binomial

Tesis presentada para optar al título de Magister en Estadística Matemática de la Universidad de Buenos Aires

Ing. Maximiliano Marufo da Silva

Director de tesis: Dr. Andrés Farall

Buenos Aires, 2023.

Índice general

Capítulo 1

Introducción

1.1. Marco teórico

La tarea de cuantificación (conocida en inglés como *quantification*) consiste en proporcionar predicciones agregadas o resumen para conjuntos de datos en vez predicciones particulares sobre los datos individuales (por ejemplo, para el caso de clasificación, predecir la proporción de clases de un conjunto en vez de la clase de cada individuo), aplicando un modelo que se ajuste usando datos de entrenamiento cuya distribución puede ser distinta a la de los datos de prueba [1].

Si bien en principio no es necesario realizar predicciones por cada individuo, muchos de los métodos se basan en obtener la cuantificación de esa manera, ya que hacer predicciones individuales suele ser un requisito de por sí de las aplicaciones prácticas, o porque ya existen en ellas sistemas que las generen. Además, cabe aclarar que si bien la aplicación más popular es con respecto a tareas de clasificación (sobre las cuales basaremos este trabajo, y en particular, sobre clasificación binaria), también se puede aplicar cuantificación a problemas de regresión, ordinalidad, etc.

Un ejemplo práctico puede ser predecir la proporción de comentarios a favor o en contra sobre un producto, servicio o candidato en una red social. En este caso, se puede utilizar un clasificador para predecir por cada comentario si la opinión es positiva (o negativa), y luego obtener la proporción de comentarios a favor contándolos y dividiéndolos por el total (este método es el más simple y es conocido como *Classify & Count* o *CC*).

Si hablamos entonces de cuantificación binaria, se tiene que por cada muestra $i \in \{1, \dots, n\}$, $(\mathbf{X}_i, Y_i, S_i)$ es un vector de variables aleatorias tal que $\mathbf{X}_i \in \mathbb{R}^d$ son las características de la muestra, $Y_i \in \{1, 0\}$ indica la clase a la que pertenece y $S_i \in \{1, 0\}$ indica si fue etiquetada (y pertenece entonces al conjunto de entrenamiento) o no. Es decir, cuando $S_i = 0$, entonces Y_i no es observable. En la cuantificación binaria, se desea estimar $\theta := \mathbb{P}(Y = 1|S = 0)$, es decir, la prevalencia de etiquetas positivas entre muestras no etiquetadas. Esta prevalencia no se asume de ser la misma que en las muestras etiquetadas, $\mathbb{P}(Y = 1|S = 1)$. Además, el estimador de θ debe depender sólo de los datos disponibles, es decir, de las características de todas las muestras y de las etiquetas que fueron obtenidas. Los supuestos que se asumen son [2]:

- $(\mathbf{X}_1, Y_1, S_1) \dots (\mathbf{X}_n, Y_n, S_n)$ son independientes
- Por cada $s \in \{0, 1\}$, $(\mathbf{X}_1, Y_1)|S_1 = s, \dots, (\mathbf{X}_n, Y_n)|S_n = s$ son idénticamente distribuidas.

- Por cada $(y_1, \ldots, y_n) \in \{0,1\}^n$, $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ es independiente de $(S_1, \ldots, S_n)$ condicionado a $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$

Si bien existen varios métodos propuestos para el aprendizaje de cuantificación [3, 4], el mismo es todavía relativamente desconocido incluso para expertos en aprendizaje automático. La razón principal es la creencia errónea de que es una tarea trivial que se puede resolver usando un método directo, como *CC*. La cuantificación requiere métodos más sofisticados si el objetivo es obtener modelos óptimos, y su principal dificultad radica en la definición del problema, ya que las distribuciones de los datos de entrenamiento y de prueba pueden ser distintas. Por ejemplo, si la diferencia entre $\mathbb{P}(Y = 1|S = 0)$ y $\mathbb{P}(Y = 1|S = 1)$ es grande, los métodos simples como *CC* suelen tener bajo rendimiento.

Un método de cuantificación muy popular en la literatura y que sí se adapta a los cambios entre $\mathbb{P}(Y = 1|S = 0)$ y $\mathbb{P}(Y = 1|S = 1)$ es el propuesto por Saerens et al. [5], conocido como *Expectation Maximization for Quantification -EMQ-* o *SLD* por las siglas de sus autores. El mismo es un método iterativo que corrige, mediante el Teorema de Bayes, las predicciones de probabilidad de pertenencia a las clases dadas por el modelo de clasificación ya ajustado (sin necesidad de reajuste), y como consecuencia estima también la proporción de clases en la muestra de prueba. Este método es una aplicación directa del algoritmo de Esperanza-Maximización *-EM-*, y se puede probar que maximiza la verosimilitud en los datos de prueba. Se ha estudiado también que el método *EMQ* mejora aún más las predicciones de cuantificación si el clasificador utilizado está calibrado [6, 7], es decir, si sus predicciones de probabilidad asociadas a las clases predichas representan la probabilidad real de pertenencia a las clases [8].

Por otro lado, muy pocos son los trabajos sobre la construcción de intervalos de confianza y predicción en cuantificación [9]. La mayoría de ellos se basan en emplear los métodos de predicción puntual junto con la técnica de *bootstrapping* [10, 11, 12], que puede ser computacionalmente costosa en este tipo de tareas, o en métodos asintóticos [2], que no funcionan bien con tamaños de muestra pequeños y además requieren estimar la varianza. Dentro de los trabajos sobre intervalos de confianza o predicción en cuantificación aplicada a la clasificación (no se encontraron de hecho trabajos sobre intervalos en otro tipo de problemas), Keith y O'Connor [13] proponen dos métodos:

- El *baseline*, al que llaman *PB-PCC*, es un método asintótico basado en la distribución Poisson-Binomial [14, 15], donde proponen utilizar la media y varianza conocida para esta distribución para calcular el intervalo mediante la distribución normal. Existen tres problemas en el método propuesto en el trabajo:

  1. Se basa en el método de cuantificación de predicción puntual conocido como *Probabilistic Classify & Count -PCC-*, que no suele tener muy buenos resultados ya que no se ajusta a las diferencias entre $\mathbb{P}(Y = 1|S = 0)$ y $\mathbb{P}(Y = 1|S = 1)$.
  2. En el trabajo no se verifica que el clasificador esté calibrado, lo que podría degradar los resultados del cuantificador.
  3. Los métodos asintóticos no son buenos con muestras pequeñas.

- Su propuesta de mejora se basa en computar la verosimilitud marginal sobre θ, la proporción de clases en la población de prueba, para obtener la distribución *a posteriori* de θ. Luego, para obtener la predicción puntual se obtiene el máximo de la distribución. Es decir, que al

igual que el método *EMQ*, se busca maximizar la verosimilitud, pero en este caso sin utilizar el algoritmo *EM* sino de forma directa. Para obtener en cambio los intervalos, se proponen dos opciones, la primera es usar la aproximación asintótica y la segunda es construir una grilla para la distribución *a posteriori* de θ. Este método es bastante efectivo. Sin embargo, hay que tener en cuenta que aquí se estima la proporción de clases en la población de prueba, y no en la muestra de prueba.

Denham et al. [16] propone un método que asume condiciones más generales de cambio en las distribuciones entre los datos de entrenamiento y de prueba, ya que no asume la tercera suposición mencionada anteriormente. El método que proponen tiene, sin embargo, peor desempeño cuando esa condición sí se cumple. No obstante, es interesante resaltar los dos métodos *baseline* que utilizan para comparar con su propuesta de mejora:

- El primer *baseline* es también basado en *PB-PCC* y similar al propuesto en Keith y O'Connor. La diferencia en este trabajo es que en vez de usar la aproximación asintótica, computan la función de distribución exacta mediante el método propuesto por Hong [17], que utiliza la transformada rápida de Fourier (*FFT*) para hacerlo de forma eficiente. Este método sigue teniendo los primeros dos problemas mencionados en 1 y 2.
- El segundo *baseline* es muy similar a la propuesta de mejora de Keith y O'Connor. La única diferencia es que usa el algoritmo de *EM* en vez de hacer el cálculo directo para obtener el máximo de la distribución *a posteriori*.

1.2. Propuesta

Se propone un método para la elaboración de intervalos de predicción de la proporción de clases en muestras de prueba sin etiquetar, a partir de un conjunto de datos con etiquetas conocidas (conjunto de entrenamiento). Los pasos son:

1. Ajustar un modelo de clasificación con los datos de entrenamiento
2. Aplicar un método de calibración al clasificador para crear buenos estimadores de probabilidad para cada individuo.
    - Se compararon todos los métodos de calibración mencionados en la bibliografía, evaluados bajo el problema de cuantificación, obteniendo los mejores resultados con los métodos propuestos en Alexandari et al. [7].
3. Aplicar el método de cuantificación de estimación puntual de preferencia.
    - Se compararon los principales métodos mencionados en la bibliografía, obteniendo los mejores resultados con los métodos de *EMQ* y *PACC*.
4. Aplicar el paso de maximización de la esperanza propuesto en Saerens et al. [5] para ajustar las predicciones de probabilidad de cada individuo en base a la predicción puntual de la proporción de clases obtenidas con el método elegido en 3.
5. Aplicar el método propuesto por Hong [17] para obtener la distribución exacta de la proporción de clases en la muestra de prueba.

6. Utilizar la distribución obtenida en 5 para elaborar un intervalo de predicción exacto de la proporción de clases en la muestra con el nivel que corresponda.

Se presenta el método tanto de forma teórica como su evaluación empírica. Se elaboraron simulaciones para su evaluación y comparación con los métodos propuestos en Keith y O'Connor [13], en Denham et al. [16] y con los intervalos obtenidos mediante *bootstrapping*.

Capítulo 2

Problema

2.1. Introducción

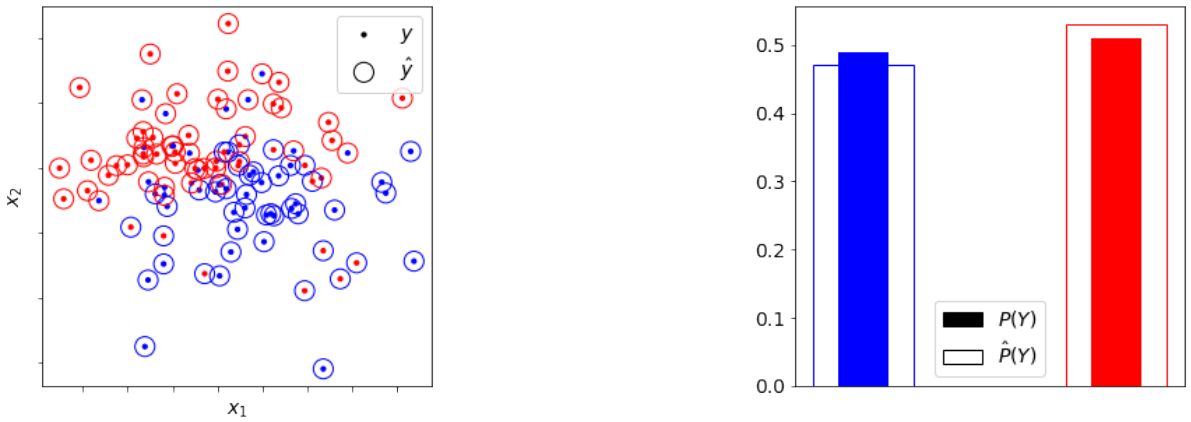
En algunas aplicaciones vinculadas a la clasificación, el objetivo final no es determinar a qué clase (o clases) pertenecen las instancias individuales no etiquetadas, sino estimar la prevalencia (tambíen llamada "frecuencia relativa' o "probabilidad prior') de cada clase en los datos sin etiquetar. En los últimos años se ha señalado que, en estos casos, tiene sentido optimizar directamente algoritmos de aprendizaje automático para este objetivo, en lugar de simplemente optimizar clasificadores para etiquetar instancias individuales.

La tarea de ajustar estimadores de prevalencia de clases a través del aprendizaje supervisado se conoce como "aprender a cuantificar' o, más simplemente, cuantificar (término acuñado por Forman [1], quien planteó el problema por primera vez). Se sabe que cuantificar mediante la clasificación cada instancia sin etiquetar a través de un clasificador estándar y luego contando las instancias que han sido asignadas a cada clase (el método *Classify & Count*) generalmente conduce a estimadores de prevalencia de clases sesgados, es decir, obtienen poca precisión en la cuantificación. Como resultado, se han desarrollado métodos que abordan la cuantificación como una tarea en sí.

Para ver la importancia de diferenciar el problema de cuantificación del de clasificación, veamos dos ejemplos. En el primero, una empresa que ofrece un servicio a sus clientes realiza una encuesta con varias preguntas para determinar el grado de satisfacción de cada persona. El objetivo de le empresa es determinar aquellos clientes que podrían no estar conformes con el servicio y ofrecerles una mejora en las condiciones para retenerlos. En el segundo ejemplo, una consultora analiza tweets para estimar el grado de aprobación de candidatos políticos. Aquí, la consultora no está interesado en predecir si un individuo específico está a favor o en contra, sino en cuántos encuestados, del número total de encuestados, aprueban al candidato, es decir, en conocer la prevalencia de la clase positiva.

Mientras en el primer escenario el interés es a nivel individual, en el último, el nivel agregado es lo que importa; en otras palabras, en el primer escenario la clasificación es el objetivo, mientras que en el segundo el verdadero objetivo es la cuantificación. De hecho, en la mayoría de las aplicaciones las predicciones que interesan no son a nivel individual sino a nivel colectivo; ejemplos de tales campos son la investigación de mercado, la ciencia política, las ciencias sociales, modelado ecológico y epidemiología.

La literatura sobre métodos relacionados con cuantificación está un tanto desconectada. Al-



(a) Clasificación **(b)** Cuantificación

Figura 2.1: En la clasificación, la predicción es a nivel individual, mientras que en la cuantificación es a nivel agregado.

gunos de los métodos que pueden usarse como cuantificadores han sido ideados para otros fines, principalmente para mejorar la precisión en clasificación cuando cambia el dominio. El desempeño de este último grupo ha sido normalmente estudiado solo en términos de mejora en las tareas de clasificación pero no como cuantificadores. Dado este escenario, y debido a la variedad de campos en los que ha surgido como una necesidad de aplicación, los algoritmos que se pueden aplicar para tareas de cuantificación aparecen en artículos que usan diferentes palabras clave y nombres, como *counting* [18], *prior probability shift* [19, 20], *posterior probability estimation* [21], *class prior estimation* [22, 23, 24], *class prior change* [25], *prevalence estimation* [26], *class ratio estimation* [27] o *class distribution estimation* [28, 29, 30], por citar solo algunos de ellos.

2.2. Tipos de Cuantificación

Aunque el estudio de la cuantificación se ha centrado principalmente en el dominio de clasificación, la cuantificación también aparece en otros tipos de problemas de aprendizaje automático, como la regresión, la clasificación ordinal, el aprendizaje sensible al costo y la cuantificación en redes.

De manera similar a la regresión, aprender a cuantificar admite diferentes problemas de interés aplicativo, basados en cuántas clases contiene Y y cuántas de las clases en Y se pueden atribuir al mismo tiempo al mismo individuo. Así, los problemas de cuantificación se dividen de esta manera:

1. Etiquetado simple *(Single-Label Quantification -SLQ-)*: cuando cada individuo pertenece exactamente a una de las clases en $Y = \{y_1, \ldots, y_{|y|}\}$.

2. Etiquetado múltiple *(Multi-Label Quantification -MLQ-)*: cuando cada individuo puede pertenecer a cualquier número de clases (cero, una o varias) en $Y = \{y_1, \ldots, y_{|y|}\}$.

3. Cuantificación Binaria *(Binary Quantification -BQ-)*:

   a) en *SLQ* con $|y| = 2$, (en este caso $y = \{y_1, y_2\}$, y cada individuo pertenece a y_1 o y_2)

b) en MLQ con $|y| = 1$, (en este caso $y = \{y\}$, y cada individuo pertenece o no a y)

4. Cuantificación Ordinal *(Ordinal Quantification -OQ-)*: cuando existe un orden $y_1 \prec \cdots \prec y|y|$ en $Y = \{y_1, \ldots, y_{|y|}\}$.

5. Cuantificación de Regresión *(Regression Quantification -RQ-)*: cuando no hay un conjunto de clases involucradas, sino que cada individuo está etiquetado con una puntuación de valor real y la cuantificación equivale a estimar la fracción de ítems cuya puntuación está en un intervalo dado $[a, b]$ con $a, b \in \mathbb{R}^d$.

2.3. Cambios en las distribuciones de los datos

En lo últimos años ha habido un interés creciente en las aplicaciones que presentan cambios en las distribuciones de datos (conocido en la blibliografía por su término en inglés *dataset shift*). Estos problemas comparten el hecho de que la distribución de los datos utilizados para entrenar es diferente a la de los datos que se usan para predecir. Al igual que para el área de la cuantificación, aquí también la literatura sobre el tema está dispersa y diferentes autores usan diferentes nombres para referirse a los mismos conceptos, o usan el mismo nombre para diferentes conceptos.

Teniendo en cuenta que en los problemas de clasificación tenemos:

- Un conjunto de características o covariables $\mathbf{X}$.
- Una variable de respuesta Y.
- Una distribución conjunta $P(Y, \mathbf{X})$.

La probabilidad conjunta $P(Y, \mathbf{X})$ puede luego se puede escribir como $P(Y|\mathbf{X})P(\mathbf{X})$ o como $P(\mathbf{X}|Y)P(Y)$. Por otro lado, cuando usamos los términos de entrenamiento (*train*) y prueba (*test*), nos referimos a las datos disponibles para entrenar al clasificador y los datos presentes en el entorno en el que se implementará el clasificador, respectivamente. Las distribuciones de datos en entrenamiento y prueba se indican como P_{tr} y P_{tst}.

El *dataset shift* aparece cuando las distribuciones conjuntas de entrenamiento y de prueba son diferentes, es decir, cuando $P_{tr}(Y, \mathbf{X}) \neq P_{tst}(Y, \mathbf{X})$. Moreno-Torres et al. [19] distingue las distintas variantes del dataset shift según qué elementos mencionados anteriormente cambian:

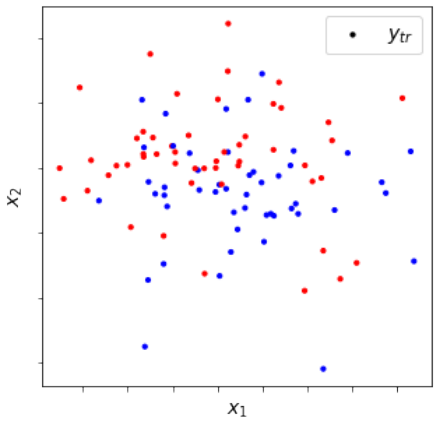
- *Covariate shift*, cuando $P_{tr}(Y|\mathbf{X}) = P_{tst}(Y|\mathbf{X})$ y $P_{tr}(\mathbf{X}) \neq P_{tst}(\mathbf{X})$
- *Prior probability shift*, cuando $P_{tr}(\mathbf{X}|Y) = P_{tst}(\mathbf{X}|Y)$ y $P_{tr}(Y) \neq P_{tst}(Y)$
- *Concept shift*, cuando $P_{tr}(Y|\mathbf{X}) \neq P_{tst}(Y|\mathbf{X})$ y $P_{tr}(\mathbf{X}) = P_{tst}(\mathbf{X})$ o $P_{tr}(\mathbf{X}|Y) \neq P_{tst}(\mathbf{X}|Y)$ y $P_{tr}(Y) = P_{tst}(Y)$

Otros tipos de dataset shift surgen cuando $P_{tr}(Y|\mathbf{X}) \neq P_{tst}(Y|\mathbf{X})$ y $P_{tr}(\mathbf{X}) \neq P_{tst}(\mathbf{X})$ y cuando $P_{tr}(\mathbf{X}|Y) \neq P_{tst}(\mathbf{X}|Y)$ y $P_{tr}(Y) \neq P_{tst}(Y)$. Sin embargo, estos tipos de cambios no se consideran generalmente en la literatura ya que aparecen mucho más raramente, o incluso porque son dificiles o imposibles de resolver.

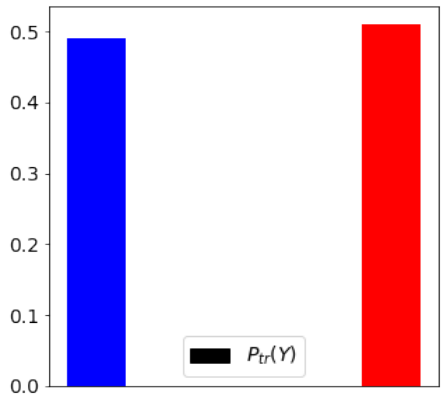
Está claro que el problema de cuantificación se trata de un caso donde $P_{tr}(Y) \neq P_{tst}(Y)$, o donde $P_{tr}(Y)$ es desconocido, ya que sino el problema sería trivial. Además, la mayoría de los métodos

de cuantificación propuestos asumen que $P_{tr}(\mathbf{X}|Y) = P_{tst}(\mathbf{X}|Y)$, por lo que estarían dentro de los casos de *prior probability shift*.

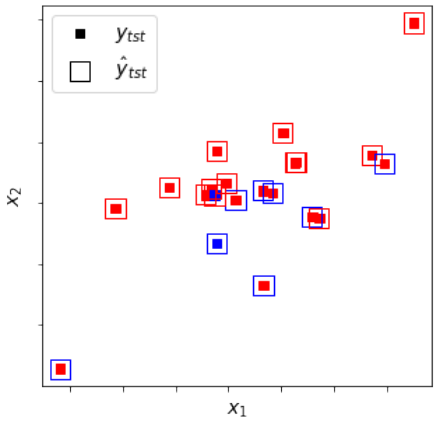
Por otro lado, en la mayoría de los casos el objetivo final de la implementación es estimar algún parámetro de $P_{tst}(Y)$. Por ejemplo, como ya mencionamos anteriormente, en la cuantificación binaria, se desea estimar $\theta := \mathbb{P}(Y = 1|S = 0)$. Es decir, en la cuantificación la tarea indirectamente suele ser aprender a aproximar una distribución desconocida (observando sólo características de una muestra) mediante una distribución conocida. En consecuencia, prácticamente todas las medidas de evaluación para la cuantificación son divergencias, es decir, medidas de cómo una distribución pronosticada difiere de la distribución real.



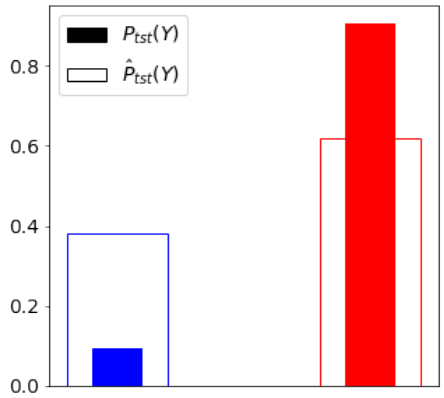
(a) Muestra de entrenamiento



(b) Prevalencia de clases en muestra de entrenamiento



(c) Clasificación en muestra de evaluación



(d) Prevalencia de clases y cuantificación en muestra de evaluación

Figura 2.2: El *prior probability shift* propio de los problemas de cuantificación puede hacer que los métodos simples de cuantificación, como *CC*, tengan grandes errores.

2.4. El problema de clasificar y contar

En ausencia de métodos para estimar los valores de prevalencia de clase de forma directa, el primer método que suele pensarse para hacerlo es *Classify & Count*, es decir, clasificar cada individuo sin etiquetar y estimar los valores de prevalencia de clase contando los individuos que fueron asignados a cada clase. Sin embargo, esta estrategia es subóptima: mientras que un clasificador perfecto es también es un cuantificador perfecto, un buen clasificador puede ser un mal cuantificador. Para ver esto, se puede ver la definición de F1, una función de evaluación estándar para la clasificación binaria, que se define como:

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{2.1}$$

donde TP, FP, FN indican el número de verdaderos positivos, falsos positivos y falsos negativos, respectivamente. Un buen clasificador puede ser un mal cuantificador ya que F1 considera buenos aquellos clasificadores que mantienen la suma FP + FN al mínimo; sin embargo, el objetivo de un algoritmo de cuantificación debe ser mantener al mínimo |FP − FN|.

Incluso un buen clasificador puede estar sesgado, es decir, puede mantener sus falsos positivos al mínimo sólo a expensas de una cantidad sustancialmente mayor de falsos negativos (o viceversa); si este es el caso, el clasificador es un mal cuantificador. Este fenómeno no es infrecuente, especialmente en presencia de datos desbalanceados. En tales casos, los algoritmos que minimizan las funciones de pérdida de clasificacion (*Hamming*, *hinge*, etc) suelen generar clasificadores con tendencia a elegir la clase mayoritaria, lo que implica un número mucho mayor de falsos positivos que de falsos negativos para la clase mayoritaria, lo que significa a su vez que tal algoritmo tenderá a subestimar las clases minoritarias.

Los argumentos anteriores indican que no se debe considerar la cuantificación como un mero subproducto de la clasificación, y debe estudiarse y resolverse como una tarea en sí misma. Hay al menos otros dos argumentos que apoyan esta idea. Uno es que las funciones que se utilizan para evaluar la clasificación no se pueden utilizar para evaluar la cuantificación, ya que estas funciones miden, en general, cuántos individuos han sido mal clasificados, y no cuánto difiere la prevalencia de clase estimada del valor real. Esto significa que los algoritmos que minimizan estas funciones están optimizados para la clasificación, y no para la cuantificación. Un segundo argumento presentado por Forman [31] es que los métodos diseñados específicamente para cuantificar requieren menos datos de entrenamiento para alcanzar la misma precisión de cuantificación que los métodos estándar basados en *CC*. Si bien esta observación es de naturaleza empírica, también existen argumentos teóricos que sustentan este hecho [32].

2.5. Cuantificadores como suplementos en clasificación

Debido a los problemas mencionados anteriormente de los clasificadores frente a cambios en las distribuciones de los datos y frente a datos desbalanceados, los algoritmos de cuantificación están cada vez más frecuentemente también siendo usados en tareas que requieren predicciones individuales. Los mismos se emplean como suplemento de clasificadores para suplir sus defectos frente a estos problemas, ya que en algunos casos no sólo predicen los valores agregados, sino que también mejoran las predicciones a nivel individual.

Por ejemplo, el *prior probability shift* puede hacer que los clasificadores performen de manera subóptima. En el caso del clasificador óptimo de Bayes, dado por:

$$h(\mathbf{x}) = \arg\max_y p(y|\mathbf{x}) = \arg\max_y \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} \tag{2.2}$$

la decisión del clasificador depende de $p(y)$, que es estimado con el dataset de entrenamiento, es decir, con $p(y) = p_{tr}(y)$. Es decir, que en caso de $P_{tr}(Y) \neq P_{tst}(Y)$, la decisión final del clasificador puede verse afectada negativamente. Para mejorar el rendimiento del clasificador frente a estos casos, en 2.5 se debería usar $p_{tst}(y)$ en vez de $p_{tr}(y)$. Pero como $p_{tst}(y)$ es generalmente desconocido, se puede usar un método de cuantificación para estimarlo [5, 21, 24, 30].

Los métodos de cuantificación pueden usarse no sólo para mejorar el rendimiento general de un clasificador, sino también para mejorar su equidad o *fairness* [33], es decir, su posibilidad de predecir resultados independientes de un cierto conjunto de variables que consideramos sensibles y no relacionadas con él (e.j.: género, etnia, orientación sexual, etc.). Por ejemplo, suponiendo que una variable S debe considerarse sensible, se puede estimar $P_{tr}(Y|S)$. Luego, si los datos de entrenamiento están sesgados, por ejemplo, con $P_{tr}(Y = 1|S = 1) \gg P_{tr}(Y = 1|S = 0)$, pero se sabe que en las muestras a inferir esto no es así, se puede optimizar el modelo de clasificación imponiento alguna penalidad basada en la estimación de $P_{tst}(Y|S)$, siendo esta última obtenido por un cuantificador.

Capítulo 3

Métodos de Evaluación

La evaluación de métodos de cuantificación es más compleja que en otros problemas. En aprendizaje supervisado, típicamente se mide el rendimiento estimando la probabilidad de predecir correctamente ejemplos individuales no observados (sin condicionar -para la exactitud o *accuracy*- o condicionando las probabilidades a las clases de pertenencia -para la exhaustividad o *recall*- o predichas -para la precisión o *precision*-). Sin embargo, en cuantificación, el rendimiento se evalúa para conjuntos de datos. Esto implica que necesitamos una colección de muestras para evaluar el rendimiento de un modelo. Dado un modelo $\overline{h}$, una función de pérdida $L(\cdot,\cdot)$, y un conjunto de muestras de evaluación $T_1,\ldots,T_s$, el rendimiento de $\overline{h}$ es:

$$Rendimiento(\overline{h}, L, T_1, \ldots, T_s) = \frac{1}{s}\sum_{j=1}^{s} L(\overline{h}, T_j) \tag{3.1}$$

Calcular la pérdida de un modelo sobre una muestra de prueba, $L(\overline{h}, T_j)$ no requiere de promediar sobre ejemplos individuales. Por ejemplo, en la cuantificación binaria, sólo la prevalencia real p y la prevalencia predicha $\hat{p}$ se comparan por cada muestra.

En cuantificación, el problema de evaluación se relaciona con el cambio en la distribución de datos entre la fase de entrenamiento y la de implementación del modelo. Se requiere una colección de muestras de prueba variada y que represente diversas distribuciones para evaluar correctamente el rendimiento del modelo y evitar sesgos. Por esta razón, la mayoría de los experimentos reportados en la literatura emplean conjuntos de datos tomados de otros problemas y se crean conjuntos de prueba con cambios en las distribuciones creados artificialmente. Este enfoque tiene la ventaja de que la cantidad del *dataset shift* se puede controlar para estudiar el rendimiento de los modelos en diferentes situaciones.

Las funciones de pérdida $L(\cdot,\cdot)$ serán elegidas de acuerdo al tipo de problema y al objetivo particular de la aplicación. Como ya se mencionó, el rendimiento de $\overline{h}$ será el promedio del resultado de la función de pérdida por cada muestra de evaluación, de acuerdo a la ecuación 3.1. Se han propuesto en la literatura distintas métricas de evaluación para problemas de SLQ. Estas también se pueden usar para BQ, ya que es un caso espacial de SLQ, y para MLQ, ya que se pueden usar para cada $y \in Y$. Esencialmente todas las medidas de evaluación que se han propuesto son divergencias, es decir, medidas de cómo una distribución difiere de otra. No se desarrollarán en esta tésis métricas para OQ, ya que no son útiles para nuestro objeto de estudio.

3.1. Propiedades

Sebastiani [34] define una serie de propiedades interesantes para medidas de evaluación en *SLQ*. Un importante resultado de este artículo es que ninguna medida de evaluación existente para *SLQ* satisface todas las propiedades identificadas como deseables; aún así, se ha demostrado que algunas medidas de evaluación son "menos inadecuadas" que otras. Aquí mencionamos brevemente las cuatro propiedades principales que habría que considerar en cada métrica M a emplear (el resto son propiedas que suelen ser satisfechas por todas las métricas).

- **Máximo (MAX)**: si $\exists \beta > 0, \beta \in \mathbb{R}$ tal que por cada $p(y), y \in Y$, (i) existe $\hat{p}(y)$ tal que $M(p, \hat{p}) = \beta$, y (ii) para ninguna $\hat{p}(y)$ se cumple que $M(p, \hat{p}) > \beta$. Si se cumple **MAX**, la imagen de M es independiente del problema, y esto permite juzgar si un valor dado significa un error de cuantificación alto o bajo. Si M no cumple cumple **MAX**, cada muestra de evaluación tendrá un peso distinto en el resultado final.

- **Imparcial (IMP)**: si M penaliza igualmente la subestimación de $p(y)$ por una cantidad a (es decir, con $\hat{p}(y) = p(y) - a$) o su sobreestimación por la misma cantidad a (es decir, con $\hat{p}(y) = p(y) + a$). Si se cumple **IMP**, la subestimación y la sobreestimación se consideran igualmente indeseables. Esto es generalmente lo deseable, a menos que exista una razón específica para no hacerlo.

- **Relativo (REL)**: si M penaliza más gravemente un error de magnitud absoluta a (es decir, cuando $\hat{p}(y) = p(y) \pm a$) si $p(y)$ es menor. Por ejemplo, predecir $\hat{p}(y) = 0.0101$ cuando $p(y) = 0.0001$ es un error mucho más serio que predecir $\hat{p}(y) = 0.1100$ cuando $p(y) = 0.1000$.

- **Absoluto (ABS)**: si M penaliza un error de magnitud independientemente del valor de $p(y)$. Mientras algunas aplicaciones requieren **REL**, otras requieren **ABS**. Si bien **REL** y **ABS** son mutuamente excluyentes, ninguna cubre el caso cuando M considera un error de magnitud absoluta a menos grave cuando $p(y)$ es menor (como en el caso de la *distancia coseno*).

3.2. Métricas

3.2.1. Sesgo

El sesgo o *bias* técnicamente no es una medida de evaluación para la cuantificación, ya que no se aplica a toda una distribución p sino solo a una etiqueta específica y, y se define como:

$$\mathrm{B}(y) = \hat{p}(y) - p(y) \tag{3.2}$$

Si se usa como un medida de evaluación para la cuantificación, un problema obvio con B es que promediar los puntajes de diferentes clases produce resultados poco intuitivos, ya que el sesgo positivo de una clase y el sesgo negativo de otra clase se anulan entre sí. El mismo problema ocurre cuando se trata de la misma clase pero se promedia entre diferentes muestras. Como resultado, esta medida se puede utilizar como mucho para determinar si un método tiene una tendencia a subestimar o sobrestimar la prevalencia de una clase específica (típicamente la clase minoritaria) en *BQ*, y no como una medida de evaluación para general usar.

3.2.2. Error Absoluto

El error absoluto o *absolute error* es una de las medidas más empleadas ya que, al ser simplemente la diferencia entre ambas magnitudes, es simple y fácilmente interpretable.

$$\mathrm{AE}(p, \hat{p}) = \frac{1}{|y|} \sum_{y \in \mathcal{Y}} |\hat{p}(y) - p(y)| \tag{3.3}$$

Como en este caso las diferencias positivas y negativas son igualmente indeseables, promediar el AE entre varias clases, o varias muestras, no es problemático. Como se muestra en [34], AE cumple **IMP** y **ABS** pero no cumple **MAX** (ni tampoco **REL**).

3.2.3. Error Absoluto Normalizado

El error absoluto normalizado *normalised absolute error*, definido como:

$$\mathrm{NAE}(p, \hat{p}) = \frac{\mathrm{AE}(p, \hat{p})}{z_{\mathrm{AE}}} = \frac{\sum_{y \in \mathcal{Y}} |\hat{p}(y) - p(y)|}{2(1 - \min\limits_{y \in \mathcal{Y}} p(y))} \tag{3.4}$$

es una versión de AE que oscila entre 0 (mejor) y 1 (peor), por lo que cumple **MAX**. A pesar de su nombre, NAE no disfruta de **ABS** (ni tampoco **REL**).

3.2.4. Error Cuadrático

El error cuadrático o *squared error*, definido como:

$$\mathrm{SE}(p, \hat{p}) = \frac{1}{|y|} \sum_{y \in \mathcal{Y}} (\hat{p}(y) - p(y))^2 \tag{3.5}$$

comparte los mismos pros y contras de AE, pero penalizando más cuanto mayor es la diferencia entre el valor real y el predicho, por lo que se usa cuando se quiere castigar los valores atípicos u *outliers*.

3.2.5. Error Absoluto Relativo

El error absoluto relativo o *relative absolute error* es una adaptación del AE que impone **REL** al hacer que AE sea relativo a p.

$$\mathrm{RAE}(p, \hat{p}) = \frac{1}{|y|} \sum_{y \in \mathcal{Y}} \frac{|\hat{p}(y) - p(y)|}{p(y)} \tag{3.6}$$

RAE cumple **IMP** y **REL** pero no cumple **MAX** (ni **ABS**, a pesar de su nombre).

3.2.6. Error Absoluto Relativo Normalizado

El error absoluto relativo normalizado *normalised relative absolute error*, definido como:

$$\mathrm{NRAE}(p, \hat{p}) = \frac{\mathrm{RAE}(p, \hat{p})}{z_{\mathrm{RAE}}} = \frac{\sum\limits_{y \in \mathcal{Y}} \frac{|\hat{p}(y) - p(y)|}{p(y)}}{|y| - 1 + \frac{1 - \min\limits_{y \in \mathcal{Y}} p(y)}{\min\limits_{y \in \mathcal{Y}} p(y)}} \tag{3.7}$$

es una versión de RAE que oscila entre 0 (mejor) y 1 (peor), por lo que cumple **MAX**. A pesar de su nombre, NRAE no disfruta de **REL** (ni tampoco **ABS**).

Tanto RAE como NRAE no están definidas cuando sus denominadores sean nulos. Para resolver este problema, se puede suavizar tanto $p(y)$ como $hatp(y)$ mediante suavizado aditivo:

$$\underline{p}(y) = \frac{\epsilon + p(y)}{\epsilon|y| + \sum_{y \in \mathcal{Y}} p(y)} \tag{3.8}$$

donde $\underline{p}(y)$ es la versión suavizada de $p(y)$ y el denominador es solo un un factor de normalización (lo mismo para los $\underline{\hat{p}}(y)$).

3.2.7. Divergencia de Kullback-Leibler

Para distribuciones de probabilidad discretas P y Q definidas en el mismo espacio muestral $\mathcal{X}$ su divergencia KL se define como:

$$\mathrm{DKL}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{3.9}$$

En cuantificación, se quiere comparar la prevalencia real p y la prevalencia predicha $\hat{p}$, y el espacio muestral corresponde a las posibles clases, con lo cuál será:

$$\mathrm{DKL}(p \parallel \hat{p}) = \sum_{y \in \mathcal{Y}} p(y) \log \left(\frac{p(y)}{\hat{p}(y)} \right) \tag{3.10}$$

que va de 0 (mejor) a $+\infty$ (peor) -por lo tanto, no cumple con **MAX**-. Si bien esta medida es una distancia, no es una métrica verdadera, ya que no obedece a la desigualdad del triángulo y no es simétrica. Además, es menos interpretable que otras métricas de rendimiento y no está definido cuando $\hat{p}$ es 0 o 1.

3.2.8. Divergencia de Kullback-Leibler Normalizada

Para suplir los problemas de DKL, se puede utilizar la función logística, quedando:

$$\mathrm{NDKL}(p \parallel \hat{p}) = 2 \cdot \frac{e^{\mathrm{DKL}(p \| \hat{p})}}{1 + e^{\mathrm{DKL}(p \| \hat{p})}} - 1 \tag{3.11}$$

que también va de 0 (mejor) a $+\infty$ (peor) -por lo tanto, si cumple con **MAX**-. Sin embargo, como se muestra en [34], ni DKL ni NDKL cumplen con **IMP**, **REL** y **ABS**, lo que hace que su uso como medidas de evaluación para cuantificación sea cuestionable, además de ser dificiles de interpretar.

3.3. Elección de la Métrica

Es evidente que ninguna de las medidas propuestas hasta ahora es completamente satisfactoria. DKL y NDKL son los menos satisfactorios y parecen fuera de discusión. Respecto a los demás, el problema es que **MAX** parece ser incompatible con **REL**/**ABS**, y viceversa.

Sebastiani [34] sostiene que cumplir con **REL** o **ABS** parece más importantes que cumplir con **MAX**, ya que reflejan las necesidades de la aplicación; si no se satisfacen estas propiedades, se

puede argumentar que el error de cuantificación que se está midiendo está vagamente relacionada a lo que el usuario realmente quiere. Si **MAX** no está satisfecho, los resultados obtenidos en muestras caracterizadas por diferentes distribuciones no serán comparables. A pesar de esto, los resultados obtenidos por diferentes sistemas en el mismo conjunto de muestras siguen siendo comparables.

Esto sugiere que AE, RAE y SE son las mejores medidas a elegir. Se debe preferir AE cuando un error de estimación de una magnitud absoluta dada debe considerarse más grave cuando la verdadera prevalencia de la clase afectada es menor. RAE debe ser elegido cuando un error de estimación de una magnitud absoluta dada tiene el mismo impacto independientemente de la verdadera prevalencia de la clase afectada. Si se quiere penalizar mayormente errores atípicos, considerando mucho más graves a los errores cuanto mayor es la diferencia entre el valor real y el predicho, entonces SE es la métrica más conveniente.

3.4. Protocolos

Mientras que en la clasificación, un conjunto de datos de tamaño k proporciona k puntos de evaluación, para la cuantificación, el mismo conjunto solo proporciona 1 punto. Evaluar algoritmos de cuantificación es por lo tanto un reto, debido a que la disponibilidad de datos etiquetados con fines de prueba es más restringido. Hay principalmente dos protocolos experimentales que se han tomado para tratar con este problema: el Protocolo de Prevalencia Natural (NPP) () y el Protocolo de Prevalencia Artificial (APP).

- NPP: Consiste en tomar un conjunto de prueba lo suficientemente grande, dividirlo en un número de muestras de manera uniformemente aleatoria, y llevar a cabo la evaluación individualmente en cada muestra.
- APP: Consiste en tomar un conjunto de datos, dividido en un conjunto de entrenamiento L y en un conjunto U de elementos sin etiquetar de manera uniformemente aleatoria, y realizar experimentos repetidos en los que la prevalencia del conjunto de entrenamiento o la prevalencia del conjunto de prueba de una clase se varía artificialmente a través del submuestreo.

Ambos protocolos tienen diferentes pros y contras. Una ventaja de APP es que permite crear muchas puntos de prueba de la misma muestra. Además, APP permite simluar distintos *Prior probability shift*, mientras que con NPP se estaría evaluando sólo con las distribuciones originales de los datos de entrenamiento y prueba. Sin embargo, una desventaja de APP es que puede no saberse cuán realistas son estas diferentes situaciones en la aplicación real, por lo que se podría estar destinando recursos a una evaluación errónea o pobre. Una solución intermedia podría ser utilizar un protocolo que utilice conocimientos previos sobre la distribución de prevalencias "probables" que se podría esperar encontrar en el dominio específico en cuestión.

Capítulo 4

Estimación Puntual

Durante los últimos años, se han propuesto varios métodos de cuantificación desde diferentes perspectivas y con diferentes objetivos. En términos generales, se pueden distinguir dos grandes clases de métodos en la literatura. La primera clase es la de métodos agregativos, es decir, métodos que requieren la clasificación de todos los individuos como un paso intermedio. Dentro de los métodos agregativos, se pueden identificar dos subclases. La primera subclase incluye métodos basados en clasificadores de propósito general; en estos métodos la clasificación de los elementos individuales realizados como un paso intermedio puede lograrse mediante cualquier clasificador. La segunda subclase se compone, en cambio, de métodos que para clasificar los individuos, se basan en métodos de aprendizaje diseñados con la cuantificación en mente. La segunda clase es la de métodos no agregativos, es decir, métodos que resuelven la tarea de cuantificación "holísticamente", es decir, sin clasificar a los individuos. Aquí de nuevo se hará tratarán métodos destinados a la cuantificación binaria, aunque como ya se mencionó la mayoría de ellos pueden luego extenderse a problemas multiclase.

4.1. Métodos Agregativos

Dentro de los métodos agregativos, algunos de ellos requieren como entrada las etiquetas de clases predichas (es decir, clasificacores duros), mientras que otros requieren como entrada las probabilidades *a posteriori* de pertenencia a cada clase (es decir, clasificacores blandos).

4.1.1. Con clasificadores generales

4.1.2. Con clasificadores específicos

4.2. Métodos No Agregativos

Capítulo 5

Estimación por Intervalos

Capítulo 6

Propuesta

Capítulo 7

Resultados

Referencias

[1] George Forman. Counting positives accurately despite inaccurate classification. In *European conference on machine learning*, pages 564–575. Springer, 2005.

[2] Afonso Fernandes Vaz, Rafael Izbicki, and Rafael Bassi Stern. Quantification under prior probability shift: The ratio estimator and its extensions. *The Journal of Machine Learning Research*, 20(1):2921–2953, 2019.

[3] Andrea Esuli, Alessandro Fabris, Alejandro Moreo, and Fabrizio Sebastiani. *Learning to Quantify*. Springer Nature, 2023.

[4] Pablo González, Alberto Castaño, Nitesh V Chawla, and Juan José Del Coz. A review on quantification learning. *ACM Computing Surveys (CSUR)*, 50(5):1–40, 2017.

[5] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

[6] Andrea Esuli, Alessio Molinari, and Fabrizio Sebastiani. A critical reassessment of the Saerens-Latinne-Decaestecker algorithm for posterior probability adjustment. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–34, 2020.

[7] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, pages 222–232. PMLR, 2020.

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[9] Dirk Tasche. Confidence intervals for class prevalences under prior probability shift. *Machine Learning and Knowledge Extraction*, 1(3):805–831, 2019.

[10] Daniel J Hopkins and Gary King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, 2010.

[11] Ashlynn R Daughton and Michael J Paul. Constructing accurate confidence intervals when aggregating social media data for public health monitoring. *Precision Health and Medicine: A Digital Revolution in Healthcare*, pages 9–17, 2020.

[12] Ashlynn R Daughton and Michael J Paul. A bootstrapping approach to social media quantification. *Social Network Analysis and Mining*, 11(1):73, 2021.

[13] Katherine Keith and Brendan O'Connor. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4575–4585, 2018.

[14] Lucien Le Cam. An approximation theorem for the Poisson binomial distribution. 1960.

[15] Yuan H Wang. On the number of successes in independent trials. *Statistica Sinica*, pages 295–312, 1993.

[16] Benjamin Denham, Edmund MK Lai, Roopak Sinha, and M Asif Naeem. Gain-Some-Lose-Some: Reliable quantification under general dataset shift. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1048–1053. IEEE, 2021.

[17] Yili Hong. On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis*, 59:41–51, 2013.

[18] David D Lewis. Evaluating and optimizing autonomous text classification systems. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 246–254, 1995.

[19] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

[20] Amos Storkey et al. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30(3-28):6, 2009.

[21] Rocío Alaíz-Rodríguez, Alicia Guerrero-Curieses, and Jesús Cid-Sueiro. Class and subclass probability re-estimation to adapt a classifier in the presence of concept drift. *Neurocomputing*, 74(16):2614–2623, 2011.

[22] Marthinus Christoffel Du Plessis and Masashi Sugiyama. Class prior estimation from positive and unlabeled data. *IEICE TRANSACTIONS on Information and Systems*, 97(5):1358–1362, 2014.

[23] Yee Seng Chan and Hwee Tou Ng. Estimating class priors in domain adaptation for word sense disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 89–96, 2006.

[24] Zhihao Zhang and Jie Zhou. Transfer estimation of evolving class priors in data stream classification. *Pattern Recognition*, 43(9):3151–3161, 2010.

[25] Marthinus Christoffel Du Plessis and Masashi Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50:110–119, 2014.

[26] Jose Barranquero, Pablo González, Jorge Díez, and Juan José Del Coz. On the study of nearest neighbor algorithms for prevalence estimation in binary problems. *Pattern Recognition*, 46(2): 472–482, 2013.

[27] Hideki Asoh, Kazushi Ikeda, and Chihiro Ono. A fast and simple method for profiling a population of twitter users. In *The Third International Workshop on Mining Ubiquitous and Social Environments*, page 19. Citeseer, 2012.

[28] Víctor González-Castro, Rocío Alaiz-Rodríguez, and Enrique Alegre. Class distribution estimation based on the Hellinger distance. *Information Sciences*, 218:146–164, 2013.

[29] Nachai Limsetto and Kitsana Waiyamai. Handling concept drift via ensemble and class distribution estimation technique. In *Advanced Data Mining and Applications: 7th International Conference, ADMA 2011, Beijing, China, December 17-19, 2011, Proceedings, Part II 7*, pages 13–26. Springer, 2011.

[30] Jack Chongjie Xue and Gary M Weiss. Quantification and semi-supervised classification methods for handling changes in class distribution. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 897–906, 2009.

[31] George Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17:164–206, 2008.

[32] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[33] Arpita Biswas and Suvam Mukherjee. Ensuring fairness under prior probability shifts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 414–424, 2021.

[34] Fabrizio Sebastiani. Evaluation measures for quantification: An axiomatic approach. *Information Retrieval Journal*, 23(3):255–288, 2020.