Proyecto de tesis para optar al título de Magister en Estadística Matemática

Estimación de proporción de clases en muestras no etiquetadas mediante modelos de cuantificación

Director: Dr. Andrés Farall

Alumno: Ing. Maximiliano Marufo da Silva

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Tema de investigación sobre el cual tratará el trabajo

La tarea de cuantificación consiste en proporcionar predicciones agregadas para conjuntos de datos, en vez de predicciones particulares sobre los datos individuales. Por ejemplo, para el caso de la cuantificación aplicada a la clasificación, se busca predecir la proporción de clases de un conjunto de individuos, en vez de la clase particular de cada individuo. 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. Sin embargo, esta estrategia es subóptima: si bien un clasificador perfecto es también un cuantificador perfecto, un buen clasificador puede ser un mal cuantificador.

En cuantificación se aplican modelos que se ajustan usando datos de entrenamiento cuya distribución puede ser distinta a la de los datos de prueba [1]. Si hablamos entonces de cuantificación binaria, se tiene que por cada muestra $i \in \{1, \ldots, n\}$, $(\boldsymbol{X}_i, Y_i, S_i)$ es un vector de variables aleatorias tal que $\boldsymbol{X}_i \in \mathbb{R}^d$ son las características de la muestra, $Y_i \in C$ con $C = \{1, 0\}$ 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. El objetivo es 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 solo 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 [2] son:

- $(\boldsymbol{X}_1, Y_1, S_1) \ldots (\boldsymbol{X}_n, Y_n, S_n)$ son independientes
- Por cada $s \in \{0, 1\}$, $(\boldsymbol{X}_1, Y_1)|S_1 = s, \ldots, (\boldsymbol{X}_n, Y_n)|S_n = s$ son idénticamente distribuidas.
- Por cada $(y_1, \ldots, y_n) \in \{0, 1\}^n$, $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ es independiente de $(S_1, \ldots, S_n)$ condicionado a $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$

Antecedentes sobre el tema

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 el de clasificar y contar. 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.

Aunque 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 modelos que las generen. Además, cabe aclarar que, si bien la aplicación más popular es con respecto a tareas de clasificación (sobre las cuales basaremos principalmente este trabajo, y en particular, sobre clasificación binaria), también se puede aplicar cuantificación a problemas de regresión, ordinalidad, etc.

La literatura sobre métodos relacionados con cuantificación está un tanto desconectada. Algunos 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* [5], *prior probability shift* [6, 7], *posterior probability estimation* [8], *class prior estimation* [9–11], *class prior change* [12], *prevalence estimation* [13], *class ratio estimation* [14] o *class distribution estimation* [15–17], por citar solo algunos de ellos.

Naturaleza del aporte proyectado

El objetivo de este trabajo es describir el problema en el que se enmarca la cuantificación, justificando las razones de por qué es necesario utilizar modelos optimizados para estos casos, y resumir el estado del arte en el área, evaluando mediante simulaciones los principales modelos propuestos.

Metodología tentativa a seguir para lograr los objetivos propuestos

En junio del 2022 el maestrando comenzó elaborando un estudio en profundidad del tema, que incluyó lectura exhaustiva de bibliografía, realización de simulaciones e incluso el inicio de la redacción del informe para la tesis. Con lo cual el presente plan se presenta luego de ya haber comenzado con el estudio y elaboración del trabajo. Los pasos que restan para lograr los objetivos propuestos, junto con sus duraciones estimadas, son:

1. Finalización de la redacción de la parte teórica del informe (2 meses).
2. Ajuste y adecuación de las simulaciones ya realizadas (2 semanas).
3. Elaboración de conclusiones, revisión final y correcciones (2 semanas).

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] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.