Regularización a partir de grafos en modelos potenciados por el gradiente

Federico Albanese[1,2,†], Esteban Feuerstein[1], and Leandro Lombardi[2]

[1]DC - FCEyN - UBA, ICC-CONICET.
[2]Instituto de Cálculo (IC), Universidad de Buenos Aires (UBA)
[†]ffalbanese@gmail.com

Mayo 2019

Palabras Claves: Aprendizaje automático; Grafos; Aprendizaje semi supervisado; términos de regularización.

Resumen extendido

En el paradigma clásico de aprendizaje supervisado, para realizar una tarea de clasificación (o regresión) tendremos acceso a un conjunto de instancias X_{label} etiquetadas previamente [1]. Sin embargo, debido al alto costo del etiquetamiento y la dificultad de obtener instancias etiquetadas, el aprendizaje semi-supervisado (SSL) utiliza también instancias $X_{unlabel}$ sin etiquetar [2]. SSL tiene como objetivo construir modelos que utilizan tanto las instancias etiquetadas como las no etiquetadas, y por los motivos anteriores, este paradigma emerge como una herramienta de vital importancia en la actualidad [3]. Una forma de encarar el problema consiste en armar una red cuyos nodos sean las instancias y generar aristas a partir de la metadata o de la similitud entre nodos [4].

Además, en el contexto del análisis de datos, existen escenarios que pueden pensarse naturalmente como grafos. Esto ocurre en situaciones donde se consideran importantes, además de las propiedades individuales, la conectividad presente entre los elementos del conjunto de datos. Por lo tanto, resulta lógico que modelos de aprendizaje automático incluyan información tanto de un nodo como de sus vecinos a la hora de realizar una predicción, si es que se asume homofilia en la red (las características de un nodo dependen de a las características de sus vecinos) [4].

A la hora de realizar una clasificación sobre datos estructurados como una red compleja, se pueden utilizar distintas estrategias [5]. Recientemente se popularizaron las redes convolucionales sobre grafos o "Graph Convolutional Networks"(GCN) [6]. Las mismas poseen una arquitectura que les permite operar

sobre una red teniendo en cuenta no solo las características de un nodo, sino también las de sus vecinos. En forma simplificada, su regla de propagación es:

$$H^{(l+1)} = \sigma(AH^{(l)}W^{(l)}) \tag{1}$$

donde $H^{(i)}$ representa los datos de la i-ésima capa de la red neuronal, $W^{(i)}$ la matriz de pesos para dicha capa, A la matriz de adyacencia y σ la función de activación.

En contraste, otra estrategia consiste en utilizar un modelo de aprendizaje automático para una tarea de aprendizaje supervisado y modificarle la función de costo, de forma tal que la función objetivo a minimizar tenga no sólo en cuenta el error de la predicción sino también la estructura de la red. Tal como se puede ver en la ecuación (2) [3], la función objetivo aplica una penalidad cuando a nodos vecinos se les predice etiquetas distintas:

$$L = \sum_{i \in L} (f(x_i) - y_i)^2 + \sum_{i,j} \alpha w_{i,j} (y_i - y_j)^2 \tag{2}$$

donde L es la función objetivo, el primer término es el error cuadrático y el segundo la diferencia entre predicciones de nodos vecinos ponderada por el peso de la arista w_{ij} entre los nodos i y j y por un hiperparámetro α.

La técnica de potenciación del gradiente utiliza ensambles de modelos predictivos para realizar la tarea de clasificación y regresión supervisada [7]. Dichos modelos predictivos luego son optimizados iteración a iteración usando el gradiente de la función de costo. En este escenario, XGBoost (XGB), una implementación particular de esta técnica, ha demostrado ser eficiente en una gran variedad de escenarios supervisados superando al resto de los modelos [8].

A partir de esta última idea, en este trabajo proponemos utilizar una función de costo con términos de regularización dependientes de la estructura del grafo combinándola con el modelo de árboles optimizados por el gradiente (XGB+SSL). De esta forma, logramos adaptar el modelo de XGB a escenarios semi-supervisados. Además, mostramos que nuestra implementación supera en performance tanto a redes profundas densas (DNN) como a las GCN.

Se utilizó el armado experimental descrito en [9] y el Cora dataset [10] con el objetivo de estudiar la performance de los distintos modelos. El mismo cuenta con 2708 publicaciones científicas clasificadas en clases que representan los nodos. Las 5429 aristas están determinadas por las citas entre publicaciones. A su vez, cada artículo científico está descripto por 1433 caracteristicas cuyos valores pueden ser 1 o 0 dependiendo de si el articulo contiene determinada palabra.

Los resultados se encuentran resumidos en la tabla 1. En la misma se reportan los resultados de la tarea de clasificación supervisada de una clase usando validación cruzada de 5 particiones.

Al realizar dicha comparación entre modelos modernos y nuestra implementación de una función de costo junto a XGBoost, se concluyó positivamente que esta última obtiene una performance superior para todas las métricas excepto por la precisión. Por ejemplo, la mejora consistió en más de un 7 % en la ac-

	DNN	GCN	XGB	XGB+SSL
Exactitud	0.560	0.814	0.876	**0.888**
Precisión	0.878	**0.930**	0.750	0.898
Valor F1	0.199	0.619	0.579	**0.627**
Exhaustividad	0.112	0.463	0.471	**0.481**
Área Bajo la Curva	0.552	0.723	0.718	**0.734**

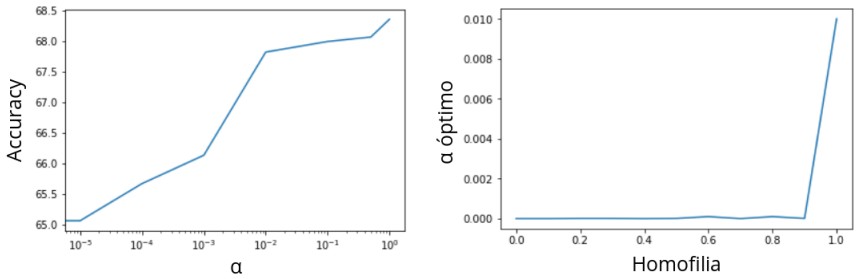
Tabla 1: Resultados de la tarea de clasificación semi-supervisada para los distintos modelos. El valor más alto se encuentra en negrita.

curacy. La importancia de la implementación realizada reside en poder adaptar eficientemente XGBoost al escenario semi-supervisado.

Luego, se buscó estudiar el método en si. Dado que el método depende de la hipótesis de homofilia, resulta de interés analizar la relación entre la accuracy al realizar una predicción y la homofilia presente en la red.

Para ello, se generaron aleatoriamente redes con distintos valores de homofilia usando el método de Erdös–Rényi [11]. Cada red cuenta con 1000 nodos y una probabilidad de que haya una arista entre dos nodos de 0,9. Las etiquetas de los nodos toman valores de 1 y 0 de forma tal de respetar la homofilia de la red.

En la figura (izquierda) se puede observar cómo para una red con una homofilia de 0,9, el accuracy aumenta al aumentar el valor del hiperparametro α (ver la ecuación 2). Dicho resultado es consistente con la idea de que en una red con homofilia, es importante que nodos conectados entre si tengan características iguales. Por otro lado, en la figura de la derecha, se puede ver el α óptimo (el que maximiza la accuracy) en función de la homofilias. De forma intuitiva, para valores bajos de homofilia, α oscila cerca del 0, ya que exigir que nodos conectados tengan predicciones iguales lleva a aumentar el error. Sin embargo, al superar un valor limite, comienza a resultar provechoso utilizar el método con términos de regularización dependientes de la conectividad de la red.



En conclusión, se logró adaptar eficientemente XGBoost al escenario semi-supervisado al implementar una función de costo con términos dependientes de la red y, de esta forma, mejorar los resultados para una tarea de clasificación semi-supervisada superando modelos de redes neuronales profundas.

Referencias

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[2] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[3] X Zhu. Semi-supervised learning literature survey, department of computer sciences, university of wisconsin at madison, madison. Technical report, WI, Technical Report 1530. http://pages. cs. wisc. edu/˜ jerryzhu/pub . . . , 2006.

[4] Andrew B Goldberg and Xiaojin Zhu. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52. Association for Computational Linguistics, 2006.

[5] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.

[6] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[7] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[8] Didrik Nielsen. Tree boosting with xgboost-why does xgboost win.^{ev}ery"machine learning competition? Master's thesis, NTNU, 2016.

[9] Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. *arXiv preprint arXiv:1603.08861*, 2016.

[10] Cecile Cabanes, A Grouazel, K von Schuckmann, M Hamon, Victor Turpin, C Coatanoan, F Paris, S Guinehut, C Boone, N Ferry, et al. The cora dataset: validation and diagnostics of in-situ ocean temperature and salinity measurements. *Ocean Science*, 9(1):1–18, 2013.

[11] Paul Erdös and Alfréd Rényi. On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.