Aprendizaje automático para clasificación anticipada en datos secuenciales [1] [2]

Doctorado en Ciencias de la Computación, Universidad Nacional de San Luis, Argentina.

Tesista: Juan Martín Loyola [1,2]
Director: Marcelo Luis Errecalde [1]
Co-Director: Esteban Gabriel Jobbágy Gampel [2]

[1] Universidad Nacional de San Luis (UNSL), Ejército de Los Andes 950, San Luis, C.P. 5700, Argentina
[2] Instituto de Matemática Aplicada San Luis (IMASL), CONICET-UNSL, Av. Italia 1556, San Luis, C.P. 5700, Argentina

Palabras claves: Detección Anticipada de Riesgos, Clasificación Anticipada de Texto.

1 Motivación

En la formulación tradicional del aprendizaje automático (supervisado) el problema es construir un clasificador que pueda predecir correctamente las clases de nuevos objetos, dados ejemplos de entrenamiento de viejos objetos. El supuesto en este caso es que los ejemplos de entrenamiento corresponden a datos aislados, e independientes entre sí, con suficiente información relevante auto-contenida como para hacer un análisis individual (clasificación) aceptable.

Sin embargo, este esquema de trabajo no se adapta a muchas situaciones del mundo real donde la efectividad del sistema de clasificación depende directamente de considerar las observaciones/datos respetando la secuencia en que se fueron generando. Tomemos, por ejemplo, un modelo del lenguaje que predice la probabilidad de ocurrencia de la siguiente letra. Si el sistema leyó una “Q”, la probabilidad de ocurrencia de una “u” será significativamente más alta que la de cualquier otra letra. De igual manera, la interpretación del significado de una palabra como “banco”, no será el mismo si previamente dije que “para comprar esta casa debo retirar dinero del” <banco>, que si hubiera dicho “me sentía cansado, por lo que decidí sentarme en el” <banco>. En ambos casos, la palabra polisémica “banco”, requiere de las secuencias previas de palabras emitidas, para eliminar cualquier ambigüedad sobre el significado que tiene en cada caso. Esta situación, que hemos ejemplificado con palabras, se repite en un sinnúmero de situaciones involucrando sonidos, imágenes y las más diversas señales sensoriales, en las cuales la correcta interpretación del dato actual de entrada sólo puede realizarse en forma realista,

[1] Video de exposición: https://www.youtube.com/watch?v=y1h6RYVXB2Q
[2] Diapositivas: https://jmloyola.github.io/files/talks/2021_encuentro_posgrado.pdf

considerando la secuencia de datos previos, e incluso en muchos casos, dependiendo de datos producidos muchos pasos hacia atrás en esa secuencia.

En este contexto, esta tesis se enmarca en el área del aprendizaje automático con datos secuenciales (AADS), es decir, asumiremos que el algoritmo de aprendizaje automático explícitamente considera que la entrada es una secuencia.

Varios autores han categorizado las aplicaciones de AADS de distintas formas, dependiendo de las características de la entrada y de la salida. En particular, Graves [1], utiliza como marco de referencia el etiquetado de secuencias (sequence labelling) cuyo objetivo es asignar secuencias de etiquetas (tomadas de un alfabeto fijo), a las secuencias de entrada. En este contexto, el tipo de tarea se vincula a las distintas restricciones que se imponen en ese proceso de etiquetado.

Cuando las secuencias de etiquetas son restringidas a tener longitud uno la tarea recibe el nombre de “clasificación de secuencia”. Si las secuencias de salida consisten en muchas etiquetas, pero los puntos de la secuencia de entrada donde estas etiquetas deben ser producidas son conocidas de antemano, las tareas son referenciadas como de “clasificación de segmentos”. Por último, el escenario que Graves llama “clasificación temporal”, no impone ningún tipo de alineamiento entre las secuencias de entrada y salida, e incluso la de salida puede ser vacía. El elemento crucial que se incorpora en este caso es que el sistema requiere de un algoritmo para decidir en qué lugar de la secuencia de entrada se debería generar la clasificación (etiqueta) correspondiente.

Esta última nomenclatura es de interés para nuestro trabajo, ya que incorpora el aspecto de la decisión de “cuándo” (en qué lugar de la secuencia de entrada) se debería tomar la decisión de generar la etiqueta (clasificación) correspondiente. Este es un aspecto fundamental en un tipo de clasificación temporal que suele ser referenciada como de “clasificación anticipada” (CA). La idea subyacente a la CA es que el clasificador debería ser capaz de poder clasificar la secuencia de entrada tan pronto tenga la información relevante necesaria para poder realizar esta clasificación de manera confiable. La clasificación anticipada suele ser un aspecto deseable, ya que puede en algunos casos evitar algún tipo de costo asociado con la lectura completa de la secuencia de entrada o bien producir una mayor utilidad/beneficio al clasificar anticipadamente el flujo de entrada.

Sin embargo, existen casos donde la CA no es sólo “deseable”, sino también “crítica” ya que existe un riesgo asociado con la demora en la clasificación de la secuencia. Estos escenarios, que serán uno de los ejes de esta propuesta de tesis, se han popularizado últimamente con el nombre de “detección anticipada de riesgos” (DAR) (en inglés “early risk detection”).

2 Objetivos y Aportes

El objetivo principal de esta tesis es el estudio, formulación y desarrollo de representaciones y métodos de aprendizaje automático para datos secuenciales. El interés principal en nuestro caso estará dado en aquellos dominios en los cuales existe una demanda concreta por clasificar las secuencias con la mayor antelación posible.

Particularmente, se busca:

- Relevar el estado del arte en enfoques de AADS y DAR.
- Construir o adaptar colecciones de datos existentes que sean adecuadas para el entrenamiento y evaluación de enfoques de AADS y DAR.
- Definir nuevas representaciones y algoritmos para AADS y DAR que sean representativas del estado del arte.

3 Estado Actual y Trabajo Futuro

La primera etapa del trabajo de tesis estuvo principalmente enfocada en el estudio del estado del arte tanto en enfoques de Aprendizaje Automático con Datos Secuenciales, como en los problemas de Clasificación Anticipada y Detección Anticipada de Riesgo. A partir de esto, se formalizó el marco de trabajo requerido para problemas de Clasificación Anticipada [2] haciendo hincapié en los dos componentes que se deben resolver: Clasificación con Información Parcial (CIP) y la Decisión del Momento de Clasificación (DMC).

Al ser un problema no abordado previamente, no existían, a mediados de 2017, conjuntos de datos con los que se pudieran evaluar los enfoques de clasificación anticipada y comparar con otros grupos de investigación. Afortunadamente, a finales de 2017 se creó el laboratorio de predicción temprana de riesgos en Internet, eRisk[3]. El objetivo del laboratorio es explorar las metodologías de evaluación, las métricas de efectividad y las aplicaciones prácticas de la detección temprana de riesgos en Internet. Todos los años, el laboratorio provee una serie de conjuntos de datos de entrenamiento donde los diferentes grupos entrenan sus modelos, y luego comparan el desempeño de todos en los respectivos conjuntos de datos de prueba.

Una particularidad de los problemas de detección anticipada de riesgo es que suelen tener un desbalance de clases considerable. Esto se debe a que, en general, los casos de riesgo son mucho menores en cantidad que los casos de no riesgo. Así, para mejorar el desempeño de los modelos propuestos, se amplió el conjunto de datos de entrenamiento utilizando información de Reddit.

En la edición del año 2021 del laboratorio eRisk nuestro grupo presentó tres tipos de modelos distintos para detección anticipada de riesgo [3]:

- EarlyModel: modelo simple basado en el marco de clasificación anticipada propuesto en [2]. El rol del CIP puede ser llevado a cabo por cualquier clasificador de texto que retorne la probabilidad de la clase predicha. Por otro lado, para la DMC se utilizó un árbol de decisión.
- SS3: modelo similar al anterior donde el rol del CIP es llevado a cabo por el modelo SS3 [4]. Por otro lado, para la DMC se utilizó una función que considera el contexto de todos los documentos siendo procesados en paralelo.

[3] https://early.irlab.org/

- EARLIEST: modelo de aprendizaje profundo *end-to-end* entrenado utilizando Aprendizaje por Refuerzo para aprender cuándo detener la lectura de la entrada y clasificar. La representación aprendida por el modelo es utilizada tanto para clasificar la entrada en riesgo o no-riesgo, como para determinar si se debe detener la lectura o no.

Como se puede ver en la tabla a continuación, los resultados obtenidos por estos modelos fueron muy alentadores, obteniendo los mejores resultados para la medida F_1 y las medidas que consideran el tiempo de clasificación [5].

team name	run id	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency\text{-}weighted\ F1$
UNSL (EarlyModel)	0	.336	.914	.491	.125	**.034**	11	.961	.472
UNSL (EARLIEST)	1	.11	.987	.198	.093	.092	**1**	**1.0**	.198
UNSL (EARLIEST)	2	.129	.934	.226	.098	.085	**1**	**1.0**	.226
UNSL (SS3)	3	.464	.803	.588	.064	.038	3	.992	.583
UNSL (SS3)	4	.532	.763	**.627**	.064	.038	3	.992	**.622**
NLP-UNED	4	.453	.816	.582	.088	.04	9	.969	.564
Birmingham	0	.584	.526	.554	.068	.054	2	.996	.551
Birmingham	2	**.757**	.349	.477	.085	.07	4	.988	.472
EFE	2	.366	.796	.501	.12	.043	12	.957	.48
BLUE	2	.454	.849	.592	.079	.037	7	.977	.578
UPV-Symanto	1	.276	.638	.385	**.059**	.056	1	1.0	.385

Queda pendiente como trabajo futuro analizar por qué el modelo EARLIEST no tuvo el desempeño esperado y proponer mejoras al modelo. Además, nos interesa determinar si agregar más información del contexto puede beneficiar el desempeño de los modelos.

Referencias

1. Graves, A.: Supervised sequence labelling. In: Supervised sequence labelling with recurrent neural networks, vol. 385, pp. 5-13. Springer, Berlin, Heidelberg (2012).
2. Loyola, J.M., Errecalde, M.L., Escalante, H.J., Montes y Gomez, M.: Learning when to classify for early text classification. In Argentine Congress of Computer Science, pp. 24-34. Springer, Cham (2017, October).
3. Loyola, J.M., Burdisso, S.G., Thompson, H., Cagnina, L., Errecalde, M.L.: UNSL at eRisk 2021 A Comparison of Three Early Alert Policies for Early Risk Detection. In Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum, Bucarest, Romania (2021, September).
4. Burdisso, S.G., Errecalde, M.L., Montes-y-Gómez, M.: A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications, vol. 133, 182-197 (2019).
5. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: Overview of erisk 2021 Early risk prediction on the internet. In Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum, Bucarest, Romania (2021, September).