

Plan de Gestión de Datos

INFORMACIÓN SOBRE EL PROYECTO

1. – Título del Proyecto

- Título del Proyecto (en castellano)

Nuevos modelos de extremo a extremo basados en aprendizaje de representaciones y redes neuronales en grafos para datos secuenciales con aplicaciones en bioinformática

- Título del Proyecto (en inglés)

End-to-end models based on representation learning and graphs neural networks for sequential data with applications in bioinformatics

-Descripción del Proyecto (en castellano) Resumen

El surgimiento del aprendizaje profundo ha potenciado la inteligencia artificial, generando además una gran transformación en diversos campos de aplicación. Los modelos profundos pueden extraer automáticamente características de textos, señales e imágenes, para luego utilizarlas en la solución a problemas específicos. Sin embargo, quedan aún muchos desafíos para su desarrollo, como su capacidad para capturar relaciones entre elementos lejanos en datos secuenciales o la capacidad de aprender a partir de pocos ejemplos etiquetados. La bioinformática en general presenta estos desafíos y provee un marco experimental bien definido para desarrollar y validar nuevos modelos y métodos de aprendizaje automático. Un tipo de secuencia de creciente interés son los ARNs no codificantes, que cumplen diversas funciones regulatorias y suelen estar involucrados en enfermedades tales como el cáncer, trastornos neurológicos y respuestas inflamatorias. Pero para poder dilucidar su función y mecanismos de acción es clave conocer su estructura, que es la que les permite interactuar con otras moléculas. La predicción de estas estructuras presenta diferentes desafíos que los métodos clásicos, basados en características termodinámicas locales y programación dinámica, aún no logran resolver. Si bien han surgido soluciones basadas en aprendizaje automático, su desempeño predictivo aún es bajo debido a los diversos desafíos computacionales involucrados en la tarea. Por un lado, desde el punto de vista del aprendizaje automático constituye un problema de alto desbalance de clases, dado que de las posibles conexiones entre sus elementos a predecir sólo unas pocas se dan efectivamente. Por otro lado, las secuencias pueden tener desde decenas hasta miles de elementos y las conexiones se pueden dar a cualquier distancia. Además, se cuenta con muy pocas estructuras validadas experimentalmente, lo que hace difícil evitar el sobreajuste durante el entrenamiento. En este proyecto se atacará el problema de aprendizaje de relaciones de largo alcance en datos secuenciales, manteniendo la capacidad de generalización a partir de datos desbalanceados y escasamente etiquetados. El proyecto incluye todas las etapas desde el curado de datos, el desarrollo de los modelos y estrategias originales de aprendizaje profundo, hasta la validación con estructuras determinadas experimentalmente. Dentro del modelado, se explorarán tres líneas principales: el aprendizaje de representaciones de ARN vía mecanismos de atención, la clasificación de secuencias largas con redes residuales y la mejora en la capacidad de generalización mediante redes neuronales en grafos. Los resultados serán de impacto tanto en el área de inteligencia artificial como en bioinformática.

-Descripción del Proyecto (en inglés) Resumen

Deep learning models can automatically extract features from text, signs, and images, then use them to solve specific problems. However, many challenges remain for its development, such as its ability to capture relationships between distant elements in sequential data or the ability to learn from few labeled examples. Bioinformatics presents these challenges and provides a well-defined experimental framework to develop and validate new machine learning models. A type of sequence of increasing interest are non-coding RNAs, which are involved in various regulatory functions and are diseases such as cancer, neurological disorders and inflammatory responses. However, in order to elucidate their function and mechanisms of action, it is key to know their structure, which is what allows them to interact with other molecules. The prediction of these structures presents different challenges that classical methods, based on thermodynamics and dynamic programming, have not yet been able to solve. Although deep learning-based solutions have emerged, their predictive performance is still low due to the various computational challenges involved in the task. On the one hand, from the point of view of machine learning it constitutes a problem of high class imbalance, given that of the possible connections between its elements to be predicted, only a few actually occur. On the other hand, sequences can have from tens to thousands of elements and connections can occur at any distance. Furthermore, there are very few experimentally validated structures, which makes it difficult to avoid overfitting during training. In this project, the problem of learning long-range relationships in sequential data will be addressed, maintaining the ability to generalize from unbalanced and few

labeled data. The project includes all stages from data curation, development of original deep learning models and strategies, to validation with experimentally determined structures. Within the modeling, three main lines will be explored: the learning of RNA representations via attention mechanisms, the classification of long sequences with residual neural networks, and the improvement in generalization capacity through graph neural networks. The results will have impact both in the area of artificial intelligence and bioinformatics.

-Palabras Clave descriptivas del Proyecto (en castellano)

Desbalance de clases Atención de corto y largo alcance Estructura secundaria de ARN

- Palabras Clave descriptivas del Proyecto (en inglés)

Class imbalance Short and long range attention RNA secondary structure

2 – Datos del Director/a del Proyecto

- Nombre y Apellido

Diego H. Milone

- Unidad Académica

FICH

- Teléfono oficial de contacto

342 4575233 int 195

-Teléfono móvil de contacto

342 5512281

-E-mail del Director/a del Proyecto

dmilone@sinc.unl.edu.ar

DATOS RESULTANTES DE LA EJECUCIÓN DEL PROYECTO

-Describa la toma de muestras / datos a realizar

Se utilizarán datos de dominio público.

– Datos: ¿Existe alguna razón por la cual los datos declarados no deban ser puestos a disposición de la comunidad / ser de acceso público? (marque X)

☒ NO

SI. Elija una de las opciones:

– Período de Confidencialidad. Marque su opción con “X”.

☐ 1 (UN) año

☐ 2 (DOS) años

☐ 3 (TRES) años

☐ 4 (CUATRO) año

☐ 5 (CINCO) años

☐ Otro.

Motivos: