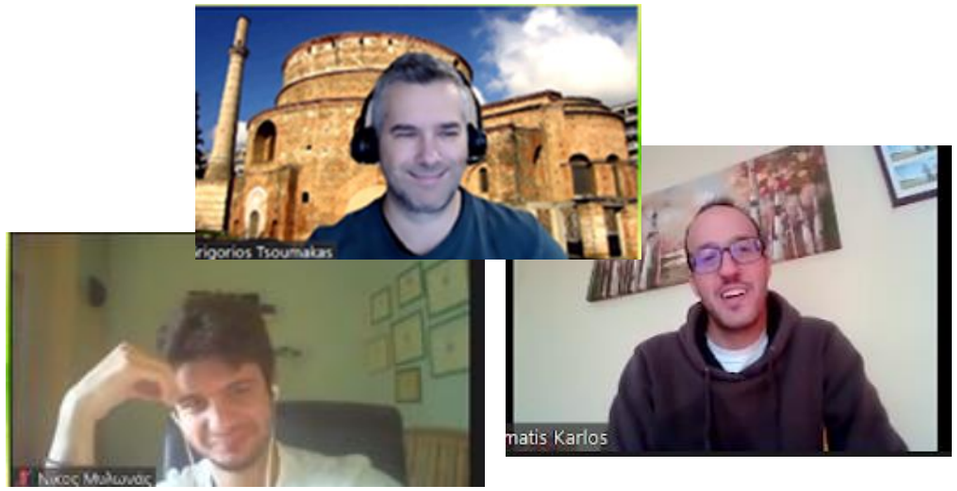
H.F.R.I.
Hellenic Foundation for Research & Innovation

Description of the funded research project

1st Call for H.F.R.I. Research Projects to Support Faculty Members & Researchers and Procure High-Value Research Equipment

Title of the research project:
Advanced Multi-Label Learning Techniques

AMULET



Principal Investigator:
Grigorios Tsoumakas

Reader-friendly title:
AMULET

Scientific Area:
Mathematics and Information Sciences

Institution and Country:
Aristotle University of Thessaloniki, Greece

Host Institution:
Aristotle University of Thessaloniki

Collaborating Institution(s):
None

Project webpage:
https://amulet.csd.auth.gr

Budget: 158,270.00 €

Duration: 36 Months

H.F.R.I.
Hellenic Foundation for Research & Innovation

Research Project Synopsis

Multi-label data consist of instances that are associated with a vector of binary target variables. The last 10 years, the topic of learning from multi-label data has witnessed enormous progress, evident by the increasing number of papers dealing with this topic, as well as by the fact that it has recently started to appear as a distinct topic in top conferences like KDD, AAAI and ICML. Despite all this amount of work, several challenges still arise when applying multi-label learning in real-world applications and industrial settings. The main goal of AMULET is to develop novel multi-label learning techniques to deal with two such key under-addressed challenges, paving the way for a wider adoption of multi-target prediction in complex real-world tasks: i) concept evolution, ii) interpretability. *Specifically, for the first challenge AMULET aims* to deal with the evolution of concepts in the output space of complicated biomedical vocabularies, which can manifest as complex changes in existing labels (e.g. merging/separation of labels, change of hierarchical structure), the addition of brand-new labels and concept drift which in such cases refers to the change of meaning in existing labels. For the second challenge, AMULET aims to develop techniques for explaining the output of multi-label models, with an emphasis on textual data. These techniques should be able to explain the binary decision per label based on the values of related labels and based on comprehensible concise representations of the text, such as interpretable embeddings and explainable keyphrases.

Project originality

Existing multi-label stream classification approaches do not fully address the complexity of changes that occur in real-world cases, such as the MeSH ontology. In AMULET, we will research into techniques to explicitly deal with the addition of new labels, with complex changes to labels (merging/splitting of labels, changes in the hierarchical structure of the labels, and changes involving the lexical terms associated with labels), as well as with automatically detecting and adapting to concept drift. A key novelty of all our research efforts in this topic is that they will heavily employ explicit information, in addition to the typical statistical concept drift detection strategies, by analyzing the time-points and semantics of MeSH label updates catalogued by the National Library of Medicine at the end of each year. Furthermore, in AMULET we will focus on interpreting the decisions of multi-label text classifiers, using as a particular use case, the classification of hate speech aspects in YouTube comments. To the best of our knowledge, no other work has studied the need for explanations in the context of moderating online speech in social media.

Expected results & Research Project Impact

The expected impact of AMULET is threefold, oriented towards science, economy and the society.

Regarding the first aspect, AMULET sets out to solve two important research challenges in the context of multi-label data that are expected to initiate further research from other researchers that will follow up our developments. Streams of multi-label data are particularly interesting today, in the context of the Internet of Things, where streams of sensor values are increasingly being recorded and need analysis, and in general due to our increasing capabilities for monitoring and recording of information. Concept drift is ever-present when monitoring real-world phenomena. Having interpretable multi-label learning approaches is very important in regulated industries, such as health, banking and insurance.

For the second aspect, AMULET aims to make multi-label learning more applicable in industrial settings. In the industry of academic publishing, dealing successfully with concept evolution is expected to improve the accuracy of automated semantic indexing systems and thus reduce the costs related to manual human annotating. This is important in other content-rich industries like news and media. In the health and banking industries understanding predictions is expected to lead to fewer mistakenly taken decisions.

Concerning the third aspect**,** semantic indexing of scientific literature is a very important multi-label learning task for the society in general, as it enables researchers to find the most relevant information among the millions of scientific articles. AMULET will improve the technology of semantic indexing by addressing the concept evolution challenge. These advancements contribute to dealing with the information overload that researchers are currently facing. In addition, online hate speech has debilitating consequences on individual victims' well-being by imposing psychological harm and damaging self-worth. AMULET will offer technologies for explaining the recommendations of machine learning systems with respect to hate speech.

The importance of this funding

HRFI funding plays a key role in making this research happen. It is not a secret that funding is the fuel that powers a research team. HRFI funding helps members of this research team keep focused on the posed research goals, allowing them to devote their full time into this project's actualization. Otherwise, separate kinds of funding would be sought by its members, which would put obstacles on being kept aligned with the state-of-the-art approaches and deliver valuable contributions to such a competitive environment. In addition, funding is important for securing the needed computational resources, as well as covering the expenses of attending highly ranked conferences around the world.

H.F.R.I.
Hellenic Foundation for Research & Innovation

COMMUNICATION

185 Syggrou Ave. & 2 Sardeon St. 2
171 21, N. Smyrni, Greece
+30 210 64 12 410, 420
communication@elidek.gr
www.elidek.gr