Título: **Feature selection with low correlated binary features for potential tax fraudsters classification**

Data: **19/06/2019**

Horário: **10:00h**

Local: **Sala de Seminários - Bloco 952**

Resumo:

Feature selection methods provides us a way of reducing computation time, improving prediction performance, and a better understanding of the data in machine learning or pattern recognition applications. It has become the focus of much research in areas of application. In this work, we use feature selection to select the most relevant features in order to improve the binary classification of potential tax fraudsters. Detecting instances of frauds from taxpayer data, with binary features, presents several challenges: firstly, taxpayer data typically have features with low linear correlation between themselves. Also, tax frauds may originate from intricate illicit schemas, which in turn requires to uncover non-linear relationships between multiple fraud indicators (features). Finally, of the fraud indicators in experiments, only a small number of them show some correlation with the targeted class.

Tax evasion represents one of the major obstacles faced by the economies of developing countries. Vast amounts of taxpayer information has been collected by fiscal agencies, thus opening up the possibility of devising novel techniques able to tackle fiscal evasion much more effectively than traditional approaches. In this work we propose ALICIA, a new feature selection method based on association rules and propositional logic with a carefully crafted graph centrality measure that attempts to tackle the above challenges while, at the same time, being agnostic to specific classification techniques. ALICIA wants to capture the intrinsic interrelation between the features in tax fraud detection. The proposed methodology is structured in three phases: firstly, ALICIA generates a set of relevant association rules from a set of fraud indicators (features). Subsequently ALICIA builds a graph, where each node represents a subset of features resulting in the association rules, while edges represent association relationships between subsets of features. Finally, ALICIA determines the most relevant features by applying a novel centrality measure, the Fraud Feature Topological Importance,

Banca:

- Prof. Dr. José Maria da Silva Monteiro Filho (MDCC/UFC - Orientador)
- Prof. Dr. José Antonio Fernandes de Macêdo (MDCC/UFC - Coorientador)
- Prof.ª Dr.ª Chiara Renso (ISTI-CNR / Itália)
- Prof. Dr. César Lincoln Cavalcante Mattos (MDCC/UFC)
- Prof. Dr. Franco Maria Nardini (ISTI-CNR / Itália)