Pulmonary Embolism

Pulmonary Embolism (PE) is a relatively common and potentially lethal condition. Most patients die in the first few hours from the event. Despite advances in diagnostics, delays in PE diagnosis are common and represent an important issue. As a cause of sudden death, massive PE is second only to arrhythmia death. Among survivors, recurrent embolism and death can be prevented with prompt diagnosis and therapy. Unfortunately, diagnosis is often missed because patients often present with nonspecific signs and symptoms. If left untreated, approximately one third of patients who survive an initial pulmonary embolism die from a subsequent embolic episode. Indeed, recent studies have demonstrated the safety of rejecting the diagnosis of PE by the combination of a low clinical probability and a normal quantitative d-Dimer test result, thereby decreasing the need for further diagnostic radiological imaging in up to 30% of patients. The best validated and therefore most widely used clinical decision rules are the Wells and the Geneva Revised scores. However, both indices have limitations.

Clinical Prediction Rules

Indeed, recent studies have demonstrated the safety of rejecting the diagnosis of PE by the combination of a low clinical probability and a normal quantitative d-Dimer test result, thereby decreasing the need for further diagnostic radiological imaging in up to 30% of patients. The best validated and therefore most widely used clinical decision rules are the Wells and the Geneva score. However, both scores have limitations. The Wells rule includes the physician’s judgement of whether an alternative diagnosis is more likely than PE. This criterion, which carries a major weight in the score, is subjective and cannot be standardized. Moreover, it has been suggested that the predictive value of the Wells rule is derived primarily from its subjective component. The Geneva score, based on 13 entirely objective variables, requires a blood gas analysis while breathing room air and has only been evaluated for patients in the emergency ward. Both scores appeared to have a comparable predictive value for PE.

Data Probing and Topological Data Analysis by Iris

Data Probing: clinical variables are strongly uncorrelated and for this reason the application of Principal Component Analysis is not recommended. The values for both classes are directly comparable except for d-Dimer and White Blood Count Cells. The results confirm that Pulmonary Embolism is difficult to diagnose.

Topological Data Analysis (TDA) is a method of analysis of multidimensional complex data primarily driven by its geometry. The three fundamental concepts of TDA are independence of coordinate systems; insensitivity to deformation as long as the shape of the data does not get broken or torn; and compressed representation. Using a mathematical concept of lenses, data can be projected onto a subspace suitable for visualization. The topological features of the subspace are then inspected with traditional statistical approaches such as Kolmogorov-Smirnov.

We used Iris for the application of TDA to the clinical data of patients suspected of high risk of PE and used shapes of the generated topological networks to identify different subgroups of patients and features that statistically significantly explain the observed differences. We used the TDA-derived features as input for artificial neural network analysis.

IRIS detected some sub-groups among healthy patients, which present other diseases to be diagnosed properly even though they are not affected by PE. It highlights a set of 6 clinical variables among the original set of 26 dimensions that can be used to discriminate the pathological and healthy classes. The new set of variables have been used for the training of the artificial neural networks.

Artificial Neural Networks

Backpropagation Algorithm: the idea of the backpropagation algorithm is to reduce the training error (difference between actual and expected results), until the Artificial Neural Network (ANN) learns the training data. The training begins with random weights, and the goal is to adjust them so that the error will be minimal. For practical reasons, ANNs implementing the backpropagation algorithm do not have too many layers, since the time for training the networks grows exponentially.

Levenberg-Marquardt: is a very simple method for approximating a function. In the ANNs field, this algorithm is suitable for small and medium-sized problems. Basically, it consists in solving the equation:

\[ (J^T J + \lambda I) \delta = J^T E \]  \hspace{1cm} (1)

The L.M. is very sensitive to the initial network weights. Also, it does not consider outliers in the data, that may lead to overfitting noise.

Conclusions

The purpose of this study was to derive a new CPR for pulmonary embolism and to study patients’ cohort characteristic with a topological approach. The new CPR has been obtained training an artificial neural network on the input formed by a set of features selected by Iris. Iris extracted new knowledge from the patients’ dataset by the application of an innovative approach for data analysis: Topological Data Analysis. Our results show that the feature selection strategy is beneficial for the performance improvement of an ANN trained on the analyzed cohort. A three-layer neural network can be trained to successfully perform the diagnostic task. In conclusion a system based on Iris and an ANN can form the basis of a CAD system to assist physicians with the right clusterization of patients. In the future we will also perform a validation of the system both increasing the number of patients in the dataset and using different cohorts, we will perform a comparison study among artificial neural networks and other classification systems.

I acknowledge Francesco Vaccarino and Giovanni Petri from ISI foundation for their introduction to algebraic topology and complex networks. The financial support of FET programme within the Seventh Framework Programme for Research of the European Commission, under the FET-Proactive grant agreement TOPDRIM (Topology-driven methods for multilevel complex systems), number FP7 ICT-318121.