SCIENCESPRINGDAY

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

DEPARTAMENTO DE INFORMÁTICA

Ubiquitous Data Mining With Artificial Neural Networks

CITI / MultiModal Systems

IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia de Setúbal

CHi CENTER FOR INFORMATICS AND INFORMATION TECHNOLOGIES

Bruno Silva

(Student)

Teacher at Escola Superior de Tecnologia de Setúbal and at the Portuguese Navy School.

Research has focused mainly on Artificial Neural Networks and in particular Self Organizing Maps.

Objectives

Ubiquitous environments pose different challenges to data mining methods. Data is scattered, generated continuously and the underlying concept may change – concept drift. Also, it is desirable that processing nodes, e.g., mobile devices, communicate with each other and share «knowledge». General data mining methods cannot be applied directly in these conditions.
Artificial Neural Networks (ANN) are a well-established set of, biologically inspired, mining algorithms and are recognized widely by their ability to discover hidden patterns, generalization capabilities and robustness to noise. However, using ANN for mining data streams is still a very unexplored path, which justifies the current research. Unsupervised ANN learning methods are best suited to learn from continuous streams of data, but must be modified to cope with the data stream model and dynamic environments.

Methodology

The most relevant and established unsupervised ANN methods are the Self-Organizing Map (SOM) and the Adaptive Resonance Theory (ART) networks. Both generate prototypes of data from observations. The SOM is a dimensionality reduction method that projects high-dimensional data onto a fixed 2D plane, over which visualization techniques are applied to get insight on existing clusters and non-linear correlations between features; ART networks are clustering methods that retain «plasticity» over time. The current proposed framework (Figure 1) uses a two-phased learning process: in the online part of the framework an ART network is used to produce data aggregations of the incoming stream; these are then used to train the SOM networks offline. This also involves the modification of the update rules used by these methods. Also, from the summarization process it is possible to evaluate concept drift. Further goals involve creating an update rule for the SOM that retains indefinite plasticity (to use in real-time) and explore Hopfield networks in these settings.

Expected Results

Some results have already been published, namely:

- Viability of the proposed ANN framework for continuous streams of data;
- Application of the framework to detect concept drift in financial markets – results show that prior to the 2008 market crash the concept changed, which can be considered a warning (Figure 2);
- The idea that, when using unsupervised ANN methods, it is relatively straightforward to exchange «knowledge» between processing nodes (devices, agents) by incorporating prototypes obtained from other node in the local model. This was shown for the SOM (Figure 3).

Other auxiliary works relating to feature clustering with the SOM and use of different metric distances have also been published.
Comparison results of the proposed ANN methods with current state-of-the-art approaches are under way and expected to be published soon.

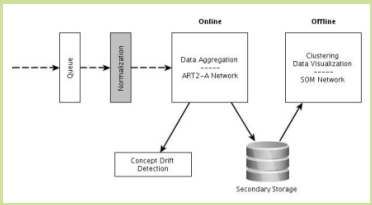

Figure 1- The proposed two-phased ANN framework for continuous data (streams).

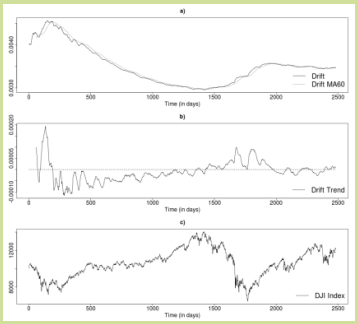

Figure 2 – Proposed framework in a simulation over the Dow Jones index and computed statistics detects concept drift prior to the 2008 crash.

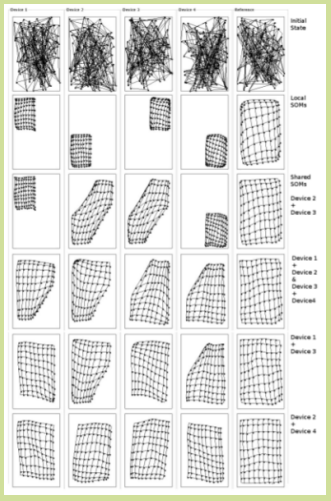

Figure 3 – Different nodes have a SOM model which are trained with separate portions of a distribution. Through a knowledge exchanging mechanism they can incorporate knowledge from the other models into their own, obtaining a model for the whole distribution.

Funding:

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR Portugal

GRANT: SFRH/BD/49723/2009/J513745O3NG0