

Self-Organising Map Techniques for Graph Data Applications to Clustering of XML Documents

Ah Chung Tsoi

Monash e-Research Centre, Monash University
Victoria 3800, Australia

ahchung.tsoi@adm.monash.edu.au

In this talk, neural network techniques based on Kohonen's self-organising map method which can be trained in an unsupervised fashion and applicable to the processing of graph structured inputs are described. Then it is shown how such techniques can be applied to the problems of clustering of XML documents.

Neural networks have been one of the main techniques used widely in data mining. There are a number of popular neural network architectures, e.g. multilayer perceptrons, self organising maps, support vector machines. However, most of these techniques have been applied to problems in which the inputs are vectors. In other words, the inputs to these neural network architectures are expressed in the form of vectors, often in fixed dimensions. In case the inputs are not suitably expressed in the form of vectors, they are made to conform to the fixed dimension vectorial format. For example, it is known that an image may be more conveniently expressed in the form of a graph, for instance, the image of a house can be expressed as a tree, with the source node (level 0) being the house, windows, walls, and doors expressed as leaves (level 1), and details of windows, walls and doors being expressed as leaves (level 2) of those leaves located in level 1, etc. These nodes are described by attributes (features, which may express colour, texture, dimensions) and their relationships with one another are described by links. Such inputs can be made to conform to a vectorial format if we "flatten" the structure and instead represent the information in each node in the form of a vector, and obtain the aggregate vector by concatenating the vectors together. Such techniques have been prevalent in the application of neural network architectures to these problems.

Recently, there have been some attempts in preserving the graph structured data as long as we can before applying the neural network technique. For example, in support vector machines, there have been some work in expressing the graph structured data in the form of string kernels, or spectrum kernels, and then use the "kernel trick" in using the support vector machine machinery to process the data. Alternatively, another way to process the data is to preserve the graph structured data format, and modify the neural network techniques to process graph structured data. In this paper, we will not consider support vector machine further, and we will concentrate only on ways to modify a classic neural network technique, self-organising maps, so that it can accept graph structured inputs.