# Chapter XII
# Outlier Detection Strategy Using the Self-Organizing Map

**Fedja Hadzic**
*DEBII Institute, Curtin University of Technology, Australia*

**Tharam S. Dillon**
*DEBII Institute, Curtin University of Technology, Australia*

**Henry Tan**
*University of Technology Sydney, Australia*

## ABSTRACT

*Real world datasets are often accompanied with various types of anomalous or exceptional entries which are often referred to as outliers. Detecting outliers and distinguishing noise form true exceptions is important for effective data mining. This chapter presents two methods for outlier detection and analysis using the self-organizing map (SOM), where one is more suitable for categorical and the other for continuous data. They are generally based on filtering out the instances which are not captured by or are contradictory to the obtained concept hierarchy for the domain. We demonstrate how the dimension of the output space plays an important role in the kind of patterns that will be detected as outlying. Furthermore, the concept hierarchy itself provides extra criteria for distinguishing noise from true exceptions. The effectiveness of the proposed outlier detection and analysis strategy is demonstrated through the experiments on publicly available real world datasets.*

## INTRODUCTION

The fact that real world datasets are often accompanied with various types of anomalous entries, introduces additional challenges to data mining and knowledge discovery. These anomalies need to be detected and dealt with in an appropriate way depending on whether the detected anomaly is caused by noise or is a true exceptional case. An anomalous entry is often referred to as an outlier. The definitions are quite similar throughout the literature and the general intent is captured by the

definition given by Hawkins (1980): "an outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism." Depending on the aim of application the term outlier detection has been commonly substituted by terms such as: anomaly, exception, novelty, deviation or noise detection. Outliers in a dataset correspond to a very small percentage of the data objects. Most data mining algorithms tend to minimize their influence or discard them altogether as being noisy data. However, eliminating outlying objects can result in loss of some important information especially for applications where the intention is to find exceptional or rare events that occur in a particular set of observations. These rare events are mostly of greater interest than the common events in applications such as: fraud detection, network intrusion detection, security threats, credit risk assessment, pharmaceutical research, terrorist attacks, and some financial and marketing applications. In general outlier detection is used in applications where the analysis of uncommon events is important and can provide extensional knowledge for the domain. On the other hand, if the outlying entries are caused by noise it is still important that they are detected and removed from that dataset in order to disallow interference with the learning mechanism. Hence, outlier detection and analysis has become an important task in data mining and knowledge discovery.

The self-organizing map (SOM) (Kohonen, 1990) is an unsupervised neural network that effectively creates spatially organized "internal representations" of the features and abstractions detected in the input space. It is based on the competition among the cells in the map for the best match against a presented input pattern. Existing similarities in the input space are revealed through the ordered or topology preserving mapping of high dimensional input patterns into a lower-dimensional set of output clusters. When used for classification purposes, SOM is commonly integrated with a type of supervised learning in order to assign appropriate class labels to the clusters. After the supervised learning is complete each cluster will have a rule or pattern associated with it, which determines which data objects are covered by that cluster. Due to its simple structure and learning mechanism SOM has been successfully used in various applications and it has proven to be one of the effective clustering techniques (Kohonen, 1990; Sestito & Dillon, 1994).

SOM has been previously used for outlier detection in (Munoz & Muruzabal, 1998), where the trained map is projected using Sammons mapping to find the initial outliers, and thereafter SOMs quantization errors are used for identifying the remaining outliers. Since SOM learns without supervision, abnormalities can be detected without knowing what to expect. This motivated many SOM applications to the problem of intrusion detection on computer networks (Girardin 1999; Heywood & Heywood, 2002; Lichodzijewski, Zincir-Labib, & Vemuri, 2002; Nuansri, Dillon, & Singh, 1997; Rhodes, Mahaffey, & Cannady, 2000; Zanero & Savaresi, 2004). SOM is also an effective tool for data exploration since the formed abstractions can be easily visualized. The work done in Vesanto, Himberg, Siponen, and Simula (1998), aims to improve SOMs visualization capabilities for novelty detection. Another common approach to outlier detection using SOM is to form the initial clusters from normal data objects and then to use a pre-defined distance measure which will indicate the outliers with respect to the clusters set exhibiting normal behavior (Gonzalez & Dasgupta, 2002; Gonzalez & Dasgupta, 2003; Ypma & Duin, 1997).

We propose a different approach to using SOM for outlier detection and analysis. The motivation behind the proposed methods comes from the results obtained in our previous works where SOM was used for classification (Dillon, Sestito, Witten, & Suing, 1993; Hadzic & Dillon, 2005; Sestito & Dillon, 1994) and for detection of frequent patterns from a transactional database