Expert Systems

A method for outlier detection based on cluster analysis andvisual expert criteria.

Juan A. Lara

Department of Computer Science, Madrid Open University, UDIMA, Engineering School, Madrid, Spain.

David Lizcano

Department of Computer Science, Madrid Open University, UDIMA, Engineering School, Madrid, Spain.

Víctor Rampérez

ETS de Ingenieros Informáticos, Universidad Politécnica de Madrid, Campus de Montegancedo, Madrid, Spain.

Javier Soria

ETS de Ingenieros Informáticos, Universidad Politécnica de Madrid, Campus de Montegancedo, Madrid, Spain.

RESUMEN:

Outlier detection is an important problem occurring in a wide range of areas. Outliersare the outcome of fraudulent behaviour, mechanical faults, human error, or simplynatural deviations. Many data mining applications perform outlier detection, oftenas a preliminary step in order to filter out outliers and build more representativemodels. In this paper, we propose an outlier detection method based on a clusteringprocess.

The aim behind the proposal outlined in this paper is to overcome the spec-ificity of many existing outlier detection techniques that fail to take into account theinherent dispersion of domain objects. The outlier detection method is based on fourcriteria designed to represent how human beings (experts in each domain) visuallyidentify outliers within a set of objects after analysing the clusters. This has an advan-tage over other clustering-based outlier detection techniques that are founded on apurely numerical analysis of clusters.

Our proposal has been evaluated, with satisfac-tory results, on data (particularly time series) from two different domains:stabilometry, a branch of medicine studying balance-related functions in humanbeings and electroencephalography (EEG), a neurological exploration used to diag-nose nervous system disorders. To validate the proposed method, we studied methodoutlier detection and efficiency in terms of runtime. The results of regression analysesconfirm that our proposal is useful for detecting outlier data in different domains,with a false positive rate of less than 2% and a reliability greater than 99%.

PALABRAS CLAVE:

Clustering, Data mining, KDD, Outlier Detection, Visual expert criteria.