Reference Paper Title :

Detection of Atypical Elements by Transforming Task to Supervised Form

Authors :
Piotr Kulczycki, Damian Kruszewski

Group : G4

Group Members:

Aditi Singh	aditi.singh@st.niituniversity.in	U101116FEC156
Hridaya Annuncio	hridaya.annuncio@st.niituniversity.in	U101116FCS046
Shreyas Sanghvi	shreyashreyash.sanghvi@st.niituniversity.in	U101116FCS122
Varuni Agarwal	Varuni.Agarwal@st.niituniversity.in	U101116FCS146

INTRODUCTION

In this paper the atypical elements considered are outliers.These kinds of elements are usually few in number in a dataset. In unsupervised learning they can prove hard to find a pattern of. Considering that the number of atypical elements is very small, to substantiate a certain pattern is challenging(less training data).
Hence, in the procedure defined in the paper, this unsupervised algorithm to identify atypical elements is brought down to a supervised classification.

PROCEDURE

1. A Random Variable X of n dimensions is defined and its density function is defined by f(x). Since a non- parametric method is used, we do not require to assume the distribution type of the data set.
   The following is the density function

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right),$$

   Here,
   m-> the number of random samples considered,
   h-> the smoothing factor,
   n-> Number of dimensions of the Random Variable X
   and
   K()-> The Kernel function

   Hence, m random samples are generated and the f(x) value for each is calculated
   We can create the same function for multidimensional data elements using multiplication.

2. 1% to 10% of the smallest valued elements, from the set of values of f(x) for each random sample, is taken. Let this number be i.

3. These i values are sorted,
   z1, z2, z3 …….. zi

4. The quantile estimator qr is calculated based on these i elements

$$\hat{q}_r = \begin{cases} z_1 & \text{for} \quad mr < 0.5 \\ (0.5 + i - mr)z_i + (0.5 - i + mr)z_{i+1} & \text{for} \quad mr \geq 0.5 \end{cases},$$

where

$$i = [mr + 0.5],$$

5. If ,
   f(x) <= qr -> classified as an Atypical element
   f(x) > qr ->classified as a Typical element
   This classification is done for every element in the dataset.

6. Usually due to fewer numbers of atypical elements compared to typical elements in the population, it is difficult to find a pattern to define them.
   Hence, based on the pattern of already classified atypical elements, more elements of the same distribution are generated so that ,
   The number of atypical elements = number of typical elements
   This is done using Von Neuman's Elimination Concept

7. A classification algorithm is applied to the entire population.

8. Testing done in 2 phases:
   a) Using Random data
   b) Using an already available dataset

DIVISION OF WORK

1. Making the model:
   Hridaya Annuncio
   Shreyas Sanghvi

2. Testing the model :
   Aditi Singh
   Varuni Agarwal