Chapter 8
Enhancing Outlier Detection in Healthcare Data through Mahalanobis Distance Metric Analysis

Santhosh Kumar Rajamani
https://orcid.org/0000-0001-6552-5578
Maharashtra Institute of Medical Education and Research, India

Radha Srinivasan Iyer
https://orcid.org/0000-0001-7387-4401
SEC Center for Indepenedent Living, India

ABSTRACT

Mahalanobis distance is a useful multivariate statistic for determining how far apart two points are from one another. It is a very helpful statistic with excellent uses in multivariate anomaly detection, one-class classification, and classification on severely unbalanced datasets.This compilation delves into the refinement of outlier detection within healthcare data by employing Mahalanobis Distance Metric Analysis as its core methodology. Using Pure Tone Audiometry threshold data as a case study, the research highlights the method's effectiveness in identifying and characterizing outliers. The emphasis is placed on the robustness and applicability of the Mahalanobis Distance Metric, showcasing its potential to enhance outlier detection methodologies across diverse healthcare datasets. This study contributes a methodological advancement that extends beyond the specific application to Pure Tone Audiometry, offering a versatile framework for improved outlier detection in

DOI: 10.4018/979-8-3693-9641-4.ch008

various healthcare domains.

INTRODUCTION

The Mahalanobis distance, introduced by Prasanta Chandra Mahalanobis in the year 1936, serves as a measure of the distance between a point and a distribution. The Mahalanobis distance is a measure of how far a point is from the mean of the distribution, considering the relationships between the variables. A point with a large Mahalanobis distance is an outlier, as it is far from the typical values of the distribution. It finds applications in various fields, including data analysis, process control, and outlier detection (Stockl, et.al.,2013).

BACKGROUND

The Mahalanobis distance is calculated using the covariance matrix of the data set. The covariance matrix is a measure of how much two variables change together. It is used to calculate the correlation between variables and is an important tool in multivariate analysis. The Mahalanobis distance is calculated by taking the difference between the mean of the data set and the data point, and then multiplying this difference by the in-verse of the covariance matrix (Mahalanobis, 1954). The result is a measure of how far the data point is from the mean of the data set, considering the correlation between variables. In practice, you might obtain a covariance matrix from real-world data using methods like sample covariance estimation or other statistical techniques. The choice of the covariance matrix is crucial as it influences the shape and orientation of the generated data distribution.

The Mahalanobis distance can also be used to classify data points. This is done by assigning each data point to the class that has the smallest Mahalanobis distance to the point. This method is known as Mahalanobis distance-based classification. The Mahalanobis distance is a useful tool for identifying outliers in multivariate data. Outliers can have a significant impact on statistical analyses, so it is important to be able to identify and remove them from the data. The Mahalanobis distance can also be used to classify points into different groups or clusters. This is because points that are close to each other in terms of their Mahalanobis distance are likely to belong to the same group or cluster. For example, if two variables are highly correlated, then a point that is far away from the mean in the direction of one variable will also be far away from the mean in the direction of the other variable. The Mahalanobis distance will take this correlation into account and will assign a large distance to the point (McLachlan, 1999).