離散変量と連続変量が混在する場合の
統計的異常検出法

2011 年 8 月

飯田 孝久

主　論　文　要　旨

報告番号	甲 ㊁ 第　　号	氏 名	飯田 孝久

主論文題目：

離散変量と連続変量が混在する場合の統計的異常検出法

（内容の要旨）

本論文では、連続変量と離散変量が混在する異常検出問題において、離散変量の値を与えたとき、連続変量が分散共分散行列が共通の正規分布にしたがうとするロケーションモデルを仮定し、分布の母数が既知の場合と未知の場合について、異常検出法を構成した。誤報率が設定値に一致する、あるいはなるべく近い値になるように棄却限界値を定める方法を与え、誤報率および検出力の性質を明らかにし、手法間の比較を行った。なお、確率が小さい離散変量の水準での異常を確実に検出することは本研究での一つの目標である。

分布の母数が既知の場合は、離散変量の値を与えて連続変量のみに基づく検定を用いる条件付法(C 法)、離散変量を連続変量と併せて求めたマハラノビス平方距離を用いるマハラノビス距離法(M 法)と、全変量を用いた尤度比検定に基づく尤度比法(L 法)を構成した。どの方法においても、異常検出統計量は連続変量のみによるマハラノビス平方距離と補正項の和として表現されることから、誤報率が正確に設定値と一致する異常検出法を構成した。誤報率ならびに検出力に関する性質を明らかにした。補正項の性質から、M 法と L 法では正常状態で確率が小さい離散変量の水準ほど異常と判定しやすくなることがわかった。2 値変量の場合についての数値計算を基に手法の比較を行った結果、L 法とM法は確率が小さい水準で高い検出力を与える方法であることが確認できた。母数の状況により最適な方法は変化するが、総合的に判断してL法が優れていると結論づけられた。

分布の母数が未知の場合は、母数が既知の場合の 3 手法の異常検出統計量に初期データによる分布の母数の推定量を代入する推定方式と、初期データに判定標本を併せた全データに対する尤度比検定に基づく検定法(T 法)を構成した。棄却限界値は、初期データについて期待値をとった期待誤報率が設定値に近くなるよう、連続変量のみに基づくマハラノビス平方距離の分布として F 分布を用いて決定し、４つの手法について、誤報率ならびに検出力の基本的性質を明らかにした。期待誤報率が設定値に一致しないため、検出力の期待誤報率に対するオッズ比を用いて手法の比較を行った。C 法は離散変量を積極的に異常検出に用いていないことから、また、M 法は期待誤報率が設定値から大きく乖離する可能性があることから、詳細な比較はL 法と T 法について行った。既知の場合と同様に、母数の状況によりその優劣は変化するが、総合的にみて期待誤報率が安定し広い範囲でオッズ比の高いT法が優れていると結論づけられた。

SUMMARY OF Ph.D. DISSERTATION

School	Student Identification Number	SURNAME, First name
		IIDA Takahisa

Title

Statistical Methods for Detecting Abnormal Items When There Exist Both Categorical and Continuous Variables

Abstract

The problem of detecting abnormal items is discussed as a hypothesis testing problem for the case when both continuous and categorical variables are observed. Assuming the location model where continuous variables are multivariate normally distributed with common covariance matrix when categorical variables are observed, detection methods are derived with the false alarm probability as near the nominal value as possible for both cases when parameter values are known and unknown.

For the case when all parameter values of the distribution for the group of normal items are known, three detection methods are constructed. Conditional (C) method is based on the conditional distribution of continuous variables when categorical variables are observed. Mahalanobis distance (M) method uses the squared Mahalanobis distance by replacing the categorical variables by their dummy variables. Likelihood ratio (L) method is based on the likelihood ratio test. For these three methods, it is shown that the test statistics are expressed as the sums of Mahalanobis distance based on continuous variables and the correction terms which are determined by the probabilities of categorical variables. So, the distribution of these test statistics for normal items is a mixture of shifted χ^2 distributions. Based on this result, critical values are determined so that the nominal value of the false alarm probability is attained.

For the case when the parameter values are unknown, two types of detection methods are considered. One is estimative method, where estimates are substituted for the unknown parameters in the test statistics of C, M, and L methods. The other is the testing method (T method), which is derived as the likelihood ratio test using all data including initial data for normal items and testing sample.

Some basic properties for these methods are shown concerning their false alarm probabilities and conditional powers given the frequencies of categorical variables in the initial data and/or the values of categorical variables of testing sample. For comparing these methods, their expected error rates for normal items and the detecting powers are numerically evaluated when one dichotomous variable is included. For the case when parameter values are known, L method has higher power for a wide range of parameter values compared to other methods. For the case when parameter values are unknown, T method performs better in the sense that it has stable false alarm probability and higher power for a wide range of parameter values.