Natto: Visualizing Mixed Type Dataset for Exploratory Data Analysis

Ryota Suzuki, Tatsuhiro Nagai and Tomoya Taniguchi

Ef-prime, Inc.

October 14, 2006

Abstract

Natto is a data visualization/analysis tool for mixed type dataset. It visualizes the associations between variables, which can be both numerical and categorical variables. Even mixed-type variables can be handled — for example, a variable `age` may consist of integer values in interval $[12, 80]$, categories "`under 12`", "`over 80`" and missing values labeled as "`NA`". While such settings are quite common in practical situations such as commercial databases, the analysis of such data has often been problematic.

Natto provides an easy, quick but efficient way to analyze such datasets. The associations between variables are visualized as directed graph. Users can interactively analyze the data by analytical tools such as cross tabulation tables and association rule mining. The strengths of associations are measured based on information theoretic approach. For numerical variables, our approach can detect not only linear relationships, but also a class of nonlinear relationships. In addition any monotonic transformation does not affect the result, so users do not need to find proper transformation to detect relationships, such as log-transformation. Furthermore, no specific probabilistic models are needed before analysis. Users even do not need to know which variables are numeric, categorical or mixed type. Natto visualizes a datafile "as it is" — it serves as a powerful tool to get the whole picture of a dataset before statistical analysis, and also an interactive data analysis tool for both expert/non-expert data analysts.

The software can be freely downloaded from our website[1]. A simple interface from statistical software R is also available.

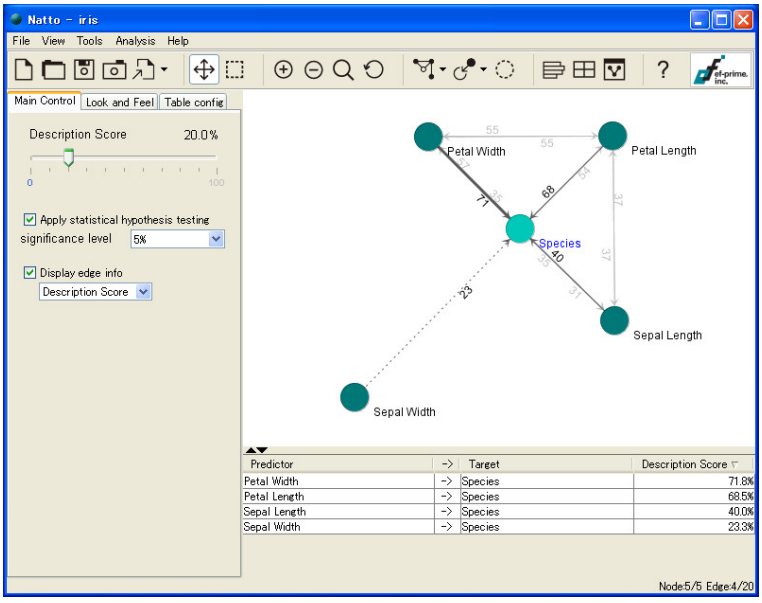

Figure 1: Natto GUI

[1] `http://www.ef-prime.com/`. Currently there is only Japanese site. English version will be available soon.

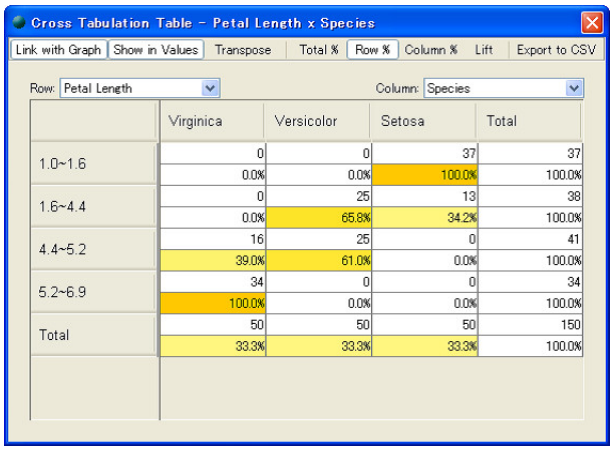

Figure 2: Cross Tabulation Table

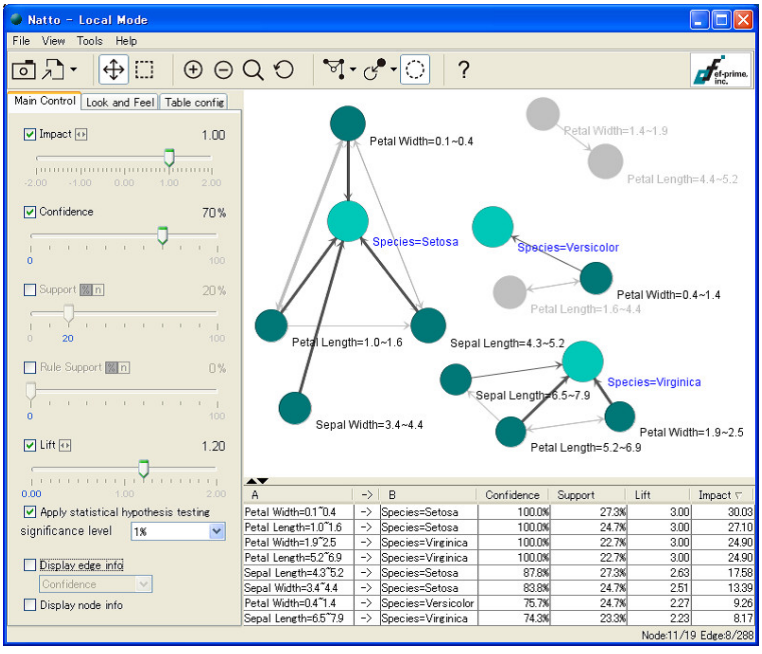

Figure 3: Association Rule Mining