Statistical Independence as Linear Dependence in a Contingency Table

Shusaku Tsumoto

Department of Medical Informatics,
Shimane University, School of Medicine
89-1 Enya-cho Izumo 693-8501 Japan
tsumoto@computer.org

Summary. A contingency table summarizes the conditional frequencies of two attributes and shows how these two attributes are dependent on each other. Thus, this table is a fundamental tool for pattern discovery with conditional probabilities, such as rule discovery. In this paper, a contingency table is interpreted from the viewpoint of granular computing. The first important observation is that a contingency table compares two attributes with respect to the number of equivalence classes. The second important observation is that matrix algebra is a key point of analysis of this table. Especially, the degree of independence, rank plays a very important role in extracting a probabilistic model from a given contingency table.

Key words: Statistical Independence, Linear Independence, Contingency Table, Matrix Theory

1 Introduction

Independence (dependence) is a very important concept in data mining, especially for feature selection. In rough sets[4], if two attribute-value pairs, say $[c = 0]$ and $[d = 0]$ are independent, their supporting sets, denoted by C and D do not have a overlapping region ($C \cap D = \phi$), which means that one attribute independent to a given target concept may not appear in the classification rule for the concept. This idea is also frequently used in other rule discovery methods: let us consider deterministic rules, described as *if-then* rules, which can be viewed as classic propositions ($C \rightarrow D$). From the set-theoretical point of view, a set of examples supporting the conditional part of a deterministic rule, denoted by C, is a subset of a set whose examples belong to the consequence part, denoted by D. That is, the relation $C \subseteq D$ holds and deterministic rules are supported only by positive examples in a dataset[8].

When such a subset relation is not satisfied, indeterministic rules can be defined as if-then rules with probabilistic information[7]. From the set-theoretical

point of view, C is not a subset, but closely overlapped with D. That is, the relations $C \cap D \neq \phi$ and $|C \cap D|/|C| \geq \delta$ will hold in this case.[1] Thus, probabilistic rules are supported by a large number of positive examples and a small number of negative examples.

On the other hand, in a probabilistic context, independence of two attributes means that one attribute (a_1) will not influence the occurrence of the other attribute (a_2), which is formulated as $p(a_2|a_1) = p(a_2)$.

Although independence is a very important concept, it has not been fully and formally investigated as a relation between two attributes.

In this paper, a contingency table of categorical attributes is focused on from the viewpoint of granular computing. The first important observation is that a contingency table compares two attributes with respect to information granularity. Since the number of values of a given categorical attribute corresponds to the number of equivalence classes, a given contingency table compares the characteristics of information granules: $n \times n$ table compares two attributes with the same granularity, while a $m \times n (m \geq n)$ table can be viewed as comparison of two partitions, which have m and n equivalence classes.

The second important observation is that matrix algebra is a key point of analysis of this table. A contingency table can be viewed as a matrix and several operations and ideas of matrix theory are introduced into the analysis of the contingency table. Especially, the degree of independence, rank plays a very important role in extracting a probabilistic model from a given contingency table. When the rank of the given table is equal to 1.0, one attribute in the table are statistically independent of the other attributes. When the rank is equal to n, which is the number of values of at least one attribute, then two attributes are dependent. Otherwise, the row or columns of contingency table are partially independent, which gives very interesting statistical models of these two attributes.

The paper is organized as follows: Section 2 discusses the characteristics of contingency tables. Section 3 shows the definitions of statistical measures used for contingency tables and their assumptions. Section 4 discusses the rank of the corresponding matrix of a contingency table when the given table is two way. Section 5 extends the above idea into a multi-way contingency table. Section 6 presents an approach to statistical evaluation of a rough set model. Finally, Section 7 concludes this paper.

This paper is a preliminary study on the concept of independence in statistical analysis, and the following discussions are very intuitive. Also, for simplicity of discussion, it is assumed that a contingency table gives a square matrix ($n \times$n).

[1] The threshold δ is the degree of the closeness of overlapping sets, which will be given by domain experts. For more information, please refer to Section 3.