Information Sciences 346–347 (2016) 73–89

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Pair-wise association measures for categorical and mixed data

Ayman Taha[a,*], Ali S. Hadi[b,c]

[a] Faculty of Computers and Information, Cairo University, 5 Dr. Ahmed Zewail St., Orman, Giza 12613, Egypt
[b] American University in Cairo, P.O. Box 74, AUC Avenue, New Cairo 11835, Egypt
[c] Cornell University, Ithaca, NY, USA

ARTICLE INFO

Article history:
Received 8 February 2014
Revised 5 November 2015
Accepted 4 January 2016
Available online 21 January 2016

Keywords:
Categorical data
Variables correlation
Attributes association
Maximal information coefficient
Correspondence analysis
Data analytics

ABSTRACT

We introduce two measures for the strength of the association between two categorical variables. The measures, denoted by η_1 and η_2, take values in the interval [0, 1]. A value of zero means there is no association between the two categorical variables, while a value of 1 means there is a perfect association (e.g., when we associate a variable with itself, we obtain $\eta = 1$). The measures are symmetric with respect to the order of variables, invariant with respect to permutations of the categories of the variables, and scalable for large number of observations. In addition, extensions of the proposed measures are presented for measuring the strength of association between pair of mixed variables, one quantitative and the other is categorical. The performance of the proposed measures compared to other association measures is investigated using simulated as well as real data.

1. Introduction

Real datasets often consist of mixed variables, that is, some variables are quantitative and others are qualitative or categorical. The statistical analysis of quantitative data has a venerable history and by comparison categorical data received less attention than quantitative data. In this paper we focus our attention on measuring association in categorical and mixed variables.

Categorical data are common in many different domains (e.g., biomedical, educational, and social sciences) [2]. Measuring association among variables is useful for clustering [29], outlier detection [9], association rule mining (e.g., [30] and [4]), and feature selection [14]. Ignoring association among variables can lead to wrong conclusions [12]. Latent Gaussian models (LGMs) are employed in modeling categorical data using Gaussian latent variables to discover and analyze hidden relationships within categorical data (e.g., [26] and [35]). A framework for exploring categorical data is presented in [10]. Measuring the variance within a categorical attribute is studied in [16]. Similarity and distance measures among categorical observations are studied, for example, in [8] and [28].

Two types of categorical variables are distinguished: ordinal and nominal. Examples of the first type include letter grade in a course (e.g., *A*, *B*, *C*, *D*, *F*), income level (e.g., high, medium, low), and a Likert scale variable (e.g., 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree). Examples of the second type include gender, nationality, and blood type. Measuring association of ordinal variables is relatively easier than measuring association of nominal data because the categories in nominal variables have no natural ordering. A comprehensive coverage of ordinal measures of association is presented in [1].

* Corresponding author. Tel.: +20 11 1830 3990; fax: +20 2 3335 0109.
E-mail addresses: a.taha@fci-cu.edu.eg, iyman.taha@gmail.com (A. Taha), ahadi@aucegypt.edu, ali-hadi@cornell.edu (A.S. Hadi).

http://dx.doi.org/10.1016/j.ins.2016.01.022
0020-0255/

Table 1
Measures of association for the three contingency tables in (1).

Table	V^2	λ		τ		U		SCA	SCOR	MIC	η	
		$\lambda_{Y.X}$	$\lambda_{X.Y}$	$\tau_{Y.X}$	$\tau_{X.Y}$	$U_{Y.X}$	$U_{X.Y}$				η_1	η_2
T_1	0.11	0.00	0.17	0.12	0.12	0.10	0.09	0.34	0.34	0.09	0.51	0.34
T_2	0.32	0.40	0.20	0.23	0.30	0.32	0.40	0.79	0.08	0.20	0.88	0.79
T_3	0.32	0.40	0.20	0.23	0.30	0.32	0.40	0.79	0.73	0.44	0.88	0.79

Agresti [2] writes "When variables in a two-way table are nominal, notions such as positive/negative association and monotonicity are no longer meaningful. It is then more difficult to describe association by a single number, and summary measures are less useful than for ordinal or interval-scale variables". Because the categories can appear in any order, the results of the analysis should be invariant to the order in which the categories are given. They should also be symmetric or invariant to the order of the variables, that is, a measure of association between X and Y is equal to the measure of association between Y and X.

Existing association measures for categorical data are presented in the Appendix. These include the following:

1. The chi-square and associated V^2 statistics
2. Goodman and Kruskal's $\lambda_{Y\,.\,X}$ and $\lambda_{X\,.\,Y}$
3. The concentration coefficients $\tau_{Y\,.\,X}$ and $\tau_{X\,.\,Y}$
4. The uncertainty coefficients $U_{Y\,.\,X}$ and $U_{X\,.\,Y}$
5. The simple correspondence analysis SCA
6. The symbolic correlation SCOR
7. The maximal information coefficient MIC.

These measures have the following limitations: First, some of the current association measures are not symmetric with respect to the order of variables (e.g., the uncertainty coefficient τ and concentration coefficient U). For example, the concentration coefficient $\tau_{X\,.\,Y} \neq \tau_{Y\,.\,X}$ and the Uncertainty coefficient $U_{X\,.\,Y} \neq U_{Y\,.\,X}$. The proposed measures are symmetric with respect to the order of variables, that is, $\eta_{XY} = \eta_{YX}$. We should note here, however, that asymmetric measures are useful if we wish to study causal relationship, that is, when one variable causes the other.

Second, some of the existing association measures (e.g., the MIC and SCOR) are not invariant with respect to the order of the categories in a categorical variable. Our proposed measures are invariant.

Third, current association measures produce very different measures when applied to the same data. Consider, for example, the following three contingency tables for two variables Y (rows) and X (columns):

$$T_1 = \begin{bmatrix} 53 & 42 \\ 10 & 40 \end{bmatrix}, \quad T_2 = \begin{bmatrix} 2 & 0 & 0 & 2 \\ 2 & 1 & 2 & 1 \\ 0 & 2 & 4 & 0 \end{bmatrix}, \quad \text{and} \quad T_3 = \begin{bmatrix} 2 & 0 & 0 & 2 \\ 1 & 1 & 2 & 2 \\ 0 & 2 & 4 & 0 \end{bmatrix}, \tag{1}$$

Note that T_3 is obtained from T_2 simply by exchanging columns 1 and 4 (that is, changing the order of which the categories of X are given). The various measures of association including the two proposed ones, η_1 and η_2 are given in Table 1.

From Table 1, we see that the current measures of association show great variations as they range from 0 to 0.34 for T_1, from 0.08 to 0.79, for T_2, and from 0.20 to 0.79 for T_3. They differ even within the same measure. For example, for T_1 $\lambda_{Y.X} = 0$ (even though the data in T_1 show a clear relationship between the two variables) and $\lambda_{X.Y} = 0.17$. We also see that the MIC and the symbolic correlation, SCOR, are not invariant with respect to the reordering of the categories. For example, for T_3, which is obtained simply by interchange the first and fourth columns of T_2, we see that the first five measures are invariant but the MIC and SCOR are not. This is a major shortcoming as it casts serious doubt about the suitability and appropriateness of MIC and SCOR as measures of association between two categorical variables.

Fourth, modern applications often have huge data sets in terms of the number of objects (observations) and the number of variables. Consequently, time complexity is a significant issue in modern applications. Some of the existing association measures (e.g., the MIC and SCOR) take very long computation time, which substantially increases with the sample size (number of observations or objects). Our proposed measures are scalable with respect to the number of observations.

Fifth, most of the current measures of association are not applicable for mixed data, that is, when one variable is categorical and the other is quantitative. This paper proposes two measures for assessing the strength of association between two categorical variables and it also extends these two measures to assessing the association between mixed variables. All proposed association measures are symmetric with respect to the order variables, invariant with respect to the order in which the categories are given and scalable for large number of objects. However, all methods including the proposed ones, are not scalable with respect to the number of categories (the dimension of the corresponding contingency table).

We should also mention here that when it comes to practical implementation of all measures of association, one may encounter two problems: (a) the presence of missing values in the data and (b) some variables may need transformation. In the presence of missing values, one can use either case-wise deletion or multiple imputations of the missing values before one computes any association measure. The reader is referred, for example, to [27] for alternative methods for the imputation of missing values.