Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 4, pp. 957~963

A Measure of Agreement for Multivariate Interval Observations by Different Sets of Raters

Yonghwan Um[1)]

Abstract

A new agreement measure for multivariate interval data by different sets of raters is proposed. The proposed approach builds on Um's multivariate extension of Cohen's kappa. The proposed measure is compared with corresponding earlier measures based on Berry and Mielke's approach and Janson and Olsson approach, respectively. Application of the proposed measure is exemplified using hypothetical data set.

Keywords : agreement measure, multivariate interval data, sets of raters

I. Introduction

Many researchers have proposed generalizations of Cohen's kappa(1960) agreement measure to high level (ordinal or interval) data among a set of two or more raters for a sample of objects. Berry and Mielke(1988) proposed an agreement measure, R, being applicable at ordinal or interval scales. They extended Cohen's kappa to several raters and one nominal variable, and also to several raters and multivariate interval or ordinal data. Their agreement measure is expressed as $R = 1 - \delta / \mu_\delta$ where δ is the observed disagreement and μ_δ is the expected disagreement, where the disagreements are measured using Euclidean distance. Janson and Olsson(2001) proposed an agreement measure, ι, for multivariate interval or nominal data by modifying Berry and Mielke's(1988) approach. Their modification is to utilize the squared Euclidean distance as disagreement measure among raters rather than Euclidean distance as for Berry

1) Associate Professor, Division of Electronic Commerce, Sungkyul University, Anyang, 430-742, Korea
Email : uyh@sungkyul.edu

and Mielke(1988). They defined it as $\iota = 1 - d_0/d_e$ where d_0 and d_e are the observed and expected disagreements, respectively. The observed disagreements(δ and d_0) represent the average distance between any two raters' observations of the same object, and the expected disagreements(μ_δ and d_e) represent the average distance between one rater's observation of a particular object and any other rater's observation of any object. Recently, Um(2004) proposed a new agreement measure, ϕ, among a set of several observers for multivariate interval data by modifying Berry and Mielke's approach. Um(2004) modified Berry and Mielke's approach by using the volume of c-dimensional simplex composed of data points as the disagreement measure. Um's(2004) agreement measure for c-variate interval data denoted by ϕ is expressed as

$$\phi = 1 - \frac{v_o}{v_e} \tag{1}$$

where v_0 is the observed disagreement representing the average, over objects and combinations of c+1 raters, of the simplex volumes among raters' observations of the same objects and v_e is the expected disagreement representing the average, over objects and combinations of c+1 raters, of the simplex volumes among raters' observations of any object. When c=1 (univariate case), the volume of simplex is the Euclidean distance between two data points and ϕ equals Berry and Mielke(1988)'s agreement measure, R.

My concern here is the case where objects are rated by different sets of raters(not necessarily equal in number). It is often of interest to evaluate the agreement measure among different sets of raters for a sample of objects. For example, if written essays are scored by different groups of raters (e.g. professors, graduate students, etc.) independently, it may be interesting to know the overall agreement measure among those groups of raters. In this article, I propose an overall multivariate agreement measure among a total of n observations of t objects, where the sets of raters who have observed the different objects are not the same and may vary in number. The proposed agreement measure is based on Um's(2004) approach and is obtained by applying an expression for expected disagreement suitable for the case with different raters and by incorporating weighting for number of raters for observed disagreement. The use of the proposed measure is exemplified with hypothetical data set and the proposed measure is compared with the corresponding measures that build on Berry and Mielke's(1988) approach and Janson and Olsson's(2001) approach, respectively.

2. Agreement Measure by Different Sets of Raters Based on Um's Approach

When the objects are rated by different sets of raters, the amount of disagreement that may be expected by chance differs from the case when one set of raters perform all ratings(Hubert 1977). When there are a total of n observations of t objects, the amount of disagreement expected by chance is the average disagreement over all possible draws of c+1 observations with replacement from n observations, disregarding which object it referred to or which rater makes the observation. Thus the expression for expected disagreement in Um's agreement measure, ϕ, is the following:

$$v_e = (n^{c+1})^{-1} \sum_{i_1=1}^{n} \cdots \sum_{i_{c+1}=1}^{n} \triangle(i_1, i_2, \cdots, i_{c+1}) \tag{2}$$

where $\triangle(i_1, i_2, \cdots, i_{c+1})$ with c-variate interval data is the volume of simplex with vertices $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_{c+1}}$ with the following value of

$$\triangle(i_1, i_2, \cdots, i_{c+1}) = \frac{1}{c!} \operatorname{abs}\left(\begin{vmatrix} 1 & 1 & \cdots & 1 \\ x_{i_1 1} & x_{i_2 1} & \cdots & x_{i_{c+1} 1} \\ x_{i_1 2} & x_{i_2 2} & \cdots & x_{i_{c+1} 2} \\ \vdots & \vdots & \cdots & \vdots \\ x_{i_1 c} & x_{i_2 c} & \cdots & x_{i_{c+1} c} \end{vmatrix} \right).$$

An expression for observed disagreement when different sets of raters rate objects must incorporate weighting for varying numbers of raters per object. Let g_s denote the number of observations (at least c+1, each made by a different rater) for the s-th of the t objects. Then observed disagreement, weighted for the number of raters per object, can be expressed as the following:

$$v_o = (n-t)^{-1} \sum_{s_1 = s_2 = \cdots = s_{c+1}} \triangle(i_1, i_2, \cdots . i_{c+1}) / g_s \tag{3}$$

where s_i denotes the object that is the origin of the i-th observation, so that $\sum_{s_1 = s_2 = \cdots = s_{c+1}}$ is the sum over all combinations of c+1 observations of same objects. Thus the amount of observed disagreement is the average amount of disagreement over all possible draws of c+1 different raters' observations of one same object, weighted inversely proportional to the number of raters for a specific

object.

In order to compare ϕ with corresponding measures that are based on Berry and Mielke's(1988) approach and Janson and Olsson's(2001) approach, respectively, I used the following expressions for the observed disagreement and the expected disagreement both in R and $\imath$.

$$\text{expected disagreement} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij} \tag{4}$$

and

$$\text{observed disagreement} = \frac{1}{n-t}\sum_{s}\sum_{i=s_j} D_{ij}/g_s \tag{5}$$

where D_{ij} is the Euclidean distance $\left(=\sqrt{\sum_{k=1}^{c}(x_{ik}-x_{jk})^2}\right)$ for R and the squared Euclidean distance $\left(=\sum_{k=1}^{c}(x_{ik}-x_{jk})^2\right)$ for $\imath$.

3. Example

Table 1 shows a hypothetical bivariate interval data example. Four different sets of raters observed height and weight of four men on the basis of photographs. Rows in Table 1 represent different raters' observations. Based on the data in Table 1, observed disagreement, v_0 , is 36.45 using equation 3 (the sum of volumes of simplexes, each divided by the number of raters for that object, is 401.0, which is divided by the total number of observations(n=15) minus the total number of objects(t =4)). The expected disagreement, v_e, is 136.75 using equation 2(the sum of volumes of simplexes of all possible combinations among observations is 461,519.8, which is divided by their number($n^3 = 3375$)). Inserting the values of v_0 and v_e into equation of $\phi = 1 - v_o/v_e$ yields an agreement measure, ϕ, of 0.733. Note that the two variables (weight and height) are incomparable metrics so that standardization of values is appropriate before calculation of agreement. But since ϕ has an affine invariance property, ϕ remains the same with respect to rotation, reflection and scale transformation of the data.

For the same data set, Janson and Olsson's agreement measure(R) and Berry and Mielke's agreement measure($\imath$) are calculated as 0.914 and 0.675, respectively (using equations 4 and 5). Janson and Olsson's agreement measure, $\imath$, is much bigger than ϕ whereas Berry and Mielke's agreement measure, R, is similar to ϕ.

Table 1. Different Raters' Observations of Weight and Height

subject	weight	height
1	53	171
1	49	169
1	51	169
2	90	194
2	87	190
2	86	185
2	87	189
2	83	185
3	88	196
3	78	192
3	82	190
4	85	174
4	77	173
4	75	178
4	83	170

4. Comparison among ϕ, R and l

The hypothetical data in Table 1 is used for the comparison of agreement measures. The comparison is made by varying the first raters' observations in each one of four sets of raters(while fixing other observations). The variation is made by adding small increments(d1 and d2, respectively) to the weight and height of first raters of four different sets so that the disagreement increases among the raters in each set. The values of (d1, d2) used as increments are (1,1), (1,2), (2,2), (2,3) and (3,3). Figure 1 shows agreement measures (ϕ, R and l) for different values of d(=d1+d2). All agreement measures decrease as the amount of increment gets large. Janson and Olsson's measure, l, is very big and uniformly bigger than ϕ and R at all values of d. Even when there is high disagreement among raters (e.g. d=6), Janson and Olsson's measure is still a big value of showing high agreement among raters (The smallest l is 0.85 when d=6). Such a high magnitude of measure of l, according to Landis and Koch(1997), is interpreted as the raters' observations are in 'almost perfect' agreement. Thus it appears that l inflates the agreement measure and is not enough to detect the high disagreement among raters. But ϕ and R show similar behavior in that they reflect the extent of disagreement properly relative to l. Such a similarity between

two measures seem to result from that D_{ij} in R and $\triangle(i_1, i_2, \cdots, i_{c+1})$ in ϕ belong to same metric space whereas D_{ij} in l belongs to non-metric space. In fact $\triangle(i_1, i_2, \cdots, i_{c+1})$ reduces to D_{ij} in R when c=1. Figure 2 shows agreement measures when increments are added to the first m(=0,1,2 and 3) raters' observations in each one of four sets of raters. As in Figure 1, ϕ and R show similar behavior as m increases and l appears to be inflated.

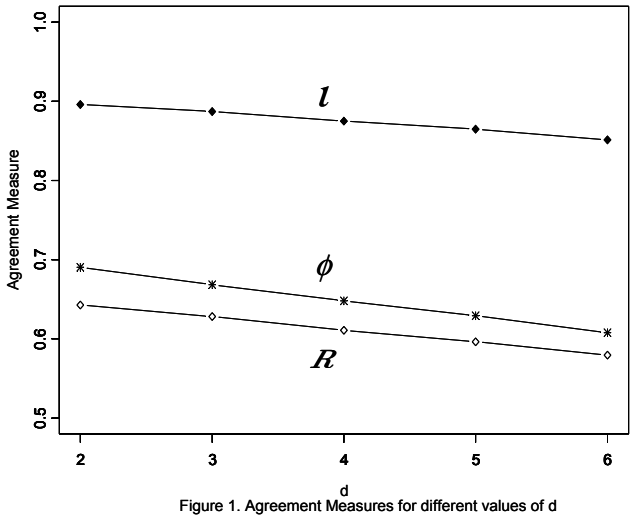

Figure 1. Agreement Measures for different values of d

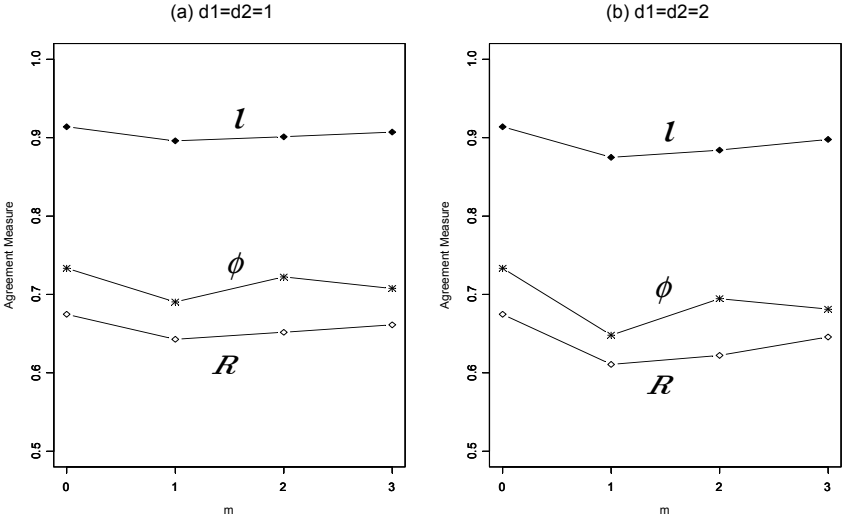

Figure 2. Agreement Measures for different values of m

5. Conclusion

A new agreement measure of multivariate interval observations by different sets of raters is proposed. It builds on Um's(2004) approach where the volume of simplex defined by data points is used as disagreement measure. The comparison study using hypothetical data set shows that the proposed measure, ϕ, performs similarly as R does and better than ι does. In other words, ϕ detects the disagreement among observations and reflect it in ϕ pretty well.

The future work will include the followings: (1) the study of ϕ for multivariate nominal or ordinal data which are not covered in this paper (2) the study of significance test for ϕ (3) the study of the variability of ϕ through bootstrapping.

References

1. Berry, K. J., and Mielke, P. W. Jr. (1988). A Generalization of Cohen's Kappa Agreement Measure to Interval Measurement and Multiple Raters. *Educational and Psychological Measurement, 48*, 921-933.
2. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement*, *20*, 37-46.
3. Janson, H., and Olsson, U. (2001). A Measure of Agreement for Interval or Nominal Multivariate Observations, *Educational and Psychological Measurement*, *61*, 2, 277-289.
4. Hubert, L.(1977). Kappa revisited. *Psychological Bulletin, 84,* 289-297.
5. Landis, J. R., and Koch, G. G. (1997). The Measurement of Observer Agreement for Categorical Data, *Biometrics, 33,* 159-174.
6. Um, Y. (2004). A New Agreement Measure for Interval Multivariate Observations, *Journal of Korean Data & Information Science Society, 15*, 1, 263-271.

[received date : Jul. 2004, accepted date : Oct. 2004]