A fully sensitive correlation measure for data mining

R. J. G. B. Campello & E. R. Hruschka
Department of Computer Sciences, University of São Paulo at São Carlos, SCC/ICMC/USP, C.P. 668, São Carlos, SP, 13560-970, Brazil

Abstract

This paper introduces a novel sequence correlation measure that is fully sensitive to both the ranks and magnitudes of the sequences under evaluation. This measure can be more appropriate than the existing ones in those application scenarios in which such a full sensitivity is desired. The applicability of the new measure in data mining tasks is motivated.
Keywords: correlation indexes, clustering analysis.

1 Introduction

A problem that appears in different contexts of data analysis is that of comparing two sequences $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_n\}$ for which there is a total order relation ($\leq$) on their elements. This problem can be addressed by means of correlation indexes, such as the well-known Pearson correlation coefficient [1, 2]. Aside from the huge applicability of such indexes in statistics [3, 4], there are also different possible scenarios for their application to data mining tasks. In this context, one may mention, for instance, the use of sequence correlation indexes for feature selection as a pre-processing step for data clustering or classification [5]. Another scenario for the application of correlation indexes to data clustering or classification is the measurement of similarities in bioinformatics data sets [6]. For example, sequences A and B can refer to the responses of a given pair of genes along a set of experiments (e.g. microarray) [7]. Since the trend of such responses plays a fundamental role to describe the function and behavior of the corresponding genes, correlation indexes have been widely used as measures of similarity when dealing with this sort of data.

WIT Transactions on Information and Communication Technologies, Vol 40, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA080041

A more traditional scenario for the application of correlation indexes in data analysis regards the validation of a clustering structure by comparing it to an expected structure [8, 9]. A typical example involves the assessment of hierarchical clustering results. In this scenario, sequence A refers to the elements of the distance matrix obtained by hierarchically clustering a given data set (cophenetic matrix). Sequence B, in its turn, refers either to the elements of the distance matrix of a baseline hierarchical clustering or to the elements of the original matrix of distances between the data objects. The larger the correlation between the data according to these (possibly ordinal) different distance matrices, the better the matching between the obtained clustering structure/hierarchy and the referential standard adopted.

The correlation indexes that are most used both in practice and in the literature are possibly the well-known Pearson Product-Moment correlation coefficient [1, 2], the Spearman's coefficient [3, 4], and the Goodman-Kruskal/Kendall's indexes [4, 10, 11]. However, these measures are either insensitive to the ranks (Pearson) or insensitive to the magnitudes (Spearman/Goodman-Kruskal/Kendall) of sequences A and B under evaluation. When both sequences are on relevant quantitative scales, it can be more appropriate to measure their correlation taking into account both the relative orders (ranks) and the relative rates of increase or decrease (magnitudes) between every pair of elements. The present paper introduces a new correlation measure that is endowed with this property. The new measure is a generalized (weighted) version of the Goodman-Kruskal index. A slightly different version is also proposed that is a weighted version of the Kendall's index.

The remaining of the paper is as follows. The Goodman-Kruskal and Kendall's indexes are briefly reviewed in Section 2. Next, the Weighted Goodman-Kruskal and Kendall's indexes are introduced in Section 3. Then, a discussion is provided in Section 4 and the final remarks are addressed in Section 5 together with some perspectives for future research.

2 Review of the Goodman-Kruskal and Kendall's indexes

The Goodman-Kruskal index measures rank correlation between two sequences $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_n\}$ in terms of the numbers of *concordant* and *discordant* pairs in A and B [10, 11]. Pairs (a_i, a_j) and (b_i, b_j) are concordant if either $a_i < a_j$ and $b_i < b_j$ or $a_i > a_j$ and $b_i > b_j$. Conversely, they are discordant if either $a_i < a_j$ and $b_i > b_j$ or $a_i > a_j$ and $b_i < b_j$. The remaining cases are deemed to be neither concordant nor discordant (neutral) (Simultaneous ties ($a_i = a_j$ and $b_i = b_j$) are also considered as neutrals from the original probabilistic perspective of the index as a measure of association for cross-classifications. In clustering or classification tasks, however, it makes more sense to count a simultaneous tie as a concordance.). The index is then defined as [8, 9]:

$$\gamma = \frac{S_+ - S_-}{S_+ + S_-} \tag{1}$$

where S_+ and S_- are the numbers of concordant and discordant pairs in A and B, respectively. Clearly, $\gamma \in [-1, 1]$.

In principle, the running time complexity of this index may seem to be necessarily proportional to the number of non-ordered pairs in sequences A and B, that is, $O(n^2)$. However, it is possible to demonstrate that the computations can actually be done in $O(n \log n)$ time by using a smarter (yet more complicate) algorithm based on a divide-and-conquer strategy.

A variant of the Goodman-Kruskal index in (1), so-called Kendall's index, is given by [4, 8]:

$$\tau = \frac{S_+ - S_-}{n(n-1)/2} \tag{2}$$

Note that the only difference is that this variant takes all the $n(n-1)/2$ non-ordered pairs into account in the denominator, no matter whether they are concordants, discordants, or neutrals. Doing so, it differs from the Goodman-Kruskal index in that it punishes neutrals, for they are included into the denominator. Indeed, the Kendall's index cannot reach its maximum ($\tau = 1$) whenever there are neutrals.

The idea behind the Goodman-Kruskal and Kendall's indexes, though very appealing, hide a possible shortcoming of these indexes, which is their insensitivity to the element *values* of sequences A and B. Indeed, only the ranks of these elements are taken into account in equations (1) and (2). This approach may not be adequate in that it gives the same (unitary) importance to any rank permutation in one sequence with respect to the other, no matter the values of the corresponding elements. In other words, any couple of concordant or discordant pairs in A and B counts the same amount in (1) and (2). When comparing two distance matrices A and B of a given data set, for instance, this makes the indexes unable to differentiate between a big and a small mistake made in the relative distances in A with respect to B. On the other hand, conventional distance measures like the squared norm can capture only the *absolute* differences between every element of A and the corresponding element of B. However, the structure of the data is represented by the *relative* differences between the elements of a given distance matrix, both in terms of their relative orders (ranks) and values (magnitudes). For this reason, a generalized (weighted) version of the Goodman-Kruskal and Kendall's indexes is proposed in the next section.

3 Weighted Goodman-Kruskal and Kendall

This section elaborates on a new index that brings together both the sensitivity of the Goodman-Kruskal index to the rankings of sequences A and B and the sensitivity of the Pearson Product-Moment index to the values of these sequences (when they are both on quantitative scales). It can be derived from a generalization of an alternative formulation for the Goodman-Kruskal index, as discussed in the sequel.

3.1 Goodman-Kruskal and Kendall's indexes revisited

The Goodman-Kruskal index in (1) can be rewritten in equivalent form as:

$$\gamma = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |w_{ij}|} \tag{3}$$

where $|\cdot|$ stands for the absolute value operator and the terms w_{ij} are defined as (Note that this definition is in conformity with the convention that a simultaneous tie is counted as a concordance.):

$$w_{ij} = \begin{cases} w_{ij}^A / w_{ij}^B & \text{if } w_{ij}^B \neq 0 \\ 1 & \text{if } w_{ij}^A = 0 \text{ and } w_{ij}^B = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

with $w_{ij}^A = \text{sign}(a_i - a_j)$, $w_{ij}^B = \text{sign}(b_i - b_j)$, and function $\text{sign}(x)$ defined such that it takes values -1, 0, and $+1$ for $x < 0$, $x = 0$, and $x > 0$, respectively. The Kendall's index in (2) can also be obtained in equivalent form by replacing the denominator in (3) with $n(n-1)/2$.

3.2 Weighted formulation

By looking at definitions (3) and (4) it turns out that the Goodman-Kruskal and Kendall's indexes are magnitude insensitive because the quantities in (4) are discrete and constrained to -1, 0, or $+1$, no matter the values of the entries in sequences A and B. In other words, pairs (a_i, a_j) and (b_i, b_j) are fully concordant or discordant no matter how much they agree or disagree. By noting that concordance and discordance are both a matter of degree, a weighted version of those indexes can be obtained by replacing the terms in the numerator of (3) with continuous versions of the weights w_{ij}, w_{ij}^A, and w_{ij}^B in (4), which can be redefined as:

$$\hat{w}_{ij} = \begin{cases} \min\{\hat{w}_{ij}^A / \hat{w}_{ij}^B, \hat{w}_{ij}^B / \hat{w}_{ij}^A\} & \text{if } \hat{w}_{ij}^A \text{ and } \hat{w}_{ij}^B \text{ have the same sign} \\ \max\{\hat{w}_{ij}^A / \hat{w}_{ij}^B, \hat{w}_{ij}^B / \hat{w}_{ij}^A\} & \text{if } \hat{w}_{ij}^A \text{ and } \hat{w}_{ij}^B \text{ have opposite signs} \\ 1 & \text{if } \hat{w}_{ij}^A = 0 \text{ and } \hat{w}_{ij}^B = 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$\hat{w}_{ij}^A = \begin{cases} \dfrac{a_i - a_j}{a_{\max} - a_{\min}} & \text{if } a_{\max} \neq a_{\min} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$\hat{w}_{ij}^B = \begin{cases} \dfrac{b_i - b_j}{b_{\max} - b_{\min}} & \text{if } b_{\max} \neq b_{\min} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $a_{\max} = \max\{a_1, a_2, \ldots, a_n\}$, $a_{\min} = \min\{a_1, a_2, \ldots, a_n\}$, $b_{\max} = \max\{b_1, b_2, \ldots, b_n\}$, and $b_{\min} = \min\{b_1, b_2, \ldots, b_n\}$.

Note that both $\hat{w}_{ij}^A$ and $\hat{w}_{ij}^B$ belong to $[-1, +1]$ and represent the (signed) percentage differences between the values of the ith and jth elements of the corresponding sequences. The weight $\hat{w}_{ij} \in [-1, +1]$, in turn, is such that: (i) it is positive if pairs (a_i, a_j) and (b_i, b_j) are concordant ($\hat{w}_{ij}^A$ and $\hat{w}_{ij}^B$ have the same sign or both are null); (ii) it is negative if (a_i, a_j) and (b_i, b_j) are discordant ($\hat{w}_{ij}^A$ and $\hat{w}_{ij}^B$ have opposite signs); and (iii) it is null in case of a neutral (either $\hat{w}_{ij}^A$ or $\hat{w}_{ij}^B$ is null).

Full concordance ($\hat{w}_{ij} = 1$) is obtained when $\hat{w}_{ij}^A = \hat{w}_{ij}^B$. This means that the relative difference between the ith and jth elements of A matches perfectly the relative difference between the corresponding elements of B. This exact match includes the sign, meaning that the relative order of the elements is also respected. Full discordance ($\hat{w}_{ij} = -1$), in its turn, is obtained when $\hat{w}_{ij}^A = -\hat{w}_{ij}^B$, that is, when the relative differences between the ith and jth elements of A and B have the same magnitude yet opposite signs (i.e. opposite relative orders). At a first glance, this may sound incorrect if the reader is misled to think that a larger difference between $|\hat{w}_{ij}^A|$ and $|\hat{w}_{ij}^B|$ should mean a greater discordance between pairs (a_i, a_j) and (b_i, b_j). That's not true. Indeed, it follows that, in relative terms, increasing the absolute value of one of these weights is equivalent to reducing the absolute value of the other. But reducing the absolute value of a given weight means driving this weight towards zero and further changing sign. This means reducing discordance towards neutrality and further turning to concordance.

Bearing the above remarks in mind, the weighted Goodman-Kruskal index can thus be defined as:

$$\hat{\gamma} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \hat{w}_{ij}}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |w_{ij}|} \tag{8}$$

where $\hat{w}_{ij}$ and w_{ij} are given by equations (5) and (4), respectively. The weighted version of the Kendall's index can be obtained in equivalent form by replacing the denominator in (8) with $n(n-1)/2$.

It can be shown that, as it is the case for the Pearson coefficient, the maximum value for the weighted versions of the Goodman-Kruskal and Kendall's indexes ($\hat{\gamma} = 1$) is obtained *iff* there exist two real-valued scalars $c > 0$ and d such that $a_i = cb_i + d$ for $i = 1, \ldots, n$, i.e., *iff* A is a linear (or affine) function of B with positive angular coefficient. Analogously, it can also be shown that the minimum value ($\hat{\gamma} = -1$) is obtained *iff* there exist two real-valued scalars $c < 0$ and d such that $a_i = cb_i + d$ for $i = 1, \ldots, n$, which means that one of the sequences becomes precisely an amplified/attenuated and/or shifted version of the other in inverse order when they are simultaneously rearranged so that one of them gets sorted.

Finally, it is straightforward to verify that the time complexity of the proposed weighted versions of the Goodman-Kruskal and Kendall's indexes is $O(n^2)$ if they are computed using (8). Whether computing these weighted indexes in less than $O(n^2)$ time is possible or not is still an open question that may be subject of future research.

4 Discussion

It is important to stress the difference between scenarios involving and not involving sequences on quantitative scales. Recalling back to the preliminary discussions in the introduction, a typical scenario for the application of sequence correlation indexes to data analysis refers to the comparison of a hierarchical clustering cophenetic matrix (sequence A) against another distance matrix obtained from the data set (sequence B). These matrices can be composed of either ordinal measurements (both A and B are not on quantitative scales) or continuous measurements (both A and B are on quantitative scales). Another usual scenario where both A and B are on quantitative scales regard, for instance, the measurement of similarity between data objects described by real-valued attributes (e.g. gene-expression data).

If sequences A and B are not on relevant quantitative scales, then either the original (non-weighted) Goodman-Kruskal/Kendall's indexes or the Spearman's coefficient should be used because neither rely on the magnitudes of the entries in A and B. Instead, if both sequences are on *relevant* quantitative scales, then either the proposed Weighted Goodman-Kruskal/Kendall's indexes or the Pearson Product-Moment correlation coefficient should be used because both rely on the magnitudes of the entries in A and B. The choice between the proposed indexes and the Pearson coefficient can be made in terms of scalability and sensitivity. If scalability is important, then the Pearson coefficient may be the best choice because it can be very efficiently computed in $O(n)$ time. If this is not an issue, then the Weighted Goodman-Kruskal or Kendall's indexes should be adopted because they are fully sensitive to both the magnitudes and rankings of A and B (at the price of an $O(n^2)$ time complexity).

5 Conclusions and future work

A Weighted formulation for the Goodman-Kruskal and Kendall's indexes have been proposed that can be more appropriate than the original (non-weighted) indexes for measuring the correlation between two sequences of numbers that are both on relevant quantitative scales. The proposed indexes can also be more appropriate than other existing correlation measures (e.g. Pearson) when full sensitivity to both the ranks and magnitudes of the sequences under evaluation is a desired property. It has been discussed in the paper that such a full sensitivity may in fact be desirable in different contexts of data analysis and data mining.

There are different research lines that started from this preliminary work and that are currently being investigated. For instance, a study about theoretical properties of the proposed indexes that can be of particular interest to the data mining community is on the way. In addition, comparisons of the proposed indexes against existing indexes have been worked out in several different aspects. The authors intend to report the results of such an ongoing work in future papers.

Acknowledgements

This work has been supported by the Brazilian National Research Council (CNPq) and by the Research Foundation of the State of São Paulo (Fapesp).

References

[1] Pearson, K., Mathematical contributions to the theory of evolution. III regression, heredity and panmixia. *Philos Trans Royal Soc London Ser A*, **187**, pp. 253–318, 1896.
[2] Casella, G. & Berger, R.L., *Statistical Inference*. Duxbury Press, 2nd edition, 2001.
[3] Hubert, L. & Schultz, J., Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology*, **29**, pp. 190–241, 1976.
[4] Kendall, M.G. & Gibbons, J.D., *Rank Correlation Methods*. Edward Arnold, 1990.
[5] Mitchell, T.M., *Machine Learning*. McGraw Hill, 1997.
[6] Baldi, P. & Brunak, S., *Bioinformatics: The Machine Learning Approach*. MIT Press, 2nd edition, 2001.
[7] Stekel, D., *Microarray Bioinformatics*. Cambridge University Press, 2003.
[8] Jain, A.K. & Dubes, R.C., *Algorithms for Clustering Data*. Prentice Hall, 1988.
[9] Everitt, B.S., Landau, S. & Leese, M., *Cluster Analysis*. Arnold, 4th edition, 2001.
[10] Goodman, L.A. & Kruskal, W.H., Measures of association for cross-classifications. *Journal of the American Statistical Association*, **49**, pp. 732–764, 1954.
[11] Goodman, L.A. & Kruskal, W.H., *Measures of Association for Cross Classifications*. Springer-Verlag, 1979.