Evaluating the robustness of goodness-of-fit measures for hierarchical clustering

A Timofeeva

Department of Theoretical and Applied Computer Science, Novosibirsk State Technical University, 20 Karl Marx Street, Novosibirsk, 630073 Russian Federation

E-mail: a.timofeeva@corp.nstu.ru

Abstract. In the presence of outliers in the data, hierarchical clustering can produce poor results. The agglomeration method for dendrogram constructing is usually chosen on the basis of the goodness-of-fit measures. But they are not robust to the presence of outliers. So that drawback is typical for cophenetic correlation. To reduce the sensitivity to outliers, it is proposed to use Spearman rank correlation coefficient between observed and cophenetic distances. But cophenetic distances have many repeated value. This results in problem of ties. So it is better to use correlation coefficients for ordered categorical variables. The correlation ratio is most appropriate. The rank modification is proposed. In addition to measures based on correlations, residuals-based measures are compared. For simulation study a sample with the strong association between internal and external validity is drawn. The outlier was randomly repeatedly added. The most stable results showed residuals-based goodness-of-fit measures. They provide the highest values of the Rand index. The proposed modifications based on the Spearman correlation and the rank correlation ratio are somewhat better than classical cophenetic correlation. They are more stable in the presence of strongly distant points.

1. Introduction

When solving problems of partitioning objects described by a set of features into homogeneous groups, a dendrogram is often used to visualize the results. It is a tree that illustrates the similarity of the objects. It is determined on the basis of various measures of similarity (dissimilarity). To construct such a tree, one can use a whole set of methods. In practice, the analyst needs to choose a measure of dissimilarity and an agglomeration method for constructing dendrograms. So in [1] standard methods of hierarchical cluster analysis were investigated based on a comparison of cophenetic correlations.

This measure compares the original matrix of pairwise distances between objects with the distance matrix calculated based on dendrogram (ultrametric distance). The authors of [1] modeled two situations (without and with outliers) and recommended the best methods of clustering for different distance measures. They were the single method and the centroid method, which provided high values of the cophenetic correlation in both situations.

Thus, most of the papers are aimed at comparing the algorithms of clustering based on the chosen goodness-of-fit measure. However, in practice, the researcher can compare the results of clustering by different methods on his own data set based on a certain goodness-of-fit measure. This requires a reliable criterion that really indicates the qualitative differences between agglomeration algorithms.

The problem is that certain goodness-of-fit measures have a number of drawbacks. For example, in the presence of outliers, the coefficient of cophenetic correlation is naturally greater than for a similar model without outliers. This feature is noted in the publications [1, 2]. It is associated with the fact that...
the dendrogram is adjusted to sharply distinguished observations, far removed from the rest of the points. Hence, at high values of the cophenetic correlation there is a danger of a false conclusion about the excellent quality of the dendrogram due to the presence of outliers.

Consequently, goodness-of-fit measures are not always reliable, especially in the presence of outliers. With just tiny bit of noise, traditional agglomerative algorithms will all fail [3]. However, studies on this issue are rather rare. An attempt is made here to make an integrated comparison of existing and proposed goodness-of-fit measures for dendrograms from the viewpoint of ensuring internal and external validity in the presence of outliers. Comparative analysis is based on how goodness-of-fit measures rank the methods of clustering. This allows us to conclude that it is justified to choose a method based on particular goodness-of-fit measure.

2. Methods of hierarchical clustering
Clustering methods, as a rule, differ in that their algorithms compute various functionals of the partitioning quality at each step. Such extreme problems make it possible to determine that quantitative criterion, following which one could prefer one partition to another. The best partition is understood as such a partition, on which an extremum (minimum or maximum) of the chosen quality functional is achieved. The choice of such a quantitative measure of the quality of the partitioning is sometimes based on empirical considerations.

The following are the seven most commonly used methods for defining cluster dissimilarity.

- Single linkage clustering defines the distance between two clusters as the distance between the two closest members (nearest neighbours). It tends to produce long and thin clusters. Further it is called "single".
- Complete linkage method considers the largest value (farthest neighbour) of pairwise dissimilarities as the distance between the two clusters. It makes spherical, more compact clusters. Further it is called "complete".
- Average linkage clustering computes the distance between two clusters as the average distance between all pairs of their members. The abbreviation UPGMA [4] is used to refer to this method. It avoids the problems of single and complete linkage. Clusters of miscellaneous shapes and outlines can be produced.
- Weighted Pair-Group Method using arithmetic Averages (WPGMA) is the modified previous. In calculations, the number of cluster members is used as a weight coefficient. Therefore, this method should be used, when unequal cluster sizes are assumed.
- Ward's method [5] says that the distance between two clusters is how much the sum of squares will increase when we merge them. In general, this method is very effective, but it tends to construct small clusters. The dissimilarities between the options "ward.D" and "ward.D2" are squared before cluster updating.
- Unweighted centroid method (UPGMC) defines the distance between two clusters as the distance between their geometric centroids.
- Weighted centroid method (median) is the modified previous. The abbreviation WPGMC is used to refer to this method.

3. Internal validity
The goodness-of-fit measures characterize how well the dendrogram describes the specified sample data. Thus, these are indicators of internal validity. According to the dendrogram, the cophenetic distances between objects can be calculated. Obviously, they should be close to real distances.

The goodness-of-fit measures can be divided into two groups. The measures of the first group are calculated as indicators of the relationship between observed and cophenetic distances. The closer the relationship, the better the goodness-of-fits. The measures of the second group are based on the discrepancies between the observed and the cophenetic distances, that is, the residuals. It is clear that the less such residuals, the better.
3.1. Measures based on correlation
The most popular goodness-of-fit measure for dendrogram is the cophenetic correlation coefficient. It is calculated as the usual Pearson correlation coefficient \( \rho_p \) between the observed and the cophenetic distances. However, it has a number of drawbacks:
- shows the closeness of only a linear relationship,
- is sensitive to outliers,
- high correlation does not guarantee that the differences between the distances are small.

To reduce the sensitivity to outliers, it is proposed to use other correlation measures. An alternative may be the Spearman rank correlation coefficient \( \rho_s \). Due to the fact that the relationship is revealed not by the original distances, but by their ranks, the coefficient is less sensitive to outliers.

Since the initial values are replaced by ranks, the effect of outliers for heavy tail distributions should be weakened due to the fact that now the distance between neighboring observation points is equalized (especially if it is a continuous value for which the probability of coincidence of values is zero and the values of ranks are different). In addition it is shown [6] that the power of the criterion when testing the hypothesis that \( \rho = 0 \), when using the Spearman rank correlation coefficient, it can be higher than using the Pearson correlation coefficient.

When using the Spearman coefficient, however, there is a problem in the presence of ties. Since the correct application of the coefficient is possible only for continuous variables [7]. If there are many ties, then this can lead to erroneous conclusions. To solve this problem, various corrections are proposed for the Spearman coefficient, taking into account repeated observations [7]. However, if the number of unique values of variables is very small, it is better to use correlation coefficients for ordered categorical features.

For cophenetic distances is characteristic that many values are repeated. The initial distances, as a rule, are different. Therefore, to identify the relationship between them, one can use the correlation ratio. When calculating it, it is assumed that one variable is continuous and the other is categorical. The root of the correlation ratio of the observed distances \( D \) from the cophenetic distances \( D' \) is determined from the expression

\[
\eta_{D,D'} = \sqrt{1 - \frac{\bar{V}_D}{\bar{V}_{D'}}}
\]

where \( \bar{V}_{D'} \) is the mean value of the conditional variance of the observed distances under the condition of the cophenetic distances, \( V_D \) is unconditional variance of the observed distances. The correlation relation shows the closeness of the non-linear relationship, but is sensitive to outliers. Therefore, it is proposed to calculate (1) not by initial distances, but by their ranks. Let's call such measure the rank correlation ratio.

3.2. Measures based on residuals
The simplest goodness-of-fit measure can be considered the delta between the real distances and the distances calculated from the dendrogram [8]

\[
\Delta_d = \left( \frac{\sum_{j,k} |d_{jk} - d'_{jk}|^{1/A}}{\sum_{j,k} (d'_{jk})^{1/A}} \right)^d
\]

where \( d_{jk} \) and \( d'_{jk} \) are observed and cophenetic distances, respectively. In equation (2) \( A \) can be equal to 1 or 0.5. Denote the matrix of observed distances \( d_{jk} \) by \( D \), cophenetic distances by \( D' \). The residual
matrix is calculated as \((D - D^*)^2\). In [9] the measure based on the largest singular value of the residual matrix was proposed. This measure is called 2-norm. It was shown in [10] that, in comparison with the cophenetic correlation coefficient, 2-norm is a more informative method for evaluating distortions generated by clustering algorithms. The described measures compare the clustering methods based on deviations from the straight line \(d = d^*\). I denote observed and cophenetic distances (values of random variables) by \(d, d^*\) to distinguish them from the random variables \(D, D^*\) introduced above. The line \(d = d^*\) is dotted in Figures 5, 6.

Interestingly, the measures based on correlations can be expressed through residuals. This allows us to illustrate the differences between the two approaches to the definition of goodness-of-fit measures. It is well known that the square of the Pearson correlation coefficient is equal to the coefficient of determination \(R^2\) in the linear regression model. Consequently,

\[
\rho^2_p = R^2 = 1 - \frac{RSS}{TSS}
\]

where \(RSS\) is the residual sum of squares, \(TSS\) is the total sum of squares. This expression is also true in the case when the response is the cophenetic distances, and the regressor is the observed distances, and vice versa. However, different agglomeration methods can give different \(TSS\) values (variations in the cophenetic distances) if the response is \(d^*\). If \(TSS\) will change depending on the method, this will not allow us to express \(\rho^2_p\) as linear function of \(ESS\). Therefore, when constructing the regression, one should take the observed distances as a response, and the covariate will be the cophenetic distance. Then in (3) \(TSS\) will be a constant. There will be an inverse linear relationship between \(\rho^2_p\) and \(ESS\).

Hence, the minimum \(ESS\) and maximum \(\rho^2_p\) will give the same result. Without taking into account the cases of \(\rho_p < 0\), the use of the cophenetic correlation coefficient is equivalent to minimizing the sum of squares of deviations from linear regression \(d\) or \(d^*\). This is illustrated in Figures 5, 6, where the regression line is represented by a solid line. Thus, the cophenetic coefficient characterizes the goodness-of-fit the regression line describing the conditional mathematical expectation \(E(D|D^*)\) as a linear function.

In contrast, the correlation ratio reveals the goodness-of-fit a nonlinear curve through the average values of the observed distances at each fixed value of the cophenetic distance. In Figures 5 and 6, this curve is shown a dashed line. Hence, the use of the correlation ratio is equivalent to minimizing the sum of squares of deviations from the nonlinear curve describing the conditional mathematical expectation \(E(D|D^*)\). Since such a curve better adapts to the data than a straight line, the correlation ratio is higher than the cophenetic correlation.

4. External validity

The external indices require a priori knowledge of partitioning \(P\). A clustering algorithm performed on the dataset \(\chi\) results in the clustering structure \(C\). The external indices are typically used to compare the structure \(C\) with partitioning \(P\) [11]. Rand index \(RI[12]\) is defined by:

\[
RI = \frac{TP + TN}{TP + TN + FP + FN}
\]

where \(TP\) (a true positive decision) is the frequency of pairs in \(\chi\) which belong to the same group in \(C\) and the same category in \(P\); \(TN\) (a true negative decision) is the number of pairs of elements in \(\chi\) that are in different subsets in \(C\) and in different subsets in \(P\); \(FP\) (a false positive decision) is the frequency of pairs in \(\chi\) which belong to the same group in \(C\), but different categories in \(P\); \(FN\) (a false negative decision) is the number of pairs of elements in \(\chi\) that are in different subsets in \(C\) and in the same subsets in \(P\).
Rand index measures the similarity between two partitions $P$ and $C$. It is obvious from (4) that index results in values in the range $[0,1]$, where 0 suggests that $C$ and $P$ are very different and 1 means highly similar partitions.

5. Simulation results

The following model example was used for the simulation study. For simplicity, it was assumed that the objects are characterized by coordinates on the plane. There are a certain number of local centers of concentration of objects. Let there be three clusters. Each of them is a circle with a base radius of 0.05. Circle data centers were chosen randomly. Area of allowable range of centersissquare $[0,1]^2$.

From the centers, the points of the cluster were spaced at a random distance uniformly distributed from 0 to 0.05 with a random angle distributed uniformly from 0 to $2\pi$.

The distances from the objects to the centers of the regions were multiplied by the coefficient $5+K$, $K$ is the ordinal number of the region. The number of objects belonging to the domain $K$ was set equal to $2+4K$. This provided the availability of areas with different density of spatial location of objects.

In order to compare the situation with the presence and absence of outliers, it was more productive, the initial sample was selected so that the internal and external validity matched with each other. This was established based on the correlation coefficients between the ranks of the clustering methods based on the goodness-of-fit measures and based on the Rand index. The random sample was used, which gave the maximum of the sum of the correlation coefficients for different fitting quality measures. The resulting clustering is shown in Figure 1.

![Figure 1. Initial scatterplot.](image1.png)

![Figure 2. Scatterplot with outlier.](image2.png)

To the initial data, the outlier was randomly added. For that, one of the coordinates of the extreme point marked in Figure 2 by a triangle was multiplied by a random uniformly distributed number from $h$ to $(h + 1)$. The greater $h$ is, the more distant the point became, and the more it should have affected the dendrogram. This is demonstrated in Figure 2.

It is interesting that for the initial situation, an excellent agreement of internal and external validity with respect to all goodness-of-fit measures was found, except for 2-norm. Figure 3 shows the relationship between the cophenetic correlation and the Rand index. It is seen that the worst method was single. The best methods are UPGMA and WPGMA. If one uses 2-norm, then these same methods are the best (Figure 4). But single is no longer the worst, it is inferior to Ward method.
Figures 3 and 4 clearly show that single linkage and Ward methods differ significantly in evaluating the goodness-of-fit based on correlations and residuals. This is due to the fact that for Ward method the cophenetic distances between objects are very different from the original ones in absolute values. This is illustrated in Figure 5. The line corresponding to the equality of the real $d$ and the reconstructed $d^*$ distances is dotted. Deviations from this line for Ward method are very large compared to single linkage (Figure 6). Hence, it turns out that the residuals are also very large. However, this does not mean that the closeness of the relationship between the distances is low. The cophenetic correlation does not depend on the scale. And, as shown above, it is equivalent to the sum of squares of deviations from the regression line. In Figures 5, 6 the regression line is solid. For single linkage method, variation around the line is higher than for Ward method. When using ranks, the difference between the measures based on correlations and on residuals is reduced. This is clearly seen in Figures 7, 8. Especially for Ward method, the regression line and the straight line $d = d^*$ are almost identical. Hence, deviations from them are almost the same.
Therefore, it is possible to improve the correspondence between the 2-norm measure and the Rand index. For this the residuals need to be calculated not based on the original distances, but based on their ranks. As a result, the ranking of the agglomeration methods will be very close to what we see in Figure 3. In terms of robustness, such a measure will be very similar to the other proposed rank measures. Therefore, these results are further omitted.

If you compare Figures 5 and 7, you can see that the use of ranks allows you to bring distances to a comparable scale. Thus, the problem of the Ward method with large values of the cophenetic distances is eliminated. Similarly, the impact of outliers should be reduced.

In order to characterize how stable the goodness-of-fit measures are with respect to the presence of outliers, 1000 generations of random outlier were conducted. For each sample, the internal and external validity of the results of the hierarchical clustering carried out by different methods is estimated. Based on the goodness-of-fit measures, the best method was chosen for each case. Table 1 shows the cases where each of the clustering methods turned out to be the best based on a given goodness-of-fit measure.

### Table 1. The relative frequency of best method chosen based on the goodness-of-fit measures.

| Method  | Cophenetic correlation | Spearman correlation | Rank correlation ratio | Delta | 2norm |
|---------|------------------------|----------------------|-----------------------|-------|-------|
|         | $h = 1$ | $h = 2$ | $h = 1$ | $h = 2$ | $h = 1$ | $h = 2$ | $h = 1$ | $h = 2$ | $h = 1$ | $h = 2$ |
| ward.D  | 0        | 0        | 0.001 | 0        | 0        | 0        | 0        | 0        | 0 |
| ward.D2 | 0.004    | 0        | 0.001 | 0        | 0        | 0        | 0        | 0        | 0 |
| single  | 0        | 0.002    | 0        | 0        | 0        | 0        | 0        | 0        | 0 |
| complete | 0.013    | 0.056    | 0        | 0        | 0        | 0        | 0        | 0        | 0 |
| UPGMA   | 0.638    | 0.543    | 0.541   | 0.819    | 0.475    | 0.821    | 0.883    | 0.952    | 0.91 |
| WPGMA   | 0.156    | 0.242    | 0.053   | 0        | 0.108    | 0        | 0.117    | 0.048    | 0.09 |
| WPGMC   | 0.009    | 0.056    | 0        | 0        | 0.106    | 0        | 0        | 0        | 0 |
| UPGMC   | 0.18     | 0.101    | 0.404   | 0.181    | 0.311    | 0.179    | 0        | 0        | 0 |
| RI      | 0.922    | 0.966    | 0.936   | 0.966    | 0.922    | 0.966    | 0.966    | 0.966    | 0.966 |

$RI = (0.83) (0.43) (0.85) (0.93) (0.83) (0.94) (0.86) (0.92) (0.86) (0.97)$
It can be seen that the classical cophenetic correlation coefficient does not provide a stable choice of the method. On the contrary, even the worst methods, in particular WPGMC, are chosen, and with a very distant outlier, even the single method can be chosen, the worst in the initial situation (Figure 3).

The proposed modifications based on the Spearman correlation coefficient and the rank correlation ratio are somewhat better. It is interesting that under conditions of strong distortion they are better than with a small distance of the outliers from other points. With \( h = 2 \), two leading methods UPGMA and UPGMC are clearly distinguished.

Residual-based measures proved to be better than measures based on correlations. With small outliers, they give quite stable results: the best are UPGMA and WPGMA. In the initial situation, these methods are also very close in terms of clustering quality (Figure 4). With a strong distortion, the measure 2normclearly selects the best UPGMA method.

Table 1 gives estimates of external validity based on the index \( RI \). The index was calculated from the results of clustering by a method that is determined to be the best by using a given goodness-of-fit measure. By its calculation, the outlier was assigned to a separate fourth class. The table shows the median values of the index; in parentheses the minimum values are given. Estimates of external validity are consistent with the conclusions about the stability of quality measures. The most stable were the residuals-based measures. They also provide the best correspondence between predicted and true classes.

Acknowledgments
The reported study was funded by Russian Ministry of Education and Science, according to the research project No. 2.2327.2017/4.6.

References
[1] Saraçli S, Doğan N, Doğan İ 2013 Comparison of hierarchical cluster analysis methods by cophenetic correlation Journal of Inequalities and Applications 2013 203
[2] Johnson R A, Wichern D W 2002 Applied Multivariate Statistical Analysis (New York: Prentice Hall)
[3] Balcan MF, Liang Y, Gupta P 2014 Robust hierarchical clustering The Journal of Machine Learning Research 15 3831-71
[4] Sneath PHA, Sokal RR 1973 Numerical Taxonomy (San Francisco: Freeman)
[5] Ward Jr J H 1963 Hierarchical grouping to optimize an objective function Journal of the American statistical association 301 236-244
[6] Fowler R L 1987 Powerand robustness in product-moment correlation Applied Psychological Measurement 11 419-428
[7] Oyeka I C A, Nwankwo C H 2014 Ties Adjusted Rank Correlation Coefficient IOSR Journal of Mathematics 10 9-17
[8] Mather P M 1976 Computational methods of multivariate analysis in physical geography (John Wiley & Sons)
[9] Mérigot B, Durbec JP, Gaertner JC 2010 On goodness-of-fit measure for dendrogram-based analyses Ecological Informat 91 1850–1859
[10] Carteron A, Jeannougin M, Leprieur F, Spatharis S 2012 Assessing the efficiency of clustering algorithms and goodness-of-fit measures using phytoplankton field data Ecological Informatics 9 64-68
[11] Xu R, Wunsch D 2009 Clustering (Piscataway, NJ: IEEE Press)
[12] Rand W M 1971 Objective criteria for the evaluation of clustering methods Journal of the American Statistical association 336846–850