Measuring the similarity of two cohorts in the n-dimensional space

Szabolcs Szekér, Ágnes Vathy-Fogarassy

Abstract: Measuring the similarity of the case and control group in clinical studies has always been an important but also difficult task. Several statistics-based methods aimed at this exist but most of them utilize dimension reduction or estimation, therefore, there are certain cases where they are not adequate. In this paper, we propose 3 dissimilarity-based measures capable of evaluating case-control group pairs without the loss of valuable information.

Keywords: similarity measure, control group, case group, non-paired matching, paired matching

Introduction

Observational studies in the field of social sciences, natural sciences and engineering are often based on comparative analysis. In these cohort-based studies, people are classified into two independent groups (case and control groups), and the conclusion is drawn on the basis of the similarities and differences between the two groups. Although the comparison-based scientific analyses significantly differ in the applied methodology and study design principles [1], due to their comparative nature they have strong scientific evidence [2]. Generally, the selection of the case group can be carried out based on the study aims, but the determination of the control group has difficulties and it raises many questions [3, 4]. Various methods have been proposed for the selection of the control group. Some methods utilize randomized sampling or stratified sampling [5, 6], while others are based on propensity score matching (PSM) [7] or on the selection of the k-nearest neighbors [8, 9]. All these methods try to form a control group which contains individuals that are similar to the individuals of the case group in certain characteristics, and in specific property or properties, they differ from them. But the measurement of the similarity of the case and control groups is not a trivial task. In this paper, we give a short overview of the applied similarity measuring approaches and propose three similarity measures to evaluate the degree of similarity of equally sized case and control groups in the n-dimensional space.

The rest of the paper is organized as follows. Section 2 briefly presents the basic approaches applied in this field and introduces our proposed measures. In Section 3, some comparative results are briefly presented. Finally, conclusions are drawn in Section 4.

Evaluation of the similarity of the case and control groups

Measuring the similarity of groups used in comparative analyses is an important task. The evaluation can happen by measuring the similarity of paired individuals from the case and control groups (paired evaluation) or by assessing the similarity of the case and control groups (non-paired evaluation). Most of the applied non-paired methods are Goodness of Fit (GoF) tests (e.g. Kolgomorov-Smirnov test, Bhattacharyya distance, Matusita distance) [10] evaluating the distribution of the two groups. Using a GoF test, it is possible to evaluate a one-dimensional distribution (that is the similarity of a certain property), but it is nearly impossible for higher dimensions [11]. However, people as the elements of the groups are characterized not by one but by many features. On the other hand, if the elements of the control group are selected by propensity score matching, the similarity of the case and control elements is measured again only in one dimension, namely as the dissimilarities of the propensity scores. As the propensity score is an estimated value, the similarity measurement is made in a lossy compressed 1-dimensional space, and not in the original feature space of the elements. Contrary to these methods, our aim is to measure the similarity of the case and control groups in the original high-dimensional feature space of the individuals. For this reason, we propose 3 dissimilarity measures: 2 for paired and 1 for non-paired evaluation. These measures can be considered as a normalized average of dissimilarities, but they differ in the interpretation of the term dissimilarity.

Paired evaluation

Nearest Neighbor Index

The first measure is called Nearest Neighbor Index (NNI) and it is quite strict. NNI checks for each attribute whether the case-control entity pairs are the closest neighbors to each other on that attribute. Pair elements are determined by the applied matching method. For categorical features the dissimilarity is 0 if the values of the attributes of the individuals are identical, otherwise 1. For continuous attributes, the dissimilarity is 0 if and only if the sample-control pair is the closest to each other pursuant to the examined attribute. Finally, the NNI is calculated as the average of the dissimilarities calculated in each dimension.

Dissimilarity for categorical features:

$$d_{ij}^{k} = \begin{cases} 0 & if \quad a_{ik} = b_{jk} \\ 1 & if \quad a_{ik} \neq b_{jk} \end{cases} \tag{27}$$

Dissimilarity for continous features:

$$d_{ij}^{k} = \begin{cases} 0 & if \quad \left|a_{ik} - b_{jk}\right| = min\left(\left|a_{ik} - b_{lk}\right|\right), \quad l = 1, 2, ..., N \\ 1 & if \quad \left|a_{ik} - b_{jk}\right| > min\left(\left|a_{ik} - b_{lk}\right|\right), \quad l = 1, 2, ..., N \end{cases} \tag{28}$$

Nearest Neighbor Index:

$$NNI(A, B) = \frac{\sum_{(a_i, b_j) \in P} \sum_{k=1}^{n} d_{ij}^{k}}{nN} \tag{29}$$

where a_i is an element of the case group ($a_i \in A$), b_j is an element of the control group ($b_j \in B$), $(a_i, b_j) \in P$ yields that they are matched case-control pairs, a_{ik} yields the value of the k-th dimension of a_i, b_{jk} analogously for the b_j element, n is the number of the characterizing features and N yields the number of individuals in either of the groups.

Global Dissimilarity Index

It is apparent that NNI checks for every dimension if the case-control pairs are closest to each other in that dimension, however, it does not consider the distance between them. The Global Dissimilarity Index (GDI) is a paired measure that is meant to account for this weakness.

GDI measures the dissimilarity for nominal features as the function of the number of different values, for ordinal features as the difference of ranks and for continuous features as the normalized distance. The calculation of GDI happens analogously as for NNI (Equation 29). The statement about paired elements still holds.

Dissimilarity for nominal features:

$$d_{ij}^{k} = \begin{cases} 0 & if \quad a_{ik} = b_{jk} \\ \frac{1}{M_k} & if \quad a_{ik} \neq b_{jk} \end{cases} \tag{30}$$

Dissimilarity for ordinal features:

$$d_{ij}^{k} = \begin{cases} 0 & if \quad a_{ik} = b_{jk} \\ \frac{\left|r_{a_{ik}} - r_{b_{jk}}\right|}{M_k - 1} & if \quad a_{ik} \neq b_{jk} \end{cases} \tag{31}$$

Dissimilarity for continuous features:

$$d_{ij}^{k} = \frac{\left|a_{ik} - b_{jk}\right|}{max\left(s_{lk}\right) - min\left(s_{lk}\right)}, \quad s \in \{A \cup B\}, \quad l = 1, ..., 2N \tag{32}$$

where M_k is the number of possible values along the k-th dimension, r yields the ordered rank for the ordinal attributes, $min\left(s_{lk}\right)$ is the minimal value and $max\left(s_{lk}\right)$ is the maximal value along the k-th dimension.

Non-paired evaluation

The above-mentioned methods measure the dissimilarity by determining the pairwise dissimilarities for each case-control pair. It is possible that only the similarity of the distributions of the characterizing features counts and the pairwise matching of the individuals is not relevant. On that score, we implemented a distribution-based measure called **Distribution Dissimilarity Index** (DDI). DDI is based on the histogram disparities of the case and control groups in each dimension. This method relies on the absolute deviation of the frequency of each property value relative to the size of the control group and the number of characterizing features. If the individuals are characterized by continuous values, the values have to be discretized before the calculation of the frequency values of the histogram.

$$DDI(A,B) = \frac{\sum_k \sum_{v \in V_k} |f^A_{kv} - f^B_{kv}|}{nN}, \quad k = 1, ..., n \tag{33}$$

where f^A_{kv} yields the frequency of the v-th value in the k-th dimension in the case group, f^B_{kv} analogously for the control group, V_k is the number of possible values along the k-th dimension.

Results of a short case study

To demonstrate the previously presented measures, we generated a benchmark dataset containing 1000 elements characterized by 8 variables (2 binary, 2 ordinal, 1 nominal and 3 continuous): 1 Bernoulli random variable with a probability value of 0.5, 1 Bernoulli random variable with a probability value of 0.3, 1 binomial variable with 3 trials and a probability value of 0.5, 1 uniform discrete variable in the range of $[0, 5)$, a uniform discrete variable in the range of $[0, 4)$, 1 uniform variable in the range of $[0, 2)$, 1 variable with normal distribution with a mean of 2 and standard deviation of 0.5 and 1 variable with normal distribution with a mean of 1 and standard deviation of 2. Additionally, a portion of the generated dataset was distorted with noise: $1\,\%$, $5\,\%$, $10\,\%$, $25\,\%$, $50\,\%$, $75\,\%$, $90\,\%$ and finally $100\,\%$ of the dataset was distorted along each dimension. In total, 9 case-control group pairs comprised our test scenario, in which dissimilarities of all pairs were evaluated by NNI, GDI, and DDI. The result of the evaluation can be seen in Table 1.

Table 1: DDI, NNI and GDI values for the 9 case-control group pairs. The presented values are dissimilarities in the range of $[0, 1]$. The smaller the value, the more similar the given case and control groups are.

Noise	0 %	1 %	5 %	10 %	25 %	50 %	75 %	90 %	100 %
NNI	0.000	0.009	0.043	0.085	0.214	0.426	0.645	0.772	0.851
GDI	0.000	0.004	0.019	0.038	0.096	0.192	0.291	0.347	0.385
DDI	0.000	0.003	0.008	0.015	0.036	0.054	0.080	0.091	0.097

As previously mentioned, NNI is the strictest measure, so it is especially sensitive to noise and dissimilar data. The 0.851 dissimilarity value is a proof of that behaviour. It is important to mention, that total dissimilarity (when the dissimilarity value is 1) is only achievable in extreme cases. These extreme cases are where the compaired values are at the opposite ends of the range of the examined variable. The statement about strict nature also holds for GDI, while DDI, the non-paired measure is noticably less sensitive, reaching only 0.097 dissimilarity when the whole dataset is distorted.

Conclusion

Control group evaluation is a non-trivial task and the selected similarity measure greatly affects the evaluation of research results. In this paper, we proposed 3 measures capable of evaluating the similarity of case and control groups as n-dimensional data in contrast to traditional methods that apply dimension reduction or estimation. All measures serve different purposes: DDI is recommended for non-paired evaluation, while NNI and GDI are recommended for paired evaluation. While in the course of scientific research analysts tend to consider only one criteria to evaluate the similarity of case and control groups, this article points to the fact that it is worth considering several aspects together.

Acknowledgements

We acknowledge the financial support of Széchenyi 2020 under the EFOP-3.6.1-16-2016-00015.

References

[1] J. W. Song and K. C. Chung, "Observational studies: Cohort and case-control studies," *Plastic and Reconstructive Surgery*, vol. 126, no. 6, p. 2234–2242, 2010.

[2] B. Everitt and C. R. Palmer, *Encyclopaedic companion to medical statistics*. Wiley, 2011.

[3] S. Wacholder, D. T. Silverman, J. K. Mclaughlin, and J. S. Mandel, "Selection of controls in case-control studies: Ii. types of controls," *American Journal of Epidemiology*, vol. 135, p. 1029–1041, Jan 1992.

[4] N. S. Weiss and T. D. Koepsell, "Epidemiologic methods," 2014.

[5] N. P. Jewell, "Least squares regression with data arising from stratified samples of the dependent variable," *Biometrika*, vol. 72, no. 1, p. 11, 1985.

[6] S. Singh, "Stratified and post-stratified sampling," *Advanced Sampling Theory with Applications*, p. 649–764, 2003.

[7] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behavioral Research*, vol. 46, no. 3, p. 399–424, 2011.

[8] S. Szekér and A. Vathy-Fogarassy, "Kontrollcsoport-generálási lehetőségek retrospektív egészségügyi vizsgálatokhoz," in *Neumann Kollokvium konferenciakiadványa, Orvosi Informatika*, p. 146, 2016.

[9] S. Szekér and A. Vathy-Fogarassy, "Novel k nearest neighbor-based control group selection methods," in *13th Miklós Iványi International PhD and DLA Symposium - Abstract Book: Architectural, Engineering and Information Sciences*, p. 124, Pollack Press, 2017.

[10] P. W. Mielke and K. J. Berry, *Permutation Methods*. Springer, 2007.

[11] G. Fasano and A. Franceschini, "A multidimensional version of the kolmogorov–smirnov test," *Monthly Notices of the Royal Astronomical Society*, vol. 225, no. 1, p. 155–170, 1987.