Computational Statistics and Data Analysis 99 (2016) 171–188

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

SMILE: A novel dissimilarity-based procedure for detecting sparse-specific profiles in sparse contingency tables☆

Mathieu Emily [a,d,*], Christophe Hitte [b], Alain Mom [c,d]

[a] Agrocampus Ouest, 65, rue de Saint-Brieuc, 35042 Rennes, France
[b] Université Rennes 1 - IGDR UMR CNRS 6290, 2 avenue du Professeur Léon Bernard, 35043 Rennes Cedex, France
[c] Université Rennes 2, Place du recteur Henri le Moal, 35043 Rennes, France
[d] IRMAR UMR CNRS 6625, 263 avenue du Général Leclerc, 35042 Rennes, France

ARTICLE INFO

Article history:
Received 9 June 2015
Received in revised form 25 January 2016
Accepted 27 January 2016
Available online 3 February 2016

Keywords:
Dissimilarity
Sparse contingency table
Single-linkage clustering
Conditional profile

ABSTRACT

A novel statistical procedure for clustering individuals characterized by *sparse-specific* profiles is introduced in the context of data summarized in sparse contingency tables. The proposed procedure relies on a single-linkage clustering based on a new dissimilarity measure designed to give equal influence to sparsity and specificity of profiles. Theoretical properties of the new dissimilarity are derived by characterizing single-linkage clustering using Minimum Spanning Trees. Such characterization allows the description of situations for which the proposed dissimilarity outperforms competing dissimilarities. Simulation examples are performed to demonstrate the strength of the new dissimilarity compared to 11 other methods. The analysis of a genomic dataset dedicated to the study of molecular signatures of selection is used to illustrate the efficiency of the proposed method in a real situation.

1. Introduction

Let us consider a two-way sparse contingency table that displays the number of occurrences of k categories for n individuals. This paper aims at detecting individuals with typical profiles of categories called *sparse-specific* profiles. *Sparse-specific* profiles, formally defined in Definition 2.1, are characterized by two main features. Firstly, the sparse profiles are those profiles for which only very few categories have non-zero counts. Secondly, specific profiles are those profiles presenting specific categories, i.e. categories that are (almost) never observed in the other individuals.

The detection of *sparse-specific* profiles is of interest in various application domains. For example, in genetics, *sparse-specific* profiles are expected to be encountered for breeds (*i.e.* a homogeneous group of domestic animals) under selection (Sabeti et al., 2006). In that context, genetic data can be summarized in a two-way contingency table for which an individual is a breed and a category is a DNA sequence, also called haplotype, of a given chromosomal region. Existing methods for detecting signatures of selection rely on strong assumptions based on population genetic theory that cannot be verified. Therefore, detecting the signatures of selection remains challenging. Alternative statistical methods that are robust to population genetic models are needed to improve the detection of selection (Wang et al., 2014).

☆ Software in the form of R code is available in the supplementary materials SMILE_CodeR.zip (see Appendix A).

* Corresponding author at: Agrocampus Ouest, 65, rue de Saint-Brieuc, 35042 Rennes, France. Tel.: +33 0 2 23 48 54 91.
E-mail address: mathieu.emily@agrocampus-ouest.fr (M. Emily).

http://dx.doi.org/10.1016/j.csda.2016.01.017
0167-9473/

The observation of *sparse-specific* profiles is also expected in other contexts such as text-mining or ecology. In text-mining, collected data are usually stored in a term-document matrix that describes the frequency of terms that occur in a collection of documents (Srivastava and Sahami, 2009). Observing a *sparse-specific* profile for a document in a term-document matrix means that the document has very few different terms that are almost not used in the other documents. In ecology, collected data can be summarized in a site by a species matrix where the abundance of different species is measured in various sites (Jongman et al., 1995). In that context, a species with a sparse-specific profile is expected to be observed in only very few sites. Those sites are assumed to host very few species, thus characterizing a low species richness. Although *sparse-specific* profiles are likely to be targeted in many applications, their detection raises the issue of detecting a non-symmetric relationship between a set of individuals and a set of categories. Furthermore, the non-symmetric relationship is well-characterized for *sparse-specific* profiles. The main challenge in the detection of *sparse-specific* profiles indeed lies in taking into account sparsity and specificity simultaneously.

In order to group individuals in sparse contingency tables, hierarchical clustering techniques are widely used. As such, the detection of *sparse-specific* profiles can be performed through a similar approach. The quality of the clustering depends on the choice of (1) a dissimilarity measure between individuals and (2) a linkage criterion for the hierarchical clustering. As quoted in Hastie et al. (2009, - p. 506), "Specifying an appropriate dissimilarity measure is far more important in obtaining success with clustering than choice of clustering algorithm". For that reason, attention was first focused on the choice of an appropriate dissimilarity to detect *sparse-specific* profiles. An abundant literature has been dedicated to improving the measure of similarity between individuals in sparse contingency tables, either by applying dimension reduction techniques or by proposing dissimilarity measures (Aggarwal and Zhai, 2012). Dimension reduction techniques, such as Latent Semantic Indexing (LSI) (Landauer et al., 1998) and Non-negative Matrix Factorization (NMF) (Lee and Seung, 2001) aim at transforming a high dimension space of features to a space of fewer dimensions using linear or non-linear combinations (Hastie et al., 2009). Applying such techniques can help selecting the most relevant categories, thus improving the quality of the clustering (Witten and Tibshirani, 2010). For instance, LSI and NMF techniques were successfully used prior to the clustering of textual data (Aggarwal and Zhai, 2012). However, these techniques does not explicitly account for *sparse-specific* profiles in the reduction of the dimensionality of the feature space. As a consequence, power to detect sparse-specific profiles for dimension reduction techniques is likely to be limited. On the one hand, the similarity between individuals can be measured with many different functions developed to deal with sparse contingency tables. In the text domain, the most well known and commonly used similarity function is the cosine similarity function (Singhal, 2001). In ecology, dedicated dissimilarities are the Bray–Curtis dissimilarity, the Jaccard dissimilarity, the d_1^2 (or Manhattan) distance, the Hellinger distance or the Gower dissimilarity (Oksanen et al., 2015).

However, these methods usually work directly with counts which might not be appropriate to detect *sparse-specific* profiles. Indeed, heterogeneity in the marginal counts of individuals gives different weights to individuals, thus leading to inappropriate conclusions. A natural way to control weights given to individuals is to focus on the conditional distribution of the categories, also known as conditional profiles. The analysis of conditional profiles is classically performed by using either the χ^2 distance or the d_2^2 distance (also known as L^2 norm). Nevertheless, capturing *sparse-specific* profile with the χ^2 distance raises some limitations since χ^2 is sensitive to profiles specificities. On the other hand, it will be shown that d_2^2 distance between two profiles is more influenced by the sparsity than the specificity.

In this paper, we propose a novel dissimilarity called d_s^2 adapted to the detection of *sparse-specific* profiles. d_s^2 is based on the comparison of conditional profiles and gives equal influence to sparsity and specificity of profiles, compared to other dissimilarities. To identify *sparse-specific* profiles, we propose a procedure called SMILE, for Statistical Method to detect sparse-specIfic profiLEs, which consists in a single-linkage hierarchical clustering (Jardine and Sibson, 1971) constructed using the d_s^2 dissimilarity. Selected profiles with the SMILE procedure correspond to the smallest subset of conditional profiles that coalesce at the final step of the hierarchical clustering.

In Section 2, we formalize the definition of a *sparse-specific* profile and give details of the SMILE procedure. Furthermore, by considering the parallel between Minimum Spanning Trees and Single-Linkage Cluster Analysis (Gower and Ross, 1969), an original characterization of the structure of the individual subset selected by the SMILE procedure is proposed.

In Section 3, we illustrate the performance of the SMILE procedure in a simulation study. For that purpose, a simple simulation algorithm generating contingency tables with respect to sparsity and specificity features was designed. Power for the SMILE procedure is compared to the power of 11 other clustering methods in simulated scenarios highlighting the highest power for the dissimilarity measure d_s^2.

Section 4 is devoted to the application of the SMILE procedure on a real dataset dedicated to the detection of molecular signatures of selection in the domestic dog (Lequarré et al., 2011). Comparing the SMILE procedure to 11 concurrent methods provides illustrative examples of the benefit of using the SMILE procedure for detecting *sparse-specific* profiles in sparse contingency tables.

2. The SMILE procedure

The SMILE procedure aims at detecting *sparse-specific* profiles. To do so, the proposed method selects the smallest subset of conditional profiles that coalesce at the final step of a single-linkage hierarchical clustering constructed with the d_s^2 dissimilarity. In this section, the approach driven by the features characterizing *sparse-specific* profiles is described.