Análisis de biarquetipos en R

Aleix Alcacer Sales[1] (aalcacer@uji.es), Irene Epifanio López[1,2] (epifanio@uji.es)
[1]Departament de Matemàtiques, Universitat Jaume I; [2]ValgrAI

Palabras clave: análisis de biarquetipos, clustering, aprendizaje no supervisado

Resumen

El análisis de biarquetipos (biAA) es una técnica estadística que extiende el análisis de arquetipos al identificar simultáneamente arquetipos de individuos y variables en una matriz de datos. A diferencia de otros métodos como el biclustering, el biAA se enfoca en puntos extremos, lo que facilita una interpretación más clara de los datos. Además, se ha creado un paquete de R para facilitar su uso.

Introducción

El análisis estadístico se ha convertido en una herramienta crucial en diversas áreas, proporcionando métodos para interpretar datos complejos y extraer patrones significativos. Dentro de este contexto, el análisis de arquetipos identifica puntos extremos en un conjunto de datos, conocidos como arquetipos. Para capturar una visión más completa de la estructura de los datos, se introduce un nuevo concepto denominado análisis de biarquetipos, que amplía el alcance del análisis de arquetipos tradicional.

Definición

El análisis de biarquetipos [1] es una extensión del análisis de arquetipos, que no solo extrae arquetipos de las filas (individuos) de una matriz de datos, sino que también lo hace para las columnas (variables) de forma simultánea. Estos arquetipos duales se denominan biarquetipos. Además, cada individuo y variable puede expresarse como combinación convexa de los biarquetipos.

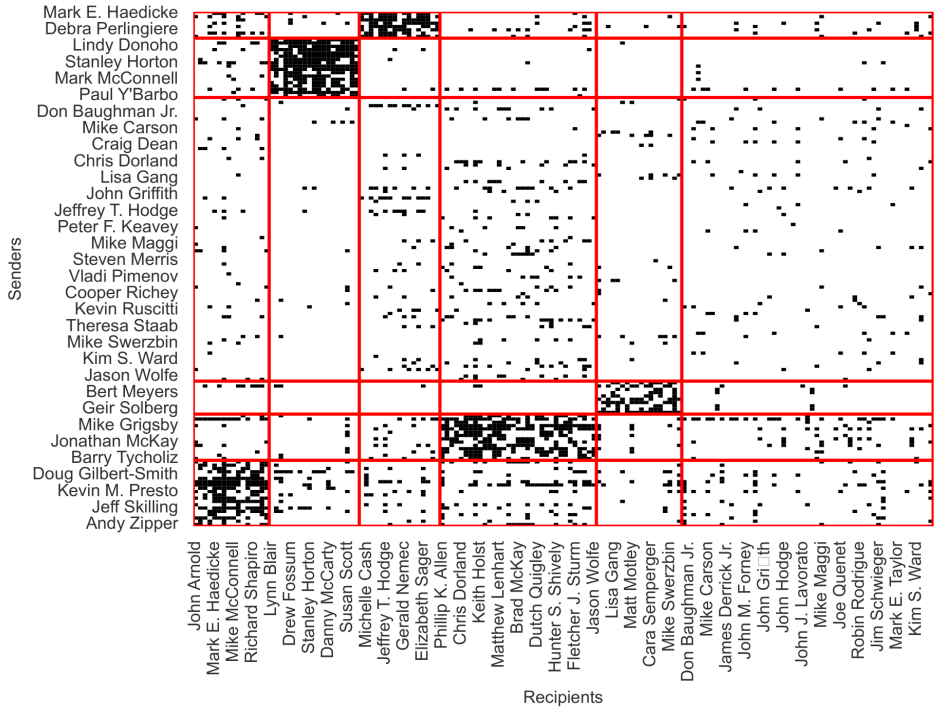
El objetivo de biAA es similar al del biclustering, pues su principal función es segmentar un conjunto de datos. Sin embargo, en lugar de obtener centroides o puntos centrales, biAA identifica arquetipos, que son puntos extremos fácilmente interpretables como combinaciones convexas de individuos y variables del conjunto de datos. Esta interpretabilidad es una ventaja significativa, ya que los extremos son más comprensibles que los puntos centrales.

Así, el análisis de biarquetipos no sólo permite interpretar a los individuos y las variables como mezclas de los biarquetipos, sino que también puede utilizarse para establecer grupos, aunque este no sea su objetivo principal. Respecto a las aplicaciones potenciales de biAA, estas abarcan diversos campos, desde la genética hasta la minería de textos y los sistemas de recomendación.

Aplicación práctica

Para ilustrar la aplicación práctica del biAA, consideremos un caso de estudio en la detección de comunidades dentro de la empresa Enron. El conjunto de datos contiene registros de correos electrónicos intercambiados entre empleados de Enron. Por ello, se creó una matriz de adyacencia donde cada valor indica si un empleado ha enviado un correo a otro (valor 1) o no (valor 0).

Al aplicar biAA con seis arquetipos y seis variables a esta matriz de adyacencia, se identificaron diversos grupos o comunidades dentro de la empresa. Por ejemplo, permitió descomponer la matriz de adyacencia en sus componentes clave, identificando grupos como el departamento Legal, ETS, Gas/Energía, West Power y la alta dirección de Enron.



Para facilitar la implementación de biAA, se ha desarrollado un paquete específico para R, denominado `biaa`, que está disponible para su descarga y uso en el siguiente repositorio: https://github.com/aleixalcacer/biaa. Este paquete permite a las personas usuarias aplicar biAA de manera eficiente a sus propios conjuntos de datos, siguiendo la metodología descrita en este ejemplo.

Conclusiones

BiAA se presenta como una herramienta poderosa para la identificación de patrones extremos en datos complejos. Su capacidad para detectar simultáneamente tanto individuos como variables biarquetípicos lo convierte en un método altamente aplicable a una amplia variedad de campos. La disponibilidad del paquete de R `biaa` simplifica la adopción de esta técnica, proporcionando al personal investigador una herramienta eficiente para aplicar biAA en sus propios conjuntos de datos.

Referencias

[1] A. Alcacer, I. Epifanio and X. Gual-Arnau, "Biarchetype Analysis: Simultaneous Learning of Observations and Features Based on Extremes," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi: 10.1109/TPAMI.2024.3400730