Análisis Multivariante
18/12/2023

UNIVERSITAT Carlemany

Actividad 4
Análisis Mamario

Katherine Carbonel
Jose David Ventura
Matias Davila
Sebastian Soto

Análisis Multivariante
18/12/2023

Introducción

En este informe hablaremos sobre lo que hicimos y lo que aprendimos en la “Actividad 4”, donde exploramos a fondo diferentes formas de analizar datos multivariados usando un conjunto de datos sobre tejido mamario. El análisis multivariado es muy importante en la bioestadística, porque nos ayuda a encontrar patrones y relaciones que no se ven a simple vista en datos complejos. Para hacer este análisis usamos el programa R, el cual es bastante flexible para hacer estadística avanzada. El conjunto de datos que usamos se llama BreastTissue.txt, y es el que nos sirve de base para este estudio. Tiene muchas variables, y eso nos da la chance de probar y comparar varias técnicas estadísticas multivariadas. Lo que queremos lograr es ver si hay grupos, patrones o correlaciones entre las variables. Así podemos entender mejor las características del tejido mamario, y también usar estos análisis para fines prácticos en la biomedicina y el diagnóstico.

Metodología

Preparación y Descripción de Datos

Para nuestro análisis, tomamos el archivo BreastTissue.txt, lleno de mediciones de tejidos mamarios variados. Lo primero fue poner todo en orden: cargamos los datos en R, observamos los datos para asi entender su estructura y les dimos una limpieza. Una columna importante, “Class”, la separamos para usarla más adelante, y preparamos el resto para el análisis.

Paquetes y Herramientas de R

usamos varios paquetes de R que nos facilitaron el análisis, con FactoMineR hicimos el análisis de componentes principales (PCA), que es una forma de simplificar los datos. Cluster fue usado para agrupar según sus similitudes. E1071 los agrupamos de forma más flexible. MVN comprobamos si nuestros datos se ajustaban a una distribución normal, que es una forma de medir la variabilidad.

Proceso Analítico

El proceso analítico de la "Actividad 4" abarcó diversas técnicas de análisis multivariante, cada una contribuyendo a una comprensión más profunda del conjunto de datos de tejido mamario. A continuación, se detallan los pasos clave de este proceso:

Distancias Mahalanobis

Para cada dato, calculamos las distancias Mahalanobis. Esta medida estadística nos dice qué tan raro o alejado es un dato de la media de todos los datos, considerando la relación entre las variables. En nuestro análisis, las distancias Mahalanobis nos sirvieron para encontrar datos que eran muy distintos de los demás, lo que podría significar que había algunos tipos de tejido mamario que tenían características especiales..

Análisis de Clustering

El análisis de clustering se inició determinando el número óptimo de clústeres. Para esto, se aplicó el método de la silueta, que evalúa cuán bien se ajusta cada observación a su clúster asignado. Este método proporciona una medida de cohesión y separación de los clústeres, siendo esencial para garantizar la validez del análisis de clustering.

Después de decidir cuántos grupos queríamos hacer, usamos el método PAM (Particionamiento Alrededor de Medoides). Este método es diferente al k-means, que es más común, porque elige datos representativos (medoides) dentro de cada grupo, y así no se confunde con los datos raros. Con este método, encontramos grupos naturales en los datos, y nos dimos cuenta de que había algunos tipos de tejido mamario que se diferenciaban por sus medidas.

PCA (Análisis de Componentes Principales)

Usamos el PCA para simplificar los datos. Esta técnica cambia las variables originales por unas nuevas, los componentes principales, que son perpendiculares entre sí y captan lo más importante de los datos. El PCA nos permitió ver y entender mejor las relaciones entre los datos y las variables. También nos ayudó a encontrar patrones escondidos y eliminar el ruido en los datos, que es muy importante cuando hay muchas variables.

Fuzzy Clustering

Usamos el fuzzy clustering, con el algoritmo c-means, para ver los datos de otra forma diferente al clustering normal. El fuzzy clustering no pone cada dato en un solo grupo, sino que permite que pertenezca un poco a varios grupos. Esto tiene sentido desde el punto de vista biológico, porque los tejidos pueden tener características que se mezclan entre varios tipos, y así podemos entender mejor los datos.

Inferencia Estadística

Al final, hicimos dos pruebas estadísticas: la de normalidad multivariante y la de homogeneidad de las matrices de covarianza. La de normalidad multivariante, que se llama prueba de Mardia, sirve para ver si los datos siguen una distribución normal, que es algo que se asume en muchos análisis estadísticos. La de homogeneidad de las matrices de covarianza (Box's M) sirve para ver si las diferentes categorías (tipos de tejido mamario) tienen matrices de covarianza parecidas, que es algo que se necesita para usar ciertas técnicas estadísticas.

Discusión

Interpretación Integral de los Resultados

En la "Actividad 4" vimos lo complejos y ricos que son los datos de tejido mamario. Con las distancias Mahalanobis encontramos datos raros, que podrían tener tejidos especiales. Con el método PAM hicimos grupos con los datos, y vimos que había tipos diferentes de tejido mamario. Estos grupos podrían mostrar diferencias biológicas importantes. Con el PCA vimos cómo se relacionan las variables, y encontramos patrones que no se veían antes. También simplificamos los datos, que eran muy complicados. Esto nos ayudó a ver y entender mejor los datos. Con el fuzzy clustering clasificamos los tejidos de otra forma, más flexible, que se parece más a cómo son las características biológicas en realidad. Con las pruebas estadísticas le dimos más seriedad al estudio, y nos aseguramos de que lo que decíamos tenía sentido estadístico.

Hallazgos Interesantes o Inesperados

Uno de los hallazgos más intrigantes fue la presencia de subgrupos claros dentro de los tipos de tejido mamario, como se evidenció en los resultados del clustering PAM. Estas subdivisiones

pueden indicar variaciones significativas en las propiedades del tejido, lo cual tiene potenciales implicaciones en la investigación biomédica y diagnóstica.

Limitaciones y Posibles Mejoras

Nuestro estudio dependía de los datos que teníamos, que podían ser mejores o diferentes. Los resultados podrían cambiar con más o distintos datos. También, aunque usamos técnicas avanzadas, siempre se puede mejorar la precisión y eficacia de los métodos estadísticos. En el futuro, se podrían probar otras formas de hacer clustering o PCA, o usar aprendizaje automático para entender mejor los datos.

Conclusiones

Lo Más Importante que Aprendimos Al analizar los datos de tejido mamario, vimos patrones y grupos importantes, que nos ayudan a entender cómo cambian las características del tejido. Con el PCA y el clustering, vimos cómo eran los datos por dentro.

Qué Significa Esto Esto significa que podemos entender mejor las diferencias en los tejidos mamarios, lo que puede ser muy importante para la biomedicina y para hacer diagnósticos más exactos.

Qué Podemos Hacer Después Después, podríamos usar más o distintos datos, o probar otras formas de analizarlos. También, sería bueno usar estas técnicas con otros datos biológicos, para ver si los resultados se mantienen.