Análisis de Datos Simbólicos: tomando en Cuenta la Variabilidad de los Datos

Prof. Paula Brito. Faculdade de Economía & LIAAD - INESC TEC, Universidade do Porto, Portugal.

Resumen

Los datos simbólicos, introducidos por E. Diday en los ochentas, se ocupan del análisis de datos con variabilidad intrínseca que debería ser tenida en cuenta. En minería de datos, análisis multivariado de datos y estadística clásica, los elementos analizados generalmente son entidades individuales para las cuales se graba un valor individual de cada variable – por ej., individuos descriptos por edad, salario, nivel educativo, etc. Pero cuando los elementos de interés son clases o grupos de algún tipo – como los ciudadanos que viven en una ciudad determinada, modelos de autos en lugar de vehículos específicos, etc. – hay variabilidad inherente en los datos. Reducir esta variabilidad mediante medidas de tendencia central – media aritmética, mediana o moda – lleva obviamente a una pérdida de información importante.

El análisis de datos simbólicos proporciona un marco que permite representar datos con variabilidad, usando nuevos tipos de variables. Además, se han desarrollado métodos que toman en cuenta adecuadamente la variabilidad de los datos. Los datos simbólicos se pueden representar usando los arreglos usuales en forma de matrices, pero en los cuales los elementos de cada celda no son valores numéricos reales individuales, como es usual, sino conjuntos finitos de valores, intervalos o, de forma más general, distribuciones.

En los últimos años surgió el término “Big Data” , refiriéndose a conjuntos de datos tan grandes y complejos que se vuelven difíciles de procesar en un tiempo razonable con aplicaciones tradicionales de análisis de datos. SDA, al ofrecer la posibilidad de agregación de datos al nivel de granularidad elegido por el usuario, mientras se mantiene la información sobre la variabilidad intrínseca, y luego analizar los arreglos de datos resultantes (simbólicos), puede desempeñar un papel importante en este contexto.

En este curso introduciremos y daremos la motivación del campo de Análisis de Datos Simbólicos, presentaremos en cierto grado de detalle los nuevos tipos de variable y lo ilustraremos con algunos ejemplos. Vamos a discutir además algunos problemas que surgen cuando se analizan datos que no siguen el modelo clásico usual, y los actuales modelos de representación de datos para algunos tipos de variables. Luego recordaremos algunos métodos que fueron desarrollados para analizar datos simbólicos. Algunos métodos presentados se pueden ilustrar usando el paquete de software SODAS.

El curso está dirigido a todos los analistas de datos potenciales que necesitan o están interesados en analizar datos con variabilidad, por ej., datos que resultan de la agregación de registros individuales en grupos de interés, o datos que representan entidades abstractas como especies biológicas en grupos de interés o datos que representan entidades abstractas como especies biológicas o regiones

como un todo. Esta metodología resulta particularmente interesante para el estudio de Economía y Gestión, Marketing, Ciencias Sociales, Geografía, estadísticas sobre datos oficiales, así como para Biología y análisis de datos Geológicos.
Se asume que los participantes dominan la estadística clásica. Para algunos de los últimos temas es conveniente conocer conceptos básicos de análisis multivariado.
El curso se dictará en castellano.

Esquema del curso:

A. El paradigma de Análisis de Datos Simbólicos

1. Introducción al Análisis de Datos Simbólicos.
1.1. Motivación. Ejemplos.

2. Fuentes de datos simbólicos: agregación (contemporaria, temporal); descripción de conceptos abstractos. Ejemplos. Alternativas al uso de medidas de tendencia central.

3. Tipos y variables y sus representaciones. Ejemplos.

4. Ejemplos de aplicaciones.

5. El paquete SODAS – presentación.
5.1. Archivos SDS y XML.
6. Visualización de datos simbólicos con SODAS

7. Interfaces: consiguiendo datos “nativos”.

8. Agregación de datos:
8.1. DB2SO – principios. Ejemplo.
8.2. Otras formas de agregación.

B. Métodos para el Análisis de Datos Simbólicos

1. Estadística Descriptiva

2. PCA
2.1. Método de los centros
2.2. Método de los vértices
2.3. Aplicación.
2.4. Referencia a otros métodos

3. Clasificación
3.1. Custering Divisivo: DIV
3.2. Clustering de particionamiento: SCLUST
3.3. Clustering Jerárquico y Piramidal: HIPYR

3.4. Otros métodos.

4. Análisis Discriminante
4.1. Árboles de Decisión :TREE
4.2. Otros Métodos.

5. Regresión
5.1. Regresión Lineal para variables valuadas en intervalos
5.2. Regresión Lineal para variables valuadas en histograma

6. Modelos Paramétricos
6.1. Principios y definiciones
6.2. Modelo Gaussiano
6.3. El paquete MAINT.DATA en R
6.4. Tests
6.5. ANOVA y MANOVA
6.6. Análisis Discriminante

7. Referencia a otros programas / paquetes

8. Bibliografía principal
8.1. Libros
8.2. Principales papers

9. La comunidad SDA y sus actividades.

Main References

Books

Bock, H.-H. and Diday, E. (2000). Analysis of Symbolic Data, ExploratoryMethods for Extracting Statistical Information from Complex Data. Springer, Heidelberg.Billard, L., Diday, E. (2007). Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley.
Diday, E. and Noirhomme-Fraiture, M. (2008). Symbolic Data Analysis and the SODAS Software. Wiley.

Papers

Brito, P. (2014): "Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics". WIREs Data Mining and Knowledge Discovery, Volume 4, Issue 4, July/August 2014, 281–295. DOI: 10.1002/widm.1133
Brito, P., Duarte Silva, A. P. (2012): "Modelling Interval Data with Normal and Skew-Normal Distributions". Journal of Applied Statistics, Volume 39, Issue 1, 3-20.
Noirhomme-Fraiture, M., Brito, P. (2011): "Far Beyond the Classical Data Models: Symbolic Data Analysis. " Statistical Analysis and Data Mining Volume 4, Issue 2, 157-170.

Brito, P. (2007): "Modelling and Analysing Interval Data". In: "Advances in Data Analysis", Decker, R., Lenz, H.-J. (Eds.), Series "Studies in Classification, Data Analysis and Knowledge Organization", Springer, Berlin, Heidelberg, New-York, 197-208.

Brito, P. (2007): “On the Analysis of Symbolic Data". In: "Selected Contributions in Classification and Data Analysis", Brito, P., Bertrand, P., Cucumel, G., De Carvalho, F. (Eds.), Series "Studies in Classification, Data Analysis and Knowledge Organization", Springer,Heidelberg, 13-22.

Duarte Silva, A. P. , Brito, P. (2006). "Linear Discriminant Analysis for Interval Data". Computational Statistics, 21, 2, 289-308.

Billard, L. and Diday, E. (2003) “From the statistics of data to the statistics of knowledge: Symbolic Data Analysis”, Journal of the American Statistical Association 98 (462), pp. 470–487.