

INTELIGENCIA ARTIFICIAL Y ESTADÍSTICA APLICADA: ALGUNAS LÍNEAS MODERNAS DE DESARROLLO

JOSÉ M.^a CARIDAD Y OCERÍN
ACADÉMICO CORRESPONDIENTE

Al referirse a una rama de la Matemática, como es la Estadística, es conveniente precisar este concepto, que a menudo es confuso para el profano:

La Estadística es la rama de la Ciencia que trata sobre la experimentación, recolección, descripción y análisis de datos. Los métodos estadísticos son herramientas para examinar y analizar datos y para tomar decisiones en ambientes de incertidumbre.

Otra cosa son las “estadísticas” o colecciones de datos tomados en una determinada población y que aportan información sobre ésta.

En general podemos diferenciar varios aspectos que dan origen a distintos enfoques de la Estadística. Al realizar un proceso de recogida de datos en un colectivo o población, se pueden pretender objetivos diversos; si se trata de describir dicho colectivo, la experimentación debe ser exhaustiva: hay que tomar datos sobre todos los elementos de la población, y los métodos matemáticos que tratan estos problemas se denominan “Análisis de Datos” o “Estadística Descriptiva”. Por otra parte, si se ha seleccionado un subconjunto del colectivo, o sea una muestra, se tratará de extrapolar los resultados obtenidos sobre ésta a toda la población; las técnicas empleadas constituyen la “Estadística matemática o inferencial”, las inferencias generales están afectadas de un nivel de incertidumbre que es cuantificado en términos probabilísticos. Así, el Cálculo de Probabilidades, que es una rama del Análisis Matemático, constituye una herramienta imprescindible para la aplicación de los Métodos Estadísticos.

Los rápidos desarrollos que ha experimentado el cálculo automático en los últimos veinte años ha cambiado radicalmente la metodología de trabajo del estadístico profesional. Por una parte, los ordenadores han permitido abordar campos antes vetados para la Estadística aplicada: el Análisis Multivariante, la teoría de series Temporales, los métodos multiecuacionales, etc.. Por otra, esta disponibilidad de cálculo ha impulsado nuevas investigaciones teóricas en méto-

dos conocidos, y ha permitido la aparición de nuevas líneas de desarrollo con la incorporación de nuevas técnicas estadísticas y de análisis de datos que están de nuevo cambiando la forma de abordar los numerosos problemas relacionados con la Estadística que surgen diariamente en las Ciencias Sociales y Experimentales.

La evolución histórica de la Ciencia Estadística puede dividirse en cinco bloques claramente definidos:

- La “prehistoria” que abarca desde el Renacimiento italiano y la aplicación de los métodos estadísticos actuariales, hasta el desarrollo del Cálculo de Probabilidades al que contribuyeron los principales matemáticos europeos de los siglos XVI a XVIII con la formalización teórica de los modelos probabilísticos durante el siglo pasado destacando, entre otras las grandes figuras de Gauss y P.S. de Laplace.
- Durante la primera mitad del siglo XX se desarrolla la Estadística Clásica con los trabajos de Karl Pearson, W. Gossett, R. Fisher, J. Neyman y otros, así como numerosos métodos estadísticos (modelos lineales, diseño de experimental, control de calidad, fiabilidad de sistemas, etc.); la Econometría surge a partir de la problemática creada por la gran crisis económica de 1929 y por la subsiguiente intervención estatal creciente; la Investigación Operativa es impulsada por las necesidades bélicas en los años cuarenta; el Cálculo de Probabilidades es sistematizado sobre una base axiomática por la escuela rusa de Análisis Matemático mediante la formulación de A.N. Kolmogorov. A partir de los años cincuenta se difunden la Teoría de la Decisión y los Métodos Bayesianos.
- En los años sesenta surgen los primeros paquetes de programas estadísticos en ordenadores de propósito general basados en proceso por lotes y en un lenguaje de control de muy alto nivel independiente del sistema operativo utilizado; las bibliotecas de rutinas científicas suponen una facilidad extraordinaria en la programación en Fortran y Algol evitando numerosos problemas de tipo numérico. El disponer de gran capacidad de cálculo potenció nuevas investigaciones teóricas entre las que podemos citar los nuevos desarrollos en Análisis Multivariante, algunos Métodos no Paramétricos como las técnicas booth-trap, modelos de optimización en ambientes probabilísticos, modelos no lineales, procesos estocásticos, análisis de datos cualitativos, etc., que los distintos paquetes han ido incorporando, poniendo a disposición del estadístico profesional, así como del usuario final, nuevas y potentes herramientas de Análisis de Datos.
- En la última década la situación de los métodos de Estadística Aplicada se caracteriza por la gran abundancia de paquetes estadísticos tanto de tipo interactivo, como para realizar proceso por lotes, muy completos y sofisticados; la gran difusión de micro y mini-ordenadores ha favorecido la aparición de nuevos programas, así como de desarrollo acelerado de métodos gráficos de análisis y representación de datos multivariantes.

Actualmente la utilización de métodos estadísticos en ordenadores se encuentra en una situación de cambios muy rápidos que se caracterizan por:

- la aparición de nuevos trabajos teóricos fundamentalmente en el campo multivariante y temporal;

- la incorporación de técnicas renovadas de especificación y contrastación de modelos;
- la generalización de métodos gráficos, incluyendo, en un futuro próximo, la aplicación de la teoría del color;
- el desarrollo de interfaces “amables” para el usuario, tanto en la manipulación y tratamiento de bases de datos estadísticos;
- la sistematización y formulación de estrategias de análisis de datos con el desarrollo de sistemas expertos y utilización de técnicas de Inteligencia Artificial en los paquetes estadísticos.

En definitiva, nos encontramos ante la tercera revolución que, en este siglo, ha conmocionado a la ciencia estadística, y a la forma de abordar el análisis de datos.

Las aplicaciones de la inteligencia artificial (I.A.) en Estadística persiguen integrar distintos contrastes, estimaciones, transformaciones y modelos para conseguir una aproximación coherente y total en Análisis de Datos, estableciendo estrategias que dirijan el proceso de modelización, de elección de técnicas y transformaciones a aplicar, y de ayuda a la interpretación de los resultados.

Es necesario poner a disposición de usuarios de Análisis de Datos, un conjunto de herramientas “inteligentes” que eviten un uso inapropiado, y desgraciadamente muy frecuente, de los paquetes estadísticos, que dirijan su utilización e interpretación, lo que conlleva a la automatización de los procesos de decisión y selección de estrategias de análisis de datos. La elaboración de un Sistema Experto requiere el desarrollo de metodologías de construcción y aplicación de estrategias de Análisis de Datos, hoy día ausentes en los clásicos paquetes BMDP, SPSS, SAS, TSP, Statgraphics, Rats y otros.

Las aplicaciones más llamativas de la I.A. en Estadística son, sin duda, en el campo de la Medicina, y en particular en problemas de diagnóstico automático, si bien hoy día existen en el mercado productos operativos para utilización de sistemas Expertos en algunos Métodos Estadísticos; por ejemplo el sistema REX para construir modelos lineales de regresión.

Se pueden clasificar estas aplicaciones en dos grandes bloques según sus destinatarios:

- sistemas de diseños y aplicaciones de estrategias de Análisis de Datos para estadísticos profesionales, y
- métodos para usuarios finales cuya formación estadística es elemental.

En la Universidad de Córdoba, y apoyado con la Comisión Nacional de Ciencia y Tecnología, el Departamento de Estadística está promoviendo la creación de varias ayudas “inteligentes” en el uso de paquetes estadísticos para usuarios no especialistas en Estadística: gestores de bases de datos muestrales, interfaces amables para manejo de datos, generadores de programas en lenguajes propios de paquetes como BMDP y SAS, y diseño de algunas estrategias útiles en el análisis de tablas de contingencia, a las que seguirán en un futuro próximo sistemas más complejos de automatización de estrategias de análisis de datos categorizados, y en particular en la especificación de modelos log-lineales, logit, y de exploración sistemática de bases de datos estadísticos nominales u ordinales que surgen, y en general aplicaciones de tipo biométrico y económico.

Cabe citar como realidad actual en este campo, el sistema S.E.A.D. recientemente presentado públicamente.

En definitiva la I.A. está contribuyendo al desarrollo tecnológico en muchas áreas científicas, entre las cuales se encuentra la Estadística; en correspondencia, entre las pocas disciplinas que pueden influir en la I.A. se encuentra precisamente la Estadística. La I.A. se ha centrado en la representación simbólica del conocimiento/información y de su uso, lo que en última instancia depende de datos experimentales que contienen una componente aleatoria (el error de medida y el muestral) lo que implica la utilización de terminología probabilística y métodos estadísticos.

Así, los razonamientos y proposiciones en sistemas de I.A. deben tener en cuenta la incertidumbre inherente en las relaciones empíricas, por ejemplo utilizando técnicas de Estadística bayesiana. La Estadística puede además contribuir en la teoría del conocimiento y aprendizaje; la formación de conceptos está relacionada con los métodos de "cluster" tanto para datos discretos como para medidas continuas.

En el campo de la Educación, desde hace varios años se han producido importantes modificaciones en los programas docentes de la Estadística Aplicada y Métodos Estadísticos. La introducción de paquetes de programas (BMDP, SPSS y otros) es algo habitual, imprescindible para alumnos de Ingeniería, Biología, Química, Medicina, Economía, etc., o sea para profesionales que utilizarán la Estadística en su actividad académica, científica o técnica. Los sistemas de ayuda: interfaces amables, gestores de datos y generadores de lenguajes es una realidad cuya utilidad es patente en el aprovechamiento docente. La introducción de sistemas expertos basados en métodos de I.A. debe ser una próxima realidad en alumnos de post-graduado, lo que obligará a replantear los sistemas docentes en beneficio de un mejor, más extenso y preciso uso de la Estadística.

En el Departamento de Estadística y Análisis de Datos de la Universidad de Córdoba, además de la actividad docente e investigadora, se realiza una intensa labor de consultoría en la utilización de técnicas estadísticas en varias áreas (Agronomía, Medicina, Economía, Biología, Control de calidad industrial, etc.) y en su aplicación en ordenador. Esta consultoría se plantea a dos niveles:

- Para investigadores y profesionales con un nivel estadístico medio o alto, en cuyo caso se proporciona un apoyo en el manejo de paquetes de programas en ordenador, en los procesos de modelización, diseño experimental y muestral, interpretación de resultados, etc. En definitiva, se realiza una cooperación intensiva en las tareas de investigación y desarrollo.
- Para científicos o técnicos con nivel bajo o medio en Análisis de Datos, estando la asesoría centrada, o bien en el uso de métodos usuales de Estadística y del logicial básico, o en una tarea de integración de un estadístico en el equipo de investigación y desarrollo para llevar a cabo todos los procesos de análisis y de modelización.

En ambos casos, la demanda de servicios de consultoría supera claramente a las posibilidades personales de los profesores de un departamento de Estadística, lo que implica la necesidad de contar con herramientas inteligentes de apoyo en estas actividades científicas de consultoría. En el Departamento de Estadística de

la Universidad de Córdoba se está desarrollando, en el ámbito del proyecto de investigación y desarrollo antes citado, un sistema experto para análisis de datos categorizados para optimizar y dirigir los procesos de tratamiento de bases de datos biomédicas y económicas, para las que existe una demanda importante en nuestra ciudad.

Bibliografía

- A. AGRETI: *Categorical Data Analysis*. Wiley, 1990.
- E.B. ANDERSEN: *Discrete Statistical Models with Social Science Applications*. North Holland, 1980.
- Y.S. BISHOP: *Discrete Multivariate Analysis: Theory and practice*.
- P. FIENBERG: *Cambridge*. M.I.T. Press, 1975.
- P. HOLLAND.
- E. BRENT: *Statistical Navigator*. Idea Works, 1988. Exys Inc.
- J.M. CARIDAD: *Sistemas expertos en el Análisis de Datos*.
- R. ESPEJO: "Categorizados". I Seminario Internacional de Sistemas.
- L. MELENDEZ: *Expertos en la Agricultura Mediterránea*. Córdoba. España. Abril de 1991.
- J.M. CARIDAD: *Estadística Aplicada a las Ciencias Sociales y Biomédicas*. S. Publicaciones. Univ. de Córdoba. 1987.
- J.M. CARIDAD: *Análisis de Datos con BMDP*. S. Publicaciones. Univ. de Córdoba. 1989.
- W.J. DIXON Ed' BMDP: *Statistical Software Manual*. Vol. 1 y 2. University of California Press. 1990.
- EUROSTAT: *Development of Statistical Expert System*. Luxemburgo. 1989.
- S.A. FIENBERG: *The analysis of cross-classified data*. Cambridge M.T.I. Press. 1980.
- R.A. FISHER: *The Desing of Experiments*. Oliver Boyd. 1935.
- W. GALE: *Artificial Intelligence and Statistics*. Addison-Wesley, 1986.
- J.M. MACDONALD: *Statistical Computation*. Wiley.
- K. MOMIROVIC, V. MILDNER (Ed): *Compstat: proceedings in computational statistics 9 Th Symposium*. Dubruvnik. Physica-Verlag Heidelberg. 1990.