FACULTAD DE INFORMÁTICA

UNIVERSIDAD NACIONAL DE LA PLATA

BÚSQUEDA DE VALORES ANÓMALOS EN BASE DE DATOS UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS **AÑO 2017**	**Carrera:** Doctorado en Ciencias Informáticas **Profesor:** Dr. Horacio Kuna **Créditos: 4** **Duración:** 70 horas

FUNDAMENTACION

La información se ha convertido en uno de los activos más importantes y es necesario garantizar la seguridad, calidad y legalidad de dicha información. A partir de este hecho, la detección de valores anómalos en Bases de Datos tiene un papel central en la prevención de riesgos relacionados con el gobierno y gestión de la tecnología de la información

OBJETIVO GENERAL

Que los alumnos desarrollen habilidades para detectar datos que por ser muy diferente a los demás pertenecientes a un mismo conjunto de datos puede considerarse que fueron creados por un mecanismo diferente, aplicando para esta detección técnicas y herramientas de Minería de Datos.

CONTENIDOS MINIMOS

Técnicas de auditoría asistidas por computadora (CATTs)
Minería de datos aplicada a la Auditoría de Sistemas
Taxonomía de metodologías de detección de valores anomalías en Bases de Datos
Comparación de métodos de detección de valores anómalos en Bases de Datos
Metodologías hibridas para la detección de valores anómalos en Bases de Datos

PROGRAMA

Unidad 1.

Técnicas de auditoría asistidas por computadora (CATTs)

Introducción. Definición de CATTs. Estándares y normas. Principales herramientas software utilizadas. Aplicación de CATTs en la Auditoría de Sistemas. Ventajas y desventajas del uso de las técnicas de auditoría asistidas por computadora. Metodología de trabajo para implementar CATTs.

Unidad 2

Minería de datos aplicada a la Auditoría de Sistemas

La auditoría de sistemas. Conceptos básicos de Minería de Datos. Auditoría continua y Minería de datos. Minería de datos para la detección de fraudes. Minería de datos para la detección de intrusos en redes de telecomunicaciones. Minería de datos para la detección de terroristas. Otras aplicaciones de la Minería de datos en la auditoría de sistemas.

Unidad 3.

Taxonomía de metodologías de detección de valores anomalías en Bases de Datos

Definición de outliers. Métodos para definir valores anómalos en bases de datos. Gráfico de bigote. Métodos univariantes y los métodos multivariantes. Métodos paramétricos y métodos no paramétricos. Métodos basados en la estadística. Métodos basados en la distancia. Métodos basados en la densidad. Métodos basados en técnicas de clustering. Métodos basados en redes neuronales. Otros métodos de detección de outliers: Métodos difusos, Métodos basados en la Teoría de la información. Métodos basados en Algoritmos Genéticos. Resumen de métodos de detección de outliers.

Unidad 4

Enfoques para la detección de valores anómalos en Bases de Datos

Enfoques para abordar la detección de valores anómalos en bases de datos. Detección de valores anómalos en Bases de Datos con aprendizaje no supervisado. Detección de valores anómalos con aprendizaje supervisado. Detección de valores anómalos en bases de datos con aprendizaje semi-supervisado. Aplicaciones de los distintos enfoques para la detección de valores anómalos en Bases de Datos. Criterios para la elección de métodos y enfoques en la detección de outliers.

Unidad 5

Metodologías hibridas para la detección de valores anómalos en Bases de Datos

Características de las metodologías hibridas para la detección de valores anómalos en Bases de Datos. Ventajas y desventajas. Combinación de algoritmos para la detección de valores anómalos en Bases de Datos. Algoritmo LOF. Algoritmo K-Means. Teoría de la Información. Redes Bayesianas. Algoritmo DBSCAN. Algoritmo PRISM. Experimentación en bases de datos de laboratorio y reales para la detección de valores anómalos en Bases de Datos. Métodos para validar los valores anómalos detectados.

ESTRATEGIAS DE APRENDIZAJE

- Clases teóricas
- Estudio de casos
- Practicas supervisadas en laboratorio
- Elaboración de Trabajos prácticos
- Elaboración y presentación de Monografías

MÉTODO DE EVALUACIÓN

80% de los trabajos prácticos aprobados
Aprobación de una evaluación final

BIBLIOGRAFÍA

Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data. *ACM Sigmod Record, Vol. 30*(2), 37-46.

Aggarwal, C. C., & Philip, S. Y. (2005). An effective and efficient algorithm for highdimensional outlier detection. *The VLDB journal*, *14*(2), 211-221.

Ben-Gal, I. (2005). Outlier detection. *Data Mining and Knowledge Discovery Handbook,* 131-146.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (1996). LOF: identifying densitybased local outliers. *ACM Sigmod Record, 29*(2), 93-104.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: identifying densitybased local outliers. *ACM Sigmod Record, 29*(2), 93-104.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *41*(3), 15.

Champlain, J. (2003). Audit Information Systems. 2nd Ed. John Wiley and Sons.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD, 96*, 226-231.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining. *MIT Press.*

Ferreyra, M. (2007). Powerhouse: Data Mining usando Teoría de la información. Recuperado de http://web.austral.edu.ar/images/contenido/facultad-ingenieria/2Data_Mining_basado_Teoria_Informacion_Marcelo_Ferreyra.pdf.

Hawkins, D. M. (1980). Identification of outliers. *London: Chapman and Hall.,* 11.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85-126.

Kirkos, S., & Manolopoulos, Y. (2004). Data mining in finance and accounting: a review of current research trends. *Proceedings of the 1st international conference on enterprise systems and accounting* (ICESAcc), 63-78.

Knorr, E. M., & Ng, R. T. (1997, November). A unified approach for mining outliers. *Proceedings of the 1997 conference of the Centre for Advanced Studies on Collaborative research,* 11.

Knorr, E. M., & Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets. *Proceedings of the International Conference on Very Large Data Bases*.

Knorr, E. M., Ng, R. T., & Tucakov, V. (2000). Distance-based outliers: algorithms and applications. *The VLDB Journal—The International Journal on Very Large Data Bases*, *8*(3-4), 237-253.

Ng, R. T., & Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. *Knowledge and Data Engineering, IEEE Transactions on*,*14*(5), 1003-1016.

Quinlan, J. R. (1993). C4.5: programs for machine learning. *Morgan Kaufmann*.

Penny, K. I., & Jolliffe, I. T. (2001). A comparison of multivariate outlier detection methods for clinical laboratory safety data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *50*(3), 295-307.

Peña, D. (2002). *Análisis de datos multivariantes*. Madrid: *McGraw-Hill/Interamericana de España*.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3-55.

Tan, P. N. (2007). Introduction to data mining. *Pearson Education India.*

Zhang, K., Hutter, M., & Jin, H. (2009). A new local distance-based outlier detection approach for scattered real-world data. *Advances in Knowledge Discovery and Data Mining*, 813-822.