CIENCIA DE DATOS APLICADA AL ESTUDIO DE LA FAUNA ÍCTICA EN LA ZONA DEL RÍO PARANÁ

Cinthia A. Cuba L., Paola V. Britos y Gladis G. Garrido

CONTEXTO

Este trabajo se lleva a cabo dentro del Proyecto de Biología Pesquera en el Laboratorio del Instituto de Biología Subtropical (IBS) de la Facultad de Ciencias Exactas, Químicas y Naturales (FCEQYN), de la Universidad Nacional de Misiones (UNaM), en el marco de un plan de trabajo final de la Maestría en Tecnologías de Información de la FCEQYN - UNaM.

LINEAS DE INVESTIGACION Y DESARROLLO

Toda la información recolectada del 1993 a la actualidad se encuentra almacenada en una base de datos relacional y puesto que una de las actividades del proyecto es brindar información a través de distintos informes, el análisis de ese gran volumen de datos dio pie a la creación de un almacén de datos, que junto a un sistema OLAP, forman parte de las nuevas herramientas de trabajo del proyecto. Si bien este set tecnológico aporta al rápido análisis de información desde distintas perspectivas, aún queda pendiente explicar e interpretar la dinámica de la comunidad íctica. Se espera responder a esta problemática aplicando técnicas de Ciencia de Datos que permitan descubrir patrones sobre la dinámica de la fauna íctica.

Finalmente, plasmar los hallazgos de manera simple, resumida y comprensible para el usuario.

RESULTADOS OBTENIDOS

Para comenzar con el proceso de extracción de conocimiento se han estudiado y comparado diferentes metodologías sobre Explotación de Información y Minería de Datos; luego las metodologías emergentes para Ciencia de Datos o la adaptación de alguna existente.

Se decidió realizar un análisis comparativo entre las metodologías ASUM y Educción de Requisitos para Proyectos de Explotación de Información. Ambas se basan en la metodología CRISP-DM y la finalidad del análisis entre ambas es tomar sus mejores cualidades. ASUM, planteada como un Método Unificado de Soluciones Analíticas se presenta como un Manual de Usuario en línea que permite navegar entre cada fase o proceso planteado e indica las tareas a realizar en cada una de ellas (orientada a su herramienta comercial SPSS). Educción de Requisitos realizada como una tesis doctoral, tiene como puntos fuertes una serie de plantillas entregables que son fáciles y comprensibles de presentar al cliente; además de mostrar ejemplos concretos sobre la aplicación de la misma.

Del análisis de metodologías se obtienen las siguientes cinco fases: 1- Definición del Proyecto. 2- Educción de Procesos de Negocio. 3- Educción de Datos de Procesos de Negocio. 4- Conceptualización del Negocio. 5- Especificación de Procesos de Explotación de Información.

De la aplicación de la fase 1 se obtuvieron cinco objetivos generales. A continuación se describen las tareas realizadas para resolver la premisa "Determinar e identificar características sobre Estructura de Peces y su relación con variables ambientales".

Para realizar las tareas prácticas se utiliza la herramienta RapidMiner Studio bajo una licencia "*Educational Edition*" en su versión 9.8.

En primer lugar, se tomó una muestra correspondiente a 1 (un) año de monitoreo con el fin de realizar las tareas de pre-procesamiento, los cuales incluyen limpieza de los datos, tratamiento de valores nulos, faltantes y calidad en general de los datos. Con la muestra tomada se trabajaron 25 variables tanto cualitativas como cuantitativas, de las cuales 8 representan datos ambientales y 17 datos bilógicos de los peces. En total la muestra se representa por 5380 registros.

Para comprender qué ocurre con los datos y cómo se relacionan entre sí, es necesario analizar una porción representativa del problema; para ello se ejecutó el proceso "Descubrimiento de Grupos" con el algoritmo *K-means*.

Una vez que la herramienta formó los grupos, se utilizó una matriz de correlación para identificar qué tan fuertemente correlacionadas se encontraban las variables y con qué atributos. Además se analizaron los pesos de cada variable para tener una noción sobre cuáles son las más representativas. Como resultado se pudo observar que las variables con mayor peso al momento de definir grupos fueron: peso y largo del pez (como principales atributos biológicos ícticos); temperatura, oxígeno y transparencia del agua (como principales atributos ambientales). Por otra parte, los pares de variables mayormente correlacionadas fueron altura-peso, altura-largo, largo-peso y temp_ambiente-temp_agua.

Teniendo este panorama general, se procedió con el análisis de cada partición (verificando si la correlación de los datos en general es correcta.). Por cada grupo conformado se identifica el atributo cluster -con el que fue catalogado en el paso anterior - y en base a este atributo se aplica el algoritmo de inducción para obtener "Reglas de pertenencia a cada grupo". Esta tarea se ejecutó con el algortimo *Decision Tree*.

Como resultado se obtuvo que las variables que definieron el agrupamiento fueron tres: peso, largo y temperatura del agua, en el orden indicado.

Actualmente la investigación para este trabajo se encuentra en la tercera fase de la metodología, donde se realizan tareas asociadas a la Educción de Datos de Procesos de Negocio. Estas tareas se subdividen en dos grandes actividades, por un lado el relevamiento de los datos del negocio y por el otro, el análisis de los repositorios de datos. En esta etapa se está analizando la incorporación de información a través de la transformación de datos existentes, por ejemplo: a partir de la fecha de pesca, transformar este dato a estaciones del año y con este nuevo valor, generar otras salidas que sirvan al cumplimiento de los objetivos del trabajo.

FORMACIÓN DE RECURSOS HUMANOS

El presente se lleva a cabo como Trabajo Final de la Maestría en Tecnologías de Información de la FCEQyN –UNaM y se conforma por tres integrantes: el maestrando, Lic. Cinthia Cuba de la FCEQyN – UnaM; la Dra. Paola Britos, perteneciente al Laboratorio de Informática Aplicada de la UNRN; y la Mgter. Gladys G. Garrido, Docente de categoría III (Sistema Nacional de Incentivo a la Investigación) perteneciente al Departamento de Biología de la FCEQYN – UNAM.