Revista Cubana de Ingeniería, 1(3), 61-65, 2010

Estimación de parámetros meteorológicos secundarios utilizando técnicas de minería de datos

Rosabel Zerquera Díaz
Correo electrónico:rzerquera@tesla.cujae.edu.cu
Ayleen Morales Montejo
Correo electrónico:ayleen@dcrhmail.cujae.edu.cu
Gil Cruz Lemus
Correo electrónico:gil@tesla.cujae.edu.cu
Alejandro Rosete Suárez
Correo electrónico:rosete@ceis.cujae.edu.cu
Instituto Superior Politécnico José Antonio Echeverría, Cujae, Ciudad de La Habana, Cuba

Resumen
El presente trabajo desarrolla un proceso de descubrir conocimiento en bases de datos (*Knowledge Discovery in DataBases*, KDD por su siglas en inglés) en el grupo de Medio Ambiente, del Instituto Superior Politécnico José Antonio Echeverría y en colaboración con el Centro de Gestión de la Información y Desarrollo de la Energía (CUBAENERGÍA) con el objetivo de obtener un modelo de datos para estimar el comportamiento de los parámetros meteorológicos secundarios a partir de datos de superficie. Se detallan algunos aspectos relacionados con la minería de datos y su aplicación en el entorno meteorológico; además, se seleccionan y describen la metodología CRISP-DM y la herramienta de análisis de datos WEKA. Se utilizan las tareas de selección de atributos y de regresión, la técnica de redes neuronales de tipo perceptrón multicapas y los algoritmos *CfsSubsetEval, BestFirst* y *MultilayerPerceptron.* Se obtienen modelos de estimación para los parámetros meteorológicos secundarios: *altura de la capa de mezcla convectiva*, *altura de la capa de mezcla mecánica* y *velocidad convectiva de escala*, necesarios para el estudio de los modelos de dispersión de contaminantes en la zona de la Cujae. Los resultados obtenidos constituyen un precedente para futuras investigaciones, así como para la continuidad de esta en su primera etapa.

Palabras clave: meteorología, estimación, parámetros meteorógicos secundarios, minería de datos, redes neuronales

Recibido: enero 2010 Aprobado: marzo 2010

INTRODUCCIÓN

En la actualidad, la automatización de las actividades de los negocios produce un flujo creciente de datos, pues incluso la información referente a acciones tan simples como una llamada telefónica o un test médico es almacenada en una computadora. Las empresas e instituciones se encuentran abrumadas por este crecimiento acelerado del tamaño y cantidad de datos. Es imprescindible convertir los grandes volúmenes de datos existentes en experiencia, conocimiento y sabiduría, formas que son útiles para la toma de decisiones y el desarrollo económico y social contemporáneo.

Tal es el caso del grupo de Medio Ambiente del Instituto Superior Politécnico José Antonio Echeverría Cujae, que cuenta con una estación meteorológica automática que recoge datos de superficie desde el 15 de abril del 2008, a los que no se les da ningún tratamiento o explotación, entre estos se encuentran: dirección y velocidad del viento, temperatura, humedad relativa, presión, entre otros. Estos datos son almacenados en soporte digital (con restricciones de capacidad), y procesados por la consola Vantage Pro2 de Davis, [1] que solo los muestra en forma numérica y predice el comportamiento de algunos de ellos para las próximas 12 horas. Mediante el software WeatherLink en

su versión 5.7 del 2006 [2] para Windows, se conecta la estación meteorológica a la computadora, posibilitando el intercambio de datos y su total almacenamiento; lo que permite plotear, analizar, exportar e imprimir los datos meteorológicos, así como configurar la estación y monitorear las alarmas.

El grupo se encuentra interesado en el análisis de la dispersión local de contaminantes gaseosos y de partículas en la zona de la Cujae; para ello es necesario poseer los valores de parámetros meteorológicos secundarios como: altura de la capa de mezcla convectiva, altura de la capa de mezcla mecánica y velocidad convectiva de escala, entre otros; actualmente no se cuenta con estos valores, pero pueden obtenerse mediante cálculos a partir de los parámetros meteorológicos primarios siguientes: dirección y velocidad del viento, temperatura exterior, humedad relativa, precipitaciones, presión barométrica y radiación solar; estos últimos son los datos de superficie que se recogen mediante la estación meteorológica automática.

Para la obtención de los parámetros meteorológicos secundarios, se propone la utilización del preprocesador meteorológico AERMET del sistema de modelos AERMOD, establecido por la Agencia de Protección Ambiental de los Estados Unidos (en inglés Enviromental Protection Agency, EPA). CUBAENERGÍA dispone del software Lakes Enviromental, una versión mejorada del AERMET de la EPA, que brinda entre otros, un potente entorno visual. Lakes Enviromental es propiedad de una empresa canadiense que limita su distribución y uso. El empleo de cualquiera de estas dos versiones de AERMET, requiere de datos de superficie y de sondeo. Las mediciones atmosféricas de sondeo de aire superior actualmente en Cuba no se realizan de forma sistemática, por lo que se desarrolló en CUBAENERGÍA la versión del AERMET: AERMET+, que simula el comportamiento vertical de la atmósfera a partir de datos de superficie (específicamente la altura de la capa de mezcla convectiva, y a partir de esta, la velocidad convectiva de escala y el gradiente de temperatura potencial por encima de la capa de mezcla). [3] Para la utilización de esta versión es necesario un fichero con los datos de superficie en un formato específico: HUSWO; este fichero se puede obtener a través de otro sistema desarrollado por CUBAENERGIA: SD_Aermet, que tiene como función la conversión de formatos.

La versión AERMET+ de CUBAENERGÍA, consume un tiempo de procesamiento considerable, una parte del cual se emplea en la preparación de los datos y la creación de ficheros de entrada para el sistema. Además, esta versión solo permite trabajar con datos horarios, por lo que si se poseen varias mediciones por hora, es necesario hacer un promedio vectorial, lo que puede implicar la pérdida de grados de exactitud en las mediciones.

Por lo expuesto anteriormente, se considera el cálculo de los parámetros meteorológicos secundarios, engorroso; y no resulta factible el procesamiento de los datos, lo cual imposibilita el análisis de la dispersión local de contaminantes. Con los valores que se obtienen mediante AERMET+, no se pueden obtener patrones de comportamiento de los datos. Además, no se cuenta con métodos y herramientas de procesamiento y análisis que le den sentido y utilidad a la información obtenida.

Como objetivo del trabajo se propone, obtener modelos de datos que permitan analizar las dependencias entre los parámetros meteorológicos, así como estimar los parámetros meteorológicos secundarios: altura de la capa de mezcla convectiva, altura de la capa de mezcla mecánica y velocidad convectiva de escala, respecto a los datos de superficie en la zona de la Cujae, utilizando técnicas de minería de datos.

MINERÍA DE DATOS (MD)

Entre las múltiples definiciones que identifican la MD se encuentra la siguiente:

Minería de datos o *data mining* es el conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar el conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con el objeto de predecir de forma automatizada tendencias y comportamientos y/o descubrir de forma automatizada modelos previamente desconocidos. [4]

En la actualidad, la aplicación de técnicas de MD en el campo de la meteorología se ha incrementado considerablemente. [5]

La estimación, es una tarea predictiva de gran importancia. Su meta es encontrar el valor numérico de una variable objetivo para objetos desconocidos. El objetivo en este caso es minimizar el error (generalmente el error cuadrático medio) entre el valor predicho y el real.

Las redes neuronales artificiales ofrecen los medios necesarios para revelar categorías comunes en los datos, debido a que son capaces de detectar y aprender complejos patrones y características dentro de los datos; así como para modelar de manera efectiva y eficiente, problemas grandes y complicados, de forma individual, o combinadas con otros métodos, para aquellas tareas de clasificación, agrupamiento y estimación.

Uno de los tipos principales de redes neuronales artíficiales existentes es el perceptrón multicapas empleado en el aprendizaje supervisado. Este constituye una red neuronal artíficial formada por múltiples capas que permiten (entre otros) modelar conjuntos de datos que son linealmente separados.

METODOLOGÍA Y HERRAMIENTA PARA ENFRENTAR EL PROCESO DE KDD

CRISP-DM (CRoss-Industry Standard Process for Data Mining) [6] es una metodología para MD que fue presentada por el consorcio CRISP-DM, encabezado por SPSS Inc. (Estados Unidos); esta se ha convertido en un estándar luego de ser liberada para su empleo y desarrollo por parte de la comunidad internacional.

WEKA, acrónimo de Waikato Environment for Knowledge Analysis (Entorno de Búsqueda de Conocimiento de Waikato), [7] constituye una extensa colección de algoritmos de máquina de conocimiento implementados en Java, útiles para ser aplicados mediante las interfaces o para embeberlos dentro de cualquier aplicación; fue desarrollada por un grupo de investigadores de la Universidad de Waikato, Nueva Zelanda. WEKA posee la licencia GPL para su libre distribución y es de código abierto.

TÉCNICAS DE MODELACIÓN

Las técnicas de modelado seleccionadas se muestran en la tabla 1, en esta se resumen los objetivos que se persiguen, las tareas planificadas y los algoritmos de WEKA que serán empleados.

Tabla 1
Técnicas de modelación

Objetivo	Tarea	Algoritmo
Obtener cuáles son los parámetros meteorológicos primarios que más influyen en los parámetros meteorológicos secundarios	Selección de atributos	CfsSubsetEval y BestFirst (métodos de evaluación y búsqueda respectivamente)
Estimar el valor de los parámetros meteorológicos secundarios a partir de los primarios	Estimación	MultilayerPerceptron

ANÁLISIS DE LA INFORMACIÓN DISPONIBLE

Los datos son recopilados mediante la estación meteorológica automática, y almacenados en formato digital por la consola Vantage Pro2 en ficheros con extensión ".wlk" que no pueden leerse directamente. Se utilizó el software WeatherLink, para exportar la información a ficheros planos en filas y columnas, con extensión ".txt". Estos ficheros se importan a un "libro" de Microsoft Excel donde se les da un posterior tratamiento.

De un total de 44 323 registros disponibles, comprendidos entre las fechas desde el 15 de abril de 2008 y 15 de abril de 2009, y 41 variables meteorológicas, y como resultado del proceso de selección, transformación y construcción de datos mediante AERMET+, así como limpieza, se obtiene un conjunto minable compuesto por 14 campos y un total de 8 784 instancias, que se resume en la tabla 2. En la referencia 5 se detalla la metodología empleada para el trabajo con AERMET+.

Los registros que tienen al menos un valor desconocido en sus atributos, han sido eliminados del conjunto, mediante la aplicación de un filtro de instancias. Debido a que los atributos pueden moverse en un rango tan amplio de valores, se eliminaron las instancias con valores extremos en sus atributos, las que presentaban una baja frecuencia de aparición y lejanía de la media numérica del grupo. Además, se normalizaron los atributos para evitar que los valores más alejados de la media numérica, pudiesen introducir errores en los modelos de estimación.

Tabla 2
Variables empleadas en el análisis

Nombre del atributo	Tipo
Día	Numérico
Mes	Numérico
Año	Numérico
Hora	Numérico
Dirección del Viento	Nominal
Velocidad del Viento	Numérico
Temperatura	Numérico
Humedad	Numérico
Presión barométrica	Numérico
Radiación solar	Numérico
Precipitación	Numérico
Altura de la capa de mezcla convectiva	Numérico
Altura de la capa de mezcla mecánica	Numérico
Velocidad convectiva de escala	Numérico

DESCRIPCIÓN DE LOS MODELOS OBTENIDOS

Para entrenar y probar los modelos de selección de atributos y de estimación, se emplean conjuntos distintos a fin de no sobrestimar su precisión. En este sentido se utiliza una validación cruzada (en inglés *cross-validation*) de 10 pliegues, la cual divide el conjunto de datos en 10 subconjuntos de forma aleatoria, y realiza 10 iteraciones, donde en cada una se reserva un grupo diferente para el conjunto de prueba y los restantes 9 para entrenar el modelo. [8]

Selección de atributos

Para la selección de atributos se realizaron 6 experimentos, 3 de ellos con los atributos normalizados y 3 sin normalizar, con los que se determinaron las variables fundamentales que influyen en los parámetros meteorológicos secundarios y su porcentaje de incidencia.

Posteriormente se seleccionaron aquellos atributos que contribuyen con la estimación entre 70 y 100 %, por considerarse un valor significativo.

En la altura de la capa de mezcla convectiva influyen en un 100 %, la hora del día, la velocidad del viento y la humedad.

También en la altura de la capa de mezcla mecánica, influyen en un 100 % el mes, la velocidad del viento y la radiación solar.

En la velocidad convectiva de escala influyen en un 100 %, la temperatura, la humedad y la radiación.

En todos los caso los resultados obtenidos coinciden con la opinión de los expertos. En la referencia 5 se muestran los resultados de los modelos de selección de atributos para todas las variables.

Estimación numérica

En el caso de la estimación numérica con las redes neuronales, se realizaron 12 experimentos, 4 por cada variable secundaria a estimar, con los atributos normalizados y sin normalizar, y teniendo o no teniendo en cuenta, la selección de atributos del proceso anterior, con fines de comparar resultados y seleccionar los modelos con los que se obtengan los mejores resultados. La tabla 3 refleja los valores del coeficiente de correlación y del error medio absoluto para la altura de la capa de mezcla convectiva en cada uno de los experimentos. En la referencia 5 se muestran los resultados para el resto de las variables meteorológicas secundarias.

Tabla 3
Estimación de la altura de la capa de mezcla convectiva

Parámetros	Con todos los parámetros primarios	Tomando en cuenta la selección de atributos
Sin normalizar	Correlación: 0,882 7 Error: 145,631 1	Correlación: 0,848 7 Error: 174,446
Normalizados	Correlación: 0,882 7 Error: 0,067 5	Correlación: 0,848 7 Error: 0,080 8

Los modelos obtenidos teniendo en cuenta la normalización reportaron resultados útiles para ser aplicados por los expertos en la estimación de las variables meteorológicas secundarias. Ambos modelos presentan un alto valor de correlación entre sus variables y un bajo error en la estimación. En dependencia de la exactitud con la que se desee trabajar, las variables meteorológicas primarias que se posean, el tiempo para el análisis con que se cuente, se puede utilizar el modelo con todos los parámetros meteorológicos primarios o, por el contrario, el modelo más simple que toma en cuenta la selección de atributos.

Los expertos de CUBAENERGÍA proponen para la evaluación de los modelos de estimación el análisis de la desviación fraccional, dada por la ecuación (1), donde se consideran aceptables los valores obtenidos si están dentro del rango [-0,67, +0,67], donde $V1$ es el valor predicho y $V2$ es el valor real.

$$\frac{V1 - V2}{v1 + v2} \cdot 2 \qquad ...(1)$$

En la tabla 4 se muestra el porcentaje de instancias que está fuera de rango para el caso del modelo de estimación con los datos normalizados, considerando la selección de atributos de la altura de la capa de mezcla convectiva; este valor es relativamente bajo, lo que indica que la mayoría de las instancias presentan valores dentro del rango permitido, pudiendo emplearse el modelo en la estimación de este parámetro meteorológico. En la referencia 5 se muestran los resultados para el resto de las variables meteorológicas secundarias.

Tabla 4
Desviación fraccional de la altura de la capa de mezcla convectiva

Parámetro	Porcentaje de instancias fuera de rango
Altura de la capa de mezcla convectiva	3,57% de las 349 2 instancias

CONCLUSIONES

Mediante la aplicación de la selección de atributos, se obtuvieron las variables meteorológicas primarias que más influyen en las variables meteorológicas secundarias: altura de la capa de mezcla convectiva, altura de la capa de mezcla mecánica y velocidad convectiva de escala. A partir de la selección de atributos se obtuvieron además modelos de regresión basados en redes neuronales del tipo perceptrón multicapas que permiten estimar los valores de los parámetros meteorológicos secundarios, los que presentan un coeficiente de correlación alto, un error cuadrático medio pequeño y con bajo porcentaje de instancias fuera de rango, siendo aceptados y válidos para la estimación.

RECONOCIMIENTOS

Los autores desean agradecer a los integrantes del Grupo de Medio Ambiente de la Cujae y a la máster Leonor Turtós Carbonell de CUBAENERGÍA por su colaboración, principalmente en temas relacionados con la meteorología.

REFERENCIAS

1. *Vantage Pro2 Console Manual*. California: Davis Instruments Corp, 2006.

2. *WeatherLink 5.7 Help. Davis Instruments Corp*. California: 2006.

3. TURTÓS CARBONELL, L. *Proyecto Programa Ramal Nuclear. Sistema de Modelos AERMOD para dispersión local de contaminantes atmosféricos. Salida 1/2007: Ampliación de la propuesta de Guía de modelación de la*

dispersión local de contaminantes gaseosos y partículas con el Sistema de modelos AERMOD. Ciudad de La Habana, Cuba: 2007.

4. **PIATETSKI-SHAPIRO, G. and FRAWLEY, W. J.** *Knowledge Discovery in Databases. AAAI/MIT Press*. 1991.
5. **ZERQUERA, R.** *Predicción de Parámetros Meteorológicos Secundarios: Altura de la capa de mezcla convectiva, altura de la capa de mezcla mecánica y velocidad convectiva de escala, en la zona de la Cujae*, utilizando técnicas de Minería de Datos. Tesis de diploma, Instituto Superior Politécnico José A. Echeverría, Ciudad de La Habana, La Habana Cuba. 2009.
6. **CHAPMAN, P.** *et al*. CRISP-DM 1.0:Step-by-step data mining guide. USA: 2000. CRISP-DM Consortium. SPSS Inc.
7. **GARCÍA MORATE, D.** *Manual de WEKA*. 2005.
8. **MOLINA LÓPEZ, J. M. y GARCÍA HERRERO, J.** *Técnicas de análisis de datos. Aplicaciones Prácticas utilizando Microsoft Excel y WEKA*. Universidad Carlos III, Madrid.

AUTORES

Rosabel Zerquera Díaz
Ingeniera Informática, Adiestrada, Vicerrectoría de Investigaciones y Posgrado, Instituto Superior Politécnico José Antonio Echeverría, Cujae, Ciudad de La Habana, Cuba

Ayleen Morales Montejo
Ingeniera Informática, Adiestrada, Dirección de Recursos Humanos, Instituto Superior Politécnico José Antonio Echeverría, Cujae, Ciudad de La Habana,Cuba

Gil Cruz Lemus
Ingeniero Químico, Doctor en Ciencias Técnicas, Profesor Auxiliar, Vicerrectoría de Investigaciones y Posgrado, Instituto Superior Politécnico José Antonio Echeverría, Cujae, Ciudad de La Habana, Cuba

Alejandro Rosete Suárez
Ingeniero Informático, Doctor en Ciencias Técnicas, Profesor Titular, Facultad de Ingeniería Informática, Instituto Superior Politécnico José Antonio Echeverría, Cujae, Ciudad de La Habana, Cuba

Estimation of Secondary Meteorological Parameters Using Mining Data Techniques

Abstract
This work develops a process of Knowledge Discovery in Databases (KDD) at the Higher Polytechnic Institute José Antonio Echeverría for the group of Environmental Research in collaboration with the Center of Information Management and Energy Development (CUBAENERGÍA) in order to obtain a data model to estimate the behavior of secondary weather parameters from surface data. It describes some aspects of Data Mining and its application in the meteorological environment, also selects and describes the CRISP-DM methodology and data analysis tool WEKA. Tasks used: attribute selection and regression, technique: neural network of multilayer perceptron type and algorithms: CfsSubsetEval, BestFirst and MultilayerPerceptron. Estimation models are obtained for secondary meteorological parameters: height of convective mixed layer, height of mechanical mixed layer and convective velocity scale, necessary for the study of patterns of dispersion of pollutants in Cujae's area. The results set a precedent for future research and for the continuity of this in its first stage.

Keywords: meteorology, estimation, secondary meteorological parameters, data mining, neural networks