LA MINERIA DE DATOS EN UN ENTORNO DE APLICABILIDAD EN LAS CIENCIAS SOCIALES[1]

Gerardo A. Colmenares L.
Instituto de Investigaciones Económicas y Sociales
Universidad de Los Andes. Mérida
gcolmen@ula.ve,
hhtp://webdelprofesor.ula.ve/economia/gcolmen

He leído en el programa del simposio que la presentación que me corresponde sería sobre LA MINERIA DE DATOS EN UN ENTORNO DE APLICABILIDAD EN LAS CIENCIAS SOCIALES y en esto se fundamentará mi charla. Antes de comenzar debo agradecer la gentileza del comité organizador por invitarme a contarles nuestras experiencias en el área de las ciencias sociales. Sin embargo, quiero enmarcarlo en el contexto de las ciencias económicas y sociales. Creo que no soy el expositor indicado para hablarles de una temática que considero son mejores sus propios profesores. Más aún, hay protagonistas en el tema con mayor experiencia que son más prolíficos en aplicaciones del área. Yo, por mi lado, más o menos desde 1997 he caído en este terreno indirectamente a través de proyectos de investigación relacionados con este tema. A propósito, en ese mismo año, un profesor de origen africano, Geoffry Okogbaa, me comentó que lo que hacía en mis trabajos era minería de datos. El término se estaba acuñando en esos años. Después de este paréntesis, comencemos.

Cualquier decisión que se tome como consecuencia de hacer uso de modelos que permitan controlar, hacer seguimiento o evaluar fenómenos físicos (sector industrial y/o manufacturero) o socio-económicos (sector financiero, bancario, laboral, médico), puede estar

[1] Presentado en el simposio de Ingeniería de Sistemas el día 03 de Noviembre de 2007.

afectada por la dinámica a que están expuestos los datos. En efecto, nuevas necesidades de información, nuevos medios de adquisición de datos, nuevos actores en un fenómeno en particular, pueden alterar la estructura y disposición de los datos al sufrir cambios, omitiendo o incluyendo series de datos sobre el mismo actor o sobre nuevos actores. Estos cambios generan nueva información que acumulada a la disponible históricamente, se torna en compleja para analizar o producir modelos de pronóstico que contribuyan a mejorar los modelos que permitan controlar, evaluar o seguir el o los fenómenos señalados anteriormente. Aunado al esfuerzo en producir nuevos datos y a la complejidad de interpretarlos en la historia, existe un costo latente que está directamente vinculado a la omisión de no usar la información histórica que se ha estado acumulando en archivos que ocupan un espacio sin aportar un valor agregado como producto de la información que no fue observada en su momento, pero que está disponible en nuevas estructuras de datos nunca vistas. Es decir, nuevo conocimiento oculto en patrones no identificados al momento de su recolección o simplemente de poco interés para el objetivo original de los datos.

Con esta premisa en mente, los medios tecnológicos actuales y el aporte científico en el área de manejo de datos, han hecho que el Instituto de Investigaciones Económicas y Sociales (IIES) haya visto la necesidad de diseñar metodologías de análisis y exploración de datos que combinadamente con la construcción de modelos de pronóstico y/o clasificación puedan contribuir en la toma de decisiones de problemas socio-económicos tales como la banca, el gasto familiar, calidad de la vivienda, situación laboral. En estos casos, permanentemente se deben evaluar las variables explicativas que inciden en la estimación de los índices cualitativos o cuantitativos para determinar tendencias, clases o proporciones que sea de sumo interés en las áreas antes señaladas. En

otras palabras, hacer Minería de Datos para reconocer patrones de comportamiento en las variables o en los resultados.

Estas metodologías, fundamentalmente, exploran los datos usando técnicas paramétricas y no paramétricas para la clasificación y predicción de los valores. De ahí, que estas mediciones y estimaciones se calculan mediante la construcción de modelos de pronóstico y clasificación, usando técnicas *estadísticas multivariantes y de inteligencia artificial,* lo que permite la consideración de nuevas variables o factores a estimar; esto es, descubrir comportamientos que claramente definen los patrones que dan respuesta a las necesidades de información. En otras palabras, construir nuevo conocimiento desde grandes volúmenes de datos representado a través de información nunca antes vista.

Se han construido modelos donde el preprocesamiento de datos es un factor determinante para la generación de variables cualitativas y cuantitativas. La identificación de estas variables y la aceptable calidad de los datos que ellas definen, son usadas en modelos híbridos de origen *paramétrico* y *no paramétrico,* conducentes al descubrimiento de conocimiento con nuevos patrones de comportamiento; tal es el caso de gestión del riesgo en el sistema bancario medido a través de categorías de calificación. Algunas de las técnicas involucradas para la conformación de los modelos híbridos son: a) Análisis de Componentes Principales como mecanismo de reducción de variables, b) Análisis Discriminante, como preprocesador de la variable dependiente, c) Análisis de Supervivencia, como método de pronóstico de crisis a través de fusiones bancarias, d) Modelos en Ecuaciones Estructurales para la construcción de variables no observables de tipo de riesgo bancario, e) Modelos en Lógica Difusa para determinar la calificación del riesgo bancario en la banca venezolana, f) Modelos de clasificación usando Máquinas de Vectores Soporte (VSM) aplicados al comportamiento del

consumo de la familia venezolana, g) Modelos de Función de Base Radial RBF para la determinación del déficit y calidad de la vivienda en Mérida, h) Modelos supervisados para pronóstico y clasificación de Redes Neuronales, fundamentalmente la función logística con propagación del error hacia atrás y la función de base radial.

Tal como lo mencioné anteriormente, nuestra experiencia en las ciencias sociales, se cristaliza apenas cinco años atrás, dando inicio de ese modo al Grupo Banca, cuyo objetivo fue el de construir modelos prototipos, usando técnicas emergentes que permitieran controlar y monitorear la banca nacional descubriendo situaciones de riesgo mediante sistemas de alerta temprana. De este modo, se iniciaron los primeros perfiles de proyectos de investigación conduciendo a la construcción de modelos híbridos, al combinar aplicaciones de análisis de datos multivariantes con modelos estadísticos o de inteligencia artificial que permitieran hacer pronóstico o clasificación. Comentaré algunos de ellos, tratando de ser lo más escueto posible y respondiendo en cada uno las preguntas: qué se hizo, cómo se hizo y para qué se hizo. Cualquier interés en alguno en particular, pueden ponerse en contacto con el IIES en la Facultad de Ciencias Económicas y Sociales.

Primero les comento sobre la aplicación de *Análisis de Supervivencia.* Este trabajo girará en torno a las eventuales crisis que se manifestaron en fusiones entre bancos comerciales y/o universales. El fenómeno de las crisis cambiarias en Latinoamérica, han afectado diversos países. Particularmente, ha estado presente en México, Brasil, Chile, Colombia, Venezuela y Argentina. Estas crisis se originaron primordialmente ante la ausencia de evaluación de las condiciones financieras de los bancos y la débil legislación en materia bancaria. A partir de los años 70 se comienza a introducir modelos útiles para la asignación eficiente de los recursos de supervisión. Dichos modelos permiten entre otras cosas, dar una pauta de la evolución de una entidad en el período de tiempo

mediante inspecciones en el sitio, determinar los principales factores de riesgo que asume la banca y realizar una mejor medición de ellos. Estos modelos se conocen como "*modelos de alerta temprana para el sistema financiero"*: *predecir la probabilidad de cambios en las condiciones de solvencia y liquidez de los bancos en estudio*. Los modelos *probit* han demostrado buen desempeño en este sentido, pero sólo permiten estimar la probabilidad de que un banco cambie de "***estado***" pero no informan sobre el tiempo que las entidades vulnerables podrían demorar en demostrar problemas. Sin embargo, la idea es conocer *cuándo* es probable que la entidad financiera cambie de ***estado***. Dicho de otra forma, la variable aleatoria de interés, es el lapso de tiempo que tarda la entidad en cambiar de estado. Específicamente interesa el tiempo qué tarda en ocurrir el cambio de estado y cuales son las variables influyentes para producir este cambio. En este estudio, para el caso venezolano, un cambio de estado se define como la fusión de entidades financieras, de una en crisis con una solvente. A partir de la crisis de 1994, la nueva Ley de Bancos promueve la fusión como medio para evitar las quiebras, es decir período de crisis.

En conclusión, determinar tempranamente la crisis manifestada por la fusión y que variables financieras influyeron; en consecuencia, descubrir el patrón de comportamiento de la crisis de presumible quiebras. La técnica que permite describir el comportamiento de datos que corresponden al tiempo o duración desde un origen bien definido hasta la ocurrencia de un cambio de estado o punto final se denomina "*análisis de supervivencia*" introducido en 1997 por Klein – Moeschberguer. Para este propósito se realizó un análisis de supervivencia clásico, una estimación por el método de Kaplan-Meier y un ajuste a través del modelo de regresión de Cox, utilizando los datos de la banca comercial venezolana entre 1996 y 2004. La fusión bancaria es el evento de interés para el cambio de estado. Se logran identificar cuatro razones de

riesgo como variables indicadoras de un patrón de comportamiento de crisis y que son consistentes con la revisión teórica de los modelos de alerta temprana. Este trabajo fue realizado por María Alejandra Ayala como tesis de grado para optar al título de Master en Economía.

El estudio ha tenido cierta continuidad al proporcionar en un nuevo proyecto una función de riesgo que permita identificar en cualquier momento la situación de solvencia o crisis de un banco en particular. Actualmente, se está haciendo ensayos donde los parámetros que componen el modelo de regresión de Cox puedan ser calculados mediante las técnicas de lógica difusa, mediante redes Neurodifusas, a los fines de capturar mayor información presente en la no linealidad inherente al dominio del problema. Este estudio se está realizando para un caso particular de producción de aluminio.

Otro proyecto, que en realidad debería habérselos comentado al principio, debido a que este marca el inicio a esta serie de trabajos, fue al que dimos por llamar ALGORITMO SEGUIDO PARA EL PRE-PROCESAMIENTO DE DATOS DE LA BANCA 1996-2004 Y CONSTRUCCIÓN DE LOS MODELOS NO LINEALES. Título ambicioso para el propósito final, pero, sin embargo, ayudó a configurar una idea global de nuestras pretensiones en el área bancaria, indicando un esquema de procura de información sobre la calificación del riesgo en la banca, a partir de los datos disponibles, mediante balances de publicación y estado de resultados de la banca nacional. La preparación de los datos de la banca comercial en el período señalado fue producto de un trabajo previo a la construcción de los modelos prototipo que originó un trabajo de reconocimiento conceptual de los datos para su posterior uso. Este trabajo de preprocesamiento de datos fundamentalmente se basó en los tres puntos siguientes: *Identificación de las Variables originales, Construcción de Variables latentes, Disposición de los Datos*. En cuanto a las variables originales, se resolvió transformarlas, debido a que sus

valores no ofrecían mayor información y resultaban de difícil lectura por la diversidad de escalas de medida presentes en ellos. Se asume, entonces, hacer la primera transformación e introducir nuevas variables denominadas razones financieras que no son más que combinaciones lineales entre las variables originales. La evaluación, consistencia y pertinencia de estas razones financieras fue fundamentada mediante referencias previas, tales como SUDEBAN, metodología de medición del riesgo usando CAMEL.

Otra alternativa de descubrimiento de nuevo conocimiento fue el que vamos a desarrollar en las siguientes líneas, mediante la aplicación de Modelos en Ecuaciones Estructurales. Las instituciones bancarias pueden verse expuestas a diversos tipos de riesgo: Crédito, Mercado, Operacional, Liquidez, Legal o Reputación. En este proyecto se construyó un conjunto de indicadores propios del Riesgo de Crédito, de Liquidez y Operacional, usando la teoría del Modelado con Ecuaciones Estructurales, demostrando a su vez, la aplicabilidad del análisis exploratorio de datos y las técnicas de clasificación, previo a la construcción de los modelos. En dichos modelos se relacionó el entorno macroeconómico con cada tipo de riesgo, lo cual permitió conocer el nivel de incidencia que tienen las principales variables macroeconómicas sobre cada uno de ellos. Estas variables fueron fundamentalmente, Tipo de cambio, Reservas Internacionales, IPC, Tipo de Cambio y Liquidez Monetaria, Tasa de Interés Activa Pasiva, Inflación, PIB. Seguidamente, se evaluó a través de técnicas de Control de Calidad Multivariado la situación de algunas de las entidades bancarias del país, en el periodo 1997-2004, respecto a cada tipo de riesgo. De este modo, se desarrolló un análisis cluster para clasificación de los bancos con el objeto de obtener dentro de la Banca Universal y Comercial, un grupo de bancos de comportamiento homogéneo en el conjunto de las variables (los indicadores), tal que sus diferencias no interfieran en los resultados de

los modelos. Por otro lado, se realizó un análisis cluster para variables, con el objetivo de obtener una primera clasificación de los indicadores, apoyado en la correlación que existe entre ellos y posteriormente agruparlos según el tipo de riesgo que les corresponda, de este modo se pueda dar inicio a la construcción del MEE. El MEE es la herramienta central de este estudio. El uso de esta técnica se justificó en el hecho de que se necesitó verificar si los indicadores financieros asignados a cada tipo de riesgo de manera teórica y mediante el análisis cluster, realmente constituyen un conjunto de medidas válidas para describirlos. Así, cada tipo de Riesgo: de Crédito, Liquidez y Operacional, puede ser considerado variables latentes, que necesita para su medición un conjunto de medidas que permitan analizarlos. Al mismo tiempo, las variables macroeconómicas, pueden considerarse también variables latentes medidas a través de un conjunto de indicadores. Se desea entonces cuantificar el nivel de incidencia sobre los tipos de riesgo.

Es así como el desarrollo de esta investigación permitió proponer una metodología alternativa para evaluar los diversos tipos de riesgo a los cuales están expuestas las entidades bancarias, enfocada en el análisis de indicadores propios de cada tipo de riesgo. Este estudio permitió probar la utilidad del análisis exploratorio de datos y las técnicas de clasificación, previo a la construcción del modelado con ecuaciones estructurales. El AED es empleado para detectar anomalías teóricas, mejorar la calidad de los datos y probar los requerimientos básicos de los MEE y el Análisis Cluster para obtener una visión exploratoria de la forma como se agrupan las variables. La utilidad de realizar estos dos análisis previos radica en obtener modelos "confiables" que no se vean alterados por comportamientos en los datos, al no responder a la realidad de la población considerada. Si se tiende a desarrollar un MEE sin la previa revisión y asignación de variables, estos podrían dar un falso ajuste, si no se conocen si los individuos que componen la muestra

responden a comportamientos similares en el conjunto de las variables, si los datos son confiables desde el punto de vista teórico, si la asignación teórica de indicadores a variables no observables se corresponde con los datos o si se cumplen los requerimientos mínimos establecidos por el modelo. Finalmente, al aplicar el Control de Calidad Multivariado ha permitido probar la utilidad que la técnica presenta al evaluar la situación de las entidades bancarias ante cada tipo de riesgo. En una situación real, dicha aplicación debe enfocarse en la evaluación continua de los indicadores, determinando si en periodos inmediatos anteriores se presenta una situación fuera de control y tomar medidas correctivas para que el suceso no se repita en periodos sucesivos. Dicha técnica podría aplicarse de forma univariante si el interés radica en evaluar una situación específica descrita por un indicador o de forma conjunta, aplicando el control de calidad multivariado, para evaluar el proceso que desencadena un conjunto de indicadores en particular. Se debe considerar además la definición apropiada de los límites de control en cada caso, lo cual representaría un estudio concreto de la situación que se evalúa. En consecuencia, los objetivos se alcanzaron mediante la aplicación del MEE ya que, como se señaló, con dichos modelos son dos las hipótesis que pueden modelarse: la que analiza las características de validez y confiabilidad de los indicadores que miden a las variables latentes y la que recogen las relaciones entre las variables latentes. Es decir, nuevo conocimiento a través de variables no observables que fueron medidas y establecidas sus relaciones. Este trabajo es una tesis de pregrado en Estadística presentada por Zuleima Andreina Durán.

Otro modelo prototipo que se ensayó con muy buenos resultados fue el que concierne al uso de las técnicas de Inteligencia Artificial (Lógica Difusa) en la Clasificación/Predicción del Riesgo Bancario. Igual que en los casos anteriores, esta investigación abarcó el estudio de los Riesgos de Crédito, Liquidez y Operacional y de igual modo, la fuente de datos

fueron los Balance de Publicación y Estado de Resultados de los bancos considerados. A través del instrumento más usado para realizar los análisis financieros como lo son las Razones Financieras, se midió su eficacia y comportamiento de la banca mediante el grado de liquidez, de rentabilidad, el apalancamiento financiero, cobertura y todo lo que tenga que ver con su actividad. De este modo las Razones Financieras fueron utilizadas como variables de entrada a los modelos difusos creados para medir y evaluar el desempeño de los bancos considerados. Por otro lado, uno de los mecanismos de supervisión tipo rating ampliamente utilizado en muchos países es conocido como metodología CAMEL, la cual proporciona un marco para evaluar cinco aspectos fundamentales de la calidad financiera bancaria, como son: Adecuación del Capital, Calidad de los Activos, Gestión Administrativa, Ganancias y Liquidez [Buniak, 2002]. La revisión de esta metodología resultó útil en el desarrollo de los modelos de calificación. Los modelos de calificación fueron creados utilizando lógica difusa, la cual se define como una lógica multivaluada tomando infinitos valores en el intervalo [0,1]. Estos modelos de calificación están representados por Sistemas de Inferencia Difusos (FIS), método que interpreta los valores en un vector de entrada y basado en un conjunto de reglas, asigna valores al vector de salida.

De los dos medios para crear el Sistema de Inferencia Difuso como lo es la adquisición de conocimiento a través de expertos para la conformación de las reglas o, el uso de algoritmos de agrupamiento difuso (Substractive Clustering) para que se construya automáticamente el conocimiento a través los agrupamientos formados por las combinaciones entrada-salida. La construcción de los modelos difusos de calificación utilizando el algoritmo de clasificación difusa “subtractive clustering”, demostró la efectividad y alta capacidad de adaptación de esta técnica al tipo de problemas que se está modelando. Sin embargo,

se pudo observar que a pesar de su capacidad para construir rápida y fácilmente sistemas de inferencia difusos de alta precisión, presentó dificultad en la revisión exhaustiva que se debe hacer de los distintos modelos creados, una vez que se comienza a variar el radio de influencia de los conglomerados. La mayoría de los resultados obtenidos en los sistemas de inferencia difusos, presentan errores muy bajos en entrenamiento y prueba, lo cual generó calificaciones similares a las de CAMEL. La herramienta formulada demostró su capacidad y confiabilidad de desempeño en la calificación y clasificación del riesgo financiero en la banca: Mostró un desempeño similar a CAMEL. Otra ventaja demostrada por este prototipo fue la de crear conocimiento no disponible al tener la capacidad de evaluar en conjunto, los indicadores del riesgo, contrario a lo realizado convencionalmente donde los indicadores son analizados individualmente sin utilizar un método de integración. Posiblemente el modelo de inferencia difusa hubiese tenido una mayor fortaleza en los resultados de la determinación del riesgo financiero si el conjunto de variables macroeconómicas hubiesen sido consideradas como entradas adicionales al sistema FIS, reagrupadas tal como se realizó en el trabajo que se mencionó anteriormente. Este aporte implicaría un nuevo prototipo que similar al construido, muy bien podría establecer medios alternativos para una mayor exploración de conocimientos: variables macroeconómicas y cambio en el enfoque constructivo del sistema FIS (Mamdani en lugar de Sugeno). Este trabajo fue una tesis de grado en ingeniería de sistemas concluida durante este año por Carlos Martínez .

Otro proyecto que se está finalizando como tesis de postgrado en el área de Estadística, es la aplicación de los datos referentes a la banca transformados totalmente en dos conjuntos nuevos: las variables de entrada reducidas mediante ACP a un grupo nuevo de variables latentes que aporten el máximo de información de la que se tiene en las variables originales al ser transformadas previamente a razones

financieras y como variable de salida una de origen categórico originada por el primer componente principal calculado también mediante ACP, para la creación de la nueva variable categórica. Esta variable de salida se construyó cambiando su condición de cuantitativa a cualitativa categórica, aplicando un criterio de partición (se seleccionaron cuatro para garantizar valores extremos y moderados) constante a cada una de las variables tal que sean tomados en cuenta desde los casos mas pesimistas hasta los más optimistas.

Los valores umbrales definitivos para cada variable serán analizados individualmente a fin de categorizar correctamente en los rangos las variables de entrada. Estos nuevos conjuntos de variables son la fuente de datos para realizar pronósticos comparativos en cada uno de los modelos no lineales híbrido ACPL-RNA(BP) como en ADL-RN(RBF). Este proyecto fue uno de los que iniciaron el enfoque de modelos exploratorios y de predicción-pronóstico. El algoritmo que está ensayando para su desarrollo definitivo es el siguiente: a) selección de las muestras de entrenamiento-prueba y de generalización. Se empleó el muestreo estratificado donde cada bloque de datos para cada año, representa un estrato. La selección se hace proporcionalmente a su tamaño de manera aleatoria y sin repetición; b) separación de los conjuntos en la muestra para la construcción del modelo y el conjunto complementario para la evaluación de la capacidad de generalización del modelo; c) construcción del modelo de redes neuronales usando el algoritmo de retro-propagación del error, los ejemplos de entrada se corresponde con la muestra seleccionada de las variables transformadas y los ejemplos de salida con la variable categórica correspondiente a cada observación de la muestra seleccionada; d) verificación de la capacidad de generalización usando ejemplos del conjunto complementario; e) medición del nivel de desempeño individual; f) replicar el modelo con nuevas muestras pero idéntica topología de red

para medir el nivel de desempeño del diseño asumido, y en consecuencia la consistencia alcanzada en el entrenamiento y en la generalización para determinar la robustez.

A raíz de de estas experiencias adquiridas con el conjunto de datos para la banca, se iniciaron otros proyectos similares pero con otros conjuntos igualmente importantes en el área socio-económica. Ellos son las encuestas EHPM y ENPF. Ambas encuestas tienen su propia periodicidad de recolección y, por ejemplo, la segunda es aprovechada fundamentalmente para calcular el IPC correspondiente al país y las ciudades más importantes. En consecuencia, se iniciaron proyectos de exploración de datos y pronóstico usando como fuente de datos estos conjuntos que igualmente, mantiene su historia. En el caso de las ENPF, se está desarrollando un proyecto para identificar patrones de consumo de los venezolanos, introduciendo el uso de las Máquinas de Vectores Soporte como técnica exploratoria de clasificación en conjunto no linealmente separables, a partir de observaciones obtenidas de la II y III Encuestas Nacionales de Presupuestos Familiares (1997 al 1998 y 2004 al 2006), en algunas regiones del país. La importancia de la investigación radica en la utilización de técnicas de análisis no lineales, las cuales se caracterizan por la incorporación del principio de Minimización de Riesgo Estructural, que en algunos casos ha demostrado ser superior al principio tradicional de Minimización de Riesgo Empírico, empleados por redes neuronales y otros métodos lineales convencionales. Se espera que el uso de las Máquinas de Vectores Soporte, como técnica para la clasificación de datos, permita una mejor precisión en la identificación de patrones de consumo de las familias venezolanas. Igual que en los casos anteriores, existe una fase de preprocesamiento para preparar los datos y adecuarlos a los requerimientos del modelo.

Consiste en trabajar con los datos a fin de detectar las características de los valores tales como los atípicos y los faltantes. Además, se organizan los registros y se hace una limpieza de los datos al hacer un reconocimiento de problemas de multivariabilidad al aplicar técnicas de ortogonalización, tal como ACP. Posteriormente, se procede a realizar un análisis descriptivo de los mismos. Una de las características de las ENPF hasta el año 2006, fue que su diseño de selección impedía hacer inferencia en subconjuntos de muestras, por ejemplo estados, debido a que su representatividad era solamente nacional. Sin embargo, para el estado Mérida, el IIES aplica un factor de expansión a la muestra original, obteniendo de este modo, una muestra representativa para el estado. Igualmente para los estados Táchira y Trujillo. Esto permitió definir un caso particular de estudio, al hacer los pronósticos respectivos.

La selección de muestras se está haciendo con base en los factores de expansión en dos conjuntos: a) Entrenamiento y pruebas, y b) Capacidad de pronóstico o de generalización de los datos. En la etapa de entrenamiento se estima la mejor función clasificadora considerando algún criterio (por ejemplo, el error de clasificación) en el conjunto de entrenamiento. En la etapa de prueba se valida la efectividad del modelo respecto de datos no utilizados en el entrenamiento. Después de una selección previa de variables que incidan sobre el consumo y adicionalmente, se haya eliminado cualquier vestigio de multicolinealidad entre las variables seleccionadas, se empleará una técnica de análisis no lineal relativamente nueva como proceso de separación no lineal y de ahí, clasificador de patrones que indican nuevo conocimiento a través de la MVS para identificar y describir los patrones de consumo de los venezolanos. La técnica como tal, se caracteriza por la incorporación del principio de Minimización de Riesgo Estructural, que, en algunos casos, ha demostrado ser superior al principio tradicional de

Minimización de Riesgo Empírico, empleado por Redes Neuronales y otros métodos lineales convencionales. Este trabajo está en desarrollo y esperamos que sirva de referencia para otros conjuntos de datos que actualmente estamos usando. Se aplicará un procedimiento que comprometa a la técnica a establecer resultados satisfactorios en el descubrimiento de nuevo conocimiento y que se respete las mínimas condiciones exigidas por la metodología. Un requisito básico para aplicar con éxito las MVS a un problema real es la elección de una función de núcleo adecuada, que debe reflejar el conocimiento a priori sobre el problema. A veces, los ejemplos de aprendizaje no son linealmente separables y focalizarse demasiado en todos los ejemplos de aprendizaje puede comprometer seriamente la generalización. Es decir un sobreentrenamiento y por tanto un sobreajuste del modelo.

Fundamentalmente, los pasos a seguir para la utilización de las MVS son los siguientes: Transformar los datos a un formato de *software* para clasificación con MVS; Elección de la función de núcleo o *Kernel;* Selección de parámetros asociados al *Kernel y* la selección del parámetro *C* el cual es conocido como el parámetro de la regularización. El ajuste del parámetro *C* puede hacer un balance entre la maximización del margen y la violación de la clasificación; Realizar el entrenamiento de las MVS con las muestras y parámetros seleccionados; Realizar las pruebas para medir el desempeño tanto para clasificación como para medir la capacidad de pronóstico. Este proyecto lo está realizando actualmente un estudiante de esta escuela de ingeniería de sistemas y esperamos que pronto obtengamos ya resultados definitivos.

De igual modo sucede con otros algunos proyectos que esperamos se cristalicen para el próximo año. Tal es el caso, de Pronóstico del Déficit de Viviendas en el Estado Mérida a través de RNA, haciendo uso de la EHPM en su componente de Vivienda. Este trabajo ya está en fase preparatoria de muestras de entrenamiento y pruebas para realizar

clasificaciones mediante el uso de las redes RBF. Debido a que las encuestas son colectadas semestralmente y no han mantenido la regularidad del encuestado presente en la encuesta y la muestra ha sufrido cambios en su tamaño, se cuenta con un archivo histórico suficientemente grande que ha originado la separación en conjuntos que faciliten el análisis de acuerdo a observaciones repetidas. Sin embargo, bajo el esquema de transversalizar los datos de acuerdo a series de datos panel, se podría describir series temporales multivariantes de observaciones que todavía no las hemos analizado. Estos datos están más orientados hacia el análisis de la información transversal (son anchos pero cortos). Un aspecto fundamental es la heterogeneidad entre las unidades analizadas. Los efectos temporales se interpretan como transiciones o cambios de estado discretos. La ventaja que ofrecen los Datos Panel está en la mayor flexibilidad para modelizar las diferencias de comportamiento entre observaciones, en nuestro caso viviendas.

Otros proyectos más recientes que se están efectuando en nuestro instituto insisten más en la estructura de los datos y su comportamiento en su propio dominio, que pretender crear otros prototipos en las áreas que corresponda. A partir de las lecciones aprendidas con la gestión del riesgo bancario, nos hemos dedicado más a la interpretación y estructura de los datos que al aprovechamiento directo de ellos en sus variables originales. Se han propuesto ideas de descubrimiento de nuevos conocimientos en haciendo uso de técnicas como árboles de decisión, búsquedas exhaustivas mediante el algoritmo A*, etc.

Por último quiero mencionar que nuestro objetivo final en cuanto a la manejo de estos grupos de datos es lograr la Disposición de un Repositorio de Datos referidos a la banca, a los hogares, a la vivienda, al jefe de familia, etc, que sea fuente permanente para nuevos modelos de exploración y enriquezcan, en definitiva, los valores originales con información construida a partir de ellos mismos.

Y para concluir, quiero hacer referencia a varios mensajes que se corresponden con la sinergia que debe existir entre las ciencias aplicadas cuando surgen nuevas vías para el perfeccionamiento del conocimiento humano. En las primeras conferencias sobre KDD (Knowledge Discovery Data), se precisan el sentido fundamental de la Minería en el campo de la estadística y de la inteligencia artificial y su contribución en la creación de nueva ciencia.

Así, podemos citar que según Fayyad 1997, la Minería de Datos *es el proceso de identificar patrones en datos que sean válidos, nuevos, potencialmente útiles y reconocibles.*

Por su lado, David Hand, 1998 en The American Statistician conceptualizó la Minería de Datos como *el proceso de análisis secundario de grandes bases de datos orientadas en encontrar relaciones no previstas las cuales son de interés a los dueños de la base de datos.*

Creo que la que se ajusta más a nuestra charla y al objeto de aplicabilidad a las ciencias sociales es la propuesta por el Grupo Gartner, que considera a la Minería de Datos como e*l proceso de descubrir nuevas correlaciones significativas (dependencias), relaciones, tendencias, patrones examinando a fondo a través de grandes cantidades de datos almacenados en repositorios, usando tecnología de reconocimiento de patrones también como técnicas estadísticas y matemáticas.*