UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FISICAS Y MATEMÁTICAS

ESCUELA DE POSTGRADO
ESCUELA DE INGENIERÍA Y CIENCIAS

METODOLOGÍA DE CLASIFICACIÓN DINÁMICA UTILIZANDO SUPPORT VECTOR MACHINE

RODRIGO ANTONIO SANDOVAL RODRÍGUEZ

PROFESOR GUÍA

SR. RICHARD WEBER HAAS

MIEMBROS DE LA COMISIÓN EVALUADORA

SR. JOSÉ MIGUEL CRUZ GONZALEZ
SR. PABLO ANDRES REY
SR. PABLO COLOMA CORREA

TESIS PARA OPTAR AL GRADO DE MAGÍSTER
EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

SANTIAGO DE CHILE
2007

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FISICAS Y MATEMÁTICAS

ESCUELA DE POSTGRADO
ESCUELA DE INGENIERÍA Y CIENCIAS

METODOLOGÍA DE CLASIFICACIÓN DINÁMICA UTILIZANDO SUPPORT VECTOR MACHINE

RODRIGO ANTONIO SANDOVAL RODRÍGUEZ

PROFESOR GUÍA

SR. RICHARD WEBER HAAS

MIEMBROS DE LA COMISIÓN EVALUADORA

SR. JOSÉ MIGUEL CRUZ GONZALEZ
SR. PABLO ANDRES REY
SR. PABLO COLOMA CORREA

TESIS PARA OPTAR AL GRADO DE MAGÍSTER
EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

SANTIAGO DE CHILE
OCTUBRE, 2007

Dedico esta tesis a mi abuelita.

Agradezco a mi familia por su constante apoyo en todo mi proceso de formación, entregándome las herramientas, consejos y motivaciones que me permiten hoy concretar esta etapa de mi vida, en especial agradezco a mi madre, por todo su cariño y preocupación durante toda mi etapa escolar y universitaria. También agradezco a mis amigos de colegio, que han sido una parte importante de mi vida, dándome su apoyo, diversión y compañía. Junto a ellos agradezco a mis amigos de la universidad con quienes compartí durante todo esta etapa de mi vida con largas jornadas de estudio y discusión, que hoy rinden fruto. Además agradezco a Richard y al data minig group, que me ayudaron con ideas y consejos para concretar esta tesis.

Finalmente agradezco a una persona muy especial, a Carla, que ha sido mi pilar durante toda esta última y futura parte de mi vida, quién me aportó grandes ideas y correcciones en esta tesis y que sin su constante apoyo y paciencia hoy no estaría finalizando esta etapa.

Resumen de Tesis para optar al Título de Ingeniero Civil Industrial y Grado de Magíster en Gestión de Operaciones

Alumno: Rodrigo Sandoval
Prof. Guía: Richard Weber
Fecha : 08/11/07

Metodología de Clasificación Dinámica Utilizando Support Vector Machine

Esta investigación se centra en el problema de clasificación, por medio de confeccionar una metodología que permita detectar y modelar cambios en los patrones que definen la clasificación en el tiempo, en otras palabras, clasificación dinámica.

La metodología desarrollada propone utilizar los resultados obtenidos en un periodo de tiempo para la construcción del modelo al siguiente periodo. Para ello se utilizaron dos modelos de clasificación distintos; el primero de ellos es *Support Vector Machine* (SVM) con el objetivo de confeccionar la metodología dinámica, que denominaremos *Dynamic Support Vector Machine* (D-SVM) y el segundo modelo de clasificación es *Linear Penalizad SVM (*LP-SVM) con la finalidad de que la metodología construida permita la selección de atributos dinámicamente. Los parámetros utilizados en el modelo de clasificación son; las ventanas de tiempo, ponderadores de relevancia, penalización de los errores y la penalización de los atributos (sólo para el modelo con selección de atributos). De los resultados obtenidos, se utiliza la ventana de tiempo que define el mejor modelo de un periodo y junto a los nuevos datos que se obtengan generan el del próximo.

Esta metodología luego fue aplicada a un caso real en una institución gubernamental chilena (INDAP), en el problema de predicción de comportamiento de pago (*credit scoring*). Para ello se analizaron 4 instancias de tiempo con 9 atributos para el modelo sin selección de atributos y 18 atributos para el modelo con selección. Luego ambos modelos fueron comparados con uno de clasificación estática, es decir, que las 4 instancias de tiempo son unidas como si fuese una data.

Los resultados obtenidos en esta aplicación son levemente superiores a la metodología estática correspondiente y en el caso de la selección de atributos el modelo utiliza una mayor cantidad.

Las conclusiones de esta investigación son que presenta la ventaja de utilizar una menor cantidad de datos a los disponibles, lo que genera modelos más rápidos y que se van adaptando a los cambios de comportamiento que se producen en el tiempo, al descartar los datos más antiguos en la construcción del nuevo modelo. Con respecto al método con selección de atributos, se destaca que no utiliza un modelo exógeno para seleccionar los atributos sino que el modelo estima los atributos necesarios para cada periodo de tiempo, por lo que se tiene un modelo más estable y generalizado; además se logra obtener información de cómo la relevancia de los atributos cambia en el tiempo. Sobre los resultados se concluye que la metodología D-SVM con y sin selección de atributos es al menos tan buena como los métodos actuales de clasificación.

TABLA DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO 1:

INTRODUCCIÓN

Esta tesis está centrada en el problema de clasificación, que se define como la búsqueda de categorías o clases, en base a atributos o relaciones comunes.

Se define un objeto como el conjunto de variables o parámetros que puede pertenecer a diferentes clases determinado por parámetros en particular.

Esta tesis es una investigación en un tema actual y novedoso, debido a que busca modelar los constantes cambios que surgen en el comportamiento de los objetos que afectan a empresas, instituciones, gobiernos y personas.

Actualmente los problemas de clasificación se resuelven para un instante de tiempo, lo cual se denominará para efectos de este trabajo Clasificación Estática, debido a que confecciona un modelo para predecir la clase de pertenencia de un objeto en un instante y este es utilizado en adelante asumiendo que el comportamiento del objeto no será variable en el tiempo.

En el último tiempo se ha formado el interés por confeccionar modelos de clasificación que capten los cambios de comportamiento de los objetos, lo que se ha traducido en el estudio y desarrollo de distintas formas de actualización de los modelos (utilizando trayectorias en base a la resolución de una ecuación diferencial [8], describiendo una manera de actualizar los modelos [9], realizando actualizaciones por medio de las perturbaciones obtenidas de los nuevos objetos que ingresan [23,28], entre otros), todos estos desarrollos se

encuentran en etapa de investigación, por lo tanto, no existen herramientas que entreguen soluciones concretas a la detección de estos comportamientos.

Con este interés se desarrolla una metodología de clasificación dinámica (detectar cambios en los patrones de comportamiento en el tiempo) caracterizada por la confección de una manera de actualizar un modelo de clasificación en particular.

Se distinguen dos familias de modelos de clasificación, los estadísticos y los de *data mining*. Los primeros se distinguen por el uso de estimadores que definen el modelo de clasificación, algunos son regresión logística, análisis discriminante, *probit*, entre otros [3,12]. Los modelos de *data mining* se distinguen por la construcción de reglas discriminantes en base al aprendizaje obtenido a partir de los datos, algunos modelos son redes neuronales, árboles de decisión, *Support Vector Machine*, entre otros [26].

De los modelos anteriormente mencionados se destaca el modelo *Support Vector Machine* (SVM) [27], que ha tenido especial atención dentro de la comunidad de *data mining* debido a sus propiedades teóricas y buenos resultados en diversas aplicaciones, razón por la que se utiliza como el modelo base para desarrollar esta tesis. Cuyo fin es construir una metodología para problemas de clasificación dinámica, es decir, que logre capturar los cambios en el tiempo con el objetivo de mejorar la predicción en la clasificación. Esta metodología se denominará Dynamic Support Vector Machine (en adelante D-SVM).

Para evaluar la metodología D-SVM se utiliza información recolectada de un caso real en una institución chilena gubernamental (Instituto de Desarrollo Agropecuario, INDAP), el que enfrenta el problema de predecir cuál es el comportamiento de pago de sus clientes (también conocido como *credit scoring*). La información de INDAP tiene la característica de que depende de la

variable temporal, es decir, que los objetos (solicitudes de crédito) dependen del tiempo en el que fueron obtenidos, lo que es necesario para construir un modelo dinámico.

Desarrollada y aplicada la metodología D-SVM, se complementa lo anterior con un modelo de clasificación con selección de atributos, para construir un modelo dinámico con selección de atributos. El que será evaluado con los datos recolectados de INDAP.

Para explicar en detalle los puntos antes mencionados, la tesis se estructura de la siguiente forma:

El capítulo 2 describe el procedimiento Knowledge Discovery in Databases, se detallan las características del modelo de *data mining* SVM y se describen conceptos involucrados en los cambios de patrones. Que son utilizados para confeccionar la metodología D-SVM.

El capítulo 3 desarrolla la metodología D-SVM sin y con selección de atributos. En la primera parte se describe la metodología sin selección, presentando la formulación del modelo, el esquema de la metodología propuesta, detallando las fases utilizadas para construir los modelos y las características de esta metodología. En la siguiente parte se presenta la metodología con selección de atributos, describiendo en primer lugar el modelo que permite hacer la selección simultáneamente a la clasificación y que es la base de la nueva formulación, además se detallan las modificaciones realizadas en la formulación, el esquema presentado en la parte anterior y las características de la metodología.

El capítulo 4 detalla la aplicación de la metodología D-SVM con datos de INDAP. En la primera parte se presenta el problema de clasificación que presenta INDAP, denominado *credit scoring*. En la segunda parte se describe los datos seleccionados y el tratamiento que reciben para ser utilizados en el

modelo. La tercera parte detalla la aplicación de la metodología D-SVM sin selección de atributos, describiendo cada fase de la metodología, los resultados obtenidos y la comparación con el modelo de clasificación estática. La última parte presenta la aplicación de D-SVM con selección de atributos, donde en primer lugar se presenta el método de clasificación estático utilizado para comparar, detallando la metodología de selección de atributos utilizada y los resultados obtenidos, también presenta cada fase de la metodología dinámica, los resultados obtenidos y finalmente la comparación de los resultados obtenidos.

El capítulo 5 concluye el trabajo en base a los resultados y observaciones de la investigación realizada y se recomiendan futuros desarrollos y análisis.

1.1 Objetivos

1.1.1 Objetivo General

Crear y desarrollar una metodología de clasificación dinámica utilizando Support Vector Machine.

1.1.2 Objetivos Específicos

i. Confeccionar una metodología de clasificación que permita utilizar la información recolectada en distintos instantes del tiempo.

ii. Extender la metodología D-SVM, para seleccionar atributos dinámicamente, al momento de realizar la clasificación.

iii. Mostrar los beneficios de la metodología D-SVM y la extensión para la selección de atributos.

iv. Analizar los atributos obtenidos por cada modelo a lo largo del tiempo, de modo de generar relaciones de relevancia para el problema estudiado.

1.2 Metodología

- Estudio del método de clasificación Support Vector Machine.
- Estudio de los conceptos utilizados en modelos de clasificación dinámica.
- Desarrollo de la metodología de clasificación dinámica, basada en la metodología KDD (Knowledge Discovery in Databases).
- Aplicación de la metodología propuesta a un problema real de clasificación.
- Comparación de la metodología propuesta con Support Vector Machine estático para el caso aplicado.
- Estudio del método de clasificación LP-SVM (Support Vector Machine con una penalización lineal sobre los atributos).
- Extender la metodología propuesta de clasificación dinámica, a una de selección de atributos dinámica.
- Aplicación de la clasificación dinámica extendida para la selección de atributos, a un problema real de clasificación.
- Comparación de la metodología de clasificación dinámica con selección de atributos con Support Vector Machine estático para el caso aplicado.

1.3 Actividades

- Revisión del método Support Vector Machine (SVM) para clasificación.
- Revisión bibliográfica de los conceptos de modelos dinámicos.
- Elaboración de la metodología D-SVM.
- Confección del prototipo de la metodología desarrollada
- Aplicación de la metodología a un problema de clasificación real.
- Comparación entre Metodología D-SVM y SVM estático.
- Revisión y análisis de los resultados obtenidos.
- Revisión del método de clasificación Support Vector Machine con una penalización lineal sobre los atributos (LP-SVM).
- Extensión de la metodología D-SVM, utilizando el método LP-SVM, de modo de obtener los atributos seleccionados en cada instante del tiempo junto con la clasificación de los objetos.
- Revisión de los métodos actuales de selección de atributos para el problema de *credit scoring*.
- Confección de un prototipo con la metodología D-SVM extendido.
- Comparación entre los modelos LP-SVM Dinámico y SVM estático utilizando un método de selección de atributos.
- Extensión del estudio y revisión de áreas de trabajo futuro.
- Preparación de resultados y conclusiones.

1.4 Resultados Esperados

i. Desarrollar una metodología para confeccionar modelos en distintos instantes del tiempo, que se retroalimente de resultados obtenidos en periodos anteriores.

ii. Extender el modelo de clasificación anteriormente desarrollado, para realizar la selección de atributos en distintos instantes del tiempo.

iii. Confeccionar un prototipo de la metodología propuesta, que será aplicado a un problema real de clasificación.

iv. Obtener información relevante en el caso de estudio específico del problema de clasificación, utilizando los atributos obtenidos en cada instante de tiempo

1.5 Alcances

Uno de los alcances que tiene la tesis es trabajar con un modelo de clasificación binaria.

Se pueden distinguir dos tipos de comportamiento para el cambio de patrones: uno de ellos es un evento anormal (que sucede sólo una vez y que no tendrá repercusiones en comportamientos futuros); y el otro una tendencia (que se observan modificaciones constantes en el comportamiento que tendrán los patrones). El modelo realizado no considera el tratamiento por separado de los casos recién expuestos, sino que genera una solución que en base a la predicción de los elementos más recientes, determine si un elemento antiguo tuvo un comportamiento anormal o tendencial.

El modelo de Support Vector Machine (SVM) utilizado en la metodología desarrollada tiene la característica de modelar relaciones lineales y no lineales (por medio de transformaciones sobre los datos y la formulación del modelo). Para esta investigación se determina realizar una formulación de clasificación dinámica sólo para las relaciones lineales de SVM.

Se utiliza la información provista por INDAP para evaluar la metodología D-SVM, debido a que contiene una variable temporal que permite identificar el momento en que se genera el objeto (característica no disponible en los datos utilizados en estudios de clasificación disponibles anteriores) y es una aplicación sobre datos reales en que se podría obtener un aporte en la entrega de créditos.

CAPÍTULO 2:

MARCO TEÓRICO

Este capítulo describe los conceptos utilizados en el desarrollo de esta tesis. En la primera parte se presenta el proceso *Knowledge Discovery in Databases* (KDD) y sus etapas, que es la base de la metodología propuesta. Luego se describe el modelo de estimación Support Vector Machine (SVM), modelo de *data mining* utilizado. Finalmente se detallan los conceptos relacionados con Clasificación Dinámica.

2.1 Proceso Knowledge Discovery in Databases (KDD)

KDD se define como el proceso no-trivial de identificar patrones desconocidos, válidos, nuevos, potencialmente útiles y comprensibles dentro de los datos [5,6,7].

El proceso se compone de 9 etapas:

i. Comprensión del contexto donde se hará la aplicación, adquiriendo el conocimiento del negocio e identificando el objetivo del KDD definido por los clientes.

ii. Selección de la data objetivo, que debe ser acorde con los objetivos planteados en el punto anterior.

iii. Limpieza de la data y operaciones básicas de preprocesamiento; remover el ruido, decidir cómo tratar los datos faltantes, etc.

iv. Reducción de las dimensiones y transformación de la data: en esta etapa se encuentran los atributos que son representativos de la data objetivo, que han sido transformados y analizados apropiadamente.

v. Encontrar él o los métodos de *data mining* afín al objetivo del KDD.

vi. Análisis y modelación exploratoria: en esta parte se elige el o los métodos de minería de datos y se seleccionan los parámetros para la búsqueda de patrones.

vii. *Data Mining*, es la búsqueda de patrones de comportamiento en la data seleccionada, la que debe ser representativa.

viii. Interpretación y/o evaluación de los patrones encontrados

ix. Acciones sobre el conocimiento encontrado

Todos estos pasos pueden incluir una iteración significativa y puede contener ciclos entre 2 pasos cualesquiera. La Figura 1 muestra los flujos básicos, aunque no muestra los potenciales ciclos e iteraciones.

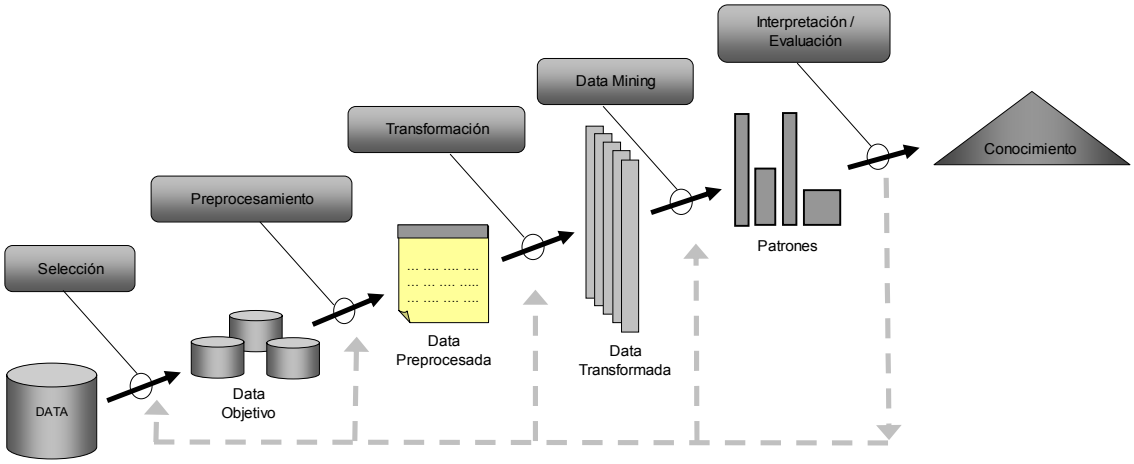


Figura 1. Proceso KDD

2.2 Support Vector Machine para Clasificación

Para comprender el modelo de SVM, se introduce los conceptos de clasificación de objetos, teoría en la que se sostiene y finalmente se explica el método.

2.2.1 Clasificación

La clasificación de objetos, es un problema ampliamente conocido dada la diversidad de aplicaciones en la que puede ser utilizada, algunos ejemplos; detección de fuga de cliente [19], identificación de tumores, *credit scoring* [1,3,15], retención de clientes [20], entre otras.

El problema de clasificación busca patrones en los objetos (definidos por un vector de atributos o *features* y una clase) que predigan la clase a la que pertenece. Para ello busca encontrar una función que, dado un cierto vector de atributos entregue como salida la clase de pertenencia del objeto.

A continuación se describe el problema de clasificación binario [27] (sólo existen dos clases de pertenencias), que es utilizado en el desarrollo de esta tesis.

Dado un objeto, definido por el par $(\bar{x}_i, y_i)$ donde $\bar{x}_i \in \Re^m$ $(\bar{x}_i = (x_{i1}, x_{i2}, ..., x_{im}))$ vector de atributos e $y_i \in \{-1,1\}$ que define la clase a la que pertenece el vector. Se busca "aprender" una función $f(\bar{x}, \alpha)$ que emule una clasificación para todos los pares.

Para construir el modelo de clasificación se requiere dividir los datos en los siguientes conjuntos:

1. Entrenamiento: Consiste en tomar un porcentaje de los datos, con el cual se construye el modelo predictivo mediante la estimación de la función $f(\vec{x},\alpha)$.

2. Validación: Este conjunto tiene otra porción de los datos distinto al conjunto anterior, se utiliza para determinar el(los) parámetro(s) requeridos en la estimación, comparando la predicción sobre este conjunto con cada uno de el(los) parámetro(s) y seleccionando aquel(los) con una mejor predicción.

3. Testeo: Este conjunto tiene la porción restante de los datos, la cual no tiene influencia en la construcción del modelo, de manera de poder conocer la efectividad de la predicción, al comparar la predicción obtenida con el modelo generado y la clase de pertenencia del objeto.

2.2.2 Teoría de Aprendizaje Estadístico

La mayoría de los métodos predictivos (redes neuronales, análisis discriminante, logit, entre otros) construyen modelos en base al principio de "Minimización del Riesgo Empírico" [21,25] (Ec. 1), que busca encontrar el mínimo error de estimación sobre los datos de entrenamiento. El problema que tiene confeccionar modelos bajo este principio, es el sobreajuste de los datos, debido a que la función aprendida se ajusta a los datos ("memoriza") con los que se confeccionó y pierde la capacidad de predecir otros datos.

$$R_{emp}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left|f(\vec{x}_i,\alpha) - y_i\right| \qquad \text{(Ec. 1.)}$$

El Riesgo Estructural [25] (Ec. 2) nace como necesidad de incorporar la capacidad de generalización de manera explícita en la construcción de un modelo predictivo y prevenir el problema del sobreajuste. Para esto propone dar

una cota superior al riesgo esperado, medida que se desea sea lo más pequeña posible. El Riesgo Estructural (Ec. 2) tiene dos componentes, la primera el Riesgo Empírico y la segunda la capacidad de generalización.

Se distinguen los siguientes parámetros:

- h: Dimensión VC (Vapnik – Chervonenkis) de la función aprendida ($f(\vec{x},\alpha)$) [25] y corresponde al mayor número de objetos que pueden ser separados en todas las formas posibles por esta función.
- n: Número total de objetos de entrenamiento.

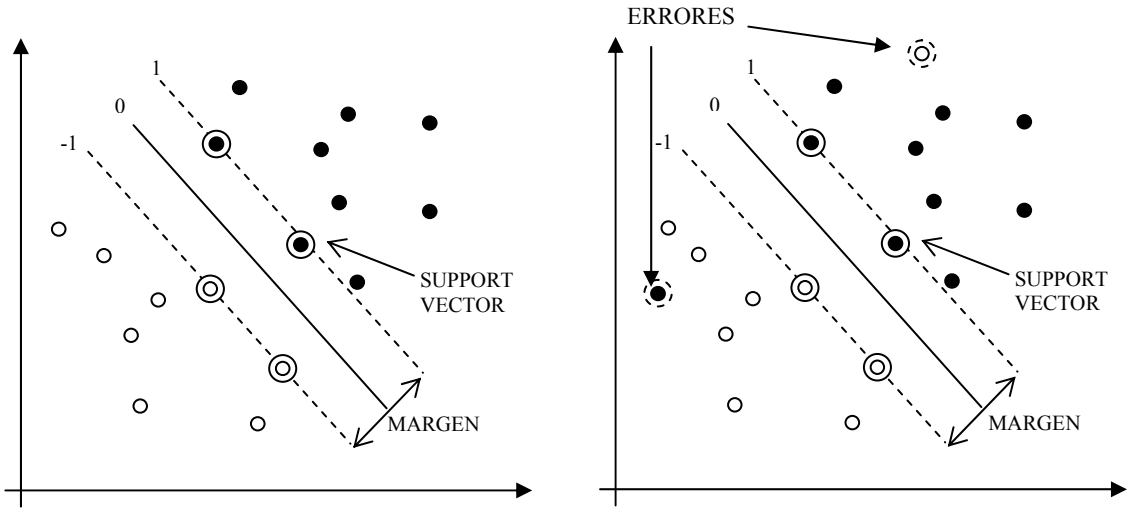
$$R(\alpha) \leq \underbrace{R_{emp}(\alpha)}_{\text{Riesgo Empírico}} + \underbrace{\sqrt{\frac{h\left(\ln\left(\frac{2n}{h}\right)+1\right)-\ln\left(\frac{n}{4}\right)}{n}}}_{\text{Capacidad de Generalización}} \qquad \text{(Ec. 2.)}$$

La teoría del aprendizaje estadístico es la base para el modelo SVM descrito en la siguiente sección.

2.2.3 Modelo SVM

El modelo SVM fue propuesto por Vapnik [2,25,26,27] y se basa en encontrar un hiperplano separador que maximice la distancia (margen) entre dos hiperplanos paralelos construidos en cada lado del hiperplano separador, donde cada una de estas regiones corresponda a una de las clases definidas. Los objetos que definen los hiperplanos paralelos reciben el nombre de *Support Vectors*.

Se distinguen dos casos en SVM. El primer caso es el llamado linealmente separable (Figura 2a), todos los objetos pertenecientes a la misma clase quedan en el mismo espacio diferenciado por el hiperplano separador. El otro caso es llamado linealmente no separable (Figura 2b), en esta situación nos encontramos con algunos elementos de una de las clases en el medio espacio de la clase contraria (Errores de clasificación).



(a) Caso linealmente separable **(b) Caso linealmente no separable**

Figura 2. Ejemplos de clasificación binaria con SVM

El objetivo de la formulación de SVM es encontrar el hiperplano de separación, que maximiza el margen, lo que lleva a la generalización del modelo.

Se presenta la formulación matemática de SVM (Ec. 3) para el caso linealmente no separable[1]. El que considera una penalización para los errores de clasificación para ajustar el modelo.

[1] Para el caso linealmente separable, los valores de $\xi_i = 0 \quad \forall i$,es decir, no hay errores de clasificación

$$\underset{\vec{w},b,\xi}{Min}\left\{\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\xi_i\right\} \qquad \text{(Ec. 3.)}$$

Sujeto a

$$y_i(\vec{x}_i \cdot \vec{w} + b) + \xi_i - 1 \geq 0 \qquad \forall i$$

$$\xi_i \geq 0 \qquad \forall i$$

- Variables de Decisión
  - $\vec{w}$, vector normal al hiperplano separador.
  - b, distancia desde el origen al hiperplano separador.
  - ξ_i, variable de holgura que define el error de clasificación del objeto i.
- Parámetros
  - $\vec{x}_i$, vector de atributos de un objeto i.
  - y_i, clase de pertenencia del objeto i.
  - C, penalización de los errores de clasificación.
- Restricciones
  - $y_i(\vec{x}_i \cdot \vec{w} + b) + \xi_i - 1 \geq 0$, restringe a cada objeto i a estar sobre el hiperplano de cada clase, permitiendo una holgura dada por los errores de clasificación. Notar que cuando la restricción es igual a cero y el error es igual a cero para un objeto i, es un Support Vector.
  - $\xi_i \geq 0$, restringe las variables de holgura (errores) a valores mayor o igual a cero.

- Función objetivo: $\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\xi_i$

  - $\frac{1}{2}\|\vec{w}\|^2$, donde $\|\vec{w}\|$ es la norma euclidiana de $\vec{w}$. Al minimizar esta expresión se logra maximizar el margen de separación ($2/\|\vec{w}\|$) de los hiperplanos, lo que permite obtener el objetivo de generalización del modelo.

  - $C\sum_{i=1}^{n}\xi_i$, esta expresión penaliza los errores de clasificación, utilizando las variables de holgura ξ_i, al minimizarla se logra el objetivo de ajuste del modelo.

A continuación se describe la formulación del problema dual a partir del Lagrangeano y posteriormente se verifican las condiciones de Karush – Kuhn – Tucker (KKT) [22], lo que lleva a describir la formulación Dual de SVM también conocido como el Dual de Wolfe [22]. Esto para introducir la transformación del espacio de atributos de origen con el fin de modelar relaciones no lineales, que es una propiedad del modelo Dual de SVM.

El Lagrangeano (Ec. 4) de la formulación Primal de SVM (Ec. 3) es el siguiente

$$L = \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\{y_i(\vec{x}_i\vec{w}+b)-1+\xi_i\} - \sum\mu_i\xi_i \qquad \text{(Ec. 4.)}$$

Las condiciones de KKT de la formulación Primal de SVM (Ec. 3) son

$$\frac{\partial L}{\partial w_j} = w_j - \sum_{i=1}^{n}\alpha_i y_i x_{ij} = 0 \qquad \text{(Ec. 5.)}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n}\alpha_i y_i = 0 \qquad \text{(Ec. 6.)}$$

$$\frac{\partial L}{\partial \varepsilon_i} = C - \alpha_i - \mu_i = 0 \qquad \text{(Ec. 7.)}$$

$$y_i(\vec{x}_i\vec{w} + b) - 1 + \xi_i \geq 0 \qquad \text{(Ec. 8.)}$$

$$\alpha_i, \xi_i, \mu_i \geq 0 \qquad \text{(Ec. 9.)}$$

$$\alpha_i\{y_i(\vec{x}_i\vec{w} + b) - 1 + \xi_i\} = 0 \qquad \text{(Ec. 10.)}$$

$$\mu_i\xi_i = 0 \qquad \text{(Ec. 11.)}$$

Con estas condiciones se obtiene

$$w_j = \sum_{i=1}^{n} \alpha_i y_i x_{ij} \qquad \text{(Ec. 12.)}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad \text{(Ec. 13.)}$$

Si $0 < \alpha_i < C$ entonces $\mu_i > 0$ por la ec. 7 y por lo tanto $\varepsilon_i = 0$ por la ec. 11, con $\alpha_i > 0$ se obtiene de la ec. 10 que $b = y_i - \vec{x}_i\vec{w}$. Con estas relaciones, el problema Dual de SVM se escribe

$$\underset{\alpha_i}{Max}\left\{\sum_i \alpha_i - \frac{1}{2}\sum_{i,s} \alpha_i\alpha_s y_i y_s \vec{x}_i \cdot \vec{x}_s\right\} \qquad \text{(Ec. 14.)}$$

Sujeto a $\quad 0 \leq \alpha_i \leq C$

$$\sum_i \alpha_i y_i = 0$$

Como el problema de SVM es un problema convexo (tanto la función objetivo como las restricciones son convexas), las condiciones de KKT son necesarias y suficientes para asegurar una solución del problema Primal de SVM (Ec. 3) que

cumpla las condiciones de KKT, es un óptimo global [22]. Entonces resolver el problema Dual de SVM es equivalente a resolver el problema Primal.

La ventaja que presenta la formulación Dual de SVM, es la capacidad de realizar una transformación del espacio de origen a uno de mayor dimensionalidad, que se puede realizar de manera directa calculando los productos punto entre los vectores de atributos. Esto se puede calcular utilizando las funciones de Kernel [22].

$$K(x,z) = \langle \Phi(x), \Phi(z) \rangle \qquad \forall x, z \in X \qquad \text{(Ec. 15.)}$$

Donde Φ es una función de proyección desde el espacio de origen X hacia el espacio de atributos F (Ec. 16), en el cual se buscan relaciones lineales en este nuevo espacio, tal como muestra la Figura 3.

$$\Phi : x \mapsto \Phi(x) \in F \qquad \text{(Ec. 16.)}$$

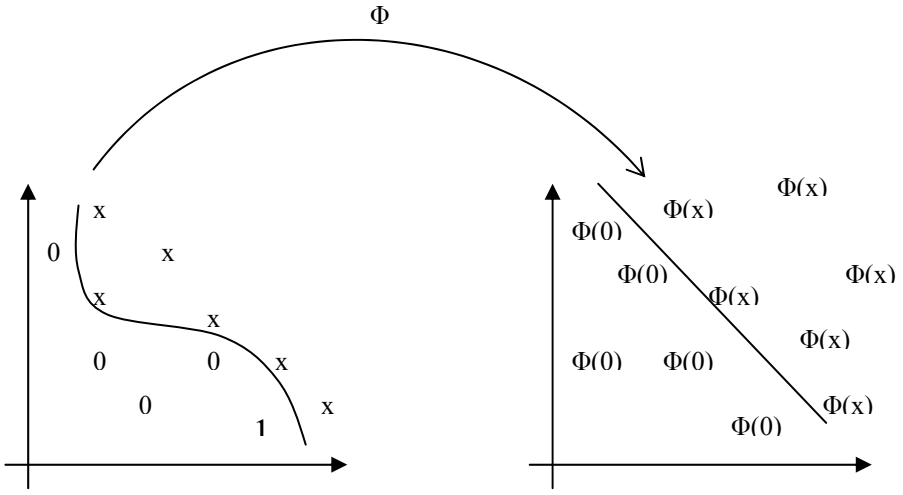


Figura 3. Transformación desde un espacio no lineal a uno lineal por medio de una función Φ.

Para que una función sea definida como *kernel*, debe cumplir con la condición de Mercer [22], que consiste en que una matriz (llamada de Gram o de *kernel*), compuesta por funciones de *kernel* de la siguiente forma:

$$\begin{bmatrix} k(x1,x1) & k(x1,x2) & \cdots & k(x1,xl) \\ k(x2,x1) & k(x2,x2) & \cdots & k(x2,xl) \\ \vdots & \vdots & \ddots & \vdots \\ k(xl,x1) & k(xl,x2) & \cdots & k(xl,xl) \end{bmatrix}$$

Figura 4. Matriz de *kernel*

cumpla las siguientes condiciones; (1) ser simétrica, (2) ser semi definida positiva (todos sus valores propios son mayores o iguales a cero).

Existen diversas funciones de *kernel*, siendo las más utilizadas para SVM las siguientes.

- Lineal: $K(x,z) = \langle x,z \rangle$
- Polinomial de grado d: $K(x,z) = (\langle x,z \rangle + 1)^d,\ d \in N$
- Radial Basis Function (RBF): $K(x,z) = e^{\frac{\|x-z\|^2}{2\sigma}},\ \sigma > 0$

Al utilizar la proyección de los objetos por medio de un Kernel, la formulación Dual de SVM queda como sigue.

$$\underset{\alpha_i}{Max} \left\{ \sum \alpha_i - \frac{1}{2} \sum_{i,s} \alpha_i \alpha_s y_i y_s K(\vec{x}_i, \vec{x}_s) \right\} \qquad \text{(Ec. 17.)}$$

Sujeto a

$$0 \le \alpha_i \le Cp_i$$

$$\sum \alpha_i y_i = 0$$

Las ventajas que presenta la técnica de SVM son [21]

i. Existe una dependencia explícita en los patrones más informativos (*Support Vectors*), lo cual permite generalizar de mejor forma frente a nuevos objetos, dado que considera el principio de "Minimización del Riesgo Estructural" propuesto en la Teoría del Aprendizaje Estructural de Vapnik (sección 2.2.2).

ii. Se construye en base a una función convexa, de modo que se asegura la obtención de un óptimo global y permite la construcción de su formulación Dual.

iii. La capacidad de modelar fenómenos no lineales mediante el uso de una transformación del espacio de origen a uno de mayor dimensionalidad.

Algunas de sus limitaciones son [21]

i. Trabaja con datos numéricos, por lo que es necesario transformar los atributos nominales a un formato numérico por medio de una codificación [4] (Ejemplo: Codificación N, N-1, Termómetro u Ordinal)

ii. El uso de distintas funciones de *kernel* puede determinar diferentes soluciones. Actualmente no hay consenso sobre qué función de *kernel* utilizar para un problema en particular.

2.3 Conceptos para la Clasificación Dinámica

Esta sección revisa los conceptos que permiten definir la metodología de Clasificación Dinámica.

2.3.1 Concept Drift

En muchos problemas la información es recolectada en un largo periodo de tiempo y como es de esperar, la distribución subyacente tiende a cambiar, este fenómeno se define como *Concept Drift* [13,14].

Un ejemplo de ello es un portal de venta de libros por Internet, donde se entregan sugerencias de libros en base al historial de compras del cliente, sin embargo, es esperable que existan cambios tanto en el interés del cliente como los contenidos de los libros.

El manejo de este efecto se basa en el supuesto que *Concept Drift* no es reversible, es decir, si un patrón cambia, éste se mantiene en el tiempo hasta que se produzca un nuevo cambio.

La información llega a través del tiempo en *batches*[2], que se definirán dependiendo de la cantidad de objetos y tipo de *Concept Drift*. El objetivo es poder predecir el *batch* t+1 en base a los objetos de entrenamiento de los *batches* 1 al t, minimizando el error de predicción.

Para identificar el *Concept Drift* a través del tiempo se utiliza generalmente una ventana de tiempo fija o una ventana adaptable sobre el conjunto de entrenamiento. Para las ventanas de tiempo fija, la elección de un "buen" tamaño de ventana tiene un *trade-off* entre elegir una rápida adaptabilidad

[2] Una cantidad de objetos que son considerados como un grupo

(ventana pequeña) o una buena generalización (ventana grande) bajo el supuesto que no hay cambios. La ventana de tiempo adaptable maneja el ajuste del tamaño de la ventana, la idea es elegir aquella que minimiza el error estimado para los nuevos datos.

La ventana de tiempo adaptable también resuelve el *trade-off* entre una ventana de tiempo grande, que provee muchos objetos para el entrenamiento lo que permite la generalización y por otro lado puede contener datos antiguos que ya no son relevantes (o incluso confusos) para el *Concept Drift* objetivo. Encontrar el tamaño correcto constituye un *trade-off* entre la calidad y el tiempo de construcción (cantidad de objetos).

En esta tesis se utilizan distintas ventanas de tiempo (ventana adaptable) y son entrenadas con el modelo SVM, eligiendo aquella ventana que tiene un mejor desempeño en los datos más recientes

2.3.2 Ponderación de la muestra

Para describir esta técnica se utiliza el ejemplo descrito anteriormente; el portal de venta de libros, en él se puede dar que, aunque haya cambios en los intereses de un cliente, este cambio sea lento. En este caso no se puede encontrar un tiempo específico donde un objeto se vuelve irrelevante y el valor de la información que se puede obtener de un cierto objeto puede decrecer sobre un largo periodo de tiempo.

El valor de la información de los objetos antiguos puede ser modelado asignando un peso c_i a cada objeto [14]. Un esquema de ponderación es seleccionar el peso de los objetos basado en la edad que tenga, utilizando una función exponencial de edad $pp_\lambda(x) = \exp(-\lambda t_x)$, donde el objeto x se encontró

hace t_x pasos y λ es la tasa con la cual decrece el valor de la información, mientras más grande el valor de λ los objetos antiguos se vuelven irrelevantes con mayor rapidez. Los casos extremos son, si $\lambda \to \infty$ sólo se aprende de los datos más recientes, si $\lambda = 0$ se aprende de todos los objetos, independiente del tiempo en que se obtuvo.

Para el desarrollo de esta tesis, se utiliza esta función de edad, con distintas parámetros λ, aunque siempre se considera el caso en que todos los objetos son iguales ($\lambda = 0$).

CAPÍTULO 3:

METODOLOGÍA DE CLASIFICACIÓN DINÁMICA

Este capítulo presenta la metodología de clasificación propuesta que se denomina Dynamic Support Vector Machine (D-SVM). Se caracteriza por definir el modo de actualización del modelo de clasificación a través del tiempo, con el fin de capturar cambios en el comportamiento del problema estudiado.

Los modelos de clasificación dinámica permiten detectar variaciones de los patrones en el tiempo y utilizar una menor cantidad de objetos ya que utiliza resultados (función de estimación, ventana de tiempo, *Support Vectors*) obtenidos en el modelo de estimación anterior. Esto se traduce en una disminución en el tiempo de construcción de los modelos (a mayor cantidad de objetos mayor el tiempo de construcción). Además debido a los cambios de patrones antes mencionados, la información del pasado puede producir ruido en el modelo y no ser de valor para la predicción.

Además de lograr un modelo de clasificación dinámico, se define una metodología para realizar la selección de atributos. Para ello se describe una modificación de la metodología D-SVM que permita realizarla al momento de clasificar los objetos. De este modo se obtiene el modelo predictivo y los atributos utilizados en distintos instantes del tiempo conjuntamente.

En la primera parte de este capítulo se describirá en detalle la metodología propuesta, presentando su formulación, esquema y sus características principales.

En la segunda parte se presenta una formulación de SVM que permite seleccionar atributos al clasificar los objetos denominado *Support Vector Machine* con una penalización lineal (LP-SVM). La ventaja de realizar la selección de atributos con este método es la simplicidad en su resolución, ya que otros métodos [1,10,17] requieren de un extenso procesamiento y análisis. En particular se evaluó el método de selección de atributos denominado "*Wrapper*", conocido como uno de los mejores métodos para el problema de *credit scoring.* Este método se basa en evaluar distintas configuraciones de atributos con un modelo de clasificación (SVM, Regresión Logística u otro), esta condición hace que el método requiera de una gran cantidad de tiempo computacional para su resolución, lo que hace impracticable en muchos casos su aplicación [10] y por lo que en esta investigación fue descartado.

3.1 Metodología D-SVM

La metodología D-SVM, se basa en el proceso KDD (Sección 2.1). Para desarrollar la nueva metodología, es necesario definir una nueva formulación de SVM (Ec. 3).

3.1.1 Formulación Matemática

La formulación matemática no difiere mucho de la formulación original del SVM lineal (Ec. 3). Se debe incorporar la ponderación del error por medio de una función de edad ($pe(t_i) = \exp(-\lambda t_i)$, sección 2.3.2).

$$\underset{\vec{w},b,\xi}{Min}\left\{\frac{1}{2}\|\vec{w}\|^2 + C\sum_i pe(t_i)\xi_i\right\} \qquad \text{(Ec. 18.)}$$

Sujeto a
$$y_i(\vec{x}_i \cdot \vec{w} + b) + \xi_i - 1 \geq 0 \qquad \forall i$$
$$\xi_i \geq 0 \qquad \forall i$$

El efecto de esta modificación, es permitir una mayor cantidad de errores para aquellos datos más antiguos (a mayor edad, menor peso) y penalizar más los recientemente recolectados.

3.1.2 Esquema de la Metodología

Fase 1. Tener la información seleccionada en un instante de tiempo T_{inicio}, preprocesada [4] (*outliers*, *missing value*, *inconsistent data* & *noisy data*) y realizada la transformación de la data (etapas ii, iii y iv del KDD, sección 2.1).

Fase 2. Definir una unidad de tiempo acorde a las características del problema (un mes, un semestre, un año, etc.), que será la medida que se utiliza para definir el tamaño de la ventana de tiempo. Además se define el periodo de actualización del modelo (mensual, semestral, anual, etc.) el cual corresponde al tiempo transcurrido entre la confección de un modelo y el siguiente.

Fase 3. Utilizando SVM lineal (Ec. 18), se debe generar los modelos con los parámetros requeridos, detallados a continuación.

a. Ventana de tiempo. Se generan distintos conjuntos de data (que generan distintos modelos de estimación). Por ejemplo, suponer que $T_{inicio} = T_4$ (según la definición impuesta en la Fase 2), se construyen 3

ventanas de tiempo: (1) $[T_1, T_4]$, (2) $[T_2, T_4]$, (3) $[T_3, T_4]$. El tamaño mínimo de una ventana de tiempo está definido por el periodo de actualización.

b. Penalización de los errores. Corresponde al parámetro de la formulación de SVM lineal, que indica cuanto cuesta un error de clasificación. Esto se debe hacer para cada conjunto de data (definido por la ventana de tiempo, sección 2.3.1), donde se prueban los distintos valores para este parámetro.

c. Ponderación de los objetos. Este parámetro indica cuánto se pondera el error (Ponderación de la muestra, sección 2.3.2). Se encuentra definida por una función de edad de la forma $\exp(-\lambda t_i)$, donde λ es el valor del parámetro y t_i corresponde al periodo de tiempo donde se encontró el dato. Esto se debe hacer para cada conjunto de data (definido por la ventana de tiempo), donde se prueban distintos valores para este parámetro.

Fase 4. Seleccionar la mejor configuración de parámetros. En este punto se obtiene el modelo predictivo a ser utilizado.

a. Confeccionar el conjunto de testeo (ver sección 2.2.1) que debe estar constituido por una fracción de los objetos del último periodo de actualización.

b. Para cada ventana de tiempo separar los datos restantes (aquellos que no se encuentran en el conjunto de testeo antes definido) en dos conjuntos; entrenamiento y validación.

c. Transformar (normalizar, escalar y codificar variables nominales) los 3 conjuntos de datos.

d. Seleccionar la mejor configuración de parámetros en base a la predicción sobre el conjunto de validación para cada ventana de tiempo.

e. Seleccionar la mejor ventana de tiempo en base a la predicción sobre el conjunto de testeo y define la efectividad del modelo predictivo.

Fase 5. Generar la data para la actualización del siguiente periodo (definido en la Fase 2).

a. Recolectar la información entre el periodo actual y el próximo.

b. Adjuntar los datos recolectados con los datos de la mejor ventana de tiempo (definido en la Fase 4).

c. Tratar los datos adjuntados según la Fase 1.

Al completarse la Fase 5 (con los datos resultantes), se debe volver a construir el modelo de clasificación que será el utilizado en el siguiente periodo (Fase 3). Ver diagrama de las fases figura 5.

Notar que existen dos parámetros externos que son definidos por el cliente; el periodo de análisis y el tiempo de actualización. Estos parámetros se pueden desprender de análisis sobre los datos.

La figura 6, muestra el esquema de actualización de la metodología propuesta. En él ΔT muestra el periodo de actualización definido, los tiempo T_0, T_1, …, T_n indica el tiempo donde se construye el modelo, los datos en T_0 indica todos los datos disponibles al momento de construir el primer modelo el cual es utilizado para predecir los nuevos datos que llegan entre T_0 y T_1 y muestra que los datos de construcción para los modelos en T_1 hasta T_n (excluye a T_0) están determinados por la ventana de tiempo (VT) encontrada en el tiempo anterior más los nuevos datos recolectados.

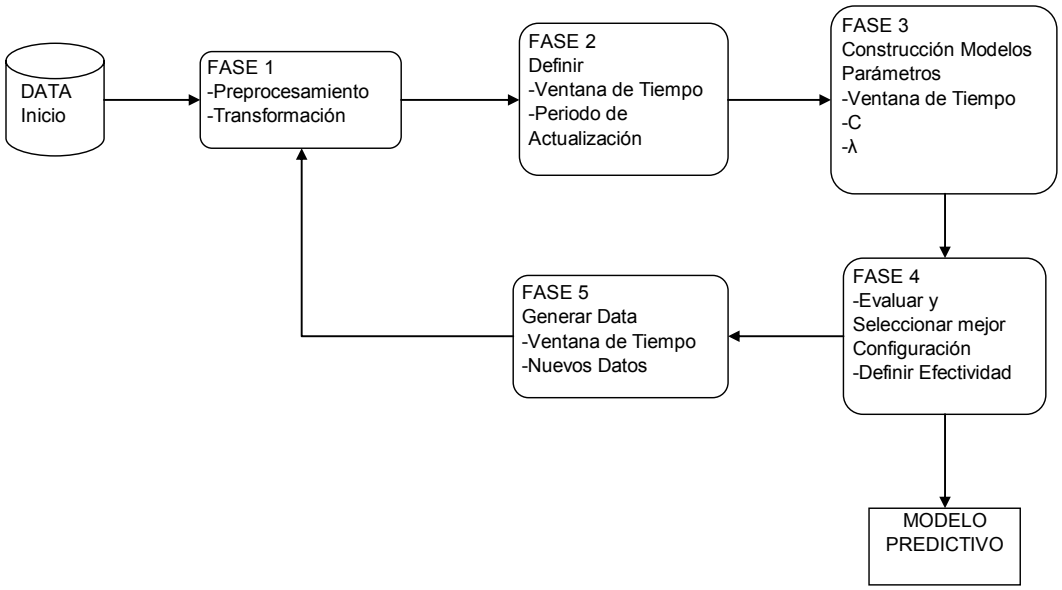


Figura 5. Diagrama Metodología D-SVM

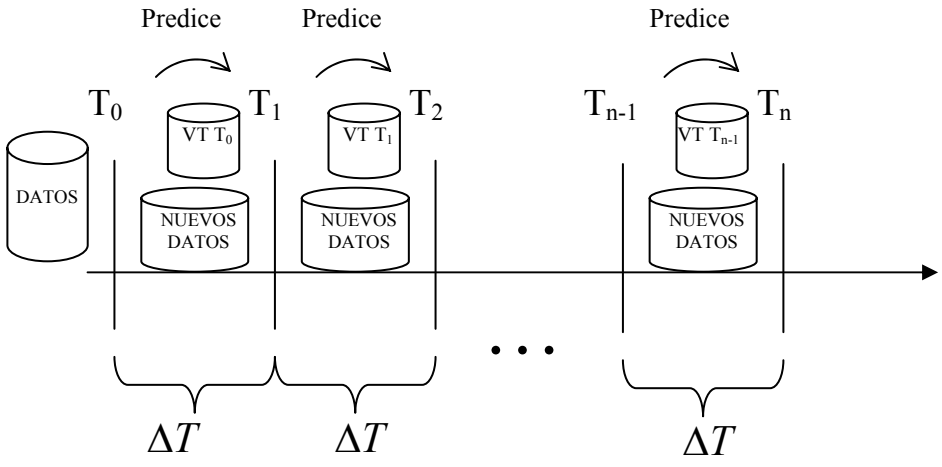


Figura 6. Esquema Actualización, Metodología de Clasificación Dinámica

3.1.3 Características de la Metodología

Como todas las metodologías presentan factores que hacen favorable su desarrollo y otros que no. Se revisan a continuación estas características.

Las ventajas de la metodología son

- Dada la estructura del problema no requiere grandes modificaciones sobre la formulación de SVM lineal, lo cual permite utilizar las herramientas existentes para resolver este problema.

- Permite captar cambios en el comportamiento de las clases a través del tiempo, aunque sin olvidar lo eventos que han ocurrido.

- Genera políticas sobre la actualización de modelos de clasificación, de modo de tener soluciones que se van adecuando a las modificaciones intrínsecas del problema de clasificación, mientras que las metodologías actuales de clasificación sólo presentan una solución que es construida bajo las condiciones presentes en ese periodo.

- Construye modelos con mayor velocidad, debido a que utiliza una menor cantidad de objetos.

Las limitaciones que tiene esta metodología son

- No poder identificar de forma clara, cuándo un evento es anormal o es tendencia. Lo que observa el modelo es cómo poder predecir de mejor forma los datos más actuales pero sin olvidar completamente los pasados, de modo de intentar captar este fenómeno. Mas no es seguro el poder obtener resultados satisfactorios.

3.2 Clasificación Dinámica para la Selección de Atributos

La formulación de *Support Vector Machine* con una Penalización Lineal (LP-SVM), permite la selección de atributos al momento de clasificar los objetos, sin tomar en consideración la variable temporal de los objetos. Después se identifican las modificaciones que requiere la formulación del modelo LP-SVM para adaptarlo a la metodología de clasificación dinámica con selección de atributos.

3.2.1 Support Vector Machine con una Penalización Lineal para la Selección de Atributos

El modelo LP-SVM [20,21] considera una modificación de la formulación lineal de SVM (formulación primal), descrito en la sección 2.2.3, para incluir la selección de atributos de manera explícita.

El modelo considera tres objetivos de manera conjunta.

i. Capacidad de generalización, mediante la minimización de la norma del vector normal al hiperplano separador.

ii. Ajuste del modelo, mediante la minimización de la penalización de los errores de clasificación.

iii. Selección de atributos, mediante la minimización del número de atributos empleados para construir el modelo de discriminación, penalizando por cada atributo utilizado.

En la formulación tradicional de SVM se consideran los dos primeros objetivos al momento de realizar la clasificación. Al considerar los tres objetivos de

manera conjunta, es posible imponer explícitamente la condición de la selección de atributos. Para controlar cada uno de los objetivos se utilizan multiplicadores (penalizadores), de modo de privilegiar el cumplimiento de un objetivo frente a los demás.

La formulación matemática del problema requiere definir dos funciones a priori

Sea $\vec{x} \in \Re^m$

Función Módulo: $$|\vec{x}|_j = \begin{cases} x_j & si \quad x_j > 0 \\ 0 & si \quad x_j = 0 \\ -x_j & si \quad x_j < 0 \end{cases}$$

Notar que $|\vec{x}| \in \Re_+^m$

Función de Paso: $$(\vec{x}_*)_j = \begin{cases} 1 & si \quad x_j > 0 \\ 0 & si \quad x_j = 0 \\ -1 & si \quad x_j < 0 \end{cases}$$

Notar que si $\vec{x} \in \Re_+^m$ $\quad (\vec{x}_*)_j \in \{0,1\} \quad \forall j$

Sea $\vec{e}^t = (1 \quad 1 \quad ... \quad 1 \quad 1)$, la formulación del modelo LP-SVM es

$$\underset{\vec{w},b,\xi}{Min} \left\{ \frac{1}{2}\|\vec{w}\|^2 + C_1 \sum_i \xi_i + C_2 \vec{e} \cdot |\vec{w}|_* \right\} \qquad \text{(Ec. 19.)}$$

Sujeto a $$y_i(\vec{x}_i \cdot \vec{w} + b) + \xi_i - 1 \geq 0 \qquad \forall i$$

$$\xi_i \geq 0 \qquad \forall i$$

donde $\vec{e} \cdot |\vec{w}|_*$ es el número de los módulos de las componentes distintas de cero del vector $\vec{w}$ y se refiere a la selección de atributos del modelo.

Con esta reformulación de la función objetivo se controla de forma explícita los tres objetivos antes descritos. El parámetro del primer objetivo (capacidad de generalización) es por defecto igual a 1, mientras que los parámetros para el segundo (Ajuste del Modelo) y el tercer objetivo (Selección de Atributos) son C_1 y C_2 respectivamente.

Se destaca que los parámetros C_1 y C_2 no solo pueden tomar valores en $\Re$ sino que también pueden ser vectores $\vec{C}_1 \in \Re^n$, de modo que se puede penalizar cada objeto por separado o por clase y $\vec{C}_2 \in \Re^m$ donde se puede penalizar de manera distinta cada atributo a seleccionar.

Sin embargo esta formulación (Ec. 19) presenta el inconveniente de no ser una función continua, ya que incorpora el módulo (función no derivable) y una función de paso (no continua). Por lo tanto no se puede resolver el problema con las herramientas para modelos no lineales [18].

Para reemplazar el módulo en la función objetivo se agrega una restricción (Ec. 20), se incluye una variable auxiliar ($\vec{v}$).

$$-\vec{v} \le \vec{w} \le \vec{v} \qquad \text{(Ec. 20.)}$$

donde $\vec{v}, \vec{w} \in \Re^m$. Esta restricción implica $\vec{v} \ge 0$, por lo que se reemplaza $|\vec{w}|$ por $\vec{v}$.

Para reemplazar la función de paso discontinua, se define la siguiente aproximación exponencial cóncava.

$$\left(\vec{e} - \exp\left(-\tau\vec{v}\right)\right) = \left(1 - \exp\left(-\tau v_1\right), \ldots, 1 - \exp\left(-\tau v_m\right)\right) \qquad \text{(Ec. 21.)}$$

, donde $\tau \in \Re$. Las propiedades de esta función son:

i. Cuando $\vec{v} = \vec{0}$ la función vale exactamente $\vec{0}$.

ii. Cuando $\vec{v} > \vec{0}$ la función vale aproximadamente $\vec{e}$ (depende del valor de τ).

iii. La concavidad de esta función permite obtener la convergencia de la minimización.

Con estas modificaciones la formulación para el método LP-SVM es

$$\begin{aligned} \underset{\vec{w},b,\xi_i,\vec{v}}{Min} & \left\{ \frac{1}{2}\|\vec{w}\|^2 + C_1\sum_i \xi_i + C_2\vec{e}^t(\vec{e} - \exp(-\tau\vec{v})) \right\} \\ \text{Sujeto a} \quad & y_i(\vec{x}_i\vec{w}_i + b) + \xi_i - 1 \geq 0 \qquad \forall i \\ & \xi_i \geq 0 \qquad \forall i \\ & -\vec{v} \leq \vec{w} \leq \vec{v} \end{aligned} \qquad \text{(Ec. 22.)}$$

3.2.2 Formulación Matemática Metodología Extendida

La modificación incorpora la ponderación del error por medio de una función de edad ($pe(t_i) = \exp(-\lambda t_i)$, sección 2.3.2). Con lo que la nueva formulación del método LP-SVM es la siguiente:

$$\begin{aligned} \underset{\vec{w},b,\xi_i,\vec{v}}{Min} & \left\{ \frac{1}{2}\|\vec{w}\|^2 + C_1\sum_i pe(t_i)\xi_i + C_2\vec{e}^t(\vec{e} - \exp(-\tau\vec{v})) \right\} \\ \text{Sujeto a} \quad & y_i(\vec{x}_i\vec{w}_i + b) + \xi_i - 1 \geq 0 \qquad \forall i \\ & \xi_i \geq 0 \qquad \forall i \\ & -\vec{v} \leq \vec{w} \leq \vec{v} \end{aligned} \qquad \text{(Ec. 23.)}$$

, donde $\vec{w},\vec{v} \in \Re^m$ y $\xi_i, b \in \Re$

El efecto es el permitir una mayor cantidad de errores de clasificación (definidos por las variables de holgura ξ_i) para aquellos datos más antiguos (definido por t_i) y penalizar más los recientemente recolectados.

3.2.3 Esquema Metodología

A diferencia de la metodología D-SVM no se requiere de un método externo para realizar la selección de atributos (necesario en la construcción de modelos de clasificación).

La figura 7 presenta un diagrama de la nueva metodología.

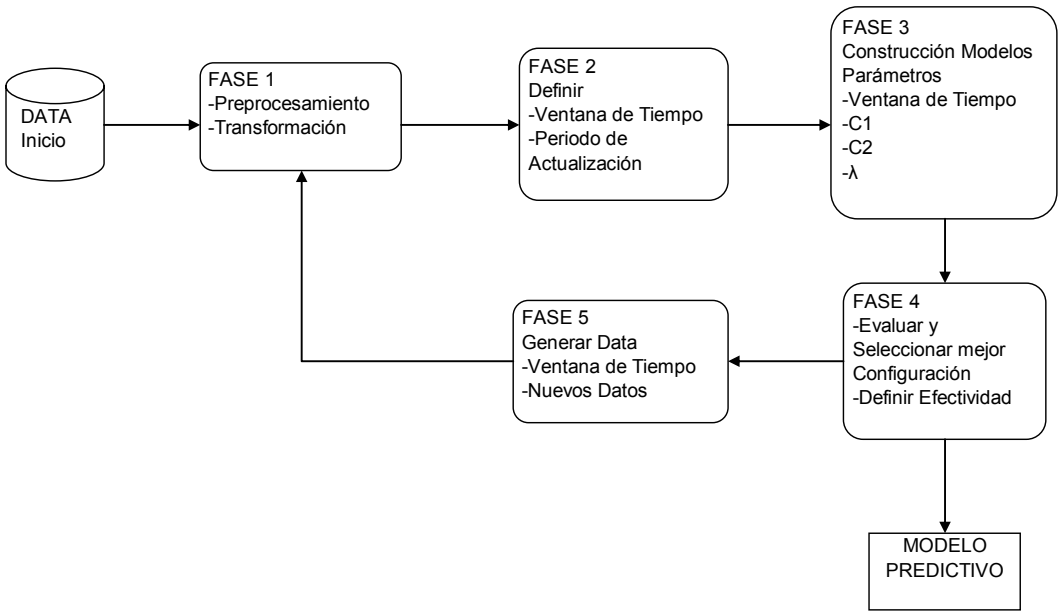


Figura 7. Diagrama Metodología de Clasificación Dinámica para la Selección de Atributos

Las modificaciones realizadas, con respecto a la metodología D-SVM están en las fases 3 y 4. La Fase 3 está determinada por la configuración de los distintos modelos a construir, se deben determinar 4 parámetros en vez de los 3 utilizados en la metodología original. Para la Fase 4 se debe definir otra manera de evaluar y seleccionar la mejor configuración de parámetro, ya que no dependerá exclusivamente del desempeño de cada modelo (medida utilizada

para evaluar la efectividad del modelo predictivo, por ejemplo, matriz de confusión [16,26]) sino que también de la cantidad de atributos utilizados.

A continuación se describe la fase 3 utilizando el esquema de la sección 3.1.2.

Fase 3. Utilizando LP-SVM (Ec. 23), de debe generar los modelos con los parámetros requeridos.

a. Ventana de tiempo, se definen del mismo modo que la metodología D-SVM.

b. Penalización de los errores, corresponde al parámetro C_1 de la formulación LP-SVM y es equivalente al parámetro C de la formulación D-SVM.

c. Ponderación de los objetos, definido de la misma manera que la metodología D-SVM.

d. Penalización de los atributos, corresponde al parámetro C_2 de la formulación LP-SVM. Indica cuanto cuesta utilizar un atributo, de este modo el modelo determina cuales son los atributos relevantes con respecto a los otros dos objetivos de la formulación (capacidad de generalización y calidad predictiva).

Fase 4. Seleccionar la mejor configuración de parámetros, en este punto se obtiene el modelo predictivo a ser utilizado.

a. Confeccionar el conjunto de testeo (ver sección 2.2) el cual debe estar constituido por una fracción de los objetos del último periodo de actualización.

b. Para cada ventana de tiempo separar los datos restantes (aquellos que no se encuentran en el conjunto de testeo antes definido) en dos conjuntos; entrenamiento y validación.

c. Transformar (normalizar, escalar y codificar variables nominales) los 3 conjuntos de datos.

d. Seleccionar la mejor configuración de parámetros en base a la predicción sobre el conjunto de validación para cada ventana de tiempo, donde la medida de evaluación debe estar determinada por la cantidad de atributos utilizados, para ello se selecciona la mejor predicción sobre el conjunto de validación al seleccionar 1, 2, 3, ...o "n" atributos y elegir aquella configuración que tenga buena calidad predictiva dado un cierto incremento marginal. Así se puede evitar casos donde un modelo con 12 atributos es levemente superior (por ejemplo un 0.05%) en predictividad a uno de 8 atributos.

e. Seleccionar la mejor ventana de tiempo en base a la predicción sobre el conjunto de testeo y que define la efectividad del modelo predictivo.

3.2.4 Características de la Metodología

Las características de esta metodología se presentan como las ventajas y limitaciones de su utilización.

Las ventajas son

- El conocer en cada instante de tiempo los atributos más relevantes en la construcción del modelo de clasificación, ya que entrega información sobre

el comportamiento del problema de clasificación, además de la predicción de los futuros objetos.

Las limitaciones son

- No existe un criterio estándar para determinar el valor del aporte marginal de un atributo, ya que este depende del problema de clasificación (un 1% más en un problema de *credit scoring* es distinto que en uno de predecir comportamiento en la compra de libros por internet) y de la medida de efectividad utilizada (matriz de confusión, ROC, MAPE, MAE, entre otras [16,26]).

CAPÍTULO 4:

APLICACIÓN A UN CASO REAL: CASO INDAP

El Instituto Nacional de Desarrollo Agropecuario (INDAP) es una institución gubernamental, que tiene por misión "promover condiciones, generar capacidades y apoyar con acciones de fomento, el desarrollo productivo sustentable de la agricultura familiar y sus organizaciones"[3].

INDAP requiere mejorar los procesos de asignación de créditos en relación a la evaluación de los clientes y la entrega de créditos. En este contexto se observan cuatro problemas.

1. Eficiencia en la evaluación y en consecuencia el tiempo de respuesta a un cliente; desde el ingreso de la solicitud de crédito hasta la respuesta del comité de crédito pueden pasar varios días.

2. El porcentaje de cobertura tiene un potencial de crecimiento de acuerdo a la población solicitante de crédito, aunque esta demanda es aún muy riesgosa. Debido al tiempo de evaluación no se puede responder a todas las solicitudes.

3. Las estimaciones de riesgo que utiliza la institución obedece a un modelo basado en la conducta de pago que es intuitivo (principalmente utiliza el comportamiento histórico de pago que ha tenido el cliente). Esto genera que no exista estimación sobre el riesgo asociado a un cliente que solicita un

[3] http://www.indap.cl

crédito por primera vez, por otro lado no reconoce la importancia de otros factores que determinan el riesgo.

4. Asignación de créditos no considera el comportamiento futuro de pago del cliente (estimación de recuperación de un crédito), por lo que el retorno esperado en el tiempo asignado no cumple las expectativas de INDAP.

Los dos primeros problemas pueden resolverse al automatizar el proceso de evaluación de crédito, utilizando una herramienta de *credit scoring* (descrito más adelante en este capítulo).

El tercer problema se resuelve eligiendo los factores adecuados a evaluar.

El cuarto problema presentado corresponde a un problema de clasificación, específicamente, determinar si un cliente paga o no el crédito solicitado (predecir el comportamiento futuro de pago). Este es un problema de clasificación binaria, sólo existen dos clases de pertenencia (se recupera el crédito o no), que predice el comportamiento de pago.

Para el desarrollo de este capítulo se define un objeto como el vector de atributos asociado a una solicitud de crédito (monto, plazo, edad cliente, etc.) cuya clase de pertenencia está definida por la recuperación del crédito (se recupera o no el monto solicitado).

En la primera parte de este capítulo, se describe en términos generales el método de *credit scoring*, con sus principales beneficios y limitaciones. A continuación se describe el preprocesamiento y selección de los datos a ser utilizados en este desarrollo. Luego se resuelve el problema planteado por INDAP en los puntos 1, 2, 3 y 4 utilizando D-SVM (sección 3.1), se presentan los resultados obtenidos con esta metodología y su comparación con un método de clasificación estático. Posteriormente este mismo problema (utilizando una mayor cantidad de atributos) es resuelto con la metodología de clasificación

dinámica para la selección de atributos (sección 3.2) y es comparado con un modelo de clasificación estático que recibe como imput los atributos seleccionados con un método ad-hoc.

4.1 Credit Scoring

Se define *credit scoring* [3,15] como el método cuantitativo que se utiliza para determinar la probabilidad que un crédito entregado se vuelva impago. Tiene por objetivo el cuantificar el riesgo financiero de cada cliente y apoyar la gestión en la entregar de un crédito.

Para construir modelos de *credit scoring* se requiere de lo siguiente.

i. Seleccionar una muestra de clientes clasificados como “bueno” o “malo” [4], dependiendo de su desempeño de pago en un periodo dado.

ii. La data es recolectada desde aplicaciones crediticias, información personal, de negocio y/o las fuentes que sean necesarias. En general se recolecta toda la información asociada al prestatario y las características del crédito solicitado.

iii. Preprocesamiento y tratamiento de los datos recolectados.

iv. Realizar análisis cuantitativos sobre la data para derivar en el modelo de *credit scoring*.

El modelo resultante contiene la suma de los pesos aplicados a las variables lo que determina el puntaje (*score*) asociado al cliente. El punto de corte, que es

[4] La definición de un crédito como “bueno” o “malo” la realiza quien requiere de la herramienta. Usualmente un crédito “bueno” es un crédito donde se recupera el monto otorgado y los intereses asociados. El crédito “malo” es aquel que no cumple las condiciones anteriores.

definido por quien requiere de la herramienta, determina si el cliente es clasificado como “bueno” o “malo”.

Las técnicas comúnmente utilizadas para *credit scoring* son métodos estadísticos tradicionales, como por ejemplo, análisis discriminante, regresión logística [3,12]. Otros métodos utilizados son algoritmos genéticos, *k-nearest neighbour* y sistemas expertos.

En el último tiempo se han utilizado métodos de *data mining* [1,15], tales como redes neuronales y árboles de decisión. Para el desarrollo de esta tesis se utiliza el modelo SVM, que entrega la pertenencia a una clase u otra y no su probabilidad, por lo que no es necesario definir un punto de corte para la probabilidad, como se tiene que realizar con otros métodos (ej. Regresión logística).

A continuación se presentan los beneficios y limitaciones de este método en particular.

Beneficios del *credit scoring* [15]

- Provee de criterios estándares para predecir el comportamiento de pago del solicitante del crédito, dado que el modelo provee un análisis objetivo. Se enfoca sólo en información relacionada con el riesgo crediticio y evita la subjetividad de quienes realicen el proceso de evaluación de créditos.
- Aumenta la velocidad de la asignación de créditos en el proceso crediticio: permite la automatización de la entrega de créditos y la cuantificación del riesgo asociado a entregar un crédito en menos tiempo (es decir no disminuye la calidad de la evaluación dada la disminución del tiempo).

Ejemplo, banco canadiense; bajó el tiempo de 9 días a 3 el tiempo de procesamiento de un crédito [15].

- Aporta información para determinar la tasa de interés que se debería cobrar a cada cliente. Se puede determinar los límites en crédito por cliente (también conocido como líneas de riesgo). Esto ayuda a manejar las cuentas de manera efectiva y rentable.

- Permite entregar créditos a clientes que tienen pocos registros crediticios y pueden ser rechazados, debido a impagos, falta de data en su historial crediticio y/o dificultad en validar sus ingresos. Debido a que un modelo de *credit scoring* no requiere de gran cantidad de variables para evaluar.

- En el largo plazo, disminuye los costos de evaluación de un crédito, ya que utiliza menor cantidad de información y no requiere la presencia de algún ejecutivo.

Limitaciones del *credit scoring* [15]

- El modelo puede ser construido a partir de una muestra de clientes estimada (no se encuentra toda la información de solicitudes de crédito, por ejemplo cuando no se guarda información sobre los créditos rechazados) y otra de clientes a los cuales se les concedió crédito. El modelo construido utilizando esta muestra puede (generalmente) no comportarse bien sobre la población completa.

- Los patrones cambian a través del tiempo. Lo que pocos modelos consideran en la actualidad.

- La omisión de variables o atributos importantes en el modelo; la información utilizada principalmente es la personal y sobre su historial crediticio, sin embargo, existen características como el empleo y las condiciones económicas actuales que no se consideran pues no son fácilmente observables.

- Existencia de errores en los reportes crediticios: los clientes pueden verse afectados a que no le entreguen crédito o a los proveedores en riesgo financiero.

- Es necesario que los individuos tengan toda la información utilizada por el modelo de *credit scoring* desarrollado, antes que un puntaje pueda ser calculado.

- Un problema en el uso se produce cuando se confía en la tecnología y se reduce el uso del juicio y de ejercer su conocimiento en casos especiales. "La tecnología está al servicio de las personas y no a la inversa".

4.2 Selección y Preprocesamiento de los datos

Esta sección corresponde a la Fase 1 de la metodología de clasificación dinámica descrito en la sección (3.1.3). Acá se explican los supuestos, preprocesamiento y los datos escogidos para la aplicación de esta metodología.

Los datos disponibles en INDAP [3], se encuentran desde enero de 1996 hasta diciembre del 2005. Las solicitudes de crédito seleccionadas para el estudio son los créditos otorgados desde enero del 2000 hasta diciembre del 2005 cuya fecha de vencimiento sea anterior al 31/12/2005. Esta elección se debe a que dada la información disponible se supuso que la historia de INDAP comienza en el año 1996 (INDAP existe desde el año 1962) y se considera razonable un

periodo de 4 años para construir el historial de comportamiento de pago de un cliente (dado que la mayoría de los créditos otorgados tienen un plazo de pago cercano a un año).

Los créditos entregados por INDAP se entregan en su mayoría a personas naturales (99,7%) por lo cual el estudio se concentra en este tipo de beneficiarios.

Luego de analizar los créditos de personas naturales, se determinan otras segmentaciones sobre la data, en base a la naturaleza de los créditos Nuevos[5] o Antiguos[6], Corto Plazo o Largo Plazo y para los de largo plazo se considera también la duración del crédito. Esto genera 6 segmentos en los datos [3].

Segmento	Descripción
NCP	Nuevos Corto Plazo
ACP	Antiguos Corto Plazo
NLP	Nuevos Largo Plazo con una duración entre 1 y 5 años
ALP	Antiguos Largo Plazo con una duración entre 1 y 5 años
NLP (+6)	Nuevos Largo Plazo con una duración mayor a 6 años
ALP (+6)	Antiguos Largo Plazo con una duración mayor a 6 años

Tabla 1. Descripción de los segmentos de los datos de INDAP

La información disponible en INDAP se distribuye en 5 fuentes [3]

- Información del cliente
- Información sobre el predio agrícola
- Rubro agrícola destino del crédito
- Características del Crédito
- Comportamiento de pago

[5] Se define un cliente Nuevo, aquel que se le otorga un crédito por primera vez, en otras palabras, aquel que no tiene comportamiento de pago, entre el año 1996 y la fecha en que solicita el crédito, con la institución.
[6] Se define un cliente Antiguo, aquel que ha tenido algún crédito con INDAP durante su historia (se supuso que la historia de INDAP comienza en 1996)

Todas las fuentes de información deben ser preprocesadas de modo de poder ser utilizadas en el desarrollo de los modelos. Algunas preprocesamientos generalmente aplicados [4,11,26] son:

- Estandarización de la información (las fuentes con el mismo formato), tener la información en la unidad de estudio 'solicitud de crédito' (para ello se agrega o resume la información de las fuentes que tienen más de un registro por crédito).

- Tratamiento de fuera de rango o *outliers* (datos que se encuentran fuera de límites normados, Ej. Año, el estudio se basa entre el 2000 y 2005).

- Valores perdidos o *missing value* (datos que no se encuentran para algunos registros, Ej. Un cliente sin estado civil) .

- Ruido o *noisy data* (datos que tienen un error o varianza en una variable numérica)

- Datos inconsistentes o *inconsistent data* (datos que son inconsistentes con los esperados, Ej. Edad de 150 años) [11].

A continuación se presenta el tratamiento específico para cada fuente de información.

Cliente

En la data se encuentran valores perdidos en objetos (solicitudes de crédito) cuya edad, estado civil o sexo es desconocido. Además se detectan datos inconsistentes en objetos cuya variable edad es menor a 18 años. Ambos tipos de datos son eliminados de la muestra.

Predio agrícola

Para esta fuente es necesario estandarizar la información, ya que el sistema donde reside la data tiene una codificación distinta a la actualmente utilizada en INDAP.

Se observa en la data más de un registro por predio agrícola. Para tratar estos datos erróneos se definen las variables relevantes que determinan un predio (rut, región, área, tenencia y tamaño) y se realiza una agrupación de los valores sobre los registros (de forma que dos registros iguales sean considerados como uno sólo), y así obtener la cantidad de predios que tiene un cliente.

En la data existen los casos en que un cliente puede tener más de un predio asociado, por lo que se requiere resumir la información de modo de tener sólo un registro para cada solicitud de crédito obteniendo la cantidad de predios que tiene un cliente y el tamaño de los predios como el resultado de la suma de ellos.

El tratamiento que reciben los valores perdidos y los datos inconsistentes presentes en esta fuente es eliminar los objetos de la muestra.

Rubro agrícola

En esta fuente se requiere estandarizar la información ya que el sistema de información actual utiliza categorías de Rubro distintas a las antiguamente utilizadas, para ello se requiere de una tabla de equivalencia de rubros provista por INDAP.

Un cliente puede tener asociado más de un rubro agrícola, por lo que es necesario determinar cual es el rubro más importante que tiene (rubro principal).

Para ello se selecciona aquel rubro que tenga asociado la mayor cantidad del dinero otorgado en el crédito.

Los objetos que tienen valores perdidos (rubro desconocido o 'FICTICIO') son eliminados de la muestra.

Características del Crédito

Los datos inconsistentes dados por objetos cuya fecha de vencimiento del crédito es anterior a su fecha de colocación (un crédito que termina antes que empiece) son eliminados de la muestra.

Se definen los siguientes rangos según el tipo de crédito

- Créditos de Corto Plazo cuya duración sea menor a 12 meses.
- Créditos de Corto Plazo cuyo monto otorgado esté entre 1.4 UF y 112 UF[7] para los créditos nuevos y entre 1.4 UF y 223 UF para los créditos antiguos.
- Créditos de Largo Plazo cuyo monto otorgado esté entre 4.2 UF y 500 UF.

Aquellos objetos que estén fuera de rango son eliminados de la muestra.

[7] Los montos se definen en base a la política de créditos vigente y a las definiciones propuestas de INDAP.

Comportamiento de pago

Se determina eliminar aquellas solicitudes que:

- Se realiza una colocación de crédito pero no hay registros de comportamiento, es decir, se registra una colocación pero no hay registros de pago, mora u algún otro tipo de movimiento.
- Se realiza la colocación pero es recuperada por medio de otros métodos (condonación, ajustes) y no por el pago de un cliente.
- Se realiza la colocación, se recupera en el mismo mes pero por un monto menor al colocado.
- Se realiza la colocación y se recupera el crédito completamente el mismo mes.

El preprocesamiento recién descrito, se realiza con el Software MS SQL Server v8.0.

La siguiente tabla muestra la cantidad de créditos y el monto entregado total para cada uno de estos segmentos, luego de realizar el preprocesamiento antes descrito.

Segmento	Cantidad	Porcentaje	Monto Total (UF)	Porcentaje
NCP	23,304	9.4%	442,323	6.9%
ACP	176,594	71.2%	4,400,600	68.7%
NLP	9,721	3.9%	359,507	5.6%
ALP	38,005	15.3%	1,170,619	18.3%
NLP (+6)	76	0.0%	4,647	0.1%
ALP (+6)	295	0.1%	26,778	0.4%
Total	247,995	100.0%	6,404,474	100.0%

Tabla 2. Caso INDAP: Cantidad de Créditos y Montos Entregados por segmento

Para el desarrollo de esta tesis se utilizará el segmento ACP para la aplicación de la metodología de Clasificación Dinámica propuesta por tres razones.

i. Dado el objetivo de esta tesis, es necesario tener objetos que tengan un tiempo de evolución normalizado (debe cumplirse el ciclo completo de asignación y retorno del crédito) para aplicar la metodología. El tiempo de retorno de un crédito es de un año para los créditos ACP, mientras que para los créditos de Largo Plazo el retorno va entre 1 y 5 años (19.2% de los créditos) y entre 5 y 10 años (0.1% de los créditos), por lo que se pierde gran cantidad de objetos si se investiga este caso.

ii. Es uno de los universos más importantes para la institución, tanto en frecuencia de créditos que otorga como en el monto entregado (ver Tabla 2).

iii. Este segmento tiene una gran cantidad y diversidad de atributos que lo hacen apropiado para la aplicación de la nueva metodología, no así el segmento NCP.

Dado lo anterior para los segmentos NCP, NLP, ALP, NLP (+6) y ALP (+6) no se construirá un modelo de clasificación. Ni podrá ser aplicable al modelo desarrollado.

Los objetos del segmento en estudio (ACP) se separaron en 9 conjuntos, de modo de tener una mayor cantidad de muestras donde aplicar y comparar la metodología D-SVM. Los conjuntos se confeccionaron según la región donde se otorgó el crédito de acuerdo a características similares como factores climáticos que determinan el nivel y tipo de producción. Los conjuntos son los siguientes.

- Norte: I, II, III, IV Región.
- Quinta: V Región.
- Sexta: VI Región.
- Séptima: VII Región.
- Octava: VIII Región
- Novena: IX Región

- Décima: X Región
- Sur: XI y XII Región
- Metropolitana: Región Metropolitana

El primer y octavo conjunto agrupa más de una región, debido a la baja cantidad de datos que tienen estas regiones particularmente.

En la tabla 3 se observan la cantidad de objetos para cada conjunto, el monto otorgado y el porcentaje de Recuperación[8].

Conjunto	Cantidad de Créditos		Monto [UF]		Recuperación
	Valor	Porcentaje	Valor	Porcentaje	
Norte	6,394	3.6%	306,173	7.0%	87.9%
Quinta	6,577	3.7%	247,775	5.6%	89.4%
Sexta	11,748	6.7%	694,945	15.8%	89.7%
Séptima	24,932	14.1%	865,035	19.7%	92.1%
Octava	29,080	16.5%	595,639	13.5%	96.2%
Novena	29,117	16.5%	649,342	14.8%	95.2%
Décima	60,543	34.3%	822,011	18.7%	98.1%
Sur	5,172	2.9%	77,442	1.8%	95.9%
Metropolitana	3,031	1.7%	142,239	3.2%	87.7%
Total	176,594	100.0%	4,400,600	100.0%	95.0%

Tabla 3. Cantidad, Monto y Recuperación de cada conjunto del Segmento ACP

Se definen 9 atributos que serán utilizados para construir el modelo dinámico y el estático. Para efectos de esta aplicación no se presenta el proceso de selección, por lo que los atributos son considerados como datos fijos[9].

[8] Porcentaje de créditos de los cuales se recupera totalmente el monto otorgado, sin haber tenido algún tipo de condonación u otro tipo de descuento sobre el monto adeudado.
[9] En la sección 4.4.1.1 se extiende la metodología para considerar la selección de atributos.

4.3 Aplicación Metodología D-SVM

Esta sección describe la metodología propuesta (sección 3.1), en particular las Fases 2 al 5, ya que la Fase 1 se describe en detalle en la sección 4.2 de este capítulo.

Se utiliza el conjunto catalogado por 'Norte', el cual tiene las solicitudes de crédito de la I, II, III y IV región, para describir cada fase. Para los conjuntos restantes el procedimiento es análogo.

La siguiente tabla muestra una descripción de los datos de este conjunto.

Conjunto	Año	Cantidad	Monto [UF]	No Recuperación
Norte	2000	1,155	52,300	16.8%
Norte	2001	1,218	52,580	12.0%
Norte	2002	1,143	50,979	11.3%
Norte	2003	1,296	60,821	9.3%
Norte	2004	1,248	64,278	11.0%
Norte	2005	334	25,215	14.4%

Tabla 4. Frecuencia, Monto y Porcentaje de No Recuperación Anual del Conjunto 'Norte'

La segunda fase de la metodología es determinar la unidad de medida asociado a la ventana de tiempo, el cual debe determinar en base a las características tanto del problema de clasificación como del entorno donde se realiza el desarrollo. De este modo para el problema de *credit scoring* en INDAP, se define la unidad de medida igual a un semestre[10]. La siguiente tabla detalla los datos de el conjunto 'Norte' dada esta unidad de medida.

[10] Aunque por razones de estacionalidad en el sector agrícola, esta unidad debiese ser un año, se ha definido en un semestre por razones experimentales.

Conjunto	Año	Semestre	Cantidad	Monto [UF]	No Recuperación
Norte	2000	Primer	778	36,031	18.4%
Norte	2000	Segundo	377	16,269	13.5%
Norte	2001	Primer	643	28,588	14.0%
Norte	2001	Segundo	575	23,992	9.7%
Norte	2002	Primer	646	33,002	13.9%
Norte	2002	Segundo	497	17,977	7.8%
Norte	2003	Primer	712	35,492	11.4%
Norte	2003	Segundo	584	25,329	6.7%
Norte	2004	Primer	690	37,536	11.2%
Norte	2004	Segundo	558	26,741	10.8%
Norte	2005	Primer	323	24,745	14.9%
Norte	2005	Segundo	11	470	0.0%

Tabla 5. Frecuencia, Monto y Porcentaje de No Recuperación Semestral del Conjunto 'Norte'

Otra definición que se requiere hacer en esta etapa, es el periodo de actualización para el modelo, el que determina el tiempo que transcurrirá entre la confección de un modelo y el siguiente. Este se fija en un año, debido a que el tiempo de retorno de un crédito ACP y como consecuencia se determina la clase de pertenencia.

La tercera fase es la construcción de los modelos de clasificación en distintos instantes del tiempo. Los que serán definidos en 4 periodos como sigue

Periodo 1. Créditos otorgados en los años 2000 y 2001, se toma dos años como primer periodo, para tener una cantidad suficiente de objetos para tener un año de historia y un año para evaluar el modelo

Periodo 2. Créditos otorgados el año 2002, los siguientes periodos están definidos por el periodo de actualización definido, es decir, un año.

Periodo 3. Créditos otorgados el año 2003.

Periodo 4. Créditos otorgados el año 2004.

El año 2005 lo excluiremos de la construcción de modelos, ya que al determinar la fecha de vencimiento del 31/12/2005 (sección 4.2) como una de las condiciones sobre los créditos observados para el estudio, no existe el tiempo suficiente para que los créditos otorgados el 2005 se hayan recuperado, por lo tanto no se está en un periodo de tiempo donde se conoce cómo se desenvolvieron (si se recupera o no) las solicitudes de crédito otorgadas.

Se realiza el detalle de esta fase para el primer periodo definido ya que para los siguientes es análogo, es decir, se está en un tiempo donde se conoce cómo se desenvolvieron (si se recupera o no) las solicitudes de crédito otorgadas desde Enero 2000 hasta Diciembre 2001. Utilizando la unidad de medida de la ventana de tiempo antes definida (Semestre) se confeccionan 3 conjuntos de datos que formaran los modelos definidos por las 3 ventanas de tiempo posible.

- Enero 2000 hasta Diciembre 2001 (2.373 datos)
- Julio 2000 hasta Diciembre 2001 (1.595 datos)
- Enero 2001 hasta Diciembre 2001 (1.218 datos)

Notar que el tamaño mínimo de la ventana es de dos unidades de tiempo (un año) debido a que es el periodo en que se conoce el comportamiento de un crédito.

Para la penalización de los errores (C), se define una grilla de valores posibles que determinarán los modelos especificados por las ventanas de tiempo. Esta grilla se define con 37 valores entre 0,01 y 100 debido a que experimentos preliminares determinaron que valores sobre 100 no tienen alguna variación en el porcentaje de acierto al igual que con valores menores a 0,01. Los valores son:

- 0,01; 0,02; 0,03; 0,04; 0,05; 0,06; 0,07; 0,08; 0,09; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; 1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 20; 30; 40; 50; 60; 70; 80; 90; 100

Para la ponderación de los objetos (λ), se define otra grilla de valores posibles, que determina la tasa con la que los errores del pasado irán disminuyendo según la función de edad exponencial definida. En este caso se utiliza una grilla hasta 0.5, porque un valor mayor es equivalente a considerar con mucha importancia los datos recientes sobre los antiguos, que es el efecto que se quiere obtener con la ventana de tiempo. Los valores son:

- 0; 0,1; 0,2; 0,3; 0,4; 0,5

Notar que el valor cero corresponde a no realizar ninguna ponderación, es decir, todos los errores, independiente de cuando ocurren, valen igual. Además se destaca que el periodo donde se encontró el dato (t_x), no está definido directamente por la fecha, sino por el semestre al cual pertenece, de este modo los valores son.

- Semestre Julio 2001 hasta Diciembre 2001, t_x toma el valor 0
- Semestre Enero 2001 hasta Junio 2001, t_x toma el valor 1
- Semestre Julio 2000 hasta Diciembre 2000, t_x toma el valor 2
- Semestre Enero 2000 hasta Junio 2000, t_x toma el valor 3

La cuarta fase es construir los distintos modelos que permiten definir la mejor configuración de parámetros para este periodo de tiempo.

Antes de comenzar con los procesos correspondientes a esta etapa, se debe definir la manera en que se van a evaluar los distintos modelos a construir. Para ello se define el siguiente procedimiento [16,26]:

Balancear[11] cada unidad de medida de la ventana de tiempo. Dado que el problema se caracteriza por tener una pequeña fracción de casos malos (no recuperación del crédito) se construyen subconjuntos de cada unidad de ventana de tiempo (para este caso son 4 unidades) donde se incluyen todos los objetos de la clase más pequeña y se incluye la misma cantidad de objetos de la otra clase en forma aleatoria. El resultado para el caso ejemplificado son 4 subconjuntos balanceados, los cuales serán reunidos según la ventana de tiempo a utilizar.

Lo anterior se realiza 10 veces[12], de modo que se obtienen 10 grupos de datos balanceados donde se evaluarán todos los parámetros y así tener una medida confiable sobre la predicción y elección del modelo resultante.

Construidos los 10 grupos de datos, se define el conjunto de testeo que se utilizará para evaluar la mejor ventana de tiempo y el modelo predictivo final. Notar que este conjunto debe ser igual para todas las ventanas de tiempo, es decir, para los 3 conjuntos definidos, además se obtiene de los últimos datos definidos por el periodo de actualización (1 año). Esta fracción se ha definido en un 30% de los datos[13] y el resto será parte del conjunto para confeccionar y calibrar el modelo.

A continuación se deben construir los conjuntos de entrenamiento y validación, el que se construye con el 70% remanente del último periodo de actualización y los datos anteriores según corresponda. De este modo los conjuntos de entrenamiento, validación y testeo para cada ventana de tiempo respectivamente son.

---

[11] Igual cantidad de objetos en cada clase de pertenencia

[12] Basado en el proceso 10-fold cross validation, el que divide la muestra en 10 subconjuntos, donde 9 de ellos son utilizados para entrenamiento y 1 de ellos para testeo, este proceso se realiza hasta que los 10 subconjuntos sean testeados.

[13] Porcentaje utilizado en diversas aplicaciones de minería de datos, para evaluar la efectividad de un modelo [26]

- Enero 2000 hasta Diciembre 2001: 414 – 178 – 88
- Julio 2000 hasta Diciembre 2001: 214 – 92 – 88
- Enero 2001 hasta Diciembre 2001: 142 – 62 – 88

En la sección 4.2 se define utilizar 9 variables de las cuales distinguimos 5 de ellas como variables nominales y las 4 restantes como variables continuas. Para utilizar el método de SVM, que es la base de la metodología de Clasificación Dinámica propuesta, es necesario realizar transformaciones a cada una de estas variables, para ellos realizamos el siguiente procedimiento según el tipo de variable [4].

Variables Nominales

Utilizar la codificación N-1 [4], que consiste en confeccionar n-1 variables donde n es la cantidad de categorías de la Variable. Por ejemplo, para la variable estado civil, los valores posibles son; Soltero, Casado, Separado y Viudo, se debe generar 3 variables a partir de los 4 valores posibles, tal como se muestra a continuación.

Valor Variable	X1	X2	X3
Soltero	0	0	0
Casado	1	0	0
Separado	1	1	0
Viudo	1	1	1

Tabla 6. Ejemplo de codificación N-1 para variables nominales

Variables Continuas

Normalizar los datos [4] (Ec. 24), de modo que las variables se encuentre en un rango numérico comparable. Se distingue x_i como el valor original de la variable, $\bar{x}_i$ el promedio de la variable en el conjunto de entrenamiento y σ_{x_i} es la desviación estándar de la variable en el conjunto de entrenamiento.

$$\frac{x_i - \bar{x}_i}{\sigma_{x_i}} \qquad \text{(Ec. 24.)}$$

Para finalizar la transformación de las variables [4], se escalan todos los atributos (codificados y normalizados) (Ec. 25) a un rango [0;1] y la variable objetivo debe estar en un rango [-1;1]. Se distingue x_i como el valor codificado o normalizado de la variable según corresponda, $Rango_{Inf}$, $Rango_{Sup}$ son el valor Mínimo y Máximo del rango a escalar, $Min(x_i)$, $Max(x_i)$ son el valor Mínimo y Máximo de la variable en el conjunto de entrenamiento.

$$(Rango_{Sup} - Rango_{Inf}) \cdot \frac{x_i - Min(x_i)}{Max(x_i) - Min(x_i)} + Rango_{\inf} \qquad \text{(Ec. 25.)}$$

La estimación de los valores de la variable se realiza con los resultados obtenidos desde el conjunto de entrenamiento (media, desviación estándar, mínimo y máximo) y estos valores son utilizados para normalizar o escalar, según la naturaleza de la variable (nominal o continua), los conjuntos de validación y testeo.

Después de las transformaciones descritas, se determina la mejor configuración de parámetros para cada ventana de tiempo en base a la predicción sobre el conjunto de validación.

La construcción de los 10 grupos, ventanas de tiempo, balance de las clases y la transformación de las variables se realiza con el Software MATLAB v7.0.

La cantidad de modelos con distintos parámetros asciende a 222 (cantidad de combinaciones entre C y λ) para las 3 ventanas de tiempo construidas y esto para los 10 grupos balanceados. El método de evaluación utilizado para determinar la mejor configuración es la matriz de confusión [16,26], descrita a continuación.

		Clase Real			
		Cantidad		Porcentaje	
		Bueno	Malo	Bueno	Malo
Clase Predecida	Bueno	a	c		Error Tipo 1
	Malo	b	d	Error Tipo 2	

Tabla 7. Matriz de confusión

, donde "a" corresponde a la cantidad de créditos clasificados como buenos y que son buenos, "b" son aquellos clasificados como malos y que son buenos (error tipo 2), "c" son los clasificados como buenos y que son malos (error tipo 1) y "d" son los clasificados como malos y que son malos. Porcentualmente el error tipo 1 y tipo 2 se calcula de la siguiente forma:

Error Tipo 1 = c / (a+c)

Error Tipo 2 = b / (b + d) (Ec. 26.)

Error Total = (b+c) / (a+b+c+d)

Para el caso de *credit scoring* el Error Tipo 1 también recibe el nombre de "riesgo crediticio" (si este error es alto, la institución está expuesta al riesgo) y el Error Tipo 2 es el "riesgo comercial" (si este error es alto la institución pierde participación de mercado) [3]. Dado que el interés de esta tesis es investigar sobre clasificación dinámica, se utiliza como medida de efectividad el Error Total, el que supone que ambos tipos de errores tienen igual ponderación, en

otras palabras, equivocarse en una unidad para el Error Tipo 1 es equivalente a equivocarse en uno para el Error Tipo 2.

Los modelos de SVM para cada ventana de tiempo y parámetros antes descritos, se confeccionan en el Software GAMS y es resuelto con el *Solver* MINOS v5.5. Para la evaluación de los resultados se utiliza MS Excel 2003.

La siguiente tabla muestra el mejor resultado promedio obtenido en el conjunto de validación para cada ventana de tiempo.

Ventana de tiempo		Parámetros		Efectividad
Inicio	Fin	C	λ	Promedio
ene-00	dic-01	100	0.5	63.1%
jul-00	dic-01	100	0.3	57.7%
ene-01	dic-01	0.6	0	61.2%

Tabla 8. Efectividad conjunto de validación: Primer periodo conjunto 'Norte'

Cada una de estas configuraciones determina la efectividad en el conjunto de testeo

Ventana de tiempo		Parámetros		Efectividad
Inicio	Fin	C	λ	Promedio
ene-00	dic-01	100	0.5	63.3%
jul-00	dic-01	60	0.3	63.3%
ene-01	dic-01	0.6	0	63.0%

Tabla 9. Efectividad conjunto de testeo: Primer periodo conjunto 'Norte'

Se selecciona aquella ventana que tiene la mejor Efectividad Promedio[14], que corresponde a la primera (ene-00 a dic-01). Este valor también determina la efectividad del primer periodo.

La última Fase corresponde a generar la data para el siguiente periodo de actualización, en el cual se recolectan los datos desde este instante de tiempo hasta el siguiente periodo (enero 2002 hasta diciembre 2002) y se adjunta la

[14] En el caso que haya dos ventanas de tiempo la misma Efectividad Promedio, se selecciona la más extensa.

ventana de tiempo resultante del periodo anterior (enero 2000 hasta diciembre 2001).

Luego esta data debe ser tratada tal como se describió en las secciones 4.2 y realizar el proceso descrito en esta sección con la nueva data.

4.3.1 Comparación Metodología de Clasificación Dinámica y Estática

En la siguiente tabla de detallan los resultados obtenidos utilizando la metodología D-SVM para cada conjunto definido por la región donde se otorgó el crédito.

Conjunto	Ventana de tiempo		Parámetros		Efectividad Promedio	
	Inicio	Fin	C	λ	Cantidad	Porcentaje
Norte	ene-00	dic-01	100	0.5	88	63.3%
	ene-01	dic-02	9	0	78	69.0%
	jul-01	dic-03	60	0	72	61.9%
	jul-01	dic-04	7	0	112	67.8%
Quinta	ene-00	dic-01	70	0.2	70	69.0%
	ene-00	dic-02	20	0	68	72.2%
	jul-00	dic-03	4	0	76	68.3%
	jul-01	dic-04	4	0	112	62.9%
Sexta	jul-00	dic-01	3	0.2	202	72.3%
	ene-01	dic-02	20	0.3	110	70.3%
	jul-01	dic-03	20	0.3	112	69.4%
	ene-02	dic-04	70	0.3	168	66.3%
Septima	ene-00	dic-01	6	0.2	300	76.4%
	jul-00	dic-02	3	0.2	194	74.2%
	ene-02	dic-03	6	0.5	140	71.0%
	jul-02	dic-04	50	0	248	69.2%
Octava	ene-00	dic-01	9	0.5	134	76.0%
	jul-00	dic-02	3	0	110	75.5%
	ene-01	dic-03	10	0	106	75.3%
	ene-01	dic-04	5	0	136	72.2%
Novena	ene-01	dic-01	3	0	90	72.7%
	ene-01	dic-02	40	0.3	158	68.5%
	ene-01	dic-03	0.7	0	130	67.5%
	jul-01	dic-04	3	0.2	100	67.8%
Décima	jul-00	dic-01	50	0.1	100	72.3%
	jul-00	dic-02	10	0	118	75.4%
	jul-00	dic-03	30	0.1	178	72.7%
	ene-04	dic-04	4	0.1	220	68.7%
Sur	ene-01	dic-01	0.4	0	16	53.8%
	ene-02	dic-02	4	0	18	72.2%
	ene-02	dic-03	60	0	12	71.7%
	jul-02	dic-04	70	0	40	63.5%
Metropolitana	ene-00	dic-01	60	0	34	69.4%
	ene-01	dic-02	0.2	0.1	32	66.3%
	jul-01	dic-03	0.5	0.3	26	64.6%
	jul-01	dic-04	40	0	64	65.0%

Tabla 10. Resultados utilizando metodología D-SVM

La tabla anterior detalla cada una de las ventanas de tiempo resultantes, la mejor configuración de parámetros para cada periodo de actualización y la cantidad de objetos en el conjunto de test y su efectividad promedio.

Para analizar los beneficios que tiene la metodología D-SVM, se comparan los resultados de la tabla anterior con los obtenidos al utilizar una clasificación estática con el método SVM lineal (Ec. 3).

Para ello se mantienen los mismos objetos en el conjunto de test de cada Conjunto (Norte, Quinta, etc.) y se divide el resto de la muestra por una proporción de 70% para el conjunto de training y 30% para el de validación. Otro factor utilizado en la comparación, como se define en la sección 4.2, es utilizar los mismos 9 atributos en ambos tipos de clasificaciones.

La siguiente tabla muestra los resultados obtenidos utilizando ambas metodologías, mostrando el Porcentaje de Efectividad promedio sobre los 10 grupos balanceados para cada metodología.

Conjunto	Metodología	
	D-SVM	SVM
Norte	65.7%	65.2%
Quinta	67.4%	67.2%
Sexta	69.7%	69.5%
Séptima	73.0%	72.5%
Octava	74.7%	73.7%
Novena	68.9%	68.9%
Décima	71.7%	70.9%
Sur	64.7%	66.6%
Metropolitana	66.2%	66.7%

Tabla 11. Comparación metodología D-SVM y SVM

De la tabla anterior se observa que para 6 de los 9 conjuntos definidos, la metodología D-SVM fue superior.

4.4 Aplicación Metodología de Clasificación Dinámica para la Selección de Atributos

Los datos utilizados para aplicar esta nueva metodología son los descritos en la sección 4.2, con la diferencia que se utilizan 18 atributos en vez de los 9 antes descritos. Estos atributos tienen información sobre características del cliente, del crédito otorgado, comportamiento de pago en créditos anteriores y del predio agrícola.

El objetivo es mostrar una aplicación de la nueva metodología y compararla con un método de clasificación estático con selección de atributos.

En la primera parte se describe el método estático utilizado y luego se describe la metodología D-SVM utilizando el modelo LP-SVM (sección 3.2.1). En ambos casos se utilizan los 9 conjuntos de datos descritos en la sección 4.3 y se realizan las transformaciones pertinentes a cada conjunto.

4.4.1 Clasificación Estática con Selección de Atributos

Para resolver este tipo de problema se comienza seleccionando los atributos que serán utilizados en la confección del modelo de clasificación. Se hace notar que la metodología consta de dos procesos independientes; Selección de atributos y Construcción del modelo de clasificación.

4.4.1.1 Selección de atributos

Se utiliza como técnica la regresión logística [3,12] por su confiabilidad estadística. Esta técnica describe la relación entre una variable “objetivo”

(también llamada variable “dependiente” o variable “respuesta”) y una o más variables “explicativas” (también llamadas variables “independientes” o covariables). Es frecuente encontrar casos en que la variable “respuesta” es discreta, tomando dos o más valores posibles. En particular, en los sistemas de *credit scoring* se generan modelos que permitan discriminar entre clientes “buenos” y “malos” (sección 4.1).

Dado que se pronosticará un evento dicotómico (y_i) en base a la información de ‘m’ variables independientes $(x_1,...,x_m)$, la regresión logística busca determinar la probabilidad de ocurrencia del evento dicotómico en función de la información contenida en las variables independientes, asumiendo una relación funcional (Ec. 27). La probabilidad de ocurrencia del evento que se estudia (denotado por $\pi(x)$) es función de los valores de las variables independientes $x = (x_1,...,x_m)$. De esta manera, cuando se quiere ajustar un modelo de regresión logística a un conjunto de observaciones (x_i, y_i) con i = 1,…, n, lo más común es estimar el valor de los coeficientes $\beta = (\beta_0,...,\beta_m)$ de acuerdo al método de *máxima verosimilitud*. En términos generales, el método de máxima verosimilitud encuentra los valores de los parámetros desconocidos que maximizan la probabilidad de obtener el conjunto de datos observados. De esta manera, se encuentran estimadores $\hat{\beta} = (\hat{\beta}_0,...,\hat{\beta}_m)$ de los parámetros, y con ello se genera el modelo predictivo buscado.

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + ... + \beta_m x_m}}{1 + e^{\beta_0 + \beta_1 x_1 + ... + \beta_m x_m}} \qquad \text{(Ec. 27.)}$$

Los datos son separados en dos conjuntos de datos[15]; entrenamiento (70%), conjunto con el cual se confecciona el modelo para seleccionar variables y validación (30%), conjunto con el cual se evalúa el desempeño del modelo y por consiguiente la selección de atributos.

Se utilizaron tres medidas para determinar los atributos a seleccionar.

1. Comparar distintos enfoques de selección de variables el que permite especificar cómo se introducen las variables independientes en el análisis. Utilizando distintos métodos se pueden construir diversos modelos de regresión a partir del mismo conjunto de variables. Los métodos son [24]:

    a. Selección hacia adelante - Razón de verosimilitud (FORWARD): Método de selección por pasos que contrasta la entrada de variables basándose en la significación del estadístico de puntuación y contrasta la eliminación basándose en la probabilidad del estadístico de la razón de verosimilitud, que se basa en estimaciones de la máxima verosimilitud parcial.

    b. Eliminación hacia atrás - Razón de verosimilitud (BACKWARD): Método de selección por pasos hacia atrás. El contraste para la eliminación de variables se basa en la probabilidad del estadístico de la razón de verosimilitud, que se basa a su vez en las estimaciones de máxima verosimilitud parcial.

2. Comparación del mejor promedio de acierto en las clases ‘recuperada’ y ‘no recuperada’ en el conjunto de entrenamiento, evaluándose su desempeño además en el conjunto de validación.

3. Comparar la complejidad del modelo en términos del número de variables involucradas y el sentido lógico de las relaciones indicadas por el modelo,

[15] Porcentaje de separación recomendado [26].

prefiriéndose modelos más simples. Por ejemplo, si un modelo de 10 variables tiene un desempeño levemente inferior a otro modelo que tiene 12 variables, se prefiere el modelo de 10 variables debido a su simplicidad.

Se utiliza el Software SPSS v11.5 en español para resolver el problema recién descrito.

La siguiente tabla resume los resultados obtenidos en el conjunto “Norte” en cada uno de los 10 grupos balanceados.

	METODO						
	FORWARD		BACKWARD				
NORTE	Acierto [%]	Cantidad Atributos	Acierto [%]	Cantidad Atributos	Diferencia Acierto [%]	Método Seleccionado	Cantidad Atributos Seleccionados
Grupo 1	65.0%	10	65.0%	10	0.0%	Ambos	10
Grupo 2	63.1%	9	63.6%	11	0.5%	Forward	9
Grupo 3	63.6%	10	63.6%	10	0.0%	Ambos	10
Grupo 4	68.1%	9	68.9%	9	0.8%	Backward	9
Grupo 5	60.6%	9	62.2%	10	1.6%	Backward	10
Grupo 6	62.8%	8	62.5%	9	0.3%	Forward	8
Grupo 7	64.2%	8	65.0%	11	0.8%	Forward	8
Grupo 8	64.4%	10	65.6%	11	1.2%	Backward	11
Grupo 9	66.4%	10	66.4%	10	0.0%	Ambos	10
Grupo 10	63.1%	8	63.1%	8	0.0%	Ambos	8

Tabla 12. Resultados Selección de Atributos conjunto “Norte” por Grupo

En ella se observa el porcentaje de acierto y la cantidad de atributos entregados por cada método de selección de atributos. La columna “Diferencia Acierto [%]” presenta el valor absoluto de la diferencia entre los aciertos obtenidos por un método u otro, que es utilizado para evaluar los atributos a seleccionar. La columna “Método Seleccionado” que tiene un valor “Ambos” cuando los dos métodos entregan los mismos atributos y como consecuencia una Diferencia de Acierto igual a 0.0%, en caso contrario aparece el nombre del método seleccionado. Finalmente se presenta la “Cantidad de Atributos Seleccionados” que es una copia de la cantidad de atributos del Método elegido.

Para ejemplificar la metodología y criterios de selección de atributos, se explica el procedimiento utilizado en los grupos 2, 4 y 8.

- Grupo 2: De la tabla se observa una diferencia de acierto asciende de 0.5% (el mayor acierto es del método Backward) y que la cantidad de atributos del método Forward es de 9 mientras que el del otro es de 11. Utilizando sólo el criterio de acierto, se seleccionaría el método Backward pero implica utilizar 2 atributos más, tomando en cuenta que no se pierde un porcentaje importante de acierto al elegir el otro método se elige utilizar una menor cantidad de atributos ya que disminuye la complejidad del modelo.

- Grupo 4: En este caso la diferencia de acierto asciende a un 0.8% (el mayor acierto es del método Backward) y la cantidad de atributos es igual para ambos métodos. En este caso la elección se basa sólo en el porcentaje de acierto, con lo que el método seleccionado es el Backward.

- Grupo 8: Acá la diferencia de acierto asciende a un 1.2% (el mayor acierto es del método Backward) y la cantidad de atributos del método Forward es de 10 y el del otro es 11. Es fácil ver que el método Forward es el que tiene una menor cantidad de atributos, pero la pérdida de acierto que se obtiene es de un 1.2%, lo cual es considerable, por lo tanto se eligen los atributos seleccionados con el método Backward.

Con el fin de detallar y ejemplificar los atributos seleccionados en cada grupo, se presenta el siguiente gráfico que muestra la frecuencia de selección de cada atributo[16].

[16] Notar que la frecuencia máxima a obtener por un atributo es de 10, que es la cantidad de grupos confeccionados para cada conjunto.

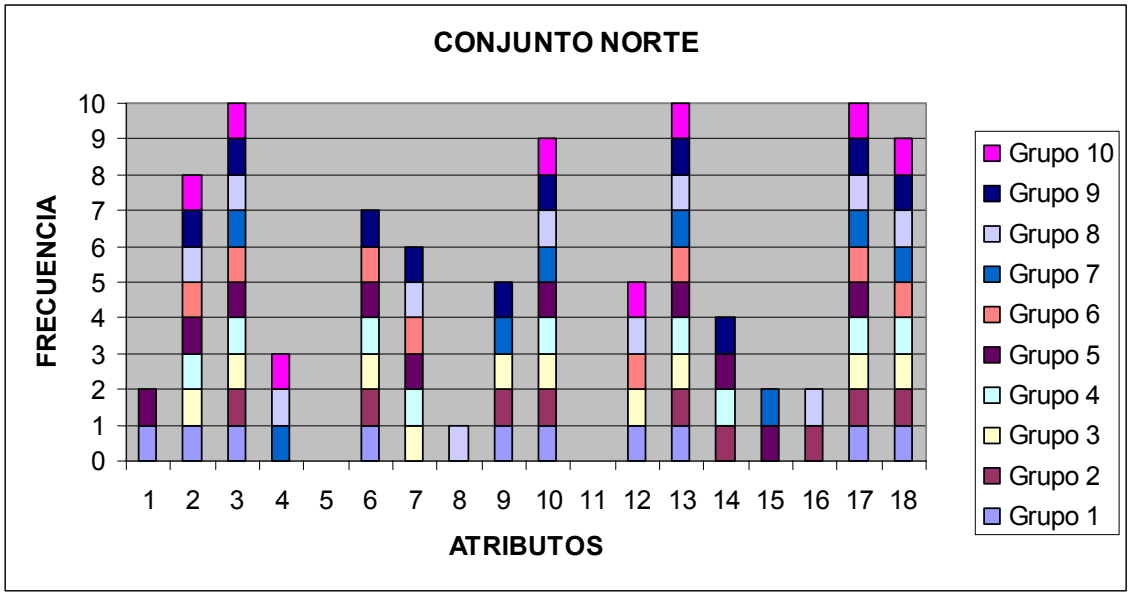


Figura 8. Frecuencia Acumulada para cada Atributo en el conjunto "Norte"

Se observa que los atributos 5 y 11 no fueron seleccionados por ninguno de los grupos mientras que los atributos 3, 13 y 17 están presentes en todos. Estas frecuencias muestran la relevancia de cada atributo en la confección del modelo de clasificación.

Los resultados de los otros conjuntos se encuentra en el Anexo 1, presentando en primer lugar la tabla que permite seleccionar el método que define los atributos a utilizar y luego el gráfico con la frecuencia acumulada por atributo.

4.4.1.2 Modelo de Clasificación Estático

Definidos los atributos a utilizar en cada grupo, se procede con el siguiente paso, que corresponde a confeccionar el modelo SVM.

Se define para la penalización de los errores (C), una grilla de 28 valores posibles que determinan el parámetro requerido por el modelo SVM. Se

escogen los valores entre 0,1 y 100 debido a que experimentos preliminares sobre estos datos determinaron que valores sobre 100 no tiene alguna variación en el porcentaje de acierto al igual que con valores menores a 0,1. La grilla utilizada es la siguiente:

- 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; 1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 20; 30; 40; 50; 60; 70; 80; 90; 100

El Software utilizado para construir los distintos modelos es MATLAB v7.0.

La tabla 13, muestra el porcentaje de acierto (o efectividad del modelo), dado por el promedio de acierto de los 10 grupos para cada conjunto de datos, de la mejor configuración (columna Parámetro) obtenida según el porcentaje de acierto en el conjunto de validación y la cantidad de atributos seleccionados, como el promedio de los atributos seleccionados en el paso anterior.

Conjunto	Acierto [%]	Parámetro [C]	Cantidad Atributos Seleccionados
Norte	66.74%	2.0	9.3
Quinta	65.89%	80.0	9.4
Sexta	69.43%	3.0	9.8
Séptima	72.07%	0.7	11.6
Octava	74.57%	30.0	12.0
Novena	68.60%	10.0	10.2
Décima	72.65%	1.0	9.8
Sur	63.72%	5.0	6.2
Metropolitana	64.68%	60.0	7.4

Tabla 13. Resultados Modelo de Clasificación Estático

4.4.2 D-SVM con Selección de Atributos

Esta sección detalla la metodología D-SVM con selección de atributos, descrita en la sección 3.2.

La primera fase es el preprocesamiento de los datos, en el que se escogen los datos que serán utilizados en la confección del modelo (sección 4.2). Se determina eliminar de la muestra aquellos datos cuyos atributos de edad, sexo, estado civil, predio, rubro y comportamiento de pago tengan un valor desconocido o inconsistente También se definen los montos (mayor a 1.4 UF para los créditos de corto plazo, con un límite máximo de 112 UF para los nuevos y 223 UF para los antiguos; en los créditos de largo plazo se definen un límite inferior de 4.2 UF y uno superior de 500 UF) y plazos válidos (duración menor a 12 meses para uno de corto plazo) para un crédito. Además se determina que el conjunto de datos a utilizar son los créditos ACP (Antigüos Corto Plazo), debido a sus características de plazo y a la importancia que tiene para INDAP. Además se define seccionar estos datos en 9 conjuntos.

La segunda fase corresponde a especificar la unidad de medida de la ventana de tiempo y el periodo de actualización, que al igual que en la sección 4.3, es de un semestre y un año respectivamente (dejando siempre como tamaño mínimo para una ventana igual al periodo de actualización).

La tercera fase detalla las diferencias existentes con respecto a la metodología original D-SVM y se hará un resumen de aquellas etapas donde el procedimiento es idéntico al presentado en la sección 4.3.

El objetivo de esta fase es determinar que parámetros son utilizados en la construcción de los modelos de clasificación y el desarrollo de los mismos en los distintos instantes del tiempo (4 periodos).

Se definen los siguientes parámetros.

- Ventana de tiempo, definida por la cantidad máxima de conjuntos posibles de construir con los datos disponibles en un periodo de tiempo, considerando una ventana mínima de 1 año (definida por el periodo de

actualización). Por ejemplo si los datos disponibles son de ene-01 a dic-03, se pueden construir 5 ventanas de tiempo, ene-01 a dic-03, jul-01 a dic-03, …, ene-03 a dic-03.

- Se define la siguiente grilla para la penalización de los errores (C_1), en base a pruebas preliminares sobre el conjunto de datos.
    - 1; 3; 5; 7; 9; 10; 30; 50; 70; 80; 90
- Se define la siguiente grilla para la penalización de los atributos (C_2).
    - 0,1; 0,3; 0,5; 0,7; 0,9; 1; 3; 5; 7; 9; 10; 20
- Ponderación de los objetos (λ), se define la siguiente grilla, que corresponde a la tasa en que disminuye la importancia del error del pasado según la función de edad exponencial definida.
    - 0; 0,1; 0,2; 0,3; 0,4; 0,5

La cuarta fase corresponde a la construcción de los modelos, definidos por las combinaciones dadas de los parámetros recién descritos (528 en total).

Se destaca que antes de comenzar con el desarrollo de los modelos, los 9 conjuntos deben ser balanceados (de modo de obtener 10 grupos para cada conjunto y cada periodo) y se definen los conjuntos de entrenamiento, validación y testeo (al igual que en la sección 4.3, se define un 30% de los datos del último año para testeo y el 70% remanente más los anteriores definen el conjunto de entrenamiento y validación, en una proporción de 70% y 30% respectivamente).

Definido los conjuntos se procede a la transformación de los datos [4] que para las variables nominales (13 de 18) se utiliza la codificación N-1 y las variables

continuas (5 de 18) son normalizadas. Finalmente todas los atributos son escalados en un rango [0;1] y la variable objetivo en un rango [-1;1].

Todo el procesamiento y transformaciones descritos se realizan con el Software MATLAB v7.0.

El siguiente paso de esta fase corresponde a la elección de la mejor configuración. Para ejemplificar este procedimiento utilizaremos el mismo ejemplo que en la sección 4.3[17], es decir, primer periodo del conjunto "Norte". Las ventanas de tiempo y la cantidad de objetos en los conjuntos de entrenamiento, validación y testeo respectivamente, están descritos a continuación:

- Enero 2000 hasta Diciembre 2001: 414 – 178 – 88
- Julio 2000 hasta Diciembre 2001: 214 – 92 – 88
- Enero 2001 hasta Diciembre 2001: 142 – 62 – 88

Al considerar los atributos como parte del modelo de clasificación, el criterio de selección en el periodo es distinto al utilizado en la sección 4.3, que solo utiliza el porcentaje de acierto sobre el conjunto de validación. Para ejemplificar este criterio, la siguiente tabla muestra 3 de las 528 iteraciones realizadas, en ella se considera el mejor porcentaje de acierto encontrado, la iteración seleccionada y otra que tiene una menor cantidad de atributos pero que no fue escogida.

[17] Los otros conjuntos y periodos siguen un procedimiento análogo

Iteración	Parámetros			Efectividad Promedio	Cantidad de Atributos Promedio
Inicio	C1	C2	λ		
48	7	0.9	0	67.2%	13.6
59	7	1	0	66.7%	12.3
190	5	1	0.1	66.3%	10.8

Tabla 14. Resultados de 3 iteraciones del conjunto de validación en la ventana de tiempo "ene-00 a dic-01" del conjunto "Norte"[18]

La tabla anterior muestra que la iteración con mejor porcentaje de acierto es la 48 con un 67.2% y 13.6 atributos, pero si analizamos la iteración 59 observamos que la diferencia en efectividad de de un 0.5% y que la cantidad de atributos utilizados disminuye en 1.3, por lo que sin perder mucho en efectividad, se puede utilizar una menor cantidad de atributos. Ahora si comparamos la iteración 190 con la iteración seleccionada hasta ese momento, también se tendría una perdida en efectividad (0.4%), pero siempre la variación de efectividad debe ser medida con respecto a la mejor, es decir, que la perdida de efectividad ascendería al 0.9% lo que es considerable aunque se disminuyan los atributos en 2.8.

La siguiente tabla detalla los resultados sobre el conjunto de validación obtenida para cada ventana de tiempo en el primer periodo del conjunto "Norte".

Ventana de tiempo		Parámetros			Efectividad Promedio	Cantidad de Atributos Promedio
Inicio	Fin	C1	C2	λ		
ene-00	dic-01	7	1	0	66.7%	12.3
jul-00	dic-01	1	0.7	0	64.5%	6.3
ene-01	dic-01	1	0.5	0.1	66.9%	5.3

Tabla 15. Resultados conjunto de validación: Primer periodo conjunto "Norte"

Estas configuraciones determinan los resultados obtenidos sobre el conjunto de testeo, detallados en la siguiente tabla.

[18] Los promedios son calculados en base a los 10 grupos construidos para cada ventana de tiempo.

Ventana de tiempo		Parámetros			Efectividad Promedio	Cantidad de Atributos Promedio
Inicio	Fin	C1	C2	λ		
ene-00	dic-01	7	1	0	66.3%	12.3
jul-00	dic-01	1	0.7	0	64.4%	6.3
ene-01	dic-01	1	0.5	0.1	66.1%	5.3

Tabla 16. Resultados conjunto de testeo: Primer periodo conjunto "Norte"

Esta tabla se analiza análogamente al realizado con la tabla 14. De modo que la configuración seleccionada es:

- ventana de tiempo "ene-01 a dic-01"
- C1 "1"
- C2 "0.5"
- Lambda "0.1"

La ventana de tiempo "ene-01 dic-01" tiene una diferencia de efectividad promedio con el mayor de un 0.2% y la diferencia de atributos utilizados es de 6, por lo que claramente esta ventana de tiempo es la escogida para definir el siguiente periodo. Con respecto a la otra ventana de tiempo la efectividad promedio caería y aumentaría el número de atributos utilizado, por lo tanto es descartada como solución.

La siguiente tabla muestra los resultados obtenidos con la metodología D-SVM con selección de Atributos para cada conjunto y en cada ventana de tiempo.

Conjunto	Ventana de tiempo		Parámetros			Efectividad Promedio		Cantidad Promedio Atributos
	Inicio	Fin	C1	C2	λ	Cantidad	Porcentaje	
Norte	ene-01	dic-01	1.0	0.5	0.1	88	66.1%	5.3
	ene-01	dic-02	5.0	0.3	0.1	78	65.3%	14.8
	ene-02	dic-03	7.0	0.9	0.1	72	66.8%	13.4
	jul-02	dic-04	50.0	7.0	0.5	112	67.6%	11.2
Quinta	ene-00	dic-01	30.0	5.0	0.1	70	66.7%	13.7
	jul-01	dic-02	9.0	1.0	0.3	68	68.5%	13.1
	jul-02	dic-03	3.0	0.5	0.1	76	69.5%	13.6
	jul-02	dic-04	3.0	1.0	0.0	112	63.3%	12.6
Sexta	ene-00	dic-01	1.0	1.0	0.1	202	72.8%	9.0
	jul-00	dic-02	7.0	5.0	0.0	110	69.2%	9.4
	jul-00	dic-03	7.0	5.0	0.0	112	71.1%	10.9
	jul-01	dic-04	1.0	0.7	0.0	168	63.8%	12.5
Septima	jul-00	dic-01	10.0	7.0	0.0	300	75.3%	10.0
	jul-00	dic-02	7.0	5.0	0.0	194	71.9%	12.2
	jul-02	dic-03	1.0	0.3	0.3	140	70.2%	12.2
	jul-02	dic-04	7.0	3.0	0.0	248	70.9%	12.5
Octava	jul-00	dic-01	9.0	3.0	0.1	134	75.1%	11.8
	ene-01	dic-02	7.0	3.0	0.0	110	74.9%	11.5
	jul-01	dic-03	3.0	0.7	0.5	106	75.2%	13.0
	ene-02	dic-04	10.0	5.0	0.1	136	73.3%	12.1
Novena	jul-00	dic-01	7.0	1.0	0.3	90	72.6%	13.7
	jul-01	dic-02	10.0	3.0	0.0	158	68.6%	12.0
	ene-02	dic-03	5.0	5.0	0.0	130	66.2%	8.6
	jul-02	dic-04	7.0	7.0	0.1	100	68.4%	8.1
Décima	jul-00	dic-01	30.0	5.0	0.0	100	73.0%	11.3
	ene-01	dic-02	70.0	10.0	0.5	118	75.4%	13.0
	ene-02	dic-03	10.0	3.0	0.1	178	73.4%	11.5
	ene-02	dic-04	3.0	1.0	0.5	220	72.0%	12.0
Sur	ene-01	dic-01	5.0	0.7	0.3	16	60.0%	4.5
	ene-02	dic-02	50.0	1.0	0.5	18	72.2%	13.2
	ene-02	dic-03	10.0	0.5	0.5	12	69.2%	12.2
	jul-02	dic-04	50.0	10.0	0.0	40	62.0%	11.2
Metropolitana	jul-00	dic-01	0.3	7.0	0.3	34	68.2%	12.7
	jul-01	dic-02	3.0	0.3	0.3	32	69.1%	8.2
	jul-01	dic-03	70.0	1.0	0.5	26	59.6%	10.3
	jul-02	dic-04	50.0	5.0	0.0	64	66.1%	9.0

Tabla 17. Resultados Modelo de Clasificación Dinámica con Selección de Atributos

Al igual que en el caso estático se puede analizar qué atributos tienen una mayor relevancia en la construcción del modelo de clasificación y además se obtiene información adicional: el cambio que cada atributo tiene en el tiempo. En la figura 9 se muestra para cada uno de los 18 atributos su frecuencia acumulada (cantidad de veces que un atributo es utilizado en un modelo) en el conjunto “Norte”, que puede llegar a un valor máximo de 40, definido por las 4

ventanas de tiempo (seleccionadas en función de su efectividad y cantidad de atributos) y los 10 conjuntos generados al balancear la muestra. De este modo para cada atributo se distingue su relevancia en función a la frecuencia acumulada y el cambio en la relevancia de cada atributo en el tiempo en función de la frecuencia en cada ventana de tiempo.

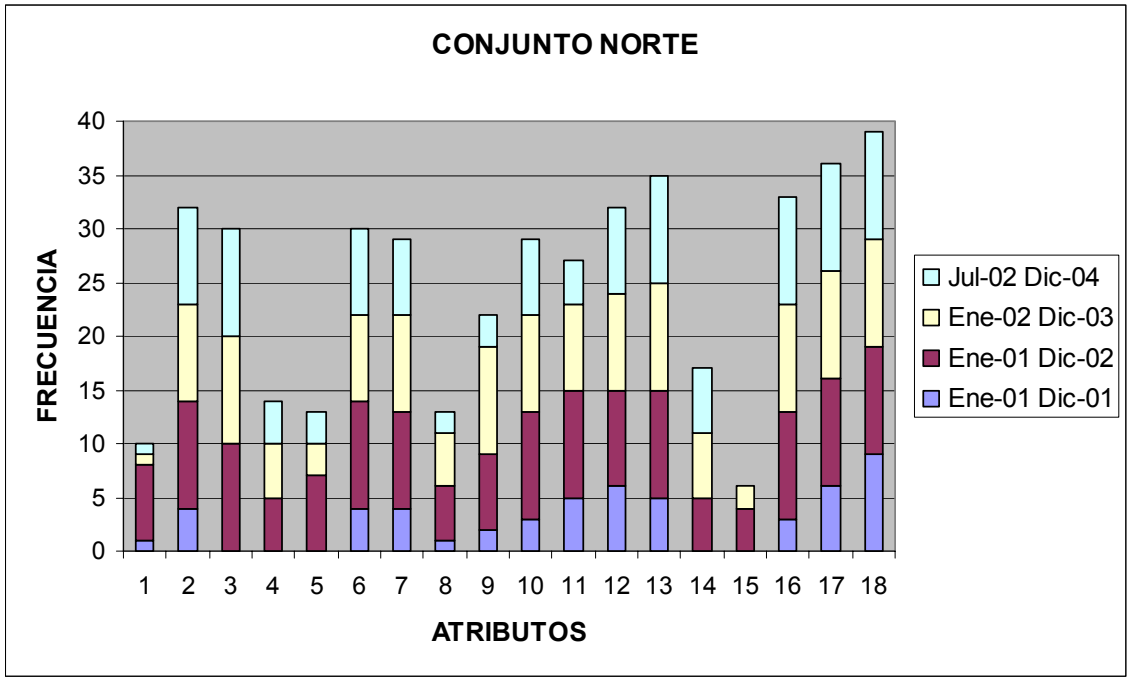


Figura 9. Frecuencia Acumulada por grupo y periodo para cada Atributo en el conjunto "Norte"

En este gráfico se observan los atributos que han presentado una relevancia en la construcción de los modelos de clasificación que en orden creciente son: 16, 2, 13, 17 y 18.

Una de las ventajas que se observan de esta metodología radica en el determinar movimientos de relevancia en los distintos periodos de tiempo. En este gráfico se observa la consistencia en la importancia de los atributos antes mencionados. También se puede determinar que los atributos 1, 4, 5 y 15, son

atributos que no tienen una relevancia consistente, en particular el atributo 15 no es considerado en dos periodos de tiempo.

Con estos movimientos en la relevancia de los atributos se pueden realizar acciones para mejorar la captación de nuevo créditos, ya que se dispone de los atributos relevantes y como afectan al modelo de clasificación.

4.4.3 Comparación Metodología de Clasificación Dinámica y Estática con Selección de Atributos

Esta sección presenta la comparación entre los modelos descritos en las dos secciones anteriores.

Se comparan los dos aspectos ya mencionados: la efectividad de los modelos construidos y la cantidad de atributos utilizados.

Efectividad

La tabla 18 contiene los resultados de efectividad obtenidos del modelo de clasificación dinámico y del estático.

La efectividad del modelo D-SVM es medido como la ponderación entre el porcentaje de acierto obtenido y la cantidad de objetos en el conjunto de testeo correspondiente.

Conjunto	Metodología	
	D-SVM	SVM
Norte	66.5%	66.7%
Quinta	66.6%	65.9%
Sexta	69.2%	69.4%
Séptima	72.5%	72.1%
Octava	74.6%	74.6%
Novena	68.7%	68.6%
Décima	73.2%	72.6%
Sur	64.8%	63.7%
Metropolitana	66.1%	64.7%

Tabla 18. Comparación en Efectividad Metodología D-SVM y SVM con Selección de Atributos

Se observa que en los conjuntos Norte, Sexta, Séptima, Octava y Novena se presenta una diferencia aproximada de 0,5 % entre los métodos,, en cambio, en los conjuntos Quinta, Décima, Sur y Metropolitana se observa que la metodología D-SVM fue levemente superior al modelo estático.

Cantidad de Atributos

La tabla 19 describe los atributos utilizados por cada modelo, en el caso de la metodología D-SVM la cantidad de atributos calculada es el promedio de los utilizados para cada periodo de tiempo.

Conjunto	Metodología	
	D-SVM	SVM
Norte	11.2	9.3
Quinta	13.3	9.4
Sexta	10.5	9.8
Séptima	11.7	11.6
Octava	12.1	12.0
Novena	10.6	10.2
Décima	12.0	9.8
Sur	10.3	6.2
Metropolitana	10.1	7.4

Tabla 19. Comparación en Cantidad de Atributos Metodología D-SVM y SVM con Selección de Atributos

Se observa que los conjuntos Sexta, Séptima, Octava y Novena, la cantidad de atributos utilizada no presentan una diferencia relevante (menos de un atributo)

entre los métodos y en el resto de los conjuntos de datos la metodología SVM utilizó menor cantidad de atributos que la D-SVM

Los análisis de la factibilidad y cantidad de atributos de ambos modelos presentan resultados similares, lo que asegura que el modelo D-SVM entrega al menos el mismo grado de confiabilidad que el SVM, que hasta el momento es el más utilizado y probado en la clasificación de clases de pertenencia.

CAPÍTULO 5:

CONCLUSIONES

5.1 Resultados Observados

La Metodología de Clasificación Dinámica presentada en esta tesis (D-SVM) permite determinar el problema de clasificación en distintos instantes del tiempo. A diferencia de los métodos usuales que proponen la construcción del modelo en un tiempo fijo y utilizarlo hasta que la predicción no cumpla con las expectativas de la organización, condición difícil de determinar y que genera costos en la organización. A continuación se presentan las conclusiones en los análisis de los métodos sin selección de atributos y de los métodos con selección de atributos, finalmente se concluirá las diferencias generales de ambos métodos.

Métodos sin selección de Atributos

De los resultados obtenidos en la aplicación y comparación entre la metodología D-SVM y SVM estático, se obtiene que la metodología propuesta es consistentemente mejor (6 de los 9 conjuntos) en los aspectos de efectividad y cantidad de atributos. A pesar que la diferencia es levemente superior, 0.54% en promedio, sugiere la utilización del modelo en los casos en que las características del problema tengan un grado alto de variabilidad en el tiempo, dado que cambiar el modelo estático podría ser muy costoso.

Método con selección de atributos

Utilizar la metodología D-SVM con selección de atributos presenta la ventaja de realizar la clasificación utilizando los atributos que el modelo estime necesarios para confeccionar la regla de discriminación, de modo que se tienen modelos más estables y generalizados. La metodología estática, requiere de un método exógeno que permita realizar la selección de atributos, lo que hace que la calidad del modelo de clasificación sea dependiente del método escogido.

De los resultados de efectividad obtenidos en la aplicación la metodología D-SVM hubo una diferencia menor a un 0.5% en 5 de los 9 conjuntos en la efectividad del modelo, en los otros 4 esta metodología fue superior en un 0.9% promedio al método estático. En relación a la cantidad de atributos utilizados, la metodología D-SVM tuvo en 4 de los 9 conjuntos una cantidad similar a la del modelo estático y en 5 utilizó una mayor cantidad de atributos, en promedio 2.92 atributos más, estos resultados muestran que el modelo D-SVM selecciona atributos con un buen porcentaje de acierto, sin embargo, requiere de nuevas investigaciones que aseguren una superioridad sobre el uso de una metodología estática.

En general, aunque los resultados obtenidos con la metodología D-SVM con y sin selección de atributos, no son concluyentes en la superioridad de la metodología, esta investigación muestra que el método es al menos igual de bueno que los métodos actuales de clasificación. Una posible causa de la similitud de los resultados, es que los datos estudiados presenten un comportamiento homogéneo en el tiempo, lo que produciría que el modelo D-SVM no marcase la diferencia con respecto al método de clasificación estática utilizado respecto a dichas variaciones, por lo que para elegir el tipo de metodología a utilizar es de primera importancia el análisis de los datos a nivel

de homogeneidad en el tiempo. Esta característica es el factor que determinará la decisión de utilizar una metodología estática o dinámica,

La metodología D-SVM presenta ventajas sobre SVM ya que no incurre en costos de investigación y desarrollo en actualizar modelos, donde los cambios de patrones en el tiempo son frecuentes o bien donde condiciones externas y las condiciones del negocio y clientes afecten al modelo. Sin embargo se recomienda realizar estudios detallados de los costos y beneficios en cada caso particular a evaluar, para la elección final del modelo.

5.2 Futuros Trabajos

Es necesario presentar de los resultados observados las variables que podrían ser determinantes en las diferencias y ventajas del método D-SVM v/s SVM. Los proyectos recomendados para agregar valor al estudio son:

- Confeccionar una metodología complementaria a D-SVM, que permita determinar el tamaño de la unidad de la ventana de tiempo y el periodo de actualización a partir de los datos entregados.

- Desarrollar un método que permita determinar si un nuevo objeto que entra, es muestra de una tendencia o es una anomalía.

- Desarrollar y confeccionar la metodología D-SVM en su formulación no lineal.

- Desarrollo de nuevos experimentos con datos controlados, de modo de confirmar la superioridad del método D-SVM en la construcción de modelos de clasificación. Los datos deben considerarse para comprobar la detección

de patrones que cambian en el tiempo, el detectar patrones sobre anomalías y el uso adecuado de los datos del pasado.

CAPÍTULO 6:

BIBLIOGRAFÍA

[1] CHENG-LUNG HUANG et al. 2006. “Credit scoring with a data mining approach based on support vector machines”. Expert Systems with Applications. doi:10.1016/j.eswa.2006.07.007

[2] CHIH-CHUNG CHANG, CHIH-JEN LI, CHIH-WEI HSU. 2003. “A Practical Guide to Support Vector Classification”. Department of Computer Science and Information Engineering, National Taiwan University, Taiwan.

[3] P. COLOMA, J. GUAJARDO, J. MIRANDA, R. WEBER. 2006. “Modelos analíticos para el manejo del riesgo de crédito”. Trend Management 8, Nov. 2006, 44-51.

[4] S.F. CRONE, S. LESSMANN, R. STAHLBOCK. 2005. “The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing”. European Journal of Operation Research. doi:10.1016/j.ejor.2005.07.023

[5] U. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMYTH. 1996. “From Data Mining to Knowledge Discovery in Databases”. AI Magazine, 17(3):37-54.

[6] U. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMYTH. 1996. “Knowledge Discovery and Data Mining: Towards a Unifying Framework”. Proceedings of the Second International Conference on Knowledge discovery and Data Mining, 82-88.

[7] U. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMYTH, R. UTHURUSAMY. 1996. “From Data Mining to Knowledge Discovery: An Overview”. Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1-34.

[8] R. GROSSMAN, S. BAILEY. 1998. “An Overview of Dynamic Classification: Mining Collection of Trajectories”. Proceedings of the Section on Physical and Engineering Sciences, American Statistical Association, Alexandria, Virginia, 24-28.

[9] J. GUAJARDO, J. MIRANDA, R. WEBER 2006. “A Forecasting Methodology Using Support Vector Machine and Dynamic Feature Selection”. Journal of Information & Knowledge Management 5, No 4, 329-335.

[10] M. HALL, G. HOLMES. 2000. “Benchmarking Attribute Selection Techniques for Data Mining”. Department of Computer Science, University of Waikato, Hamilton, New Zealand.

[11] J. HAN, M. KAMBER. 2001. “Data Mining – Concepts and Techniques”. Morgan Kaufmann, San Francisco.

[12] D. HOSMER, S. LEMESHOW. 2000. “Applied Logistic Regression”. Second Edition, Wiley Series in Probability and Statistics.

[13] R. KLINKENBERG, T. JOACHIMS. 2000. “Detecting Concept Drift with Support Vector Machines”. Proceedings of the Seventeeth International Conference on Machine Learning (ICML), 487-494, San Francisco, CA, USA.

[14] R. KLINKENBERG. 2004. “Learning drifting concepts: Example selection vs. example weighting”. Intelligent Data Analysis 8, 281-300.

[15] KOH HIAN CHYE, TAN WEI CHIN. 2004. “Credit Scoring Using Data Mining Techniques”. Singapore Management Review, Volume 26 No 2, 25-47.

[16] Y. LIU. 2002. “The evaluation of classification models for credit scoring”. Georg-August University, Institut for Economical Informatics, Göttingen, Germany.

[17] Y. LIU, M. SCHUMANN. 2005. “Data mining feature selection for credit scoring models”. Journal of the Operational Research Society, Volume 56, Number 9, 1099-1108.

[18] B. MURTAGH, M. SAUNDERS, P. GILL, R. RAMAN. MINOS Manual.

[19] J. MIRANDA, P. REY, R. WEBER. 2005. “Predicción de Fugas de Clientes para una Institución Financiera mediante Support Vector Machines”. Revista Ingeniería de Sistemas, Volumen XIX, 49-68.

[20] J. MIRANDA, R. MONTOYA, R. WEBER. 2005. “Linear Penalization Support Vector Machines for Feature Selection”. Lecture Notes in Computer Science 3776, 188-192.

[21] R. MONTOYA MOREIRA. 2002. “Programación matemática para data mining: utilización de Support Vector Machines para selección de atributos”. Tesis, Universidad de Chile.

[22] J. SHAWE-TAYLOR, N. CRISTIANINI. 2004. “Kernel Methods for Pattern Analysis”. First Edition, Cambridge University Press, Cambridge University.

[23] H. SHIMODAIRA et al. 2001. “Dynamic Time-Alignment Kernel in Support Vector Machine”. Advances in Neural Information Processing Systems 14, NIPS2001, 2:921-928.

[24] SPSS Modelos de Regresión 12.0, SPSS Inc., 2003

[25] V. VAPNIK. 1998. “Statistical learning theory”. John Wiley and Sons, New York.

[26] R. WEBER. 2005. Apuntes del curso IN60E, “Aplicaciones de Base de Datos en la Empresa”. Departamento de Ingeniería Industrial, Universidad de Chile.

[27] G. WIKSTRÖM. 2005. “Data Classification using Support Vector Machines”. Department of Physics, Stockholm University, Stockholm, Sweden.

[28] YINGXU YANG. 2005. “Adaptive Credit Scoring with Kernel Learning Methods”. SHS Information Systems, Credit Scoring & Control IX Conference, University of Edinburgh Management School, Credit Research Centre.

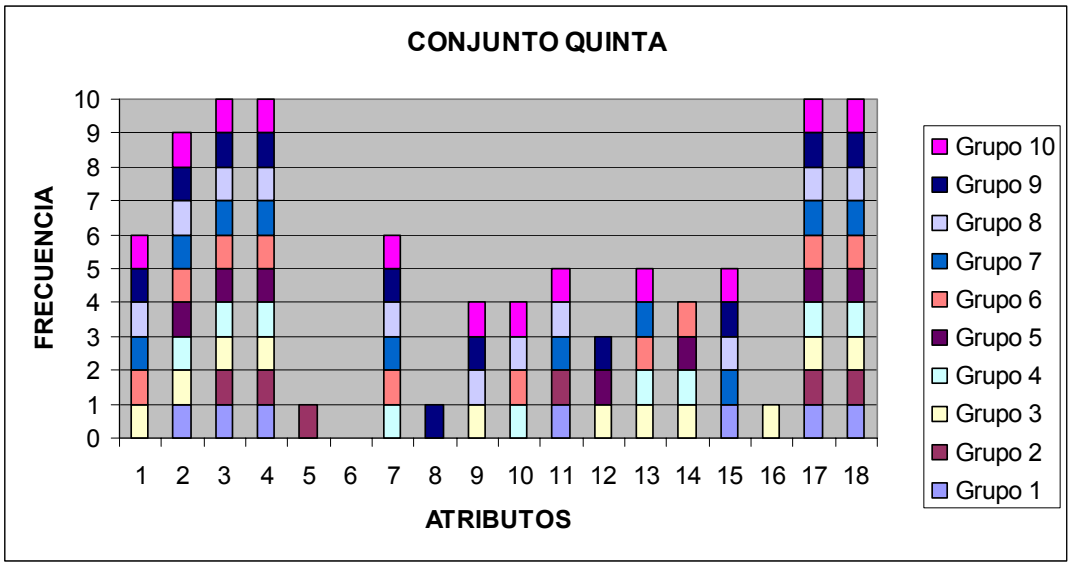
CAPÍTULO 7:

ANEXOS

ANEXO 1. Resultados Selección de Atributos, Método de Clasificación Estático.

	METODO						
	FORWARD		BACKWARD				
QUINTA	Acierto [%]	Cantidad Atributos	Acierto [%]	Cantidad Atributos	Diferencia Acierto [%]	Método Seleccionado	Cantidad Atributos Seleccionados
Grupo 1	67.2%	7	67.8%	7	0.6%	Backward	7
Grupo 2	65.0%	6	65.0%	6	0.0%	Ambos	6
Grupo 3	65.3%	9	66.6%	11	1.3%	Backward	11
Grupo 4	64.4%	9	63.4%	10	1.0%	Forward	9
Grupo 5	62.2%	7	62.2%	7	0.0%	Ambos	7
Grupo 6	71.3%	10	71.3%	10	0.0%	Ambos	10
Grupo 7	69.4%	10	69.7%	10	0.3%	Backward	10
Grupo 8	71.9%	11	70.9%	12	1.0%	Forward	11
Grupo 9	64.1%	9	66.6%	11	2.5%	Backward	11
Grupo 10	64.7%	8	66.6%	12	1.9%	Backward	12

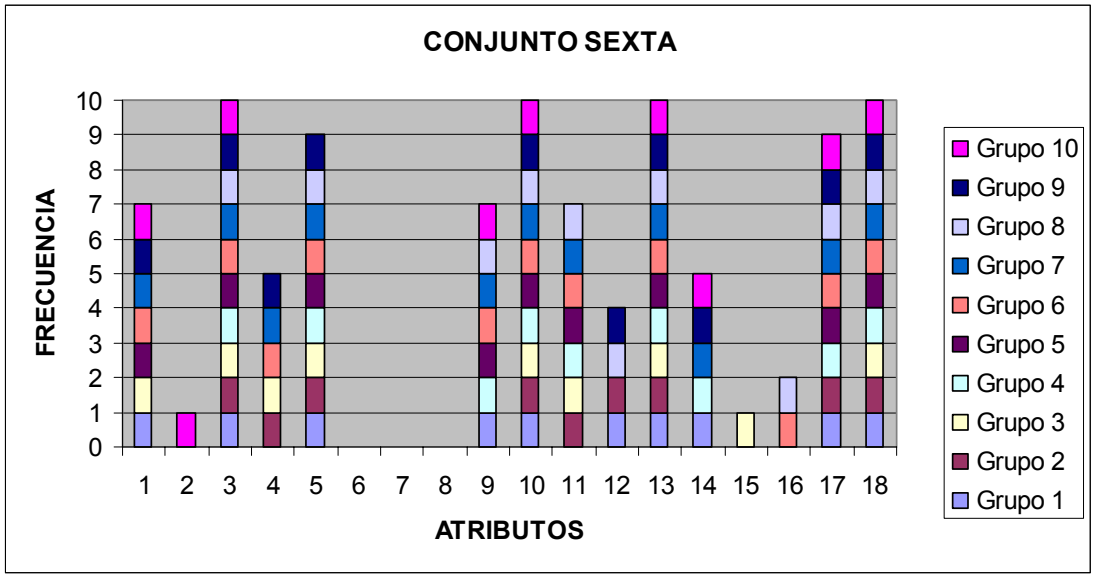
Resultados Selección de Atributos conjunto "Quinta" por Grupo



Frecuencia Acumulada para cada Atributo en el conjunto "Quinta"

	METODO						
	FORWARD		BACKWARD				
SEXTA	Acierto [%]	Cantidad Atributos	Acierto [%]	Cantidad Atributos	Diferencia Acierto [%]	Método Seleccionado	Cantidad Atributos Seleccionados
Grupo 1	67.3%	10	66.6%	11	0.7%	Forward	10
Grupo 2	67.0%	9	66.6%	13	0.4%	Forward	9
Grupo 3	67.3%	10	67.7%	10	0.4%	Backward	10
Grupo 4	72.4%	9	72.3%	10	0.1%	Forward	9
Grupo 5	67.2%	9	67.5%	10	0.3%	Forward	9
Grupo 6	67.0%	11	67.0%	11	0.0%	Ambos	11
Grupo 7	67.3%	11	67.3%	11	0.0%	Ambos	11
Grupo 8	71.2%	10	70.8%	11	0.4%	Forward	10
Grupo 9	68.6%	10	68.4%	11	0.2%	Forward	10
Grupo 10	69.3%	9	69.3%	10	0.0%	Forward	9

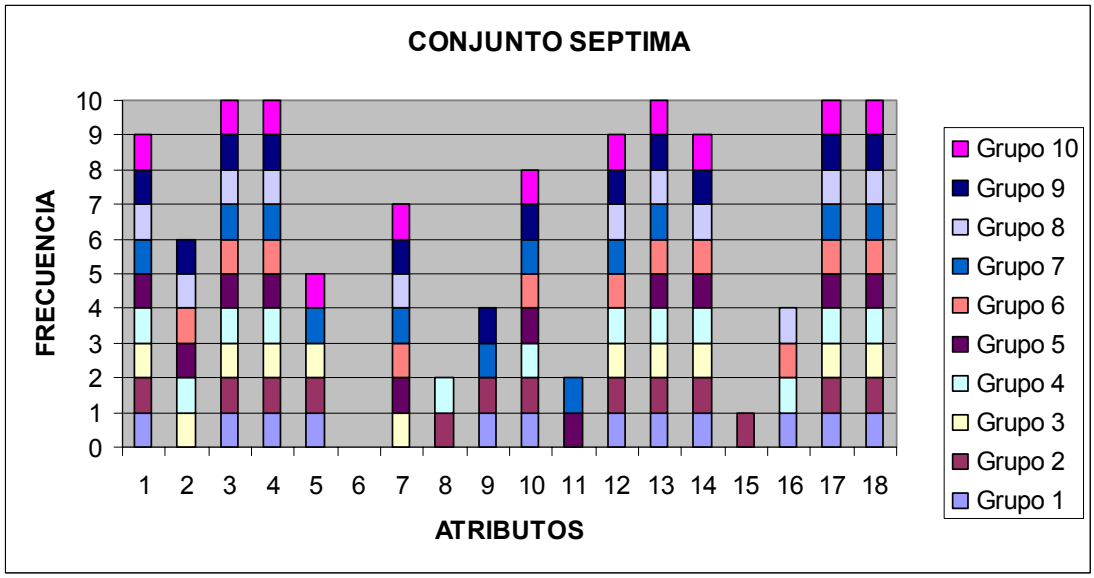
Resultados Selección de Atributos conjunto "Sexta" por Grupo



Frecuencia Acumulada para cada Atributo en el conjunto "Sexta"

	METODO						
	FORWARD		BACKWARD				
SEPTIMA	Acierto [%]	Cantidad Atributos	Acierto [%]	Cantidad Atributos	Diferencia Acierto [%]	Método Seleccionado	Cantidad Atributos Seleccionados
Grupo 1	72.4%	12	72.4%	12	0.0%	Ambos	12
Grupo 2	75.1%	13	75.2%	13	0.1%	Backward	13
Grupo 3	73.4%	11	73.7%	13	0.3%	Forward	11
Grupo 4	73.9%	12	73.9%	12	0.0%	Ambos	12
Grupo 5	71.5%	11	71.8%	12	0.3%	Forward	11
Grupo 6	74.8%	11	73.9%	12	0.9%	Forward	11
Grupo 7	74.2%	12	74.2%	12	0.0%	Ambos	12
Grupo 8	72.5%	11	72.8%	12	0.3%	Forward	11
Grupo 9	72.2%	12	73.7%	12	1.5%	Backward	12
Grupo 10	73.3%	11	73.7%	13	0.4%	Forward	11

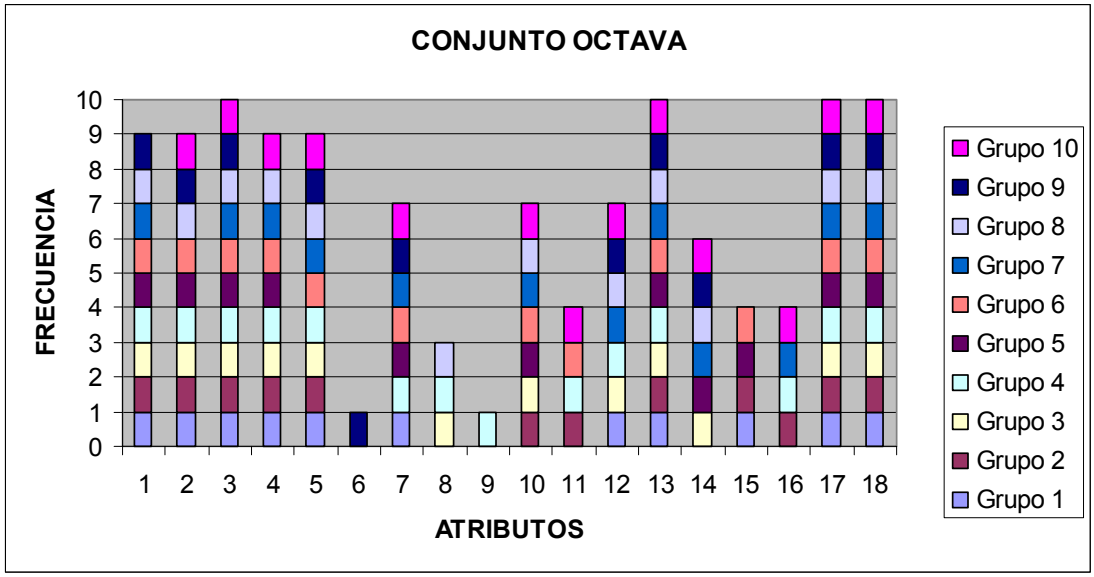
Resultados Selección de Atributos conjunto "Séptima" por Grupo



Frecuencia Acumulada para cada Atributo en el conjunto "Séptima"

	METODO						
	FORWARD		BACKWARD				
OCTAVA	Acierto [%]	Cantidad Atributos	Acierto [%]	Cantidad Atributos	Diferencia Acierto [%]	Método Seleccionado	Cantidad Atributos Seleccionados
Grupo 1	70.5%	11	71.5%	14	1.0%	Forward	11
Grupo 2	69.8%	12	70.2%	12	0.4%	Backward	12
Grupo 3	75.9%	12	75.9%	12	0.0%	Ambos	12
Grupo 4	73.0%	14	73.0%	14	0.0%	Ambos	14
Grupo 5	77.2%	11	75.9%	12	1.3%	Forward	11
Grupo 6	73.0%	12	72.6%	13	0.4%	Forward	12
Grupo 7	73.2%	12	73.2%	12	0.0%	Ambos	12
Grupo 8	76.8%	12	75.9%	13	0.9%	Forward	12
Grupo 9	76.4%	11	75.7%	13	0.7%	Forward	11
Grupo 10	69.6%	13	70.0%	14	0.4%	Forward	13

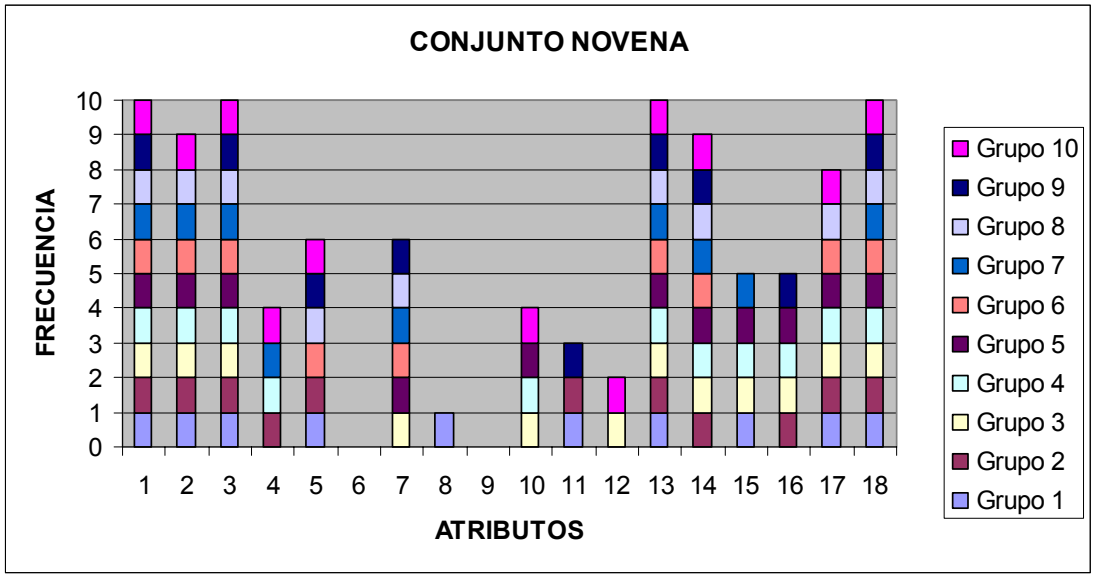
Resultados Selección de Atributos conjunto "Octava" por Grupo



Frecuencia Acumulada para cada Atributo en el conjunto "Octava"

	METODO						
	FORWARD		BACKWARD				
NOVENA	Acierto [%]	Cantidad Atributos	Acierto [%]	Cantidad Atributos	Diferencia Acierto [%]	Método Seleccionado	Cantidad Atributos Seleccionados
Grupo 1	67.6%	10	67.1%	11	0.5%	Forward	10
Grupo 2	67.4%	11	67.3%	11	0.1%	Forward	11
Grupo 3	68.2%	12	68.2%	12	0.0%	Ambos	12
Grupo 4	69.9%	11	69.9%	11	0.0%	Ambos	11
Grupo 5	68.6%	11	67.6%	13	1.0%	Forward	11
Grupo 6	70.1%	9	70.1%	9	0.0%	Ambos	9
Grupo 7	71.7%	9	72.5%	12	0.8%	Forward	9
Grupo 8	68.9%	9	69.0%	11	0.1%	Forward	9
Grupo 9	71.6%	9	71.0%	11	0.6%	Forward	9
Grupo 10	68.3%	11	68.3%	11	0.0%	Ambos	11

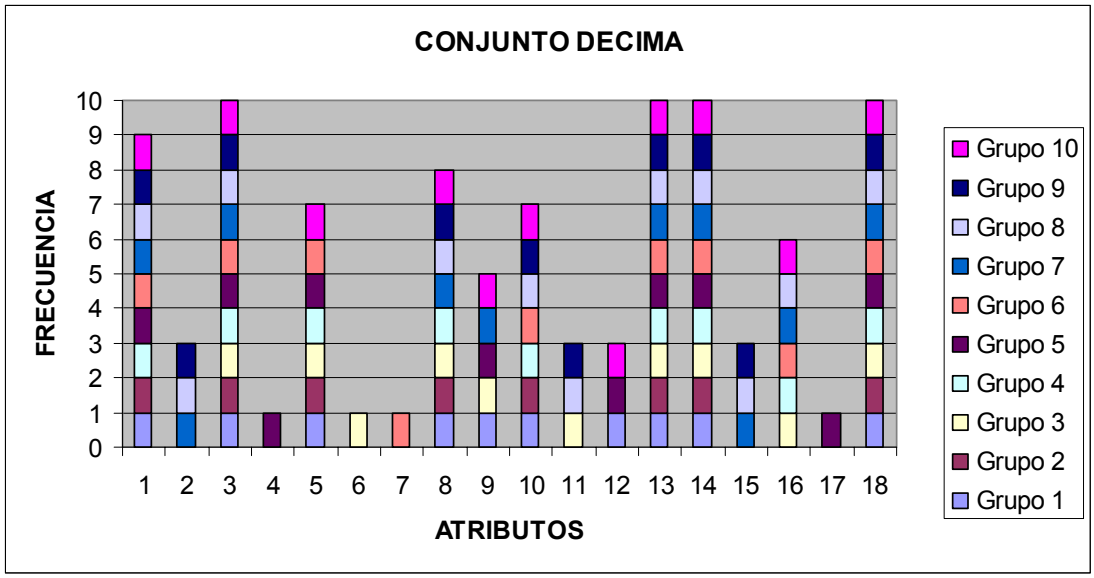
Resultados Selección de Atributos conjunto "Novena" por Grupo



Frecuencia Acumulada para cada Atributo en el conjunto "Novena"

	METODO						
	FORWARD		BACKWARD				
DECIMA	Acierto [%]	Cantidad Atributos	Acierto [%]	Cantidad Atributos	Diferencia Acierto [%]	Método Seleccionado	Cantidad Atributos Seleccionados
Grupo 1	74.0%	10	74.0%	10	0.0%	Ambos	10
Grupo 2	71.5%	8	71.3%	12	0.2%	Forward	8
Grupo 3	71.5%	10	71.5%	12	0.0%	Forward	10
Grupo 4	69.1%	9	69.7%	11	0.6%	Forward	9
Grupo 5	75.8%	10	74.0%	13	1.8%	Forward	10
Grupo 6	70.1%	9	70.3%	11	0.2%	Forward	9
Grupo 7	72.4%	10	71.5%	12	0.9%	Forward	10
Grupo 8	72.0%	11	72.4%	13	0.4%	Forward	11
Grupo 9	71.1%	10	70.1%	12	1.0%	Forward	10
Grupo 10	69.6%	10	70.8%	11	1.2%	Backward	11

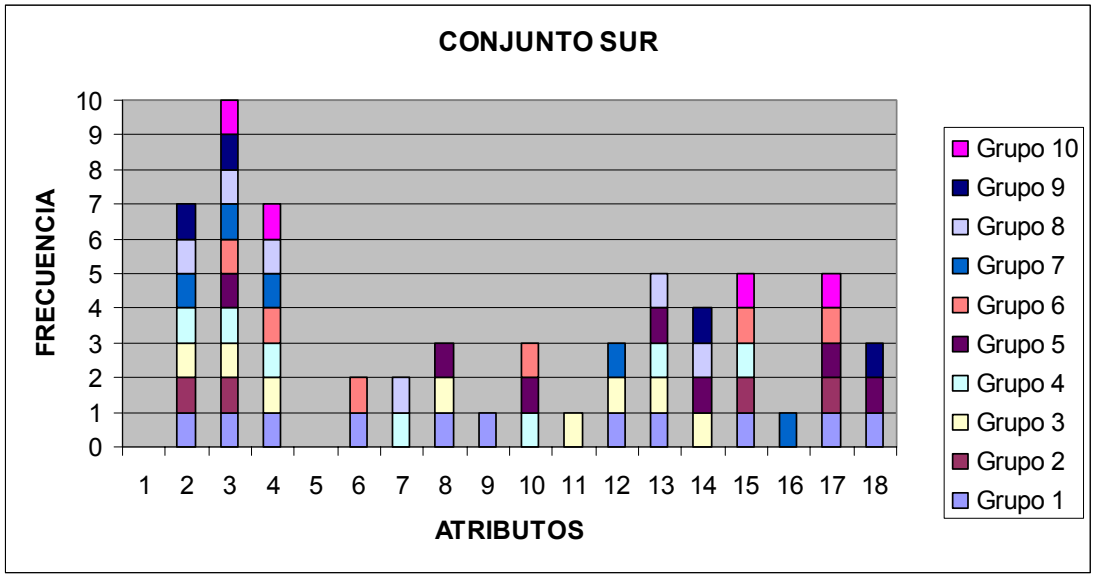
Resultados Selección de Atributos conjunto "Décima" por Grupo



Frecuencia Acumulada para cada Atributo en el conjunto "Décima"

	METODO						
	FORWARD		BACKWARD				
SUR	Acierto [%]	Cantidad Atributos	Acierto [%]	Cantidad Atributos	Diferencia Acierto [%]	Método Seleccionado	Cantidad Atributos Seleccionados
Grupo 1	66.7%	8	64.7%	11	2.0%	Backward	11
Grupo 2	54.5%	4	51.5%	5	3.0%	Forward	4
Grupo 3	57.8%	7	59.8%	8	2.0%	Backward	8
Grupo 4	59.8%	6	62.7%	7	2.9%	Backward	7
Grupo 5	57.4%	3	60.4%	7	3.0%	Backward	7
Grupo 6	62.7%	4	63.7%	6	1.0%	Backward	6
Grupo 7	63.7%	3	71.6%	5	7.9%	Backward	5
Grupo 8	67.6%	6	64.7%	11	2.9%	Forward	6
Grupo 9	68.6%	4	67.6%	6	1.0%	Forward	4
Grupo 10	65.7%	4	62.7%	9	3.0%	Forward	4

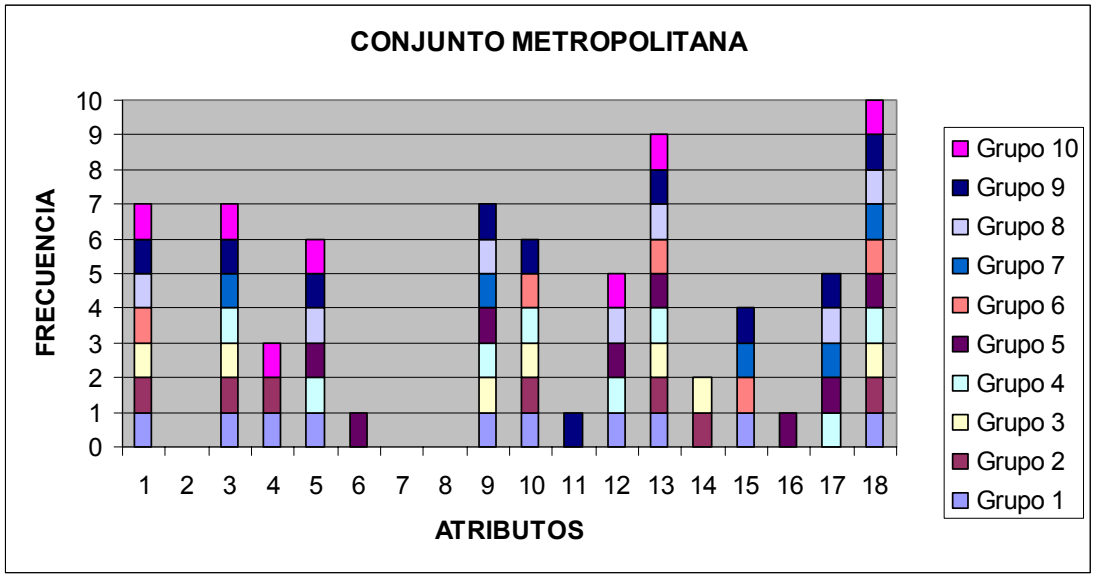
Resultados Selección de Atributos conjunto "Sur" por Grupo



Frecuencia Acumulada para cada Atributo en el conjunto "Sur"

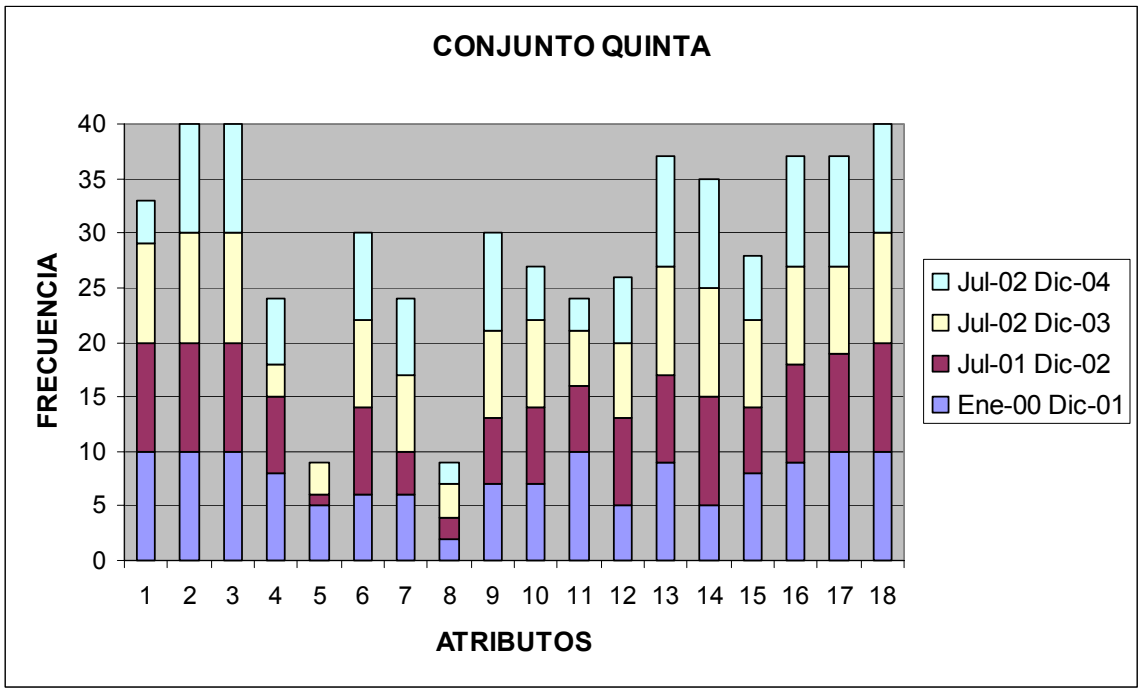
	METODO						
	FORWARD		BACKWARD				
METROPOLITANA	Acierto [%]	Cantidad Atributos	Acierto [%]	Cantidad Atributos	Diferencia Acierto [%]	Método Seleccionado	Cantidad Atributos Seleccionados
Grupo 1	71.9%	10	68.5%	13	3.4%	Forward	10
Grupo 2	60.7%	7	60.7%	7	0.0%	Ambos	7
Grupo 3	69.1%	7	67.4%	9	1.7%	Forward	7
Grupo 4	65.7%	7	66.9%	8	1.2%	Backward	8
Grupo 5	62.4%	7	63.5%	8	1.1%	Backward	8
Grupo 6	62.9%	5	62.4%	11	0.5%	Forward	5
Grupo 7	65.2%	5	65.2%	8	0.0%	Forward	5
Grupo 8	67.4%	7	68.0%	8	0.6%	Forward	7
Grupo 9	65.2%	8	66.3%	10	1.1%	Backward	10
Grupo 10	65.2%	7	62.4%	8	2.8%	Forward	7

Resultados Selección de Atributos conjunto "Metropolitana" por Grupo

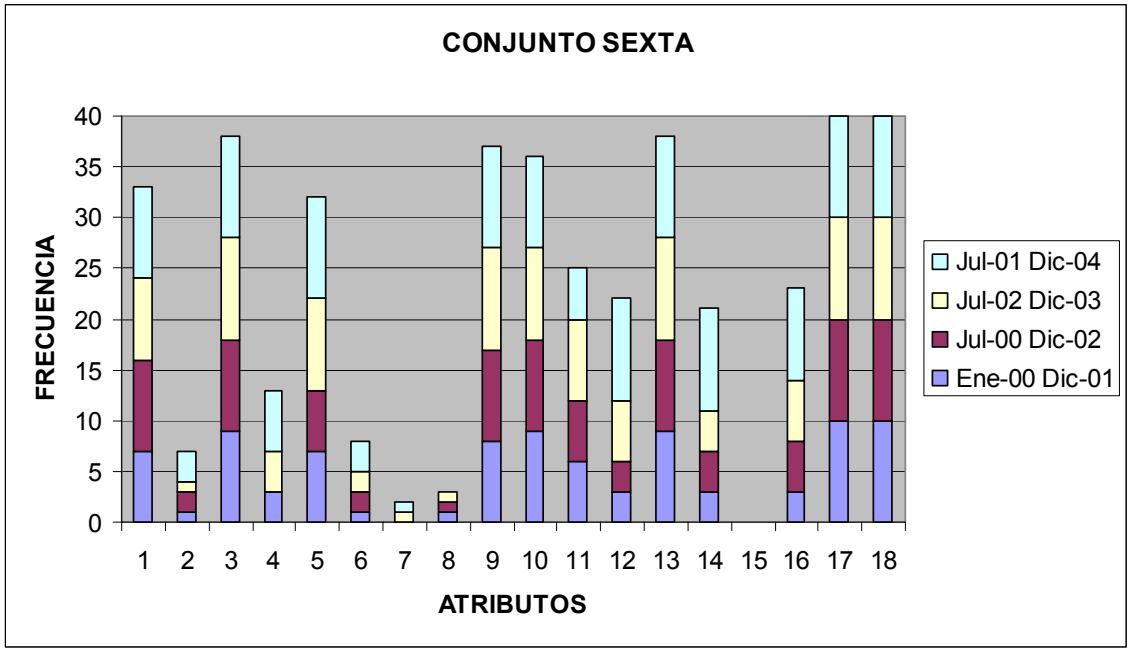


Frecuencia Acumulada para cada Atributo en el conjunto "Metropolitana"

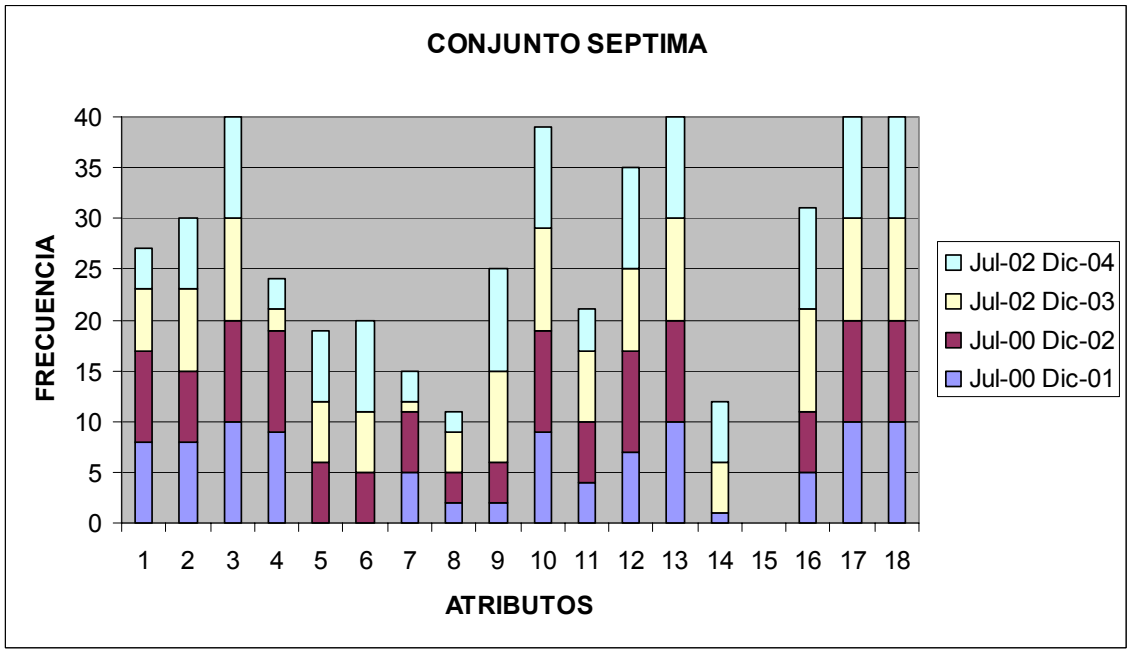
ANEXO 2. Resultados Selección de Atributos, Metodología de Clasificación Dinámica D-SVM



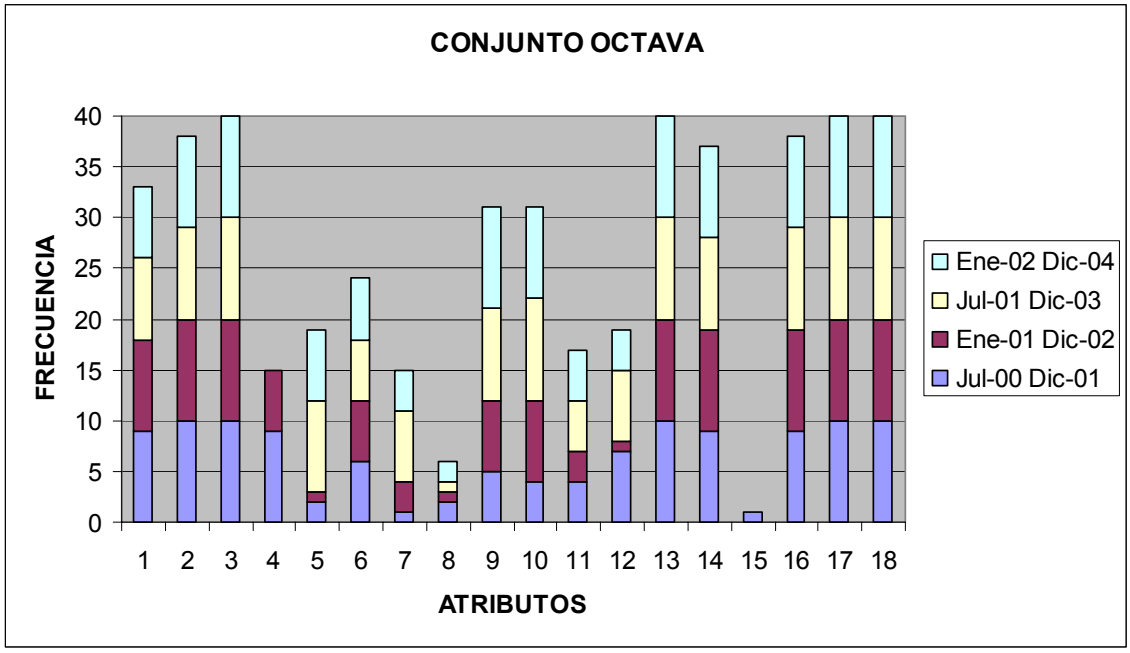
Frecuencia Acumulada por grupo y periodo para cada Atributo en el conjunto "Quinta"



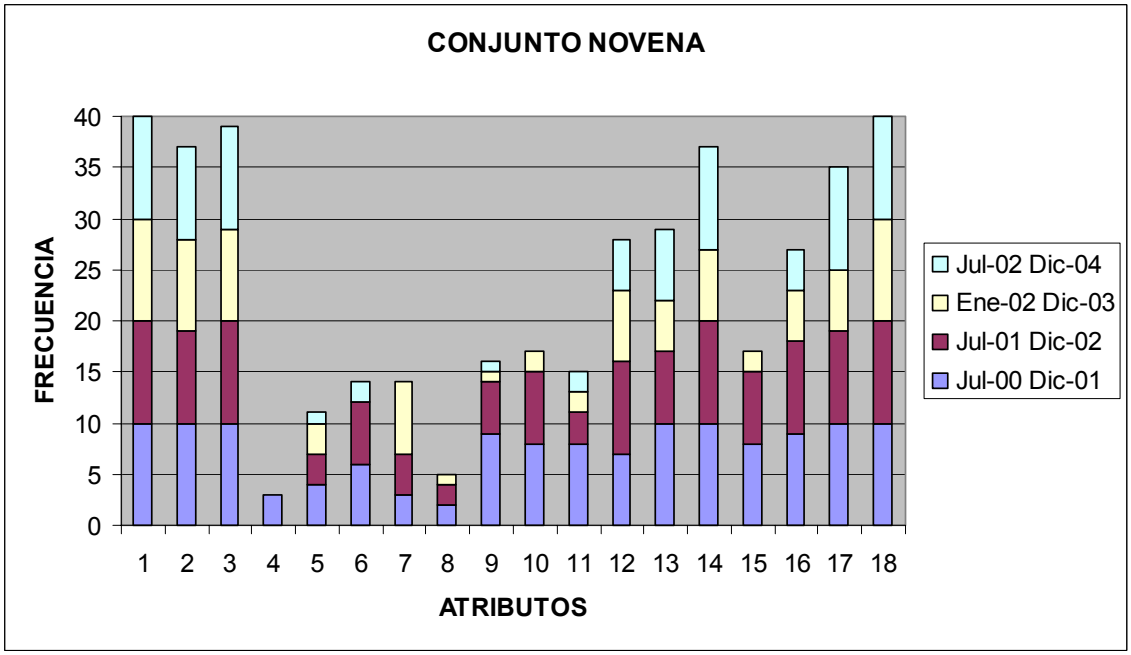
Frecuencia Acumulada por grupo y periodo para cada Atributo en el conjunto "Sexta"



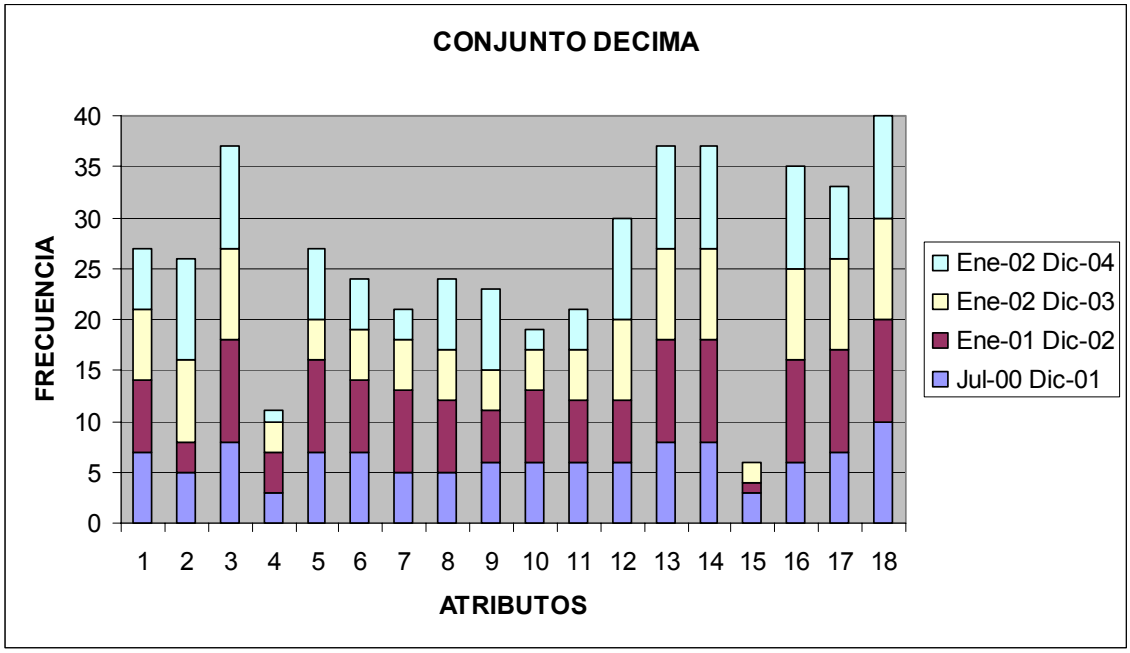
Frecuencia Acumulada por grupo y periodo para cada Atributo en el conjunto "Séptima"



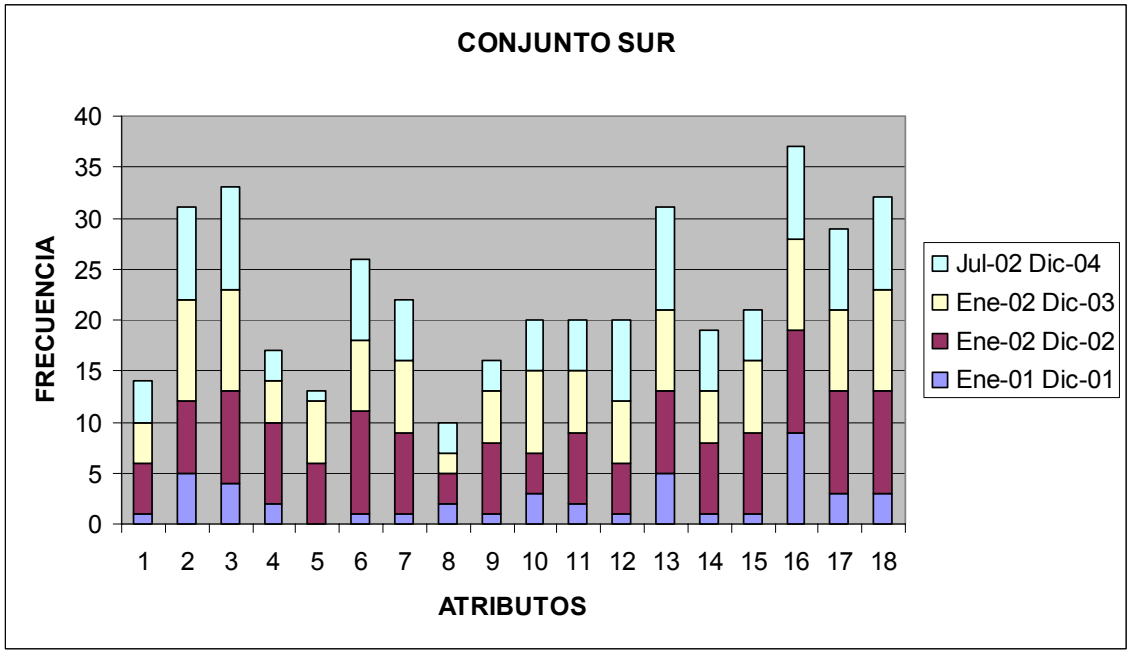
Frecuencia Acumulada por grupo y ventana de tiempo para cada Atributo en el conjunto "Octava"



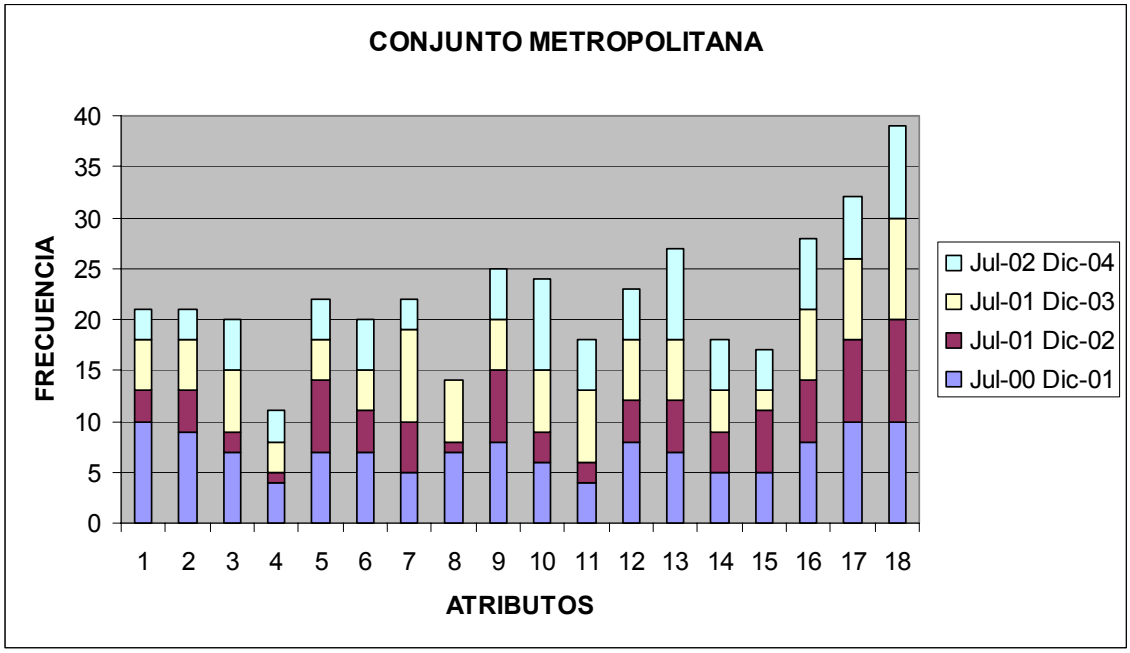
Frecuencia Acumulada por grupo y periodo para cada Atributo en el conjunto "Novena"



Frecuencia Acumulada por grupo y periodo para cada Atributo en el conjunto "Décima"



Frecuencia Acumulada por grupo y periodo para cada Atributo en el conjunto "Sur"



Frecuencia Acumulada por grupo y periodo para cada Atributo en el conjunto "Metropolitana"