Análisis de Accidentes de Tránsito con Máquinas de Soporte Vectorial LS-SVM

Cecilia Montt
Escuela de Ingeniería de Transporte, Pontificia Universidad Católica de Valparaíso
Av. Brasil 2950 - Teléfono 32- 2273742 - Fax 32-218854 - Valparaíso – Chile
cmontt@ucv.cl

Félix Castro y Nibaldo Rodríguez
Escuela de Ingeniería Civil Informática, Pontificia Universidad Católica de Valparaíso
Av. Brasil 2950 - Teléfono 32- 2273742 - Fax 32-218854 - Valparaíso – Chile
felixmagno@gmail.com; nibaldo.rodriguez@ucv.cl

RESUMEN

El objetivo del trabajo es desarrollar y evaluar un modelo de clasificación utilizando máquinas de soporte vectorial con algoritmos de optimización de enjambre de partículas, para clasificar el grado de severidad en el cual resultan las personas involucradas en accidentes de tránsito en la V región. Dentro de la minería de datos existen diversas técnicas, como las Máquinas de Soporte Vectorial (SVM), que reconocen patrones basada en la metodología de aprendizaje supervisado. Es complejo trabajar con SVM, por el ajuste de sus parámetros de entrada, por lo que se utiliza el algoritmo de Optimización por Enjambres de Partículas (PSO), para estimar los mejores parámetros para la maquina clasificadora y así clasificar el grado de severidad (lesionado o ileso) con el que resultan las personas involucradas en un accidente. Se crean tres modelos de clasificación utilizando una reformulación de SVM (LS-SVM), estos son: PSO con adaptación Dinámica; IPSO (Improved Particle Swarm Optimization) y factor de Inercia Lineal. A partir de los análisis de ellos, se concluye que PSO con factor de inercia lineal obtuvo mayor rendimiento en la exactitud con un promedio de 82,04%, de un total de 12 muestras de datos. Esto quiere decir que se clasificaron correctamente un 82,04% las personas que resultan en el estado ileso y lesionado, por lo que las SVM son capaces de presentar un buen grado de generalización para clasificar el estado en el cual resultan las personas.

Palabras claves: máquinas de soporte vectorial (SMV), optimización por enjambre de partículas (PSO), datamining, accidentes de tránsito.

Número de Palabras 5.813.

ABSTRACT

The objective of this study is to develop and evaluate a classification model using support vector machines with optimization algorithms swarm of particles, to classify the degree of severity in which people are involved in traffic accidents in the V region. Inside there are several data mining techniques such as Support Vector Machines (SVM), which recognize patterns based learning methodology supervisad. It is complex work with SVM, by adjusting the input parameters, so the algorithm used Swarm Optimization for Particle (PSO), to estimate the best parameters for machine sorting and thus classify the degree of severity (injured or uninjured) with which people are involved in an accident. Creates three classification models using a reformulation of SVM (LS-SVM), these are: PSO with adaptive dynamics; IPSO (Improved Particle Swarm Optimization) and linear inertia factor. From these analyzes, we conclude that PSO with inertia factor linear outyielded the accuracy with an average of 82.04%, a total of 12 data samples. This means that 82.04% correctly classified individuals in the state are uninjured and injured, so that the SVM are able to present a good degree of generalization to classify the state in which people are.

Keywords: support vector machine (SMV), particle swarm optimization (PSO), data mining, traffic accidents.

1. INTRODUCCIÓN

Según la Organización Mundial de la Salud (OMS), todos los años fallecen más de 1,2 millones de personas en las vías de tránsito del mundo, y entre 20 y 50 millones sufren traumatismos no mortales, siendo la novena causa de mortalidad mundial y se estima que para el año 2030 escale hasta la quinta posición. En Chile la situación no es menor, los siniestros de tránsito se han posicionado como una verdadera epidemia social, alcanzando cifras alarmantes y complejas de abordar. Según la Comisión Nacional de Seguridad de Tránsito (CONASET), solamente en 2009 se contabilizaron un total de 56.330 accidentes, siendo la colisión el tipo de siniestro más recurrente. La región de Valparaíso es la segunda región con más siniestro a nivel nacional y el 2009 contribuyó con un total de 6.662 siniestros, en los cuales se perdieron 134 vidas humanas (CONASET 2009). La información de accidentes de tránsito de la V región, utilizada en este trabajo es de los años 2003 a 2009, obtenida de CONASET.

Muchos investigadores han propuesto modelos predictivos acerca de las circunstancias en las cuales es más probable que ocurra algún siniestro de tránsito, sin embargo, ninguna ha sido absolutamente certera para poder controlarlos o reducirlos. Mediante técnicas de Minería de Datos o Data Mining, es posible obtener patrones o identificar las variables más significativas que ayuden y faciliten a esclarecer las condiciones en que ocurren los accidentes, y así poder controlar y reducir los niveles de accidentes de tránsito.

Dentro de la minería de datos existen diversas técnicas para el análisis. Una de ellas corresponde a las Máquinas de Soporte Vectorial o Support Vector Machines (SVM), la cual es una técnica de reconocimientos de patrones basada en la metodología de aprendizaje supervisado y usado para regresión y clasificación. Sin embargo, SVM presenta algunas desventajas importantes como la resolución de un sistema dual con programación cuadrática y el problema combinatorial en la estimación y ajuste de los parámetros de entrada de su modelo. Generalmente, dichos parámetros deben ser ajustados por el investigador en base a su experiencia, lo que no siempre conlleva a asignar valores óptimos afectando directamente al rendimiento del modelo.

El método de los mínimos cuadrados para SVM o Least Squares Support Vector Machines (LS-SVM), elimina la desventaja de la resolución de un sistema dual con programación cuadrática, vía un sistema lineal de ecuaciones.

Para salvar el problema de los parámetros se ha escogido este algoritmo de Optimización por Enjambres de Partículas o Particle Swarm Optimization (PSO), el cual es una metaheurística evolutiva y de búsqueda basada en el comportamiento social de enjambres de insectos.

Dado lo anterior el objetivo de este trabajo es desarrollar y evaluar un modelo de clasificación utilizando máquinas de soporte vectorial con algoritmos de optimización de enjambre de partículas, para clasificar el grado de severidad en el cual resultan las personas involucradas en accidentes de tránsito de la región de Valparaíso.

Se destaca que el alcance de este trabajo es solo para información de accidentes de la Región de Valparaíso. Por ello es que se pretende crear un modelo de clasificación binario para determinar el estado en el cual resultan las personas, estos son lesionados o ilesos.

A continuación en el capítulo 2, se entrega el estado del arte del uso de DATAMINING en accidentes de tránsito, en el capítulo 3 se describe la teoría utilizada en SVM y LS-SVM de manera de introducir los fundamentos básicos para su comprensión, y la metaheurística de optimización por enjambres de partículas PSO. En el capítulo 4 se muestra la metodología utilizada, en el capítulo 5 se implementa el modelo y se obtienen los resultados, para posteriormente concluir sobre lo trabajado.

2. ESTADO DEL ARTE

Se han analizado accidentes, por Dia, Rose 1997, donde utilizaron datos del mundo real para comparar la técnica red neuronal MLP, con un modelo heurístico de detección de accidentes en la autopista de Melbourne. Los resultados mostraron que el modelo de la red neuronal era más fiable, ya que podía detectar más rápido los accidentes sobre el modelo que estaba en operación de las autopistas. En Yang *et al* 1999, se utilizaron redes neuronales para la detección de patrones seguros de conducción que tuviesen menos probabilidad de causar la muerte y lesiones cuando se produce un accidente de tráfico. Concluyeron que mediante el control de una sola variable (velocidad de conducción o condiciones de luz), podrían reducir las muertes y lesiones hasta en un 40%.

Por su parte, Mussone, *et al*, 1999, utilizaron redes neuronales para analizar accidentes ocurridos en intersecciones en Milán, Italia. Eligieron las técnicas de feedforward (MLP) con Back Propagation (BP). Evanco, 1999 realizó un análisis estadístico multivariado, para determinar la relación entre las muertes y los tiempos de notificación de accidentes.

En Bedard *et al*, 2002, aplicaron una regresión logística para determinar la independencia entre la contribución del conductor, choque y características de los vehículos. Los autores encontraron que la prevención de fatalidades radicaba en el mayor uso del cinturón de seguridad, en la reducción velocidad y la reducción de impactos del lado del conductor. Por otra parte Ossiander *et al*, 2002 usaron modelos de regresión de Poisson para analizar la asociación entre el índice de accidentes mortales (accidentes mortales por cada milla recorrida del vehículo) y el aumento en el límite de velocidad. Concluyeron que al aumentar el límite de velocidad, este se asociaba con un mayor índice de accidentes mortales, por ende aumentaban las muertes en las carreteras del estado de Washington.

Por otra parte, Abdelwahab *et al*, 2003, estudiaron los accidentes de tránsito en Florida. El análisis se centró en los accidentes de vehículos producidos en intersecciones con semáforos. Se comparó el desempeño de las técnicas de Perceptron Multicapa (MLP) con Fuzzy ARTMAP, concluyendo que la precisión de clasificación de MLP era superior. Por otra parte Sohn *et al*, 2003, aplicaron clustering para mejorar la precisión de clasificadores individuales para dos categorías de gravedad (lesiones corporales y daños a la propiedad). Para ello utilizaron redes neuronales y árboles de decisión. Concluyeron que los algoritmos de clustering trabajan mejor para clasificar accidentes de tránsito.

En Chong, *et al* 2003, se utilizaron redes neuronales, árboles de decisión y un modelo hibrido, para construir modelos que pudiesen predecir la gravedad de las lesiones. Además incluyeron una breve aplicación con Máquinas de Soporte Vectorial. Concluyeron que para los casos en que las personas resultan ilesas o posiblemente lesionadas, se comportaba mejor el modelo híbrido que las redes neuronales, y para el caso no lesionado y posiblemente lesionado se pudo modelar mejor haciendo uso de árboles de decisión, en este caso el mayor porcentaje de exactitud obtenido fue de un 60,3%.

Por otra parte, se ha trabajado en el análisis de accidentes de tránsito de las principales regiones de Chile mediante técnicas estadísticas y algunas técnicas de minería de datos. Por ejemplo en Montt *et al* 2006, se analizaron los siniestros mediante estadísticas descriptivas y árboles de decisión, sin embargo, no se utilizó una metodología que ameritase un trabajo con gran cantidad de datos. Más tarde, Montt *et al*, 2010 trabajaron con técnicas de minería de datos basadas principalmente en estructuras de redes bayesianas. Procedieron a aplicar algoritmos de aprendizaje paramétrico y propagación de evidencia.

En Montt, *et al* 2009 se utilizaron métodos de agrupamiento e índices de Calinski y Harabasz para encontrar agrupaciones que representaran mejor la información. Lograron caracterizar los distintos tipos de accidentes, como choques con objeto, volcadura, caídas y los atropellos, encontrando características particulares en cada uno de los tipos de accidentes.

El último informe de seguridad vial realizado por la OMS en el 2009, sostuvo como conclusión que la tasa más alta de letalidad por cien mil habitantes correspondían a países de ingresos bajos y medios. Sin embargo, este estudio no utilizó herramientas de minería de datos sino que solo técnicas y herramientas estadísticas.

A través de la recolección de la literatura de análisis de accidentes de tránsito en el mundo y en Chile, se puede inferir que la minería de datos ha aportado bastante ayuda y conocimiento al análisis en los accidentes de tránsito. Redes Neuronales, Árboles de Decisión, Redes Bayesianas y Clustering son técnicas predominantes en este contexto, sin embargo, las técnicas de clasificación de SVM y LS-SVM no han sido muy utilizadas por el hecho de que es una técnica más nueva y compleja. No obstante, tras la buena utilidad que ha prestado la minería de datos en accidentes de tránsito se deduce que es posible utilizar la técnica de clasificación LS-SVM.

3. MARCO TEÓRICO

La clasificación de datos en el caso de SVM es descrita como una técnica de aprendizaje supervisado. Un proceso de clasificación supervisado incluye dos fases: entrenamiento y prueba. En la fase de entrenamiento o training un conjunto de datos inicial es usado para decidir que parámetros deberán ser ponderados y combinados con el objetivo de separar las clases y de esta manera

construir un clasificador para la fase siguiente. El aprendizaje intenta descubrir una representación óptima a partir del conjunto de datos cuya etiqueta de clase es conocida por el investigador. En la fase de prueba o testing, el clasificador determinado en la fase de entrenamiento es aplicado a un conjunto de datos u objetos (conjunto de prueba) cuyas etiquetas de clase se desconoce. De esta forma se clasifican los elementos y se comparan con los datos reales para determinar la efectividad del modelo.

3.1 Definición de SVM

Las Máquinas de Soporte Vectorial (SVM) son una técnica de reconocimientos de patrones basada en la metodología de aprendizaje, generando resultados robustos y satisfactorios. Fueron desarrolladas como una herramienta robusta y sólida para regresión y clasificación en dominios complejos, por Vladimir Vapnik y su equipo en los laboratorios AT&T. Su proceso de aprendizaje es supervisado, es decir, del ámbito predictivo. En clasificación supervisada los casos pertenecientes al conjunto de datos tienen asignada una clase o etiqueta a priori, siendo el objetivo encontrar patrones o tendencias de los casos pertenecientes a una misma clase. (Vapnick, 1995).

Las Máquinas de Soporte Vectorial se diferencian de otras técnicas como redes neuronales y programación genética, ya que las SVM no son afectadas por el problema de los mínimos locales, debido a que su entrenamiento se basa en problemas de optimización convexa.

En una SVM, el hiperplano óptimo es determinado para maximizar su habilidad de generalización. Pero, si lo datos de entrenamiento no son linealmente separables, es decir, no se pueden separar las clases dentro del espacio de solución original, el clasificador obtenido puede no tener una alta habilidad de generalización, aun cuando los hiperplanos sean determinados óptimamente. Por este motivo, para poder maximizar el espacio entre clases (hiperplano óptimo), el espacio de entrada original es transformado dentro de un espacio de mayor dimensión llamado "espacio de características", como se muestra en figura 1. A esto se le llama SVM no lineal.

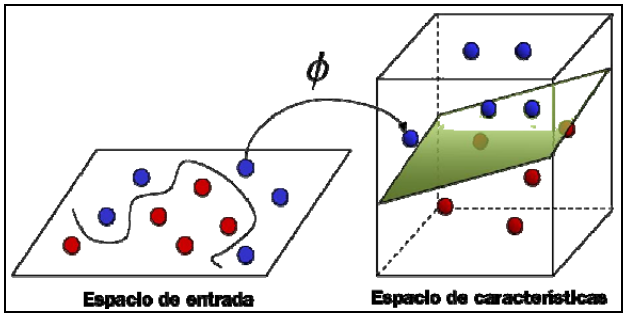

FIGURA 1: SVM no Lineal

Se tiene que si se considera el problema de separar el conjunto de vectores de entrenamiento $(x_1, y_1), ..., (x_l, y_l), \in R^n$ pertenecientes a dos clases separadas $y_i = \{+1, -1\}$.

En este caso el vector de entrenamiento es definido por las siguientes características, donde cada vector x_i presenta 12 atributos, estos son:

- Comuna
- Urbano/Rural
- Estado Atmosférico
- Hora
- Causa
- Tipo Calzada
- Estado Calzada
- Condición Calzada
- Tipo de Accidente
- Calidad
- Sexo
- Edad

Por otro lado los y_j, corresponde a la etiqueta de la clasificación, "-1" lesionado o "1" ileso.

En este problema el objetivo es separar los vectores de entrenamiento en dos clases mediante un hiperplano:

$$(w \cdot x) + b = 0, w \in R^n, b \in R \tag{1}$$

Donde w y b son parámetros que se inducen a partir de la función de decisión:

$$f(x) = signo(w \cdot x + b) \tag{2}$$

Existen muchos posibles clasificadores lineales que pueden separar los datos, sin embargo, hay uno solo que maximiza el margen de separación. El hiperplano $(w \cdot x_i) + b = 0$ satisface las siguientes condiciones:

$$\begin{aligned} (w \cdot x_i) + b > 0, y_i = +1 \\ (w \cdot x_i) + b < 0, y_i = -1 \end{aligned} \tag{3}$$

Combinando las dos expresiones de (3) y escalando w y b, con un factor apropiado, una superficie de decisión equivalente se puede formular como aquella que satisface la siguiente restricción:

$$y_i \left[(w \cdot x_i) + b \right] \geq 1, i = 1, ..., l \tag{4}$$

Además, se puede demostrar que el hiperplano que separa en forma óptima los datos en dos clases es aquel que minimiza la función:

$$\tau(w) = \frac{1}{2} \|w\|^2 \tag{5}$$

Por lo tanto, el problema de optimización cuadrático puede ser reformulado como un problema de optimización no restringida, usando multiplicadores de Lagrange y su solución estaría dado por la identificación de los puntos de silla del funcional de Lagrange:

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{l} \alpha_i \left\{ yi \left[(w \cdot x_i) + b \right] - 1 \right\} \tag{6}$$

Por lo que derivando se llega a la siguiente expresión:

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha y_i y_j (x_i x_j) \tag{7}$$

Donde los α_i son los multiplicadores de lagrange y los x_i, e y_j son el vector de características nombrado anteriormente.

En SVM no lineal un punto clave es la función kernel a utilizar, la cual realiza el mapeo no lineal en el espacio de características, a un espacio de mayor dimensión, de esta forma, encontrar un hiperplano óptimo que separa las clases (lesionado o ileso).

En SVM no lineal se puede trabajar con las siguientes funciones kernel:

- Kernel Lineal
- Kernel Sigmoidal
- Kernel Polinomial

- Kernel Función de Base Radial, utilizado en este trabajo, y tiene la siguiente expresión:

$$\varphi(x, x_k) = \exp\left\{\frac{-\|x - x_x\|^2}{\sigma^2}\right\}$$

Por su parte, el clasificador que implementa el hiperplano de separación óptima en el espacio de característica está dado por:

$$f(x) = signo\left(\sum_{i=1}^{l} y_i \alpha_i k(x_i \cdot x) + b\right) \quad (8)$$

De esta forma, aplicando la función kernel, la ecuación (8) es representada de la siguiente forma:

$$f(x) = signo\left(\sum_{i=1}^{l} y_i \alpha_i \varphi(x_i \cdot x) + b\right)$$

3.2 Características Claves de SVM

Una de las principales desventajas que presenta esta técnica, corresponde al alto costo de procesamiento que se produce al utilizar programación cuadrática para resolver la ecuación 7 del punto anterior. En base a esto, algunas técnicas han sido propuestas para salvar esta dificultad. En este caso se utilizó la técnica de los mínimos cuadrados SVM o Least Squares Support Vector Machine (LSSVM), la cual fue propuesta y creada por (Suykens, Vandewalle, 1999). Esta técnica es una reformulación de las SVM y soluciona el problema nombrado anteriormente a través de un conjunto de ecuaciones lineales. De esta forma, este método puede tratar una cantidad más considerable de datos de entrenamiento y a la vez el costo de procesamiento de los datos de entrenamiento disminuye considerablemente. Una característica particular de SVM es el hecho de que la mayoría de los multiplicadores de Lagrange, asociados a los vectores de entrenamiento, son nulos. Esto no ocurre en LS-SVM, los valores se distribuyen, sin predominancia de valores nulos (Vapnick, 1995).

3.3 Estimación de Parámetros con PSO

Para estimar los parámetros que se utilizarán en SVM, en este trabajo se utiliza. Optimización por Enjambres de Partículas (PSO) que es una meta heurística evolutiva y de búsqueda basada en el comportamiento social de enjambres de insectos.

La optimización por enjambre de partículas llega a una solución próxima a la global, en lugar de converger hacia soluciones locales como puede suceder con otros métodos tradicionales. Muchos estudios han optado por utilizar PSO, debido a las sencillas características que presenta el modelo, estas son:

1. Pocos parámetros a ajustar.
2. Normalmente trabaja con poblaciones pequeñas.
3. Una convergencia más rápida.

Por lo anterior, se desarrolla y evalúa un modelo de clasificación utilizando SVM con algoritmos de optimización de enjambre de partículas, para clasificar el grado de severidad (lesionado o ileso) en el cual resultan las personas involucradas en un accidente. Para ello, se crean tres modelos de clasificación utilizando variantes de PSO, que son PSO con adaptación Dinámica DAPSO; IPSO (Improved Particle Swarm Optimization) y PSO con factor de Inercia Lineal.

4. METODOLOGÍA

4.1 Descripción del Modelo General

El modelo representado en la Figura 2, representa el modelo base para la estimación de los parámetros de LS-SVM mediante el uso de PSO. Específicamente abarca los siguientes puntos:

1. Estimar los parámetros de LS-SVM mediante la meta heurística PSO (Compatible con cualquier variante de PSO), para lograr un mayor grado de generalización del modelo, es decir, que al momento de clasificar un nuevo elemento sea capaz de diferenciar correctamente a que clase corresponde.
2. Construir un clasificador para el estado en el cual resultan las personas involucradas en accidentes de tránsito. Específicamente clasificar si una persona termina en estado Lesionado o Ileso, dada las características que definen el suceso.

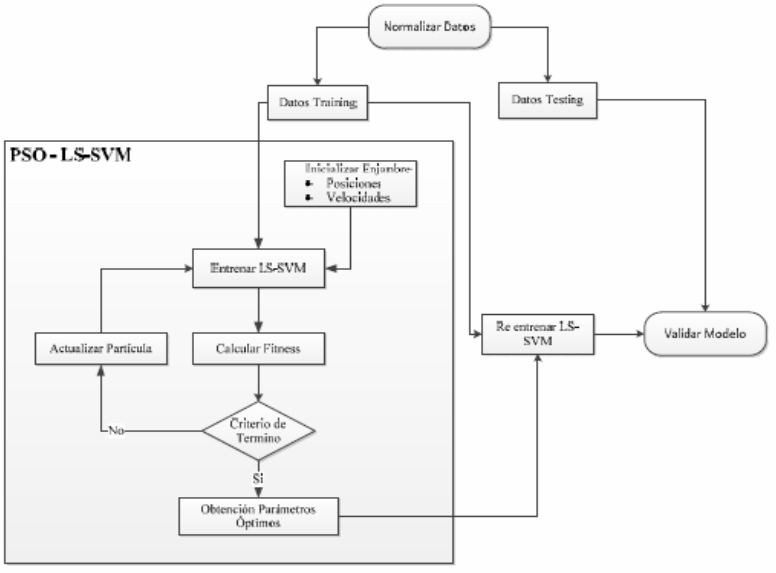

FIGURA 2: Modelo General de la Solución

Normalizar Datos. En primera instancia los datos son normalizados en un rango específico y son divididos para cada etapa: entrenamiento y validación.

Inicializar Enjambre. Corresponde a inicializar las posiciones y velocidades de las partículas, por ende se debe saber a priori cómo representar o codificar los parámetros en la meta heurística PSO. Esta codificación debe incluir el parámetro de penalización de errores y los parámetros del kernel a utilizar, ya que se pretende buscar la combinación de parámetros que proporcione un mejor rendimiento al clasificador.

De esta forma, la codificación de los parámetros para ser representados con el algoritmo de optimización por enjambre de partículas depende del tipo de kernel escogido para el entrenamiento, que en este caso es representada por el siguiente vector: (C, σ^2), donde C el parámetro de penalización de errores y σ^2 es la función del Kernel tipo RBF(radial basis function).

Actualizar Partículas. En este paso se actualizan las posiciones y velocidades de las partículas dependiendo del tipo de PSO a utilizar. Luego se seleccionan los parámetros que serán ingresados a la etapa de entrenamiento de la máquina. Dependiendo de la t-ésima iteración, se escogen las partículas correspondientes y se entregan a LS-SVM para entrenar y calcular el fitness que indican los datos mal clasificados.

El total de información con la que se cuenta para el estudio es alrededor de 70.000 datos, donde cabe recalcar que las muestras se toman aleatoriamente del conjunto de datos de accidentes comprendidos entre los años 2003 al 2009 para la V región.

Para entrenar el modelo, con los parámetros seleccionados anteriormente por el kernel, se utilizan 3000 accidentes al azar, de los 70.000 que se tienen, se ajustan los parámetros de LS-SVM con PSO, y se retorna el valor fitness de cada partícula en la t-ésima iteración, de manera de obtener el mínimo costo que indica los datos mal clasificados. En cada caso se realizan comparaciones en base a la función de desempeño, la cual debe minimizar el costo de entrenamiento de los datos. Una representación de dicha validación se hace a continuación:

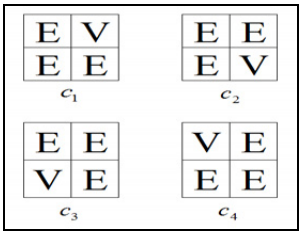

FIGURA 3: Esquema para Validación Cruzada para 4 Subconjuntos

E = Datos de Entrenamiento

V = Datos de Validación (se conocen estos datos a priori, si son accidente con resultado ileso o lesionado)

c_i = Costo de elementos mal clasificados del subconjunto i

La suma de todos los costos c_i divididos por el total de subconjuntos, da lugar al costo promedio utilizado como fitness en el modelo, que en este caso es la medida de datos mal clasificados.

Criterio de término. Si no se ha alcanzado el número total de iteraciones o no se ha alcanzado algún valor óptimo definido a priori, se actualizarán las posiciones y velocidades de cada partícula y se entrenará nuevamente. Por el contrario si se ha llegado a algún valor óptimo el cual se haya definido a priori, o se haya alcanzado el número total de iteraciones el ciclo de búsqueda de los parámetros finaliza.

Una vez encontrado los parámetros óptimos, el modelo es reentrenado con dichos parámetros.

Validar Modelo. Finalmente el modelo es validado con la data de testing. Para ello, se utilizan diversas métricas, como exactitud, sensibilidad y especificidad.

4.2 Medidas de Rendimiento para Evaluar el Clasificador

Para construir un modelo de clasificación se debe pasar por dos etapas: entrenamiento y prueba. Cada una de estas etapas debe ser medida para conocer el grado de exactitud de la clasificación. Si la etapa de entrenamiento logra un alto porcentaje de elementos clasificados, se dice que el modelo obtuvo un buen aprendizaje. Por otra parte, si la etapa de prueba obtiene un alto porcentaje de elementos bien clasificados, se dice que el modelo de clasificación generaliza bien, de lo contrario se dice que el modelo no logra clasificar bien los nuevos elementos.

En este trabajo se consideran las métricas de exactitud, sensibilidad y especificidad. Dichos métricas son creadas en base a los siguientes errores de prueba:

- Verdaderos Positivos (VP): número de éxitos. En este contexto corresponden al número de personas detectadas lesionadas correctamente.
- Verdaderos Negativos (VN): número de rechazos correctos. En este contexto corresponde a las personas detectadas ilesas correctamente.
- Falsos Positivos (FP): número de falsas alarmas. En este contexto corresponden al número de personas detectadas lesionadas, siendo que en realidad resultaron ilesas.
- Falsos Negativos (FN): En este contexto corresponde al número de personas detectadas ilesas, siendo que en realidad resultaron lesionadas.

Los VP, VN, FP y FN son resumidos en la siguiente matriz de confusión o tabla de contingencia:

TABLA 1: Matriz de Confusión

		Resultado Real	
		Lesionados	Ilesos
Resultado obtenido	Lesionados	VP	FP
	Ilesos	FN	VN

Métricas:

- Exactitud: corresponde al total de personas bien clasificadas, ya sea con lesión o sin lesión, dentro del total de personas de personas clasificadas:

$$Exactitud = \left(\frac{VP + VN}{VN + VP + FN + FP} \right)$$

Sensibilidad: corresponde a la probabilidad de que una persona realmente lesionada sea detectada como tal por la prueba:

$$Sensibilidad = \left(\frac{VP}{VP + FN} \right)$$

- Especificidad: corresponde a la probabilidad de que una persona ilesa sea detectada como tal por la prueba:

$$Especifidad = \left(\frac{VN}{VN + FP} \right)$$

- Valor Predictivo Positivo: corresponde a la probabilidad de padecer la lesión si se obtiene un resultado positivo en el test. Por ejemplo si se tiene un VPP de 80%, significa que el 80% de las personas que se detecten como lesionadas van a estar realmente lesionadas:

$$VPP = \left(\frac{VP}{VP + FP} \right)$$

• Valor Predictivo Negativo: corresponde a la probabilidad de que una persona con un resultado negativo en la prueba esté realmente ileso. Por ejemplo si se tiene un VPN de 90%, significa que el 90% de las personas que se detecten como ilesos van a estar realmente ilesos:

$$VPN = \left(\frac{VN}{VN + FN} \right)$$

5. IMPLEMENTACIÓN DE MODELOS Y RESULTADOS

5.1 Clasificación LS-SVM con PSO

A continuación se presenta la clasificación de los distintos modelos propuestos e implementados, en base a las variantes del algoritmo de optimización por enjambre de partículas. Para ello, se escogió y utilizó el tipo de kernel Función de Base Radial (RBF), ya que según la literatura y trabajos relacionados presenta mejores resultados.

Las variantes de PSO implementadas fueron las siguientes:

- PSO con Factor de Inercia Lineal
- IPSO (Improved Particle Swarm Optimization)
- DAPSO (Dynamic Adaptation Particle Swarm Optimization)

Por otra parte, la elección de la función de costo o fitness en PSO, juega un papel fundamental en la generalización del modelo. En base a esto, se utilizó el siguiente fitness:

- Costo de elementos mal clasificados utilizando validación cruzada para 10 subconjuntos.

El total de la población con la que se cuenta para el estudio es alrededor de 70.000 datos, de los cuales en primera instancia se extraen 3000 para cada modelo de PSO. Estos se dividen en 2.000 para la etapa de entrenamiento y los 1000 restante para testing. En segunda instancia se extraen distintas muestras para volver a testear el modelo y así comprobar el comportamiento del clasificador con distintos tamaños de muestra, es decir, advertir que tan preciso es el clasificador. Cabe recalcar que las muestras se toman aleatoriamente del conjunto de datos, comprendidos entre los años 2003 al 2009.

Por otra parte, el kernel del tipo RBF requiere solo un parámetro a ajustar, este corresponde a σ^2. También se debe ajustar el parámetro C que controla el trade-off entre la maximización del margen y la minimización del error de entrenamiento. De esta forma, para cualquier variante de PSO la representación de la partícula para este tipo de kernel, es representada por el siguiente vector: (C, σ^2):

Antes de iniciar el algoritmo se realizó un proceso de integración, selección y codificación de datos.

5.2 Integración de Datos

Los datos recolectados desde el año 2003 al 2009 fueron entregados en formato Excel, los cuales fueron traspasados a un motor de base de datos MySQL para su posterior uso. Para ello, se crearon las entidades representadas en la Figura 4. De esta forma, se pudieron rescatar los datos mediantes simples consultas sql para luego ser guardados en Matlab.

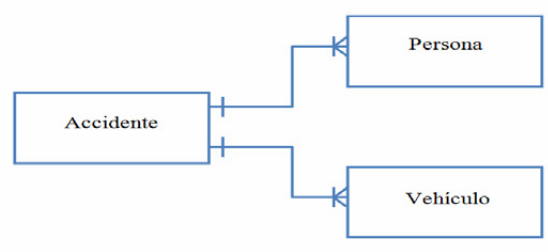

FIGURA 4: Entidades Representadas

5.3 Selección y Codificación de los Datos

Como el modelo a clasificar corresponde al estado en el cual resultan las personas (lesionadas o ilesas) involucradas en los accidentes de tránsito, se procedió a identificar aquellas características más representativas. En base a esto, se analizaron y escogieron los atributos propuestos en Montt *et al*, 2010. Para ello se rescataron de la base de datos la información con los siguientes atributos, indicados en el punto 2.1.

Estos atributos sirven para identificar y formar las dos clases a utilizar (personas Lesionadas y personas ilesas), como se muestra en la tabla 2.

TABLA 2: Formación de Clases

Clase	Atributos	Etiqueta
Persona lesionada	Grave, Menos grave y leve	1
Persona ilesa	Ileso	-1

5.4 Implementación de los Modelos

Para cada modelo se utilizó el fitness del tipo costo de valuación cruzada para 10 subconjuntos. Se explica en detalle el modelo PSO con Factor de Inercia Lineal. El factor de inercia lineal se caracteriza por ir decreciendo en un intervalo preestablecido, a medida que aumentan las iteraciones. El intervalo escogido para dicho factor es entre [0.9-0.4].

Para el Fitness del tipo Costo Validación Cruzada, en primera instancia se utilizó la muestra de 3.000, los que se dividieron en 2.000 para training y 1.000 para testing. El tiempo total de estimación de parámetros con esta variante fue exactamente de 32.247 segundos, equivalentes a 8,95 hrs aprox. Los parámetros óptimos obtenidos fueron:

C =167.8857 y σ^2 =7.9708, en tanto, el mejor fitness fue de: 0.1795. El resumen general se resume en la tabla 3.

TABLA 3: Resultados PSO F. Inercia Lineal, Fitness Costo Validación Cruzada

ETAPA	Exactitud	Sensibilidad	Especificidad	Mal Clasificados
Training	90,15	88,7	92,22	197
Testing	85,9	85	87,39	141

Como se muestra en la tabla anterior, esta variante obtuvo un 85,9% de exactitud en la etapa de testing, esto quiere decir que el modelo ha clasificado de manera correcta en un 85,9% las personas que resultan en el estado lesionado o ileso.

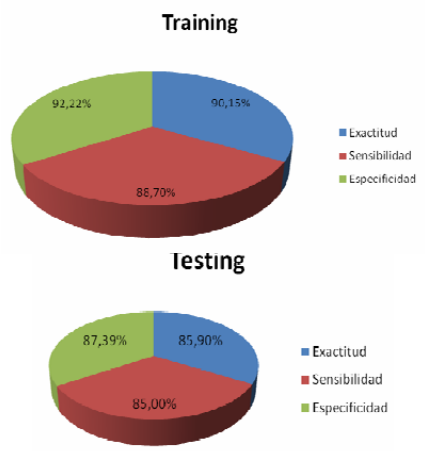

FIGURA 5: Resultados PSO F. Inercia Lineal, Fitness Costo Validación Cruzada, en Training y Testing

Por otra parte, la sensibilidad obtenida fue de un 85%, cifra correspondiente al porcentaje de personas bien clasificadas en el estado lesionado. Asimismo, la especificidad resultó ser un 87,39% correspondiendo al porcentaje de personas bien clasificadas en el estado ileso. En cuanto a la seguridad del resultado, se tiene un valor predictivo positivo de 91,8%, este significa que un 91,8% de las personas detectadas como lesionadas están realmente lesionadas, en tanto, el valor predictivo negativo fue de 72,8%, este significa que un 72,8% de las personas detectadas como ilesas están realmente ilesas.

En segunda instancia se procedió a testear el modelo con distintos tamaños de muestras a partir del modelo entrenado anteriormente, para comprobar el comportamiento del clasificador. En la figura 6 se aprecian los resultados.

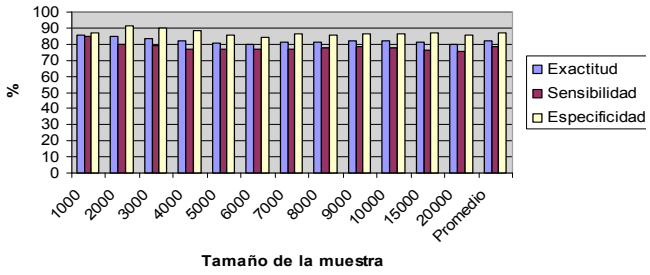

FIGURA 6: Resultados para Varias Muestras PSO F. Inercia Lineal, Fitness Costo Validación Cruzada

A través de la figura anterior se aprecia que al variar el tamaño de la muestra para testing, la exactitud del clasificador se mantiene entre el rango 80.1% y 85.9%, obteniendo como promedio un 82%, si bien va bajando a medida que se aumenta la muestra debido a que el entrenamiento se realizo siempre con 2000 datos. Asimismo ocurre con las demás métricas, las cuales se mantienen dentro de un rango aceptable.

5.5 Resultados

En este capítulo se discuten los modelos implementados, en base a los resultados obtenidos por cada `modelo. Es por esto, que en primer lugar se contrastan los modelos a partir de las tres variantes de PSO utilizadas.

Aunque lo primordial es analizar la etapa de testing para comprobar el rendimiento del modelo, es conveniente tratar de analizar la etapa de entrenamiento, ya que cada variante de PSO trabaja directamente con la fase de training. Cabe recordar que dicha etapa fue entrenada con una cantidad de 2.000 datos. La siguiente tabla resume dichos resultados:

TABLA 4: Comparación Training LS-SVM con Variantes PSO

Variantes	Exactitud	Sensibilidad	Especificidad	Mal Clasif.	Tiempo Entrenamiento
PSO F.I.L	90,15	88,7	92,22	197	8,96 hr
IPSO	86,9	85,21	89,3	262	11,13 hr
DAPSO	88,2	82,3	94,1	236	9,57 hr

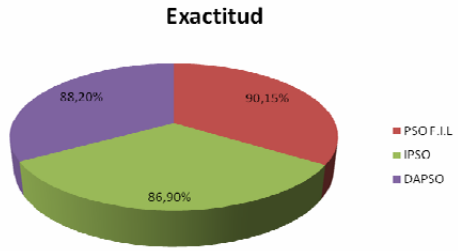

FIGURA 7: Gráfico Exactitud LS-SVM, Variantes de PSO

Al contrastar los resultados de los tres modelos se aprecia que no hubo una gran diferenciación entre ellos. Aunque entre los modelos de PSO con factor de inercia lineal y DAPSO hubo una diferencia mínima, PSO con factor de inercia lineal es la que obtuvo un mayor rendimiento en la exactitud, es decir, el mejor porcentaje de personas bien clasificadas en el estado lesionado e ileso.

6. CONCLUSIONES

Gracias al análisis y estudio realizado sobre LS-SVM y PSO se logró diseñar la estructura del modelo de clasificación, el cual fue base para la implementación de estimación de parámetros de cada modelo. En base a esto, se utilizaron tres variantes de PSO, en donde cada una de ellas se evaluó en dos tipos de fitness, de los cuales el costo de la validación cruzada para 10 subconjuntos obtuvo mejores resultados en comparación con el error absoluto medio. Esto ocurrió en cada variante de PSO implementada.

De lo anterior podemos deducir que es posible analizar cualquiera de los atributos del punto 2.1, con los datos analizados. Es decir, con los datos de las muestras utilizadas se pueden analizar por ejemplo las colisiones, y el resultados de las personas, además, de edad, sexo y causa del accidente, obteniéndose reglas, que podrían ayudar a evitar los accidentes, así mismo para la causa en estado de ebriedad, lo que será trabajo de próximas investigaciones.

Asimismo, al comparar los resultados de la mejor variante obtenida en este trabajo de investigación con los resultados obtenidos en [22] y [24] se establece que PSO con factor de inercia lineal superó ampliamente a dichos trabajos, en exactitud, sensibilidad y especificidad. Es de importancia destacar que las comparaciones fueron realizadas en base al promedio de distintas muestras de los resultados obtenidos en PSO con factor de inercia lineal, no así en los trabajos relacionados que no utilizan un promedio de varias muestras, sino más bien, el mejor resultado para una muestra en particular. Por ende, las diferencias podrían ser aún más elevadas con el modelo obtenido en este trabajo.

Finalmente se concluye que las máquinas de soporte vectorial son capaces de presentar un buen grado de generalización para clasificar el estado en el cual resultan las personas (lesionadas o ilesas) en cualquier región del País,

presentándose el modelo PSO con factor de inercia lineal con un 82,04% de exactitud en rendimiento.

Con otro tipo de máquina con capacidad industrial se podrían utilizar un mayor conjunto de datos para clasificar los accidentes y así obtener un modelo mas precisos y verídico, es decir para cada región se podrían utilizar 4 o 5 años de estadísticas de accidentes de tránsito.

7. BIBLIOGRAFÍA

Abdelwahab, M y Abdel, Aty. Analysis and Prediction of Traffic Fatalities Resulting From Angle Collisions Including the Effect of Vehicles' Configuration and Compatibility. s.l.: Accident Analysis, 2003.

Bedard, M., Guyatt, G. H., Stones, M. J., & Hireds, J. P. The Independent Contribution of Driver, Crash, and Vehicle Characteristics to Driver Fatalities. s.l.: Accident analysis and Prevention, 2002. págs. 717-727.

Chong, Miao. Ajit,Abraham. Paprzycki, Marcin. Accident Data Mining Using Machine Learning Paradigms. Computer Science Department. Oklahoma State University, USA: s.n., 2003.

Comisión Nacional de Seguridad de Tránsito. Tipos de Transito 2000-2009. CONASET. Santiago: s.n., 2009. http://www.conaset.cl/portal/portal/default/estadisticas_generales.

Dia, H y Rose, G. Development and Evaluation of Neural Network Freeway Incident Detection Models Using Field Data. s.l.: Transportation Research C, 1997. págs. 313-331.

Evanco, W. M. The Potential Impact of Rural Mayday Systems on Vehicular Crash Fatalities. s.l.: Accident Analysis and Prevention, Vol. 31, 1999. págs. 455-462.

Montt C. (2006) "Definición de variables que afectan al individuo en su comportamiento respecto a la seguridad vial", Proyecto de Investigación Interno PUCV, DGPI Nº 288.733/2006

Montt, Cecilia, Musso, Reynaldo y Chacón, Max. "Análisis de Accidentes de Tránsito con Métodos de Agrupamiento" Actas del VIII Congreso Chileno de Investigación de operaciones, OPTIMA 2009, Universidad del Bi- Bio, Concepción, Chillan.

Montt, Cecilia, Zúñiga, Alejandro y Chacón, Max Identificación de factores determinantes en accidentes de tránsito que afecten a las personas mediante redes bayesianas, Actas de XVI PANAM, July 15-18, 2010 Lisboa, Portugal.

Mussone, L, Ferrari, A y Oneta, M. An analysis of urban collisions using an artificial intelligence model. Milán: Accident Analysis and Prevention, 1999. págs. 705-718.

Organización Mundial de la Salud. Informe sobre la situación mundial de seguridad vial: es hora de pasar a la acción. Ginebra: s.n., 2009. http://www.who.int/violence_injury_prevention/road_safety_status/2009.

Sohn, So Young y Lee, Sung Ho Lee. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. s.l.: Safety Science, 2003.

Suykens, J.A.K y Vandewalle, J. Least Squares Support Vector Machine Classifiers. s.l.: Neur.Proc.Lett, 1999.

Vapnik., V. The nature of statistical learning theory. New York: Springer-Verlag: s.n., 1995.

Yang, W.T, Chen, H. C y Brown, D. B. Detecting Safer Driving Patterns by A Neural. 1999.