Hybrid Support Vector Machines to Classify Traffic Accidents in the *Región Metropolitana de Santiago*

Marcelo Farías Concha
Programa de Magíster en Ingeniería Informática,
Pontificia Universidad Católica de Valparaíso,
marcelo.fariasc@gmail.com

Abstract. This work proposes a method to classify the traffic accidents, especially in the territorial unit with greater number of people and vehicles of Chile: Metropolitan region. It used Support Vector Machines (SVM), tools which given a set of training samples as examples, allow to classify and thus train the SVM to build a model that predicts the class of a new sample. This technique despite being robust, it also has weaknesses, which are presented as a combinatorial problem in estimating and adjusting their input parameters. Obtaining good results depends on the intrinsic characteristics presented by SVM also the correct choice of the Kernel function and the input parameters. The choice and adjustment of parameters was performed with an evolutionary algorithm of Particle Swarm Optimization (PSO). Finally, to solve the problem different models were developed used SVM with PSO algorithms, which sought to classify the degree of severity of the people who are involved in traffic accidents: uninjured or injured. Searching better results, variations of PSO where used, generating different models, comparing the results obtained with this to make the best choice for optimal results in the classification. Therefore, the best results were obtained for Puente Alto, with 94% accuracy, 100% sensitivity and 83% specificity.

Keywords: Traffic Accidents, Classification, Support Vector Machine, Particle Swarm Optimization.

1.- Problema

Los siniestros de tránsito son eventos complejos y aleatorios que involucran una variedad de factores que se conjugan para su ocurrencia; entre las que destacan los factores humanos, del entorno, del estado del vehículo y del tránsito, entre otros. Los expertos también coinciden en que estos siniestros ocurren en gran medida porque no se respetan las reglamentaciones y normas existentes [1].

Los accidentes de tránsito a nivel mundial constituyen uno de los principales problemas sociales que han surgido en los últimos años. La proyección que se maneja de su influencia dentro de la salud de la población es bastante importante, puesto que pasa de estar dentro de las diez primeras causas de mortalidad a nivel mundial en el año 2004, a ubicarse entre las cinco de mayor relevancia para el año 2030. Estos accidentes influyen directamente en un alza considerable del gasto público y privado en las urbes del mundo. Esta situación también se repite en Chile tal como lo confirma la Comisión Nacional de Seguridad de Tránsito (CONASET) [2].

Al respecto, no disponer de una clasificación pertinente que considere la complejidad involucrada, entorpece la toma de decisiones; especialmente cuando la ocurrencia de estos accidentes se registra en metrópolis donde el transporte urbano ya en sí es un problema. En este contexto presente estudio se realiza en la región Metropolitana de Santiago, la cual concentrada la mayor población del país, lo

que trae como consecuencia que también se encuentre el mayor parque automotriz.

2.- Objetivos

El presente trabajo pretende clasificar los accidentes de tránsito ocurridos en la región Metropolitana mediante la utilización de Máquinas de Soporte Vectorial con Algoritmos Evolutivos.

Estas herramientas presentan buenas características para manejar datos espaciales y su aplicación permitiría generar información para una intervención adecuada del sistema de transporte.

3.- Estado del Arte

3.1.- Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial (del inglés *Support Vector Machine*) fueron desarrolladas en 1995 por Vladimir Vapnik y están basadas en la teoría de aprendizaje estadístico [3], que a su vez corresponden a la familia de los clasificadores lineales. A diferencia de las Redes Neuronales Artificiales, que utilizan durante la fase de entrenamiento el principio de Minimización del Riesgo Empírico (ERM de sus siglas en inglés, *Empirical Risk Minimization*), las SVM se basan en el principio de Minimización del Riesgo Estructural (SRM de sus siglas en inglés, *Structural Risk Minimization*), el que ha mostrado un mejor desempeño que el ERM, ya que las SVMs buscan minimizar la probabilidad de una clasificación errónea sobre nuevos ejemplos, a diferencia del ERM que minimiza el error sobre los datos de entrenamiento [4]. O sea, en palabras simples, lo que persigue esta herramienta es el aprendizaje a partir de los datos de entrada, los que pueden presentar características bastantes dispersas, tal como los datos existentes en presente estudio, además éstos son separados en dos grandes conjuntos o clases. Luego, el aprendizaje se logra mediante la búsqueda de alguna dependencia funcional entre un conjunto de vectores con los datos de entrada y de salida, permitiendo así encontrar un espacio lo más amplio posible con el cual se pueda separar los datos pertenecientes a una clase u otra.

Una SVM es un método de aprendizaje supervisado basado en Kernel (funciones núcleo), usados tanto para problemas de clasificación como de regresión. En el caso de la clasificación, las funciones de Kernel se utilizan usualmente para transformar los datos de entrada a un espacio de características de dimensión mayor en el cual los datos de entrada se vuelven más separables en comparación con el espacio de entrada original, para luego encontrar el hiperplano que los separe, y maximice el margen *m* entre las clases, tal como se puede apreciar en la Figura 1.

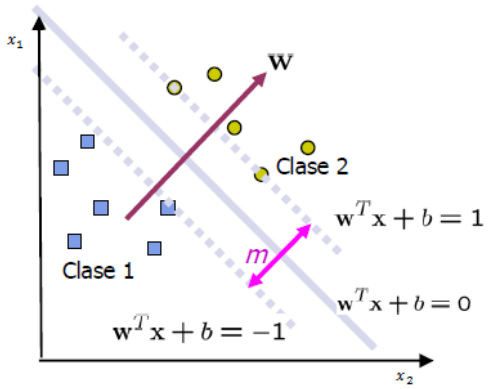

Figura 1: Ejemplo de hiperplano de separación entre las clases, en un problema linealmente separable.

La solución del hiperplano óptimo puede ser escrita como la combinación de unos pocos puntos de entrada que son llamados vectores de soporte. En un principio a los datos utilizados para hallar el hiperplano de decisión se les llama vectores de entrenamiento o aprendizaje. Algunas de las razones por las que este método ha tenido éxito es que no padece de mínimos locales y el modelo sólo depende de los datos con más información, los cuales son los vectores de soporte

Posterior a la fase de aprendizaje o entrenamiento, se comprueba el error cometido tomando otra muestra de datos

(denominados conjunto de test o validación) y se compara la salida que se obtiene con la clase original.

Las SVMs han sido desarrolladas como una técnica robusta para clasificación aplicada a grandes conjuntos de datos complejos con ruido; es decir, con variables inherentes al modelo que para otras técnicas aumentan la posibilidad de error en los resultados, pues resulta difícil poder cuantificarlas y observarlas. Además de sus sólidos fundamentos matemáticos en la teoría de aprendizaje estadístico, las SVMs han demostrado un rendimiento altamente competitivo en un amplio número de aplicaciones de la vida real, tales como bioinformática, minería de texto, reconocimiento facial y procesamiento de imágenes. Estas ventajas han establecido a las SVMs como una de las herramientas de última generación en máquinas de aprendizaje y minería de datos, junto con otras técnicas tales como Redes Neuronales y Sistemas Difusos [5].

Cabe destacar que existen aplicaciones en las que las SVMs han demostrado tener mejor desempeño que las técnicas tradicionales como las Redes Neuronales [6] y han sido introducidas como herramientas poderosas para resolver problemas de clasificación. Además, las SVMs se diferencian de las otras técnicas anteriormente mencionadas ya que no son afectadas por el problema de los mínimos locales, debido a que su entrenamiento se basa en problemas de optimización convexa. En resumen, algunas de las fortalezas de las SVMs son [7]:

- El entrenamiento es relativamente fácil.
- No hay óptimo local como en las Redes Neuronales Artificiales.
- Se escalan relativamente bien para datos en espacios dimensionales altos.
- El compromiso entre la complejidad del clasificador y el error puede ser controlado explícitamente.
- Datos no tradicionales como cadenas de caracteres y árboles pueden ser usados como entrada a la SVM, en vez de vectores de características.

Dentro de las debilidades se encuentra que, se necesita una buena función Kernel: es decir, se necesitan metodologías eficientes para sintonizar los parámetros de inicialización cualquier SVM. Por lo mencionado anteriormente, para este caso los parámetros serán estimados mediante la utilización de algoritmos genéticos tal como la Optimización por Enjambre de Partículas (PSO) y algunas de sus variaciones.

3.2.- Optimización por Enjambre de Partículas

La Optimización por Enjambre de Partículas (PSO de sus siglas en inglés, *Particle Swarm Optimization*) es una metaheurística evolutiva y de búsqueda [8], que fue desarrollada por Kennedy y Eberhart en 1995 [9]. La PSO simula el comportamiento social de organismos presentes en la naturaleza, tal como las bandadas de pájaros, los cardúmenes de peces o los enjambres de abejas; ésto con la finalidad de describir un sistema de evolución de forma automática. Cada candidato único a solución, tal como un ave individual de la bandada, es una partícula en el espacio de búsqueda, y cada partícula utiliza su memoria individual y conocimiento adquirido mediante el enjambre en su conjunto para encontrar la mejor solución [10]. Todas las partículas tienen valores *fitness* (medida de la calidad de la solución), que son evaluados por funciones *fitness* para ser optimizados, y tienen velocidades que dirigen el movimiento de las partículas en el sentido que corresponda a las mejores soluciones.

Durante el movimiento, cada partícula ajusta su posición de acuerdo a su propia experiencia, y al mismo tiempo de acuerdo a la experiencia de las partículas vecinas, haciendo uso de la mejor posición encontrada por sí mismo y por su vecino. Las partículas se mueven a través del espacio del problema siguiendo a las partículas óptimas actuales.

El enjambre inicial es generalmente creado de tal manera que la población de partículas se distribuye aleatoriamente sobre el espacio de búsqueda. En toda iteración, cada partícula es actualizada mediante dos mejores valores, llamados *pbest y gbest*.

Cada partícula realiza un seguimiento de sus coordenadas en el espacio del problema, las que se asocian con la mejor solución (*fitness*) que la partícula ha alcanzado hasta el momento. Este valor de *fitness* es almacenado, y es llamado *pbest*. Cuando la partícula toma la población completa como su vecino topológico, el mejor valor es un valor global y éste es llamado *gbest*.

Los valores funcionales adaptativos están basados en los datos de las características de las partículas que representa la dimensión característica. Esta data es clasificada mediante Máquinas de Soporte Vectorial para obtener precisión en la clasificación, la SVM sirve como evaluador de la función *fitness* de PSO.

3.3.- Variantes de PSO

El modelo tradicional de PSO ha experimentado modificaciones y generado algunas variantes tales como Quantum PSO y Improved PSO. A continuación se presentan las variaciones de PSO utilizadas y que además serán comparadas con la versión tradicional de PSO con el fin de evaluar los resultados y el desempeño de cada una con las distintas configuraciones.

La Quantum PSO (QPSO) es una versión de inspiración cuántica (quantum) del algoritmo PSO propuesta relativamente hace no mucho tiempo [11]. El algoritmo QPSO permite a todas las partículas tener un comportamiento cuántico, en lugar de la dinámica clásica que tenía la versión anterior. Así, en lugar del movimiento aleatorio, una especie de movimiento cuántico se aplica en el proceso de búsqueda. Cuando QPSO es probado frente un *set* de funciones de evaluación comparativa, se ha demostrado un rendimiento superior comparado a la versión clásica de PSO, pero bajo la condición de grandes tamaños de población [11]. Una de las características más atractivas de este nuevo algoritmo es que reduce el número de parámetros de control. Entonces estrictamente hablando, existe solo un parámetro que debe ser ajustado en QPSO.

La Improved PSO (IPSO en inglés Improved Particle Swarm Optimization) es una variante de PSO propuesta por Bin, Zhigang y Xingsheng [12]. En ella se plantea un peso de inercia dinámico con el cual se realiza la búsqueda, y que va disminuyendo de acuerdo a como van aumentando las iteraciones. Posee dos valores de entrada: $w_{inicial}$ y μ, que son definidos desde el principio.

4.- Método

4.1.- Clasificador Vectorial de Soporte

Para resolver el problema planteado se utilizan Máquinas de Soporte Vectorial de mínimos cuadrados (LS-SVM) las que permiten simplificar algunos aspectos de las SVM tradicionales sin perder sus ventajas. Además, para la selección y estimación de los parámetros de la SVM se utiliza PSO, QPSO e IPSO. Se presenta el modelo que será la base para construir el Clasificador Vectorial de Soporte propuesto.

- Modelo General

El modelo que se presenta a continuación tiene como fin optimizar el parámetro γ, que es el encargado de regularizar la máquina vectorial, además de encontrar los parámetros asociados al Kernel. En este estudio es utilizado principalmente el Kernel Gaussiano y como ya se mencionó, para la obtención de los parámetros óptimos del modelo se utiliza PSO.

- Parámetros de Entrada

En PSO, ajustar los parámetros permite optimizar el rendimiento del algoritmo, pero

para el primer procesamiento se definen parámetros iniciales que sirven de partida para el modelo. Una buena elección de los parámetros C_1 y C_2 puesto que pueden producir una rápida convergencia del algoritmo y evitar mínimos locales, es $C_1 = C_2 = 2$.

Para la velocidad, se debe limitar un máximo y mínimo para que los movimientos de las partículas no se acerquen a valores que no presentan utilidad, por lo tanto se define $V_{max} = V_{min} = \pm 1,6$. Finalmente, el coeficiente inercial $w = 0,8$, donde éste a medida que aumentan las iteraciones va disminuyendo lo que provoca un cambio desde un modo de exploración, donde se generan evaluaciones en regiones distantes del espacio de búsqueda, a un modo de explotación en el cual se evalúan soluciones en regiones acotadas y pequeñas con respecto al espacio de búsqueda.

- Métricas de Evaluación

Para evaluar los distintos modelos y comparar los resultados obtenidos. es necesario utilizar métricas que permitirán conocer el grado de exactitud de la clasificación que se obtenga. Fueron seleccionadas las siguientes métricas: Exactitud, Sensibilidad, Especificidad y área de la curva ROC. Estas métricas se forman considerando los siguientes conceptos:

- Verdaderos Positivos (VP): número de éxitos. O sea, corresponde al número de personas que se clasificaron con daños.
-Verdaderos Negativos (VN): número de rechazos correctos. Corresponde a las personas que fueron detectadas ilesas correctamente.
- Falsos Positivos (FP): número de falsos lesionados. En este contexto corresponden al número de personas detectadas como lesionadas, pero que realmente resultaron ilesas.
-Falsos Negativos (FN): número de falsos ilesos. Corresponde al número de personas detectadas como ilesas, pero que realmente resultaron lesionadas.

Los valores mencionados anteriormente, VP, VN, FP y FN se presentan en la Tabla 1, la que se denomina matriz de confusión, donde cada fila de la matriz representa el número de resultados obtenidos en cada clase, mientras que cada columna representa las instancias en la clase real.

Tabla 1 : Matriz de Confusión

		RESULTADO REAL	
		POSITIVO	NEGATIVO
RESULTADO OBTENIDO	POSITIVO	**VP**	**FP**
	NEGATIVO	**FN**	**VN**

A partir de los Verdaderos Positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP) y Falsos Negativos (FN) se construyen los siguientes ratios, los cuales corresponden a las siguientes métricas:

Exactitud: corresponde al total de personas bien clasificadas, ya sea con lesión o sin lesión, dentro del total de personas, siendo representada de la siguiente forma:

$$Exactitud = \left(\frac{VP + VN}{VP + VN + FN + FP}\right) \quad (1)$$

Sensibilidad: corresponde a la probabilidad de que una persona realmente lesionada sea detectada como tal por la prueba, siendo representada por la siguiente ecuación:

$$Sensibilidad = \left(\frac{VP}{VP + FN}\right) \quad (2)$$

Especificidad: corresponde a la probabilidad de que una persona ilesa sea detectada como tal por la prueba, siendo representada por la siguiente ecuación:

$$Especificidad = \left(\frac{VN}{VN + FP}\right) \quad (3)$$

Valor Predictivo Positivo (VPP) o Precisión: corresponde a la probabilidad de padecer la

lesión si se obtiene un resultado positivo en el test, siendo representada por la siguiente ecuación:

$$VPP = \left(\frac{VP}{VP + FP}\right) \quad (4)$$

Valor Predictivo Negativo: corresponde a la probabilidad de que una persona con un resultado negativo en la prueba esté realmente ilesa. Es representada por la siguiente ecuación:

$$VPN = \left(\frac{VN}{VN + FN}\right) \quad (5)$$

Fracción de Falsos Positivos (FFP o "1-Especificidad"):

$$FFP = \left(\frac{FP}{FP + VN}\right) o\ 1 - Especificidad \quad (6)$$

Otra de las métricas más usadas en clasificación binaria corresponde a la curva ROC (*Receiver Operating Characteristic*). La curva ROC es un gráfico en el que se observan todos los pares sensibilidad/especificidad resultantes de la variación continua de los puntos de corte en todo el rango de resultados observados. Se define por FFP y FVP como ejes x e y respectivamente. Representa los intercambios entre verdaderos positivos y falsos positivos. Dado que FVP es equivalente a la "sensibilidad" y FFP es "1-especificidad", el gráfico ROC se llama a veces la representación de (1-Especificidad) frente a la Sensibilidad. Cada resultado de la clasificación de una instancia de la matriz de confusión representa un punto en el espacio ROC. Cada punto de la curva representa un par S/1-E correspondiente a un nivel de decisión determinado. Una prueba con discriminación perfecta, sin solapamiento de resultados en las dos poblaciones, tiene una curva ROC que pasa por la esquina superior izquierda, donde Sensibilidad y Especificidad toman valores máximos (igual a 1). El Área Bajo la Curva (UAC): A partir de la curva ROC se deriva el "Área Bajo la Curva" (UAC). El área bajo la curva ROC es siempre mayor o igual a 0,5. El rango de valores se mueve entre 1 (discriminación perfecta) y 0,5 (no hay diferencias en la distribución de los valores de la prueba entre los 2 grupos).

- Representación y Explicación de los datos utilizados

Se describen los datos de los accidentes de tránsito que fueron considerados para el desarrollo del estudio. En primera instancia se obtienen a partir del registro que genera Carabineros de Chile luego de concurrir al lugar de un accidente de tránsito. Se obtiene la información de forma manual en planillas que se encuentran predefinidas.

Posteriormente estos datos se entregan a la CONASET, entidad encargada de almacenarlos en sus bases de datos para generar estadísticas y estudios que permitan obtener información que sea de utilidad.

En esta oportunidad, los datos fueron entregados por la Escuela de Ingeniería de Transporte de la Pontificia Universidad Católica de Valparaíso, quienes trabajan directamente con la CONASET y es por esto que tienen acceso a esta información. Existían datos desde el 2003 al 2009, los cuales a su vez se encuentran divididos en 3 sub planillas, estas son Accidentes, Personas y Vehículos, las cuales son explicadas a continuación.

4.2.- Desarrollo del Clasificador de Soporte Vectorial

Se presenta el proceso de desarrollo de los modelos realizados, donde se considera la preparación correspondiente de los datos, puesto que no venían en el formato necesario para ser introducidos en el Clasificador de Soporte Vectorial. Por lo tanto, en primera instancia fue necesario realizar una limpieza de la data, para luego codificarla tal como se muestra más adelante. Posteriormente se presenta la función de Validación Cruzada, la cual fue la función *fitness* utilizada para evaluar los modelos que fueron desarrollados.

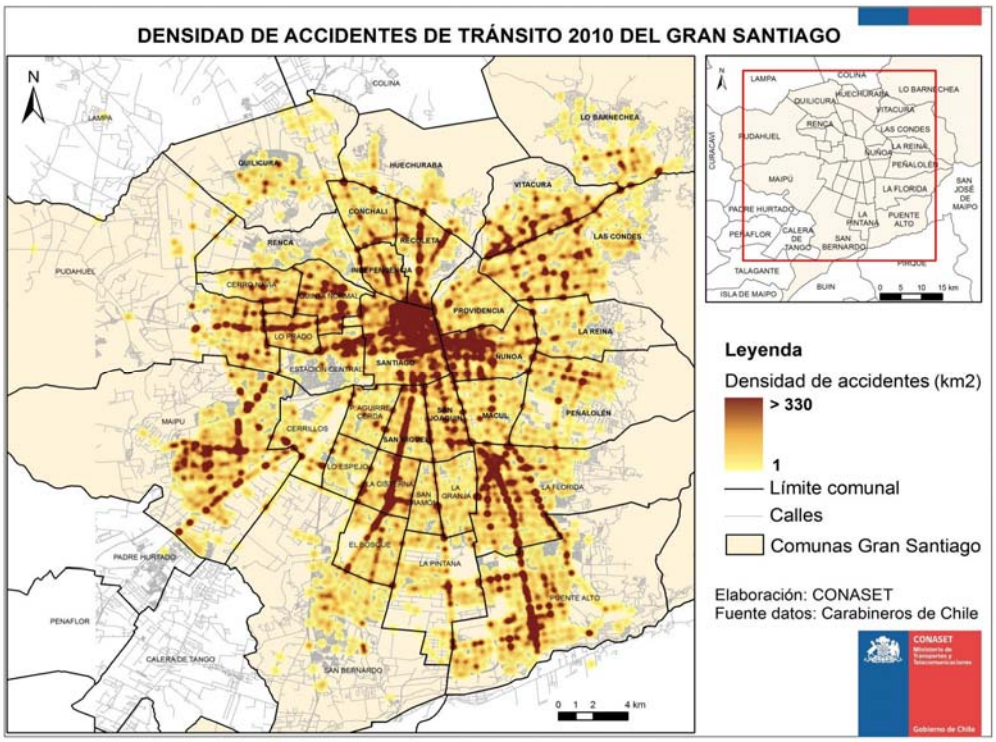

Figura 2: Densidad de Accidentes de Tránsito Región Metropolitana.

- Pre-Procesamiento de los Datos

Anterior a la implementación, es necesario realizar un trabajo previo con los datos; es decir, prepararlos para que estos puedan ser utilizados por la SVM. Cabe destacar que de la cantidad de datos existentes de unos 75.000 registros entre el 2003 y 2009, se consideran específicamente los datos para la región Metropolitana ya que ésta es la ciudad con más habitantes, por lo tanto con mayor cantidad de vehículos y en consecuencia con mayor cantidad de accidentes del país.

En primer lugar fueron eliminados los datos en los que existían atributos con valores nulos puesto que estos no presentan utilidad. Luego, se filtran por región, puesto que se deben considerar solo los datos correspondientes a las comunas de la región Metropolitana ya que esta es el foco del estudio.

Posteriormente fueron eliminados los datos en los que se encontraron anomalías como por ejemplo que la edad de las personas excediera el límite lógico, puesto que hubo casos en que la edad era de 999 años.

Selección de Datos y Codificación

Dados los antecedentes presentados por CONASET en su sitio web y lo observado en los datos existentes [2], la mayor cantidad de accidentes de la Región Metropolitana ocurren en las siguientes comunas: Santiago, Puente Alto, La Florida, tal como se puede apreciar en el gráfico de la Figura 2.

También tomando en cuenta que el modelo a clasificar corresponde al estado en el que resultan las personas que se vieron involucradas en los accidentes de tránsito, el que puede ser, con daños o ilesos. Se deben tomar en cuenta los atributos más relevantes y que permita representar de la mejor manera el modelo.

Se debe tener en cuenta que para el trabajo es necesario utilizar los datos codificados, puesto que para poder trabajar con la SVM se debe contar con datos cuantitativos mientras que los datos originales son cualitativos. Por lo tanto, a continuación cada atributo es codificado.

De estos 13 atributos que se consideraran, cabe destacar que el número 13, o sea, el atributo Resultado será considerado como objetivo. Éste está compuesto por las siguientes posibilidades:

- Muertos = 1
- Graves. = 1
- Menos Graves. = 1
- Leves = 1
- Ilesos = -1

Todo Resultado con el código 1 será considerado como un solo resultado, puesto que se quiere clasificar el estado en el que resultan las personas luego del accidente, o sea, con daño o ilesos. Por lo tanto a una "Persona con Daños" se le asigna el código 1 mientras que a una "Persona Ilesa" se le asigna el código -1.

Función de Costo o Fitness

La función fitness corresponde a la función utilizada para evaluar la calidad de la solución que obtiene el Clasificador Vectorial de Soporte. En este caso en particular se utilizará la Validación Cruzada, función con la cual serán evaluados los 3 modelos que serán presentados a continuación. La función fitness antes mencionada, consiste básicamente en dividir los datos de entrenamiento en forma aleatoria en L partes. En la $i-ésima(i=1,\dots,l)$ iteración la $i-ésima$ parte de los datos es usada para realizar la validación y las $l-1$ son asignados como datos de entrenamiento, donde se mide el costo de los elementos mal clasificados. De la misma forma, se itera hasta la última parte del conjunto de datos, rescatando el costo en cada iteración. Finalmente se realiza un promedio de los costos de las distintas iteraciones y este se retorna al algoritmo PSO (o sus variaciones). Cabe destacar que en un principio del estudio se utilizó también otra función de fitness, MSE (error cuadrático medio) pero ésta en ningún caso presentó mejores resultados que la Validación Cruzada, por lo cual se descartó del trabajo final.

5.- Resultados

5.1.- Clasificaciones

Modelos LS-SVM

Debido a la necesidad de encontrar los parámetros para cada uno de los tres modelos realizados, fue necesario realizar pruebas en cada uno de los casos. Por lo tanto, durante la etapa de entrenamiento de los tres modelos se realizó un gran número de iteraciones con la finalidad de buscar el conjunto óptimo de parámetros para la etapa posterior de prueba o testing, para lo que se utilizaron combinaciones de los siguientes parámetros.

Tabla 2: Parámetros de Entrenamiento

Parámetros de Entrenamiento	
n° de Partículas	10-20-30
Número iteraciones	80-100-150
% de Entrenamiento	70% a 90%

Con los resultados obtenidos en esta etapa, se logró encontrar los parámetros que serán utilizados en los 3 modelos para el resto del trabajo.

Por otro lado, es necesario destacar que luego de las primeras pruebas se determinó que ejecutar el algoritmo con todos los registros correspondientes a un año no era posible, puesto que la herramienta y la máquina no eran capaces de realizar este

trabajo por la gran cantidad de información que se debía manejar además del manejo de matrices de gran dimensión. Lo anterior generaba problemas de memoria y la ejecución no se podía realizar con éxito. Por lo tanto la decisión de considerar solamente las comunas más representativas de la región Metropolitana presentó grandes beneficios puesto que el trabajo se podía realizar con mayor rapidez ya que el número de datos se reduce en gran medida.

Modelo General LS-SVM PSO

Se utiliza el algoritmo PSO tradicional para la optimización de los parámetros necesarios, por lo que la partícula del enjambre está compuesta por dos parámetros que se inicializan aleatoriamente: γ y σ^2.

Se realizaron iteraciones en etapas de training de la máquina vectorial para poder encontrar los parámetros óptimos propios de PSO con los que se realizaría a continuación la etapa de *testing*. Dentro de todas las ejecuciones se obtuvo que los parámetros óptimos son los que se muestran a continuación. Además el criterio de término corresponde al número máximo de iteraciones.

Tabla 3: Parámetros PSO.

Parámetros PSO	
Número de partículas	20
Número iteraciones	100
Coeficiente inercial (w)	0,8
Componente cognitiva (c1)	2
Componente social (c2)	2
Velocidad máxima	1,6
% de entrenamiento	90%

Modelo General LS-SVM QPSO

La diferencia entre PSO y QPSO está en el algoritmo de búsqueda de los parámetros γ y σ^2, puesto que en QPSO se realiza la optimización basándose solamente en la posición de la partícula. Es decir, la velocidad de PSO es dejada de lado. Pero ambos modelos se asemejan, ya que utilizan la misma configuración de sus partículas; es decir, se componen de γ y σ², inicializadas aleatoriamente. A continuación se muestran los parámetros óptimos con los que se ejecutarán las pruebas:

Tabla 4: Parámetros QPSO.

Parámetros QPSO	
Número de partículas	20
Número iteraciones	100
Alfa 1	0,6
Alfa 2	1
% de entrenamiento	90%

Modelo General LS-SVM IPSO

IPSO posee dos valores de entrada, los que fueron inicializados con los siguientes valores, $w_{inicial} = 0,8$ y $\mu = 1,0012$.

Además, tiene por característica un peso de inercia dinámico; es decir, el peso de inercia va disminuyendo conforme se incrementa la generación iterativa. Se debe considerar que un valor alto del peso de inercia privilegia la exploración global y un peso de inercia más pequeño privilegia la exploración local. En IPSO, todas las partículas comparten información mutuamente a nivel global y se favorece con descubrimientos y experiencias previas de las partículas aledañas durante el proceso de búsqueda. También utiliza los parámetros γ y σ², los cuales se inicializan aleatoriamente, y los parámetros son los siguientes:

Tabla 5: Parámetros IPSO.

Parámetros IPSO	
Número de partículas	20
Número iteraciones	100
Coeficiente inercial (w)	0.8
μ	1.00012
Componente cognitiva (c1)	2
Componente social (c2)	2
Velocidad máxima	1,6
% de entrenamiento	90%

LS-SVM PSO para Santiago
El mejor resultado obtenido con una exactitud de 84% es presentado a continuación.

Tabla 6: Matriz de Confusión del Mejor Resultado PSO para Santiago.

	POSITIVO	NEGATIVO	Total		
POS	48	5	53	PRECISIÓN	0,906
NEG	17	68	85	VPN	0,80
Total	65	73	138		
	SENSIBILIDAD	ESPECIFICIDAD			
	0,74	0,93			

Se clasifican correctamente el 84% de las personas que resultan lesionadas o ilesas. La sensibilidad muestra que un 74% de personas fueron bien clasificadas y que resultaron lesionadas. La especificidad de un 93% correspondiente al porcentaje de personas ilesas bien clasificadas. La precisión indica que del porcentaje de personas detectadas lesionadas un 90,6% de estas realmente resultaron lesionadas, y el valor predictivo negativo fue de 80%, lo que significa que de las personas detectadas como ilesas el 80% de estas están realmente ilesas. Los mejores valores obtenidos para los parámetros γ y σ^2 son 1104,8 y 193,76 y el tiempo de ejecución es de 13 minutos aproximadamente.

LS-SVM QPSO para Santiago
El mejor resultado con una exactitud de 82% es presentado a continuación. Para este valor se obtuvo que los mejores parámetros obtenidos fueron γ=944,4 y σ^2=43,68 con un tiempo de ejecución de 13 minutos.

Tabla 7: Matriz de Confusión del Mejor Resultado QPSO para Santiago.

	POSITIVO	NEGATIVO	Total		
POS	55	10	65	PRECISIÓN	0,846
NEG	15	58	73	VPN	0,79
Total	70	68	138		
	SENSIBILIDAD	ESPECIFICIDAD			
	0,79	0,85			

Se clasifican correctamente el 82% de las personas que resultan en el estado lesionado o ileso. La sensibilidad de un 79%, corresponde al porcentaje de personas bien clasificadas y que resultaron lesionadas. La especificidad de un 67% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso. La precisión indica que del porcentaje de personas detectadas como lesionadas un 84,6% de estas realmente resultaron lesionadas, y finalmente, el valor predictivo negativo fue de un 79%, lo que significa que de las personas detectadas como ilesas el 79% resulta realmente ilesas.

LS-SVM IPSO para Santiago
El mejor resultado se obtuvo con una exactitud de 83%. Y los mejores valores para los parámetros γ y σ^2 fueron 440,26 y 37,11 respectivamente y se demoró aproximadamente 15 minutos.

Tabla 8: Matriz de Confusión del Mejor Resultado IPSO para Santiago.

	POSITIVO	NEGATIVO	Total		
POS	49	9	58	PRECISIÓN	0,845
NEG	15	65	80	VPN	0,81
Total	64	74	138		
	SENSIBILIDAD	ESPECIFICIDAD			
	0,77	0,88			

Se han clasificado correctamente el 83% de las personas que resultan en el estado lesionado o ileso. La sensibilidad, un 77%, indica que esta cifra corresponde al porcentaje de personas bien clasificadas y que resultaron lesionadas. La especificidad es de un 88% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso. Luego, relativo a la seguridad del resultado, la precisión indica que del porcentaje de personas detectadas como lesionadas un 84,5% de estas realmente resultaron lesionadas, y finalmente, el valor predictivo negativo fue de un 81%, lo que significa que de las personas detectadas como ilesas el 81% resulta realmente ilesas.

LS-SVM PSO para Puente Alto

El mejor resultado obtenido con una exactitud de 89,18% es presentado a continuación.

Tabla 9: Matriz de Confusión del Mejor Resultado PSO para Puente Alto.

	POSITIVO	NEGATIVO	Total		
POS	74	9	83	PRECISIÓN	0,892
NEG	3	25	28	VPN	0,89
Total	77	34	111		
	SENSIBILIDAD	ESPECIFICIDAD			
	0,96	0,74			

Se han clasificando correctamente el 89,18% de las personas que resultan en el estado lesionado o ileso, la sensibilidad nos muestra con este 96% que a esta cifra corresponde el porcentaje de personas bien clasificadas y que resultaron lesionadas. La especificidad resultó ser un 74% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso. Relativo a la seguridad del resultado, la precisión del 89,2% indica que del porcentaje de personas detectadas como lesionadas el 89,2% de estas realmente resultaron lesionadas, y el valor predictivo negativo fue de 89%, lo que significa que de las personas detectadas como ilesas el 89% de estas están realmente ilesas. Para este modelo se puede apreciar también que los mejores valores obtenidos para los parámetros γ y σ^2 son 38,36 y 169,56 respectivamente, y el tiempo aproximado de ejecución fueron 7 minutos.

LS-SVM QPSO para Puente Alto

El mejor resultado obtenido con una exactitud de 94% es presentado a continuación.

Tabla 10: Matriz de Confusión del Mejor Resultado QPSO para Puente Alto.

	POSITIVO	NEGATIVO	Total		
POS	70	7	77	PRECISIÓN	0,909
NEG	0	34	34	VPN	1,00
Total	70	41	111		
	SENSIBILIDAD	ESPECIFICIDAD			
	100	0,83			

Se clasifica 94% de las personas que resultan lesionadas o ilesas. La sensibilidad del 100% indica que todas las personas fueron bien clasificadas y estas resultaron lesionadas. La especificidad resultó ser un 83% correspondiente al porcentaje de personas bien clasificadas en el estado ileso. Relativo a la seguridad del resultado, la precisión indica que del porcentaje de personas detectadas como lesionadas un 90,9% realmente resultaron lesionadas, y el valor predictivo negativo fue de un 100%, lo que significa que de las personas detectadas como ilesas el 100% resultó ileso. Los mejores valores para los parámetros γ y σ^2 fueron 641,27 y 51,51 respectivamente. El tiempo que demoró esta ejecución fue de 7 minutos aproximadamente.

LS-SVM IPSO para Puente Alto

El mejor resultado obtenido con una exactitud de 88% es presentado a continuación. Además los mejores valores para los parámetros γ y σ^2 fueron 482,92 y 98,34 respectivamente y se demoró aproximadamente 7 minutos.

Tabla 11: Matriz de Confusión del Mejor Resultado IPSO para Puente Alto.

	POSITIVO	NEGATIVO	Total		
POS	76	9	85	PRECISIÓN	0,894
NEG	4	22	26	VPN	0,85
Total	80	31	111		
	SENSIBILIDAD	ESPECIFICIDAD			
	0,95	0,71			

La exactitud fue de un 88%, lo que quiere decir que el modelo ha clasificado de manera correcta el 88% de las personas que resultan en el estado lesionado o ileso. La sensibilidad fue un 95%, que corresponde al porcentaje de personas bien clasificadas y que resultaron lesionadas. A continuación se presenta la especificidad, y esta resultó ser un 71% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso. Relativo a la seguridad del resultado la precisión fue de un 89,4%, y esto significa

que de este porcentaje de personas detectadas como lesionadas, estas realmente resultaron lesionadas, y finalmente, el valor predictivo negativo fue de 85%, lo que significa que de las personas detectadas como ilesas el 85% de estas están realmente ilesas.

LS-SVM PSO para La Florida

El mejor resultado obtenido con una exactitud de 89%, demoró aproximadamente 10 minutos en ejecutarse y los mejores parámetros γ y σ^2 fueron 718,79 y 137, 87 respectivamente.

Tabla 12: Matriz de Confusión del Mejor Resultado PSO para La Florida.

	POSITIVO	NEGATIVO	Total		
POS	77	10	87	PRECISIÓN	0,885
NEG	4	31	35	VPN	0,89
Total	81	41	122		
	SENSIBILIDAD	ESPECIFICIDAD			
	0,95	0,76			

Se ha clasificando de manera correcta el 89% de las personas que resultan en el estado lesionado o ileso, por otra parte la sensibilidad nos muestra que el 95% corresponde al porcentaje de personas bien clasificadas y que resultaron lesionadas. La especificidad resultó ser un 76% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso. Relativo a la seguridad del resultado, la precisión indica que del porcentaje de personas detectadas como lesionadas un 88,5% de estas realmente resultaron lesionadas, y finalmente, el valor predictivo negativo fue de 89%, lo que significa que de las personas detectadas como ilesas el 89% de estas están realmente ilesas.

LS-SVM QPSO para La Florida

El mejor resultado obtenido con una exactitud de 84% es presentado a continuación.. Además los mejores valores para los parámetros γ y σ^2 fueron 897,49 y 47,17 respectivamente y se demoró aproximadamente 8 minutos.

Tabla 13: Matriz de Confusión del Mejor Resultado QPSO para La Florida.

	POSITIVO	NEGATIVO	Total		
POS	67	12	79	PRECISIÓN	0,848
NEG	8	35	43	VPN	0,81
Total	75	47	122		
	SENSIBILIDAD	ESPECIFICIDAD			
	0,89	0,74			

Se ha clasificado correctamente el 84% de las personas que resultan en el estado lesionado o ileso. La sensibilidad fue de un 89%, lo que indica que esta cifra corresponde al porcentaje de personas bien clasificadas y que resultaron lesionadas. La especificidad de un 74% corresponde al porcentaje de personas bien clasificadas en el estado ileso. Relativo a la seguridad del resultado, la precisión indica que del porcentaje de personas detectadas como lesionadas un 84,8% de estas realmente resultaron lesionadas, y el valor predictivo negativo fue de un 81%, lo que significa que de las personas detectadas como ilesas el 81% resulta realmente ilesas.

LS-SVM IPSO para La Florida

El mejor resultado obtenido con una exactitud de 83%. Además los mejores valores para los parámetros γ y σ^2 fueron 482,92 y 98,34 respectivamente y se demoró aproximadamente 8 minutos.

Tabla 14: Matriz de Confusión del Mejor Resultado IPSO para La Florida.

	POSITIVO	NEGATIVO	Total		
POS	60	8	68	PRECISIÓN	0,882
NEG	13	41	54	VPN	0,76
Total	73	49	122		
	SENSIBILIDAD	ESPECIFICIDAD			
	0,82	0,84			

La exactitud es de un 83%, por lo que el modelo ha clasificado de manera correcta el 83% de las personas en estado lesionado o ileso. La sensibilidad obtenida fue 82%, que corresponde al porcentaje de personas bien clasificadas. La especificidad de un 82%, corresponde al porcentaje de personas bien clasificadas en el estado ileso. La precisión

de 88,2%, es el porcentaje de personas detectadas como lesionadas que realmente resultaron lesionadas, y el valor predictivo negativo fue de 76%, que significa que de las personas detectadas como ilesas el 76% de estas están realmente ilesas.

5.2.- Comparación Modelos.

Comparación por Exactitud

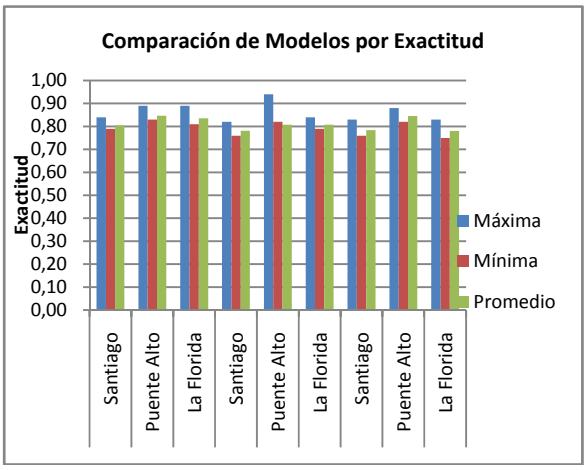

Figura 3: Comparación de los Modelos por Exactitud.

De gráfico de la Figura 3 se puede decir que los mejores resultados fueron obtenidos con PSO, puesto que presenta los mejores valores en general, los máximos y mínimos. Mientras que el mejor valor individual obtenido fue para el modelo QPSO en la comuna de Puente Alto con un 94%, pero hay que tener en cuenta que para este caso el valor mínimo era el que presentaba la mayor diferencia del máximo, esta diferencia era de 12 puntos porcentuales.

Comparación por Tiempo

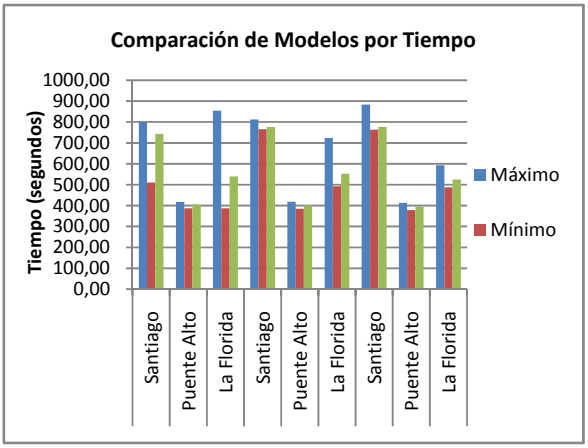

Figura 4: Comparación de Modelos por Tiempo.

En el gráfico de la Figura 4 se presentan los tiempos de ejecución por cada modelo y comuna analizada. Entonces, se observa que PSO es el más costoso en relación al tiempo máximo que se presenta, y luego, se puede ver que IPSO es el modelo que en promedio es menos costoso.

6.- CONCLUSIONES

Se ha logrado establecer el estudio del Estado del Arte en general de los componentes básicos con los que se debe contar para realizar el desarrollo de un modelo basado en Máquinas de Soporte Vectorial (SVM) para la clasificación de accidentes de tránsito, además realizando la optimización de los parámetros de esta SVM mediante Algoritmos Evolutivos, tal como PSO y algunas de sus variaciones.

Luego, se ha desarrollado el modelo del clasificador propuesto, y con esto se da inicio a la correspondiente implementación. Lo dicho anteriormente relacionado con el desarrollo del modelo se basa puntualmente en el estudio realizado sobre LS-SVM y PSO con sus variaciones, permitiendo la construcción de los modelos que permitieron el desarrollo del trabajo planificado y la obtención de resultados, que permitieran realizar una comparación para la obtención

del mejor clasificador dentro de los distintos modelos que fueron generados.

Los resultados obtenidos luego de evaluar los modelos, se encuentran en un buen nivel considerando las tres alternativas que fueron consideradas, presentando leves variaciones entre sí. El mejor resultado puntual fue presentado por el modelo LS-SVM QPSO, mientras que el modelo LS-SVM PSO presentó los mejores resultados en promedio para las tres comunas.

Con estos resultados que fueron presentados, se puede apreciar que las Máquinas de Soporte Vectorial permiten obtener buenos resultados para realizar clasificación de datos, pero si no fuera por la estimación de los parámetros mediante la utilización de PSO, y sus variaciones; los resultados obtenidos no hubiesen sido tan alentadores. Por lo que se puede concluir que los Algoritmos Evolutivos, particularmente los que fueron utilizados para este trabajo, presentan gran utilidad y alto desempeño al momento de ser utilizados en problemas de optimización, con lo que se demuestran las potencialidades de los modelos desarrollados, ya que con la obtención de buenos parámetros se lograrán buenos resultados.

El mejor resultado obtenido fue para el modelo LS-SVM QPSO en la comuna de Puente Alto con un 94% de exactitud, 100% de sensibilidad y 83% de especificidad, con lo que se puede concluir que las Máquinas de Soporte Vectorial, utilizando Algoritmos Evolutivos tal como PSO para la obtención de los parámetros, presentan gran utilidad para la clasificación de datos.

Comparado con otros sistemas clasificadores, las SVMs son muy eficientes desde diversas perspectivas. El proceso de aprendizaje es un proceso matemático definido que permite obtener el mejor clasificador, no tan solo un buen clasificador como se obtiene en muchos entrenamientos de redes neuronales, claramente hay que tener presente que esta mejor solución es dependiente del Kernel utilizado y de los parámetros que sean escogidos, los cuales pueden variar dependiendo de la experiencia de quién realice el estudio y también de los resultados obtenidos en pruebas previas. Por otra parte, una vez obtenido el modelo, es muy fácil implementarlo en diferentes sistemas. Además, se debe destacar que el tiempo de entrenamiento es relativamente corto y que posee una alta velocidad de ejecución en la clasificación de grandes conjuntos de datos.

Este tipo de estudio podría ser de bastante utilidad, en primer lugar a nivel estadístico puesto que permitiría contrastarlo con los datos que existen en la actualidad, y en segundo lugar aún más importante, para que las autoridades tomen decisiones sobre cambios o mejoras que realizar en las calles del país con las cuales se puedan disminuir los accidentes o la gravedad de los mismos.

Finalmente, cabe mencionar que existen variadas opciones para desarrollar nuevos proyectos relacionados a este tema, entre los cuales destacan, la utilización de datos actualizados que permitan validar los resultados obtenidos y además considerar nuevos puntos que puedan presentar mayor cantidad de accidentes. También, se podría realizar un análisis de los datos iniciales antes de ser filtrados, de manera que se pueda determinar qué características permiten que la clasificación sea más eficiente, con lo que se tendría mayor claridad en las causas que determinan la ocurrencia de los accidentes. Otra alternativa relacionada a este tema, sería utilizar otro tipo de herramientas para las distintas fases del desarrollo del trabajo, como por ejemplo para la obtención de los parámetros, con las que se puedan obtener nuevos resultados que permitan realizar comparaciones y así poder generar aún mejores clasificadores. O también, realizar un cambio más drástico y variar la herramienta de fondo, y utilizar por ejemplo Redes Neuronales Artificiales que permitan tener modelos paralelos con los cuales se

pueda conseguir una comparación general del resultado de los distintos clasificadores.

REFERENCIAS

[1] Organización Mundial de la Salud. Los Accidentes de Tránsito Entre las Principales Causas de Muerte Entre los Jóvenes. Disponible vía web en http://www.who.int/mediacentre/news/releases/2009/adolescent_mortality_20090911/es/index.html, 2009.

[2] Comisión Nacional de Seguridad de Tránsito. Estadísticas Generales, Cantidad de siniestros de tránsito y de víctimas. Disponible vía web en http://www.conaset.cl/portal/portal/default/estadisticas_generales, 2000 – 2010.

[3] Vapnik., V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

[4] Vapnik, V. The Nature of Statistical Learning Theory (2nd Edition). Springer, 2000.

[5] Tsoukalas y Uhrig. Fuzzy and Neural Approaches in Engineering. John Wiley and Sons, N.Y., 1997.

[6] Burges, C. A Tutorial on Support Vector Machine for Pattern Recognition. Data Mining and Knowledge Discovery. 1998.

[7] Betancourt, Gustavo. Las Máquinas de Soporte Vectorial (SVMs). Scientia Et Technica, vol. XI, núm. 27, abril, 2005, Universidad Tecnológica de Pereira.

[8] Glover, F, y G.G Kochenberger. Handbook of Metaheuristics. Kluwer Aca-demic Publishers, 2003.

[9] Eberhart, R y Kennedy, J. A new optimizer using particle swarm theory. Proceedings of the sixth International Symposium on Micro Machine and Human Science, 1995.

[10] Venter, G. and Sobieszczanski-Sobieski, J. Proceedings of the 43rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference. Denver, 2002.

[11] Jun Sun, Bin Feng, and Wenbo Xu. Particle swarm optimization with particles having quantum behavior. in Proc. Cong. Evolutionary Computation, 2004.

[12] Jiao, Bin, Lian, Zhigang y Gu, Xingsheng. *A dynamic inertia weight particle swarm optimization algorithm.* China, 2006.

Autor principal

Marcelo Farías Concha es Magister en Ingeniería Informática (PUCV, 2013) e Ingeniero Civil Informático (PUCV, 2013). Actualmente se desempeña como Analista de Sistemas en la Subgerencia de Fidelización en Entel. Sus áreas de intereses son: Inteligencia de Negocios y temas relacionados a las Tecnologías de Información. También con interés por la actualización personal y el trabajo en equipos multidisciplinarios para obtener un conocimiento general de distintos temas.

Paper Info

Fecha de recepción: julio 2012.
Fecha de aceptación: julio 2012.
Revisores: 3.
Cantidad de revisiones consolidadas: 2.
Total de observaciones: 5.
Índice de Novedad: 0,71.
Índice de Utilidad: 0,89.