UAEM | Universidad Autónoma del Estado de México

Oficio No. DIR/043/2016.
Zumpango, Estado de México, a 08 de febrero del 2016.

DRA. EN EST. LAT. ANGELES MA. DEL ROSARIO PÉREZ BERNAL
SECRETARIA DE INVESTIGACIÓN
Y ESTUDIOS AVANZADOS, UAEM
PRESENTE

Anticipando un cordial saludo me dirijo a usted, para presentarle el Reporte Técnico y el Documento donde se especifican los productos generados del Proyecto de Investigación con Clave UAEM 3790/2014/CID, cuyo responsable técnico es el Dr. Asdrúbal López Chau, y los colaboradores son el M.T.I. Jorge Bautista López y el M.T.I. Carlos Alberto Rojas Hernández. Ambos documentos han sido aprobados por los H. H. Consejos Académico y de Gobierno del Centro Universitario UAEM Zumpango, en la Sesión Ordinaria del 29 de enero de 2016.

Se anexa copia del acta del H. Consejo de Gobierno de la Sesión Ordinaria del mes de enero del año en curso.

Sin más por el momento, le reitero mis más altas y distinguidas consideraciones.

A T E N T A M E N T E
PATRIA, CIENCIA Y TRABAJO
"2015, Año del Bicentenario Luctuoso de José María Morelos y Pavón"

CENTRO UNIVERSITARIO UAEM ZUMPANGO
DIRECCIÓN

DR. EN E.J. RODOLFO TELLEZ CUEVAS
ENCARGADO DEL DESPACHO DE LA DIRECCIÓN
DEL CU UAEM ZUMPANGO

www.uaemex.mx

c.c.p. Archivo
RTC/lmrs*

Centro Universitario UAEM Zumpango
•Km. 3.5 Camino Viejo a Jilotzingo, Valle Hermoso, Zumpango, Méx. C.P. 55600, Tel: 01(591)917-41-39, 01(591)917-41-40
rtellezc@uaemex.mx

Hoja de presentación de Informe Final del Proyecto de Investigación

Título y clave del proyecto:
MÉTODOS INTELIGENTES DE PRE PROCESAMIENTO PARA MEJORAR EL DESEMPEÑO DE ALGORITMOS DE CLASIFICACIÓN, clave 3790/2014/CID

Nombre y firma del responsable:
Asdrúbal López Chau

Nombre y firma de los participantes:
Jorge Bautista López
Carlos Alberto Rojas Hernández

Fuente de Financiamiento UAEM

Anexos: (indicar número)

Libros (0) Artículos (3) Tesis (2) Ponencias (2) Otros (3)

Vo. Bo.
Líder del Cuerpo Académico
Mtro. Valentín Trujillo Mora

ESTADOS UNIDOS MEXICANOS
UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO
CENTRO UNIVERSITARIO UAEM ZUMPANGO
DIRECCION

Vo. Bo:
Coordinador de Investigación
Mtro. Lucio Navarro Sánchez

UAEM | Universidad Autónoma del Estado de México

ACTA DE LA SESIÓN ORDINARIA DEL H. CONSEJO GOBIERNO
CORRESPONDIENTE AL 29 DE ENERO DE 2016
Lugar: **SALA-GALERÍA "HORACIO ZÚÑIGA ANAYA" EDIFICIO "B"**
Hora de inicio: **14:30** Hora de término: **16:10**
Integrantes al inicio: 09 Integrantes al término: 10

1. **Lista de asistencia.**

2. **Aprobación del orden del día.**

Acuerdo: Aprobado por Unanimidad.

3. **Aprobación del acta de la sesión ordinaria del mes de diciembre 2015.**

Acuerdo: Aprobado por Unanimidad.

4. **Prórroga de Pasantía.**

Se solicita otorgar prórroga de pasantía a favor de:

NO.	NO. CUENTA	NOMBRE	PE	Tiempo solicitado
1	0625239	Saúl Báez Triunfante	LCPyAP	09 Meses

Acuerdo: Aprobado por unanimidad de votos, por la prórroga solicitada por el tiempo solicitado.

5. **Evaluaciones Profesionales.**

Se da a conocer la relación de aspirantes a presentar evaluación profesional del 18 al 26 de febrero del año en curso, solicitando al pleno, otorgar Visto Bueno a fin de poder llevar a cabo dichas actividades.

	NOMBRE	LIC	MODALIDAD	FECHA	HORA	GENERACIÓN
1	BECERRIL TESILLO SALVADOR REY	LTU	TESINA	18/02/2016	10:00	2009-2013
2	BARREDA COVARRUBIAS FERNANDO ESTEBAN	LAM	APROV	19/02/2016	12:00	2011-2015
3	CHOREÑO ESCALONA EDGAR	LAM	APROV	19/02/2016	12:00	2011-2015
4	GARCÍA LÓPEZ ILTZE IVETTE	LAM	APROV	19/02/2016	12:00	2011-2015
5	GUZMÁN OROPEZA AURORA	LAM	APROV	19/02/2016	12:00	2011-2015
6	HERNÁNDEZ GÓMEZ JANETH JOCELIN	LAM	APROV	19/02/2016	12:00	2011-2015
7	MENA GARCÍA MERARI SARAI	LAM	APROV	19/02/2016	12:00	2011-2015
8	MENDOZA NICOLÁS LIZBETH	LAM	APROV	19/02/2016	12:00	2011-2015

UA del CU UAEM Zumpango PE LDE <origen>	Calif.	Clave UAEM	UA del CU UAEM Zumpango PE LAM <destino>	Calif.	Crd.
Inglés C1	8.6	L00062	Inglés C1	8.6	6
Inglés C2	8.5	L00070	Inglés C2	8.5	6
Introducción al Estudio del Derecho	8.5	AC3003	Fundamentos de Derecho	8.5	10
Lectura y Redacción	8.5	L00725	Redacción y comunicación	8.5	8
Derecho Laboral	8.0	L30096	Derecho Laboral	8.0	10
Actos, contratos y sociedades Mercantiles	8.0	AC3006	Derecho Mercantil	8.0	10
Derecho Fiscal	9.8	L30097	Teoría general de la Tributación	9.8	7

Acuerdo: Se solicitará a comité técnico su opinión.

d) A través del oficio del Mtro. Francisco Platas López se pide la aprobación para la finalización del proyecto "Propuesta Metodológica en apoyo a grupos vulnerables de Guatemala y México como medida de prevención ante sismos", lo anterior por requerimientos administrativos de la Secretaría de Investigación.

Acuerdo: Pendiente hasta que presente oficio de liberación por la fuente financiadora donde indique que el proyecto ha concluido satisfactoriamente.

e) Lectura del oficio del Dr. Asdrúbal López Chau donde se pide el aval para el reporte del proyecto de investigación "Métodos inteligentes de pre-procesamiento para mejorar el desempeño de algoritmos de clasificación", con clave 3790/2014/CID.

Acuerdo: Aprobado unanimidad.

f) Se da lectura del oficio de Mtra. Ma. Lourdes Vargas Santillán, en la que solicita el aval de los HH. Consejos para su informe de actividades correspondiente al ciclo escolar agosto 2015 B de los estudios en Doctorado en Ciencias de la Salud, además solicita prórroga de seis meses con goce de sueldo para dar seguimiento al trabajo de tesis y publicaciones.

Acuerdo: Aprobado por unanimidad de votos su informe 2015B y licencia con goce de sueldo para el periodo 2016A.

Universidad Autónoma del Estado de México

CU UAEM ZUMPANGO

PE DE INGENIERO EN COMPUTACIÓN

Reporte Técnico del proyecto UAEM 3790/2014/CID

"MÉTODOS INTELIGENTES DE PRE PROCESAMIENTO PARA MEJORAR EL DESEMPEÑO DE ALGORITMOS DE CLASIFICACIÓN"

Dr. en C. en Computación Asdrúbal López Chau
PTC CU UAEM Zumpango, alchau@uaemex.mx
Responsable técnico

M.T.I. Joge Bautista López
TATC CU UAEM Zumpango, jbautistal@uaemex.mx
Colaborador

M.T.I. Carlos Alberto Rojas Hernández
TATC CU UAEM Zumpango, carojash@uaemex.mx
Colaborador

Cuerpo Académico: Tecnologías Computacionales Aplicadas

Línea de generación y aplicación del conocimiento: Sistemas Software y Hardware**.**

Enero 2016

CONTENIDO

1 Introducción

Una de las tareas más importantes en inteligencia artificial es la clasificación (**Han2005**). Las aplicaciones del mundo real en las que se usa la clasificación son diversas y numerosas, por ejemplo: la identificación de usuarios mediante el reconocimiento de su rostro o huella digital, diagnóstico de fallas en máquinas eléctricas, predicción del estado del tiempo e identificación de correo no deseado (**Witten2005)**.

En la tarea de clasificación, una computadora “aprende” (genera un modelo) a partir de ejemplos que se le presentan en forma de vectores (cada vector representa características de objetos). El objetivo del modelo es ser usado para predecir la categoría o clase de objetos que no han sido previamente presentados a la computadora, es decir, el modelo debe ser capaz de generalizar el conocimiento.

Existen actualmente varios métodos de clasificación, entre los que destacan las redes neuronales, las máquinas de soporte vectorial, los árboles de decisión y las técnicas Bayesianas. Todos estos métodos, y otros no mencionados, son aplicados exitosamente en muchas aplicaciones del mundo real. Sin embargo, pese a los grandes avances logrados, todavía existen varios retos en los cuales la comunidad científica continúa trabajando. Algunos de los retos en el área de clasificación son los siguientes: adaptación de métodos para datos a gran escala (Big Data) y para flujos de datos el ultra alta velocidad, aplicaciones en dominios específicos (datos no balanceados, datos no completamente etiquetados, datos de

dimensionalidad variable) y creación de esquemas novedosos en los que se incorpore conocimiento previo en los modelos.

En este trabajo se muestran algunos de los resultados obtenidos en el proyecto UAEM con clave 3790/2014/CID, denominado "**Métodos inteligentes de pre-procesamiento para mejorar el desempeño de algoritmos de clasificación**", en el que se diseñó un método para árbol de decisión C4.5 y un sistema para generación de datos de usuarios reales, para realizar pruebas con métodos de clasificación.

El resto del documento está conformado como se indica a continuación. En la subsección 1.1 se presentan los antecedentes del problema de clasificación con conjuntos de datos no balanceados. La definición del problema es expuesta en la subsección 1.2, seguida de la justificación y objetivos en las subsecciones 1.3 y 1.4, respectivamente. En la sección 2, se presenta el marco de referencia teórico metodológico, en el cual se describe el método propuesto y el sistema desarrollado. En esa misma sección se presentan los resultados obtenidos. El trabajo finaliza con las conclusiones generales y con las referencias utilizadas.

1.1 Antecedentes de la investigación

El efecto de desbalance en los conjuntos de datos sobre el desempeño de los métodos de clasificación ha sido ampliamente estudiado durante los últimos años. Uno de los primeros trabajos sobre ello fue el descrito en (**Japkowicz2002**), en el cual se usaron conjuntos de datos generados sintéticamente para realizar una exploración sobre la relación que existe entre la complejidad de los conceptos

subyacentes en los datos, el tamaño de los conjuntos de datos de entrenamiento y el nivel de desbalance entre las clases. Los métodos usados en (**Japkowicz2002**) fueron arboles de decisión, redes neuronales y máquinas de soporte vectorial. En (**Drummond2003**) y (**Luengo2009**), se llevaron a cabo algunos experimentos aplicando el árbol de decisión C4.5 con conjuntos de datos no sintéticos. En (**Batista2004**) fue realizado un estudio de los algoritmos especialmente diseñados para la limpieza de los datos, y balance de los conjuntos de datos. Recientemente, en (**Hulse2009**), fueron revisados los efectos del ruido en los datos sobre varios métodos de clasificación. Algunas conclusiones sobre el problema de desbalance en los datos, que se han obtenido son las siguientes: el desbalance entre clases perjudica el desempeño de los métodos de clasificación, pero este no es el factor más importante (**Weiss2004**); el desbalance interno tiene un impacto más negativo que el desbalance entre clases (**Gong2009**,**Japkowicz2002**); el ruido en los datos y los casos raros son perjudiciales para los clasificadores (**Hulse2009**,**Seiffert2008**). Probablemente, el algoritmo más famoso, diseñado específicamente para atender el desbalance de los conjuntos de datos es Synthetic Minority Oversampling Technique (SMOTE) (**Chawla2002**). Este sobre-muestrea la clase minoritaria tomando cada muestra de la clase minoritaria e introduce nuevos objetos a lo largo de los segmentos de línea que unen a la instancia con los K vecinos más cercanos. Un problema con SMOTE es que tiende a sobre-generalizar, y también introduce traslape entre clases, esto es porque las nuevas muestras son generadas sin considerar la distribución de los datos. Una revisión más amplia sobre clasificación con

conjuntos de datos no balanceados puede encontrarse en (**He2009**). Uno de los algoritmos propuestos recientemente, y que utiliza un árbol de decisión es el presentado en (**Boonchuay2011**), en el cual se presenta el concepto de entropía minoritaria. Sin embargo, encontramos que existe un error en la definición de dicho concepto, por lo que hacemos una corrección y proponemos un método basado para mejorar el desempeño del algoritmo de clasificación árbol de decisión en conjuntos de datos no balanceados.

1.2 Definición del problema

Los conjuntos de datos con dos clases, son representados como se muestra en la ecuación (1):

$$X = \left\{(x_i, y_i) |\ x_i \in R^d, y_i \in \{+1, -1\}\right\}_{i=1}^{N} \text{----- (1)}$$

Donde y_i es la clase del objeto x_i, d es el número de atributos de cada vector y N es el número de vectores en el conjunto de datos.

Un conjunto de datos se dice que está desbalanceado si la relación (llamada tasa de desbalance) entre objetos de la clase con menos elementos, es notablemente menor a los de clase con más elementos. En la literatura, a los objetos que son minoría en un conjunto de datos se les asigna la clase y= +1 (clase minoritaria), mientras que a los otros la clase y=-1 (clase mayoritaria).

Los métodos de clasificación presentan problemas para predecir con precisión nuevos objetos de clase minoritaria **(Orriols2005, Gu2008, Koknar-Tezel2009,**

Huang2014). Es por esto que en los últimos años se han propuesto varios métodos para poder enfrentar dicha dificultad.

En este proyecto, el problema a resolver es mejorar el desempeño de algoritmos de clasificación para conjuntos de datos no balanceados.

1.3 Justificación

La clasificación en conjuntos de datos no balanceados es un problema actual e importante (**Khan2012**). Ejemplos de aplicaciones del mundo real en las que se generan este tipo de datos son, por ejemplo, diagnósticos médicos de enfermedades tales como cáncer o diabetes mellitus, detección de errores en código fuente, ataques informáticos en servidores e identificación de transacciones bancarias fraudulentas. En este tipo de escenarios, es sumamente importante detectar casos que generalmente ocurren con poca frecuencia con respecto a la cantidad de datos obtenidos, ya que representan un riesgo que hay que observar oportunamente. Para los ejemplos de aplicaciones mencionadas anteriormente, estos casos poco frecuentes serían los pacientes enfermos de cáncer o diabetes, los fragmentos de código fuente con errores, los accesos inválidos a servidores y las transacciones ilícitas. Sin embargo, como se mencionó anteriormente, los algoritmos de clasificación, presentan problemas con conjuntos de datos no balanceados **(He2009)**.

Es debido a esta importancia, que recientemente la comunidad científica del área de inteligencia artificial se ha enfocado en atacar este problema, generando diversos tipos de algoritmos, entre ellos, los basados en la selección de datos (pre

procesamiento). En este proyecto, se diseñaron nuevos métodos de pre-procesamiento y se realizaron pruebas en conjuntos de datos disponibles públicamente en Internet, con el objetivo de mostrar su eficiencia.

1.4 Objetivos

El objetivo general de este trabajo es diseñar nuevos métodos de pre-procesamiento para mejorar el desempeño de algoritmos de clasificación cuando estos son aplicados a conjuntos de datos no balanceados. Para el logro de este objetivo, se presentan los siguientes objetivos específicos:

1. Realizar una investigación documental sobre el estado del arte de métodos para el pre-procesamiento de datos para conjuntos de datos no balanceados.
2. Diseñar nuevos métodos inteligentes orientados a mejorar el desempeño de algoritmos de clasificación en conjuntos de datos no balanceados.
3. Analizar los códigos fuentes de algoritmos del estado del arte similares a los métodos propuestos, con el objetivo de realizar comparaciones de desempeño.
4. Realizar pruebas y comparar desempeño del algoritmo propuesto con otros del estado del arte.
5. Documentar resultados.

2 Marco de Referencia Teórico-Metodológico

A continuación, se presentan dos métodos para mejorar el desempeño de algoritmos de clasificación con conjuntos de datos no balanceados, y una aplicación para generar datos de este tipo.

El método mostrado está basado en un árbol de decisión C4.5, en el que se modifica la medida de impureza de las particiones realizadas recursivamente durante la etapa de entrenamiento.

2.1 Entropía minoritaria

El árbol de decisión presentado en (**Boonchuay2011**), realiza tres particiones basándose en el concepto de entropía minoritaria. La idea básica es utilizar dos puntos de separación en el atributo seleccionado, de modo que permitan capturar la mayor cantidad de elementos de clase minoritaria en la partición central. Esta idea resulta atractiva para atacar el problema de clasificación sobre conjuntos de datos no balanceados, ya que se evita que el clasificador ignore a las instancias de clase minoritaria, tal como ocurre comúnmente con los algoritmos tradicionales.

La entropía minoritaria es presentada y definida en (**Boonchuay2011**) como se muestra en la ecuación siguiente:

$$MinorityEntropy(n) = -\sum_{i=1}^{N} \frac{n_i}{n} \left(\log \left(\frac{n_i}{n} \right) \right)$$

Donde N representa la cantidad de particiones (mínimo 1 y máximo 3), n_i representa la cantidad de instancias de clase minoritaria en la i-esima partición y n

representa la cantidad de elementos de clase minoritaria. Es importante observar que la ecuación anterior tiene gran similitud con la fórmula de entropía clásica. Sin embargo, la fórmula de entropía minoritaria propuesta en (**Boonchuay2011**), tiene el inconveniente de no tener como límite superior el valor 1.0, a diferencia a lo que ocurre con otras medidas de impureza no convexas. Lo que produce que las particiones del árbol de decisión puedan no ser apropiadas y por lo tanto generar modelos inexactos.

2.2 Corrección a la fórmula de entropía minoritaria y método propuesto

Descrito de manera breve, se propone seguir manteniendo a las tres particiones para la construcción del árbol de decisión, sin embargo, para el cálculo de la entropía minoritaria se propone usar únicamente dos particiones, la idea es considerar a las particiones extremas como si se tratara de una sola. El surgimiento de esta propuesta, está basado en observaciones que realizamos usando algunos conjuntos de datos sintéticos. Al analizar las condiciones para las cuales los valores de la entropía minoritaria pertenecían al intervalo cerrado [0,1], observamos que esto sólo ocurre cuando las instancias de clase minoritaria están presentes en dos de las tres particiones. En otros casos, el resultado obtenido puede estar fuera del rango indicado. La forma corregida para el cálculo de entropía minoritaria que proponemos es la mostrada en la ecuación siguiente:

$$MinorityEntropy(n) = -\left(\frac{n_1}{n}\log\left(\frac{n_1}{n}\right)\right) - \left(\frac{n_2}{n}\log\left(\frac{n_2}{n}\right)\right)$$

Donde n_1 representa la cantidad de instancias de clase minoritaria en la partición central, n_2 representa la suma de la cantidad de instancias de clase minoritaria que se encuentran en las particiones izquierda y derecha, es decir, la suma de ambas, y n representa la cantidad de elementos de clase minoritaria antes de realizar particiones.

Los árboles de decisión basados en la fórmula de entropía minoritaria corregida, son inducidos usando el algoritmo mostrado a continuación:

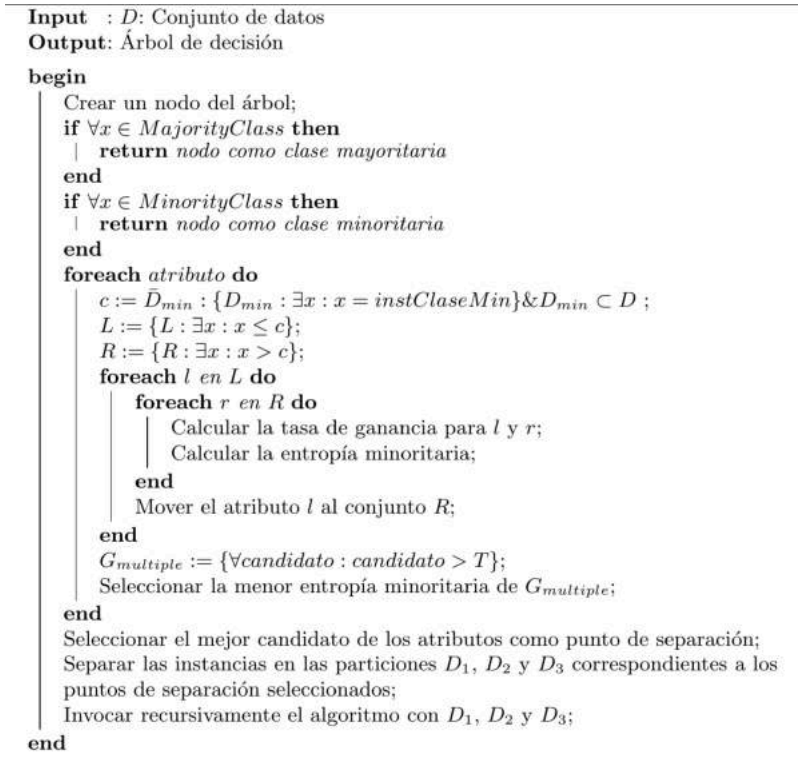

Figura 1: Algoritmo de inducción para árboles de decisión basados en entropía minoritaria.

2.3 Sistema para la generación de datos no balanceados

Típicamente, los investigadores utilizan conjuntos de datos disponibles públicamente en la Internet para probar y comparar métodos de clasificación. En la última etapa de este trabajo, se diseñó e implementó un sistema para realizar mediciones biométricas de usuarios reales, con el objetivo de generar conjuntos de datos. En específico, se usa la dinámica de tecleo en dispositivos móviles.

La dinámica del tecleo (*keystroke dynamics*) (**Yu2004**), que consiste en usar la forma o el ritmo en que un usuario escribe en un teclado para autenticar su identidad, es una de las principales técnicas de la biometría informática dinámica. Numerosos trabajos y proyectos han sido publicados recientemente (**Gaines1980**), (**Yu2004**), ya sea proponiendo algoritmos (**Obaidat1993**), implementando o aplicando la dinámica del tecleo en diferentes dispositivos. Generalmente, los estudios con dinámica de tecleo se han centrado en teclados tipo QUERTY (**Kang2015**). Sin embargo, son pocos los trabajos que se enfocan en teclados tipo numérico. En la actualidad, son varias las aplicaciones en las que se utiliza este tipo de teclado, por ejemplo, cajeros automáticos, cajas fuertes, teléfonos y para captura en hojas de cálculo.

Actualmente, es común para la mayoría de las personas contar con una computadora, teléfono inteligente o tableta electrónica. Una de las principales actividades que se realizan en estos dispositivos, es la escritura por medio del teclado tipo QUERTY, que es el predominante en este tipo de sistemas. Otro tipo de teclado que es empleado en algunas aplicaciones, es el numérico, cuya

apariencia es presentada en la Figura 2. Ejemplo de estas aplicaciones son calculadoras simples, pantallas para ingreso al sistema y pantallas para captura de números, como en la banca electrónica.

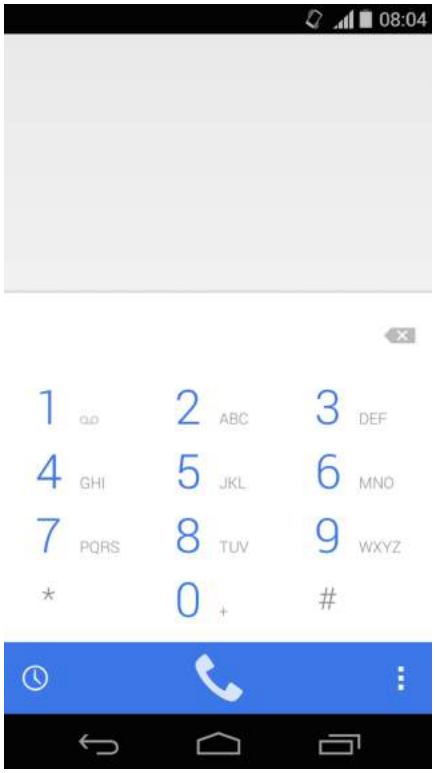

Figura 2: Teclado numérico.

Durante la interacción de un usuario con un teclado tipo numérico, se pueden medir dos tiempos:

- Tiempo de presión (*dwell-time*): Es el intervalo que transcurre cuando el usuario mantiene presionada y libera la misma tecla.
- Tiempo de cambio (*flight-time*): Se considera el tiempo que transcurre cuando el usuario suelta una tecla y presiona la tecla siguiente.

De acuerdo a (**Urtiga2011**) ambos tiempos son suficientes para distinguir un usuario de otro.

La interfaz gráfica de usuario (GUI) que se diseñó, es similar a las de acceso a sistemas informáticos, y se muestra en la Figura 3.

Figura **3:** Interfáz gráfica para medición de tiempos de presión y de cambio.

Esta GUI tiene controles para que los usuarios ingresen un nombre de usuario y una contraseña. Para preservar el anonimato de los usuarios y al mismo tiempo para distinguir entre los datos generados por cada uno, se usaron números como identificadores para los usuarios. De esta forma, el primer usuario tiene como identificador el valor 1, el segundo el 2, y así sucesivamente. Con la finalidad de que los datos generados por cada usuario puedan ser comparables con los otros, se decidió que el nombre del usuario y la contraseña fueran iguales en todos los

casos. El nombre de usuario es 94255701, mientras que la contraseña es 416850293.

El diagrama de clases del sistema implementado es mostrado en la Figura 4. La clase *Gestor de Usuarios* controla el acceso y consultas a la base de datos local. Cada usuario es identificado con un número entero. La clase *Usuario* mantiene los datos de cada usuario, como el identificador (id) y contraseña. Una de las clases más importantes es la denominada *EscucharTeclado*. Esta clase contiene los métodos para calcular los tiempos de presión y de cambio. La API (*Application Programming Interface*) de Android ofrece una serie de clases para facilitar la captura de eventos del teclado. Una combinación de los eventos *onKeyDown*() y *onKeyUp*() fue usada para capturar los tiempos requeridos.

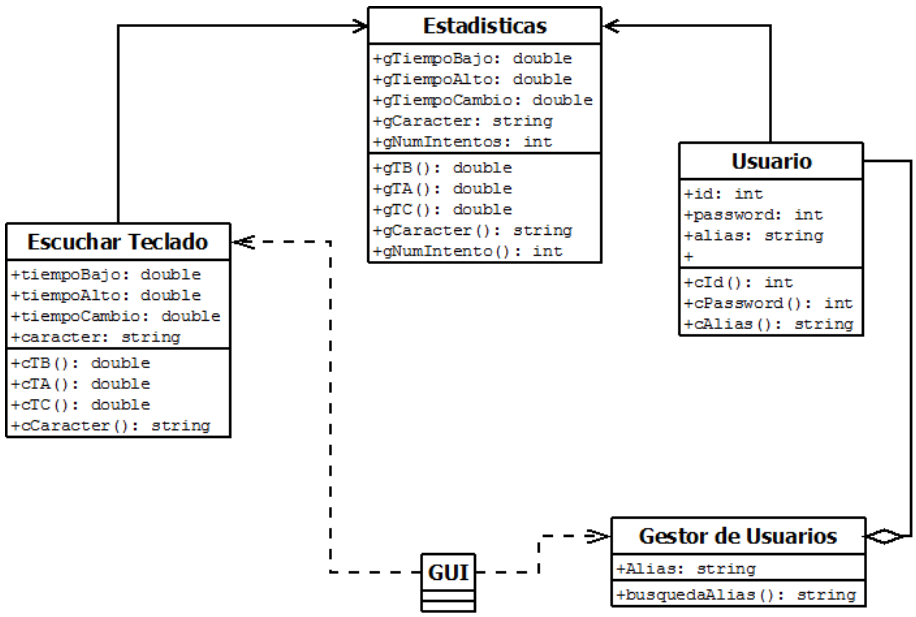

Figura 4: Arquitectura del sistema desarrollado

2.4 Resultados obtenidos

El dispositivo en el que se probó el sistema, fue un teléfono inteligente modelo XT1058, con procesador ARMv7 (v71), pantalla de 4.7 pulgadas y resolución de 720 x 1184 p. El sistema operativo es Android 4.4.4 . Se solicitó a 14 personas, que introdujeran el nombre de usuario (94255701) y la contraseña (416850293) un total de 10 veces. Para los casos en los que un usuario se equivoca, el sistema elimina ese registro y solicita que se intente nuevamente. Cada intento exitoso fue almacenado en una base de datos local en el teléfono inteligente. Un resumen de los tiempos de presión promedio al introducir es el nombre de usuario y la contraseña son mostrados en las Figuras 5 y 6, respectivamente. Los tiempos de cambio promedio, son presentados en las Figuras 7 y 8 Los datos correspondientes a todos los intentos de los 14 usuarios, han sido publicados en el sitio http://www.alchau.com/research/2015/keystrokeData-1/data.zip, y se encuentran disponibles para su descarga.

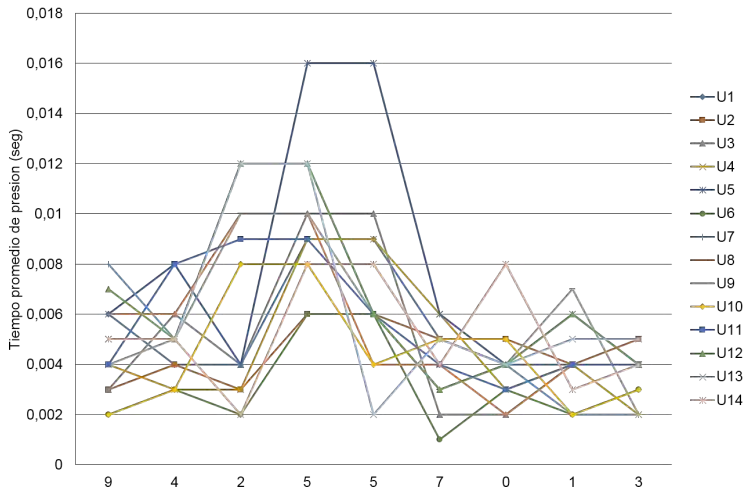

Figura 5:Tiempo de presión promedio para captura de usuario

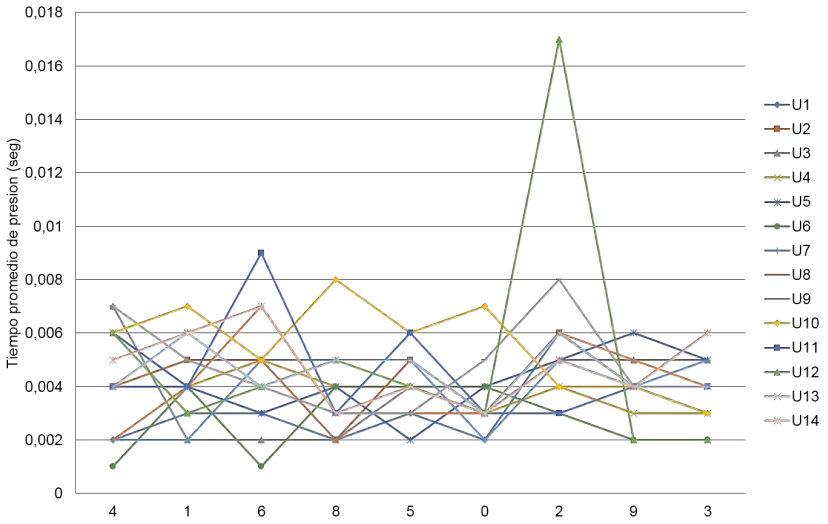

Figura 6: Tiempo de presión promedio para captura de contraseña

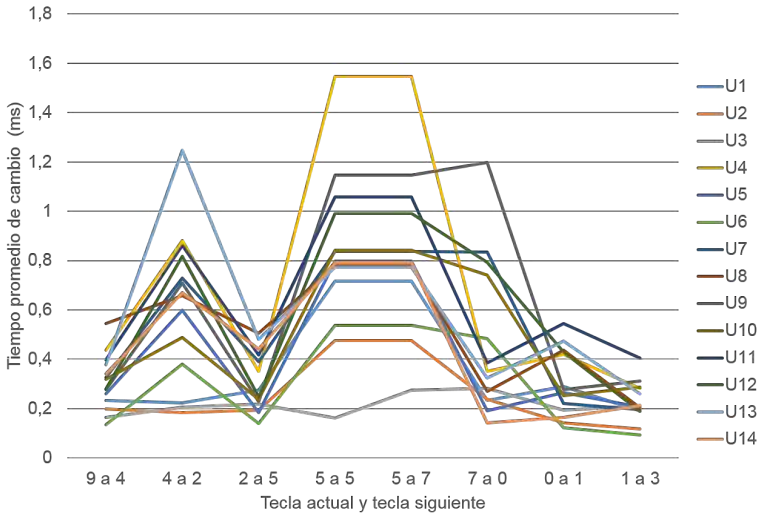

Figura 7: Tiempo de cambio promedio para captura de usuario

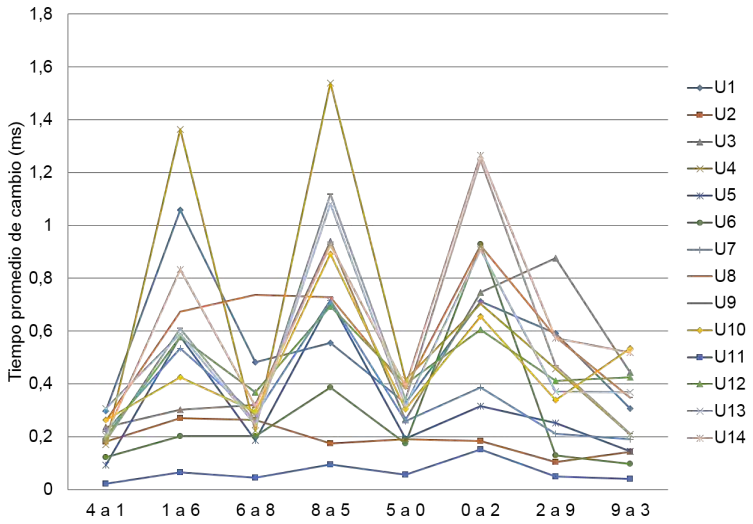

Figura 8: Tiempo de cambio promedio para captura de contraseña

3 Conclusiones

La clasificación es definida como un proceso del aprendizaje automático de las computadoras. Uno de los principales problemas en esta área, es la clasificación en conjuntos de datos no balanceados. Al aplicarse los métodos de clasificación actuales en este tipo de conjuntos de datos, se obtienen precisiones de clasificación bajas para las predicciones de los objetos de clase minoritaria. En este trabajo, se propusieron métodos de pre-procesamiento para atacar el problema mencionado, y se muestra en este documento uno de los métodos propuestos, así como una aplicación desarrollada para la generación de datos no balanceados obtenidos de usuarios reales. El lector interesado en profundizar sobre los métodos desarrollados y obtener los códigos fuentes, o las bases de

datos generadas, refiérase a los artículos publicados en donde encontrará una explicación más detallada.

3.1 Productos generados durante el proyecto 3790/2014CID

Publicaciones (artículos en revistas indizadas)

i. "Árbol de decisión C4.5 basado en entropía minoritaria para clasificación de conjuntos de datos no balanceados", Luis A. Caballero-Cruz, Asdrúbal López-Chau, Jorge Bautista-López, Research in Computing Science ISSN: 1870-4069, Vol. 92, págs 23-34, Mayo 2015. Indizada en Latin index y DBLP. En este artículo se diseñó un método para clasificación en conjuntos de datos no balanceados.

ii. "Captura de datos para análisis de la dinámica del tecleo de números para sistema operativo Android", Selene Nieto-Ruiz, Yonic A. Gómez-S ánchez, Asdrúbal López-Chau, Carlos A. Rojas, Research in Computing Science ISSN: 1870-4069, Vol. 92, págs.. 147-156, Mayo 2015. Indizada en Latin index y DBLP. En este artículo se generaron conjuntos de datos.

iii. "Data selection based on decision tree for SVM classification on large data sets", Jair Cervantes, Farid García Lamont, Asdrúbal López-Chau, Lisbeth Rodríguez Mazahua, J. Sergio Ruíz, Applied Soft Computing, ISSN: 1568-4946, Vol. 37, págs. 787-798, septiembre 2015. Indizada en JCR. . En este artículo se diseñó un método para clasificación en conjuntos de datos grandes.

Publicaciones (Capítulos de libro)

i. "Classification on Imbalanced Data Sets, Taking Advantage of Errors to Improve Performance", Asdrúbal López-Chau, Farid García-Lamont, and Jair Cervantes, Springer International Publishing Switzerland 2015 D.-S. Huang and K. Han (Eds.): ICIC 2015, Part III, LNAI 9227, pp. 72–78, 2015. DOI: 10.1007/978-3-319-22053-6_8

Formación de recursos humanos (titulados del programa educativo de ingeniería en computación del CU UAEM Zumpango)

Dos alumnos obtuvieron el título de ingeniero en computación, en el Centro Universitario UAEM Zumpango.

i. Alumno: Luis Alberto Caballero Cruz.

   Modalidad: Artículo para publicar en revista indizada.

   Título trabajo: "Árbol de decisión C4.5 basado en entropía minoritaria para clasificación de conjuntos de datos no balanceados".

   Asesor: Dr. Asdrúbal López Chau.

   Fecha examen: 20 agosto 2015

ii. Alumna: Selene Nieto Ruíz

   Modalidad: Artículo para publicar en revista indizada.

   Título trabajo: "Captura de datos para análisis de la dinámica del tecleo de números para sistema operativo Android".

   Asesor: Dr. Asdrúbal López Chau.

   Fecha examen: 20 agosto 2015

4 Referencias

(Batista2004) Batista, G. E. A. P. A., Prati, R. C. and Monard, M. C., *A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data,* SIGKDD Explor. Newsl., ACM, **2004**, Vol. 6(1), pp. 20-29.

(Boonchuay2011) Boonchuay, K., Sinapiromsaran, K. and Lursinsap, C., *Minority split and gain ratio for a class imbalance* Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on **2011**, Vol. 3, pp. 2060-2064.

(Chawla2002) Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W. *SMOTE: Synthetic Minority Over-sampling Technique,* Journal of Artificial Intelligence Research, **2002**, Vol. 16, pp. 321-357.

(Drummond2003) Drummond, C. and Holte, R. C. *C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling* **2003**, pp. 1-8.

(Gaines1980) Gaines, R., Lisowski, W., Press, S. and Shapiro, N., *Authentication by keystroke timing: some prelimary results,* Rand Corporation, **1980.**

(Gong2009) Gong, W., Luo, H. and Fan, J. *Extracting Informative Images from Web News Pages via Imbalanced Classification* Proceedings of the 17th ACM International Conference on Multimedia, ACM, **2009**, pp. 1123-1124.

(Gu2008) Gu, Q., Cai, Z., Zhu, L. and Huang, B. *Data Mining on Imbalanced Data Sets* Advanced Computer Theory and Engineering, 2008. ICACTE '08. International Conference on **2008**, pp. 1020-1024.

(Han2005) Han, J. *Data Mining: Concepts and Techniques* Morgan Kaufmann Publishers Inc., **2005.**

(He2009) He, H. and Garcia, E. A., *Learning from Imbalanced Data,* IEEE Trans. on Knowl. and Data Eng., IEEE Educational Activities Department, **2009**, Vol. 21(9), pp. 1263-1284.

(Huang2014) Huang, H., Chiew, K., Gao, Y., He, Q. and Li, Q. *Rare Category Exploration,* Expert Syst. Appl., Pergamon Press, Inc., **2014**, Vol. 41(9), pp. 4197-4210.

(Hulse2009) Hulse, J. V. and Khoshgoftaar, T. *Knowledge discovery from imbalanced and noisy data* Data & Knowledge Engineering , **2009**, Vol. 68(12), pp. 1513 – 1542.

(IanH.Witten2005) Ian H, W. and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition,* Elsevier Science, **2005.**

(Japkowicz2002) Japkowicz, N. and Stephen, S. *The class imbalance problem: A systematic study.* Intelligent Data Analysis, **2002**, Vol. 6(5), pp. 429.

(Kang2015) Pilsung Kang and Sungzoon Cho. 2015. *Keystroke dynamics-based user authentication using long and free text strings from various input devices*. *Inf. Sci.* 308, C (July 2015), 72-93.

(Khan2012) Khan, N. M., Ksantini, R., Ahmad, I. S. and Boufama, B. , *A novel SVM+NDA model for classification with an application to face recognition,* Pattern

Recogn., Elsevier Science Inc., **2012**, Vol. 45(1), pp. 66-79.

(Koknar-Tezel2009) Koknar-Tezel, S. and Latecki, L., *Improving SVM Classification on Imbalanced Data Sets in Distance Spaces,* Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on **2009**, pp. 259-267.

(Luengo2009) Luengo, J., Fernandez, A., Herrera, F. and Herrera, F. *Addressing Data-Complexity for Imbalanced Data-Sets: A Preliminary Study on the Use of Preprocessing for C4.5* Intelligent Systems Design and Applications, 2009. ISDA '09. Ninth International Conference on **2009**, pp. 523-528.

(Obaidat1993) Obaidat, M. S. and Macchiarolo, D. T., *An On-Line Neural Network System for Computer Access Security,* IEEE Transactions on Industrial Electronics, **1993.**

(Orriols2005) Orriols, A. and Bernadó-Mansilla, E. *The Class Imbalance Problem in Learning Classifier Systems: A Preliminary Study,* Proceedings of the 2005 Workshops on Genetic and Evolutionary Computation, ACM, **2005**, pp. 74-78.

(Seiffert2008) Seiffert, C., Khoshgoftaar, T. and Van Hulse, J. *Hybrid sampling for imbalanced data* Information Reuse and Integration, 2008. IRI 2008. IEEE International Conference on, **2008**, pp. 202-207.

(Urtiga2011) Urtiga, E. and Moreno, E., *Keystroke - Based Biometric Authentication in Mobile Devices,* IEEE Latin America Transactions, **2011.**

(Weiss2004) Weiss, G. M. *Mining with Rarity: A Unifying Framework* SIGKDD

Explor. Newsl., ACM, **2004**, Vol. 6(1), pp. 7-19.

(Yu2004) Yu, E. and Cho, S., *Keystroke dynamics identity verification - its problems and practical solutions,* ELSEVIER, **2004.**