Determining Electoral Preferences in Mexican Voters by Computational Intelligence Algorithms

S. Ortiz-Ángeles, Y. Villuendas-Rey, C. Yáñez-Márquez, I. López-Yáñez, and O. Camacho-Nieto

***Abstract*—In the context of political activities, electoral processes are of interest for scientists, who usually tackle their research on this field from a social sciences perspective. Computational methods have been applied to predict the electoral preferences of voters in several countries; however, this has not happened in Mexico, at least as indicated by the absence in current scientific literature of computational studies to determine voting intentions of Mexican citizens. The authors of the present work aim at reverting such absence. The proposal of this paper consists of applying Computational Intelligence methods to automatically determine electoral preferences of Mexican voters. For this, data acquired by the Secretaría de Gobernación (Secretary of the Interior), about voting intentions of Mexican citizens in the 2012 elections are used. In the voter classification stage, a modified version of the Gamma Associative Classifier (MGAC) is used, given that this is one of the relevant models of the Associative approach to Pattern Classification. Additionally, Differential Evolution is employed to guide the process of relevant features selection. Results indicate that, when compared over six data sets extracted from the information published by the Secretaría de Gobernación, our proposal exhibits the best performance in three of these data sets, outperforming some of the best similar models present in the state of the art.**

***Index Terms*—Computational intelligence, Classification algorithms, Evolutionary computation, Electoral preferences.**

I. Introducción

En los círculos políticos del mundo, es innegable la importancia que exhibe la predicción certera de las preferencias electorales de los votantes. Recientemente, se han realizado trabajos científicos para su determinación, principalmente en EEUU y en Europa [1], [2], [3]. En México, el tema electoral ha propiciado investigaciones desde el siglo pasado, destacando la de McCann y Domínguez [4], quienes estudiaron el comportamiento de los votantes mexicanos, buscando una evaluación de la opinión pública y el comportamiento electoral; el trabajo se basó en datos provenientes de estudios de opinión pública realizados durante el período de 1986 a 1995. También se destacan trabajos que analizan el comportamiento de los votantes mexicanos al escoger un presidente no perteneciente al Partido Revolucionario institucional (PRI) [5] y la visión de dichos votantes con respecto a la democracia [6], así como la influencia de las élites políticas [7] y los medios de comunicación en el proceso electoral en México [8], [9].

En la literatura especializada existe una gran cantidad de publicaciones donde se reporta que los métodos computacionales han sido aplicados con éxito en diversas esferas, tales como el análisis de riesgos [10], la seguridad [11], los deportes [12], la eficiencia energética [13], [14], los negocios [15], [16], el estudio de redes sociales [17], entre otros. Y en el tema que nos ocupa en el presente artículo, el fenómeno electoral, se han publicado diversos trabajos que hacen uso de estos métodos [18], [19], [20], [21], [22].

Sin embargo, es notable que, desde el punto de vista computacional, en México se observa un vacío, dado que en el estado del arte no aparecen estudios computacionales para determinar las intenciones de voto en ciudadanos mexicanos, a diferencia de lo que ocurre en otros países de Latinoamérica, como Perú y Brasil [23]. Es preciso revertir esta adversa situación.

La motivación expresada en el párrafo previo, justifica plenamente nuestra propuesta, donde se aplica de manera exitosa el cómputo inteligente en la determinación de las intenciones de voto de los votantes mexicanos. Así, se logra que México cuente con estudios similares a los realizados en países como Italia, Reino Unido, Holanda, Alemania, Estados Unidos, Brasil, Perú, Escocia y Chipre, entre otros.

La importancia de esto se exacerba, al considerar que las elecciones de 2012 ha sido polémicas y controversiales, lo cual motivó la ocurrencia de singulares acontecimientos que, a la postre, han sido de gran relevancia para el desarrollo de la democracia en nuestro país.

La propuesta del presente artículo consiste en aplicar modelos de cómputo inteligente para determinar, de forma automática, las preferencias electorales de los votantes

This work was supported in part by the Instituto Politécnico Nacional, the CONACyT and the Sistema Nacional de Investigadores (SNI), México, under Grant 26395.

Sonia Ortiz-Ángeles, is with Centro de Innovación y Desarrollo Tecnológico en Cómputo del Instituto Politécnico Nacional, México, D. F., México (e-mail: lizde00@yahoo.com).

Yenny Villuendas-Rey, is with Centro de Innovación y Desarrollo Tecnológico en Cómputo del Instituto Politécnico Nacional, México, D. F., México (e-mail: yvilluendasr@ipn.mx).

Cornelio Yáñez-Márquez, is with Centro de Investigación en Computación del Instituto Politécnico Nacional, México, D. F., México (e-mail: coryanez@gmail.com).

Itzamá López-Yáñez, is with Centro de Innovación y Desarrollo Tecnológico en Cómputo del Instituto Politécnico Nacional, México, D. F., México (e-mail: ilopezy@ipn.mx).

Oscar Camacho-Nieto, is with Centro de Innovación y Desarrollo Tecnológico en Cómputo del Instituto Politécnico Nacional, México, D. F., México (e-mail: ocamacho@ipn.mx).

mexicanos. Para ello, y a diferencia de propuestas existentes en la literatura [1], [2], [19], [20], [21], [22], para realizar la predicción de un votante se aplica una versión modificada de uno de los modelos importantes de enfoque asociativo de clasificación de patrones; al respecto, una de las aportaciones del presente artículo al estado del arte es la modificación que se ha realizado al Clasificador Asociativo Gamma original [11], [12], [13], a fin de que pueda manejar los datos recolectados. Una aportación adicional consiste en utilizar Evolución Diferencial para guiar un proceso de selección de rasgos relevantes [26].

Por otra parte, con respecto a trabajos existentes en la literatura científica [1], [2], [3], [18], [19], [20], [21], [22], [24], [25], la propuesta presentada en el presente artículo tiene como una de sus contribuciones relevantes que utiliza, por primera vez, información real de ciudadanos mexicanos. Además, esta propuesta tiene como objetivo predecir la intención de voto de un ciudadano en particular, no de múltiples ciudadanos a la vez, como es el caso de los trabajos publicados en las referencias [2], [19] y [21]. Por otra parte, se utilizan los datos recopilados por la Secretaría de Gobernación de México, acerca de las intenciones de voto de ciudadanos mexicanos en las elecciones del año 2012. Estos son datos provenientes de encuestas realizadas a ciudadanos mayores de edad, de diversos estados de la República, y de diferentes ámbitos socio-económicos.

Los resultados indican que nuestra propuesta exhibe una eficacia superior a la obtenida por modelos similares presentes en el estado del arte.

El resto del artículo está organizado como sigue: en la sección II se explican los bancos de datos utilizados, mientras que la sección III incluye las propuestas de esta investigación. La sección IV aborda el protocolo experimental, y la sección V contiene los resultados obtenidos y su discusión. El artículo finaliza con las conclusiones y el trabajo a futuro.

II. Bancos de Datos Utilizados

Esta sección consta de cinco subsecciones. En la subsección A se incluye la descripción de los datos crudos de las intenciones de voto de ciudadanos mexicanos en las elecciones del año 2012, tal como se recibieron de la Secretaría de Gobernación de México. En el resto de las subsecciones se describen, además, las etapas de preprocesamiento que fue preciso realizar, a fin de lograr la construcción de los conjuntos de datos en los que se aplican las dos propuestas originales del presente artículo.

A. Datos Crudos

Los datos de la presente investigación provienen de un estudio tipo panel, que realizó la Secretaría de Gobernación (SEGOB) de los Estados Unidos Mexicanos. Mediante una encuesta, se recolectaron datos acerca de las intenciones de voto de 1587 ciudadanos mexicanos procedentes de 29 estados, durante las elecciones del año 2012.

El estudio consta de 53 preguntas (además de 10 elementos generales y 30 consideraciones finales), que abordan aspectos demográficos, socioeconómicos, religiosos, de influencia mediática, y otros. Cabe señalar que varias de las preguntas tenían incisos para ofrecer respuestas detalladas, y cada registro (respuesta del cuestionario) consta de 204 rasgos.

El cuestionario de la encuesta aplicada, así como las respuestas de los encuestados, se encontraban disponibles en http://www.encup.gob.mx/es/Encup/Estudio_Panel_2012. Sin embargo, en ocasiones dicha página no se encuentra disponible. El cuestionario que respondieron los encuestados puede ser consultado en un enlace provisto por los autores[1].

Los datos se presentan crudos, pues no han sido preprocesados para la extracción de conocimiento. Es por ello que en esta investigación nos dimos a la tarea, primeramente, de realizar un preprocesamiento, el cual se explica el a continuación.

B. Eliminación de Algunos Rasgos y Registros

Se eliminaron varios de los rasgos, por considerarse redundantes o sin valor predictivo. De los elementos generales los rasgos eliminados fueron: folio del cuestionario, fecha de realización de la encuesta (día y mes), número de manzana electoral, número de encuestador, duración en minutos, hora de inicio y hora de fin de la encuesta.

De las preguntas de la encuesta, las eliminadas fueron: fecha de nacimiento (pregunta 3, dado que previamente se había registrado la edad de los encuestados en la pregunta 2), razón de voto (pregunta 9, de expresión libre en la encuesta, representada en texto plano en los datos recolectados), programa de noticias (pregunta 32) y actividades (pregunta 35, que difiere en los encuestados, puesto que la pregunta a realizar depende del folio del cuestionario).

De las consideraciones finales, se eliminaron las siguientes preguntas: nombre del encuestado, fecha de nacimiento, número telefónico, hora de terminación y duración de la entrevista, además del número de interrupciones; si estuvo o no presente otro adulto, y si estuvo o no presente un supervisor durante la entrevista. Se borraron las observaciones del encuestador, así como su nombre, código y sexo. En cuanto a los registros, inicialmente se eliminaron 337, cuyos datos estaban vacíos (estos ciudadanos no respondieron el cuestionario). Se eliminaron también seis registros que no contestaron si votaron o no en las elecciones de 2012.

C. Integración de Rasgos

Considerando la importancia de la pregunta 4 del cuestionario, ("¿Cuál diría usted que es el problema más importante que enfrenta el país hoy en día?") en la determinación de las intenciones de voto, y las respuestas ofrecidas por los encuestados, se integraron los textos que representan las respuestas, obteniéndose un conjunto de textos representativos integrados, que se muestran en la Tabla I.

Además, se integraron en un solo rasgo las respuestas a las preguntas 21 y 22 del cuestionario, puesto que la pregunta 22 solo se realiza a los ciudadanos que contestaron "Ninguno" en la pregunta 21; y se integraron las respuestas de la pregunta 37a ("¿La televisión dio un trato parejo a todos los candidatos (…) o favoreció a algunos?") y las respuestas de la pregunta

[1] https://www.dropbox.com/s/4y9lkuuhgrm6l3e/cuestionario.pdf?dl=0

37b (“¿A cuál candidato favoreció?”), en un solo rasgo.

TABLA I
INTEGRACIÓN DE RESPUESTAS A LA PREGUNTA 4

Texto integrado	Ejemplos de textos que integra
Corrupción	“Mucha corrupción”, “Fraude”, “El engaño”
Criminalidad y consecuencias	“Delincuencia”, “Violencia”, “Inseguridad”, “Asaltos en el día”, “Vandalismo”, “El narcotráfico”, “Secuestros”
Crisis política	“El conflicto entre los políticos”, “Falta de democracia”
Drogadicción	“Drogas”, “Drogadictos”, “Las drogas y el alcoholismo”
Ecología	“Cambios climáticos”, “La contaminación”
Economía	“Economía”, “Económica”, “Cada día suben más las cosas”, “Alza de los precios de la canasta básica”
Educación	“Falta de educación”, “La educación”
El gobierno	“El gobierno no cumple”, “Problemas en el gobierno”
Empleo	“No hay trabajo”, “Desempleo”, “No hay chamba”
Salud	“Salud”, “Seguros”, “La salud”
Pobreza	“Carencia”, “Mucha hambre”, “El dinero”, “La carestía”
Ninguno	Ninguno

D. Manejo de Valores en los Rasgos

El cuestionario realizado por la SEGOB incluye un conjunto de rasgos de naturaleza numérica, como por ejemplo la edad de los encuestados, mientras que otros son de naturaleza categórica, como el sexo de los encuestados. En todos los casos, se respetó la naturaleza de los rasgos, y se decidió tratarlos en consecuencia, sin discretizar los rasgos numéricos ni codificar los rasgos categóricos; esta decisión es congruente con la exitosa tendencia actual en los trabajos científicos relacionados con fenómenos sociales [27].

Por otra parte, en varias ocasiones los encuestados no contestaron, o manifestaron “no saber” sobre una determinada pregunta. Las respuestas de “No sabe” y “No contestó” se consideraron como de valor desconocido (?) para los efectos de la determinación de las intenciones de voto. De esta forma, los datos a manejar en esta investigación se consideran mezclados e incompletos, puesto que poseen simultáneamente rasgos numéricos y categóricos, así como ausencias de información en algunos rasgos.

E. Construcción de los Conjuntos de Datos

Para la determinación de las intenciones de voto de los ciudadanos mexicanos, se consideraron las preguntas 8, 10, 11 y 12, que indagan acerca de la intención de voto para los candidatos presidenciales, diputados federales, senadores y gobernadores, respectivamente. En cuanto a las intenciones de voto para presidente, se eliminaron del análisis 190 registros, que mencionan haber anulado la boleta electoral de una forma u otra, o no recuerdan haber votado. En el cuestionario, se consideraron cuatro candidatos: Josefina Vázquez Mota, Enrique Peña Nieto, Andrés Manuel López Obrador y Gabriel Quadri. El caso de Gabriel Quadri merece algunos comentarios adicionales; este candidato tuvo un total de 17 intenciones de voto, pero 9 de estos ciudadanos manifestaron haber condicionado su voto ante algún tipo de beneficios (respuestas a las preguntas 39-41), quedando así solamente 8 intenciones de voto sin condicionar a favor de Quadri. Dado que estas intenciones de voto representan sólo el 0.5% del total de encuestados, se decidió eliminar esta información.

Para la determinación de las intenciones de voto de diputados federales, se eliminaron 201 registros de ciudadanos que mencionaron haber anulado la boleta electoral de una forma u otra, o no recuerdan haber votado. Bajo el mismo análisis, en el estudio de las intenciones de voto para senadores, se eliminaron 203 registros correspondientes a boletas anuladas. Por otra parte, para las intenciones de voto para gobernadores, se eliminaron 851 registros; de ellos, 98 corresponden a boletas anuladas, y el resto a ciudadanos que viven en estados sin elecciones concurrentes, donde no procede la votación.

En todos los casos, también se eliminaron del análisis los encuestados que manifestaron haber condicionado su voto a la entrega de algún beneficio (pregunta 40), y los que respondieron haber obtenido favores, regalos o servicios a cambio de su voto (pregunta 41).

Además, en el análisis de las intenciones de voto para diputados federales, senadores y gobernadores, los votos de los partidos PVEM, PT, Movimiento Ciudadano y Panal, por considerarse muy pocos, se unieron bajo el acápite OTROS. En la Tabla II se muestra la distribución de votos obtenidos por cada partido en cada análisis.

TABLA II
DISTRIBUCIÓN DE LOS VOTOS POR PARTIDO POLÍTICO

Intención de voto	PAN	PRD	PRI	OTROS
Diputados federales	133	220	126	36
Senadores	131	231	125	26
Gobernadores	32	60	58	12

Considerando que algunos ciudadanos manifestaron haber condicionado su voto a cambio de algún beneficio, o de algún programa de gobierno (preguntas 40 y 41), se decidió determinar la influencia de estos en las intenciones de voto.

Para ello, se construyeron dos bancos de datos adicionales, correspondientes a los ciudadanos que condicionaron, o no, su voto a cambio de un beneficio o de un programa de gobierno, respectivamente. En la Tabla III se muestra una descripción de los seis conjuntos de datos obtenidos, que fueron utilizados en la presente investigación.

TABLA III
DESCRIPCIÓN DE LOS CONJUNTOS DE DATOS

Bancos de datos	Registros	Clases	Razón de desbalance
vn_diputados	515	44	6.11
vn_gobernadores	162	4	5.00
vn_presidente	521	3	2.21
vn_senadores	513	4	8.88
v_beneficio	1241	2	16.00
v_programa	1243	2	1.09

En cada caso, los registros (ya depurados) fueron representados por 14 rasgos numéricos y 150 categóricos. La razón de desbalance se calcula como la cantidad de registros de la clase mayoritaria entre la cantidad de registros de la clase minoritaria. Por ejemplo, para el banco de datos vn_diputados, se tiene que la clase mayoritaria corresponde al partido PRD, con 220 intenciones de voto, mientras que la clase minoritaria corresponde al partido OTROS, con 36 votos. Así, la razón de desbalance en este conjunto es de 6.11.

El manejo de bancos de datos desbalanceados constituye un reto para la mayoría de los algoritmos de clasificación, puesto que éstos tienden a sesgarse hacia la clase mayoritaria [28].

III. Propuesta

Esta sección consta de dos subsecciones. Cada una de las subsecciones incluye una de las dos contribuciones originales del presente trabajo de investigación. En la subsección A se presenta la parte conceptual central: el Clasificador Asociativo Gamma Modificado (CAGM); se incluyen ambas fases: la fase de entrenamiento y la fase de clasificación, además de los parámetros y su significado. En la subsección B se presenta la selección de rasgos mediante Evolución Diferencial, que permite al CAGM exhibir desempeños notables.

La primera propuesta original se presenta en la subsección A, y consiste en modificar el Clasificador Asociativo Gamma (CAG) [29], [30], [31], mediante la sustitución del operador de similitud Gamma por un operador Gamma Híbrido de Similitud, que se presenta por primera vez como una de las aportaciones relevantes de la presente investigación. El nuevo Clasificador Asociativo Gamma Modificado (CAGM) es capaz de manejar datos híbridos e incompletos, con las ventajas que ello representa en la clasificación de patrones.

La segunda propuesta original del presente artículo se presenta y describe en la subsección B; consiste en aplicar la metaheurística Evolución Diferencial (ED) para guiar un proceso de selección de rasgos [26].

A. Clasificador Asociativo Gamma Modificado (CAGM)

El Clasificador Asociativo Gamma Modificado (Fig. 1) consta de dos fases: entrenamiento y clasificación. La fase de entrenamiento incluye el almacenamiento del conjunto de aprendizaje y la determinación de varios parámetros para el CAGM (Tabla IV), cuyos valores sugeridos aparecen en [26]. Consideramos en este artículo que los conjuntos de entrenamiento X y de prueba P, son conjuntos de datos de un universo U, donde cada objeto $x \in X,\ p \in P$ está descrito por un conjunto de rasgos o atributos $A = \{A_1, A_2, \cdots, A_m\}$; cada rasgo A_i tiene asociado un dominio de definición $dom(A_i)$, que puede ser numérico o categórico. Como un caso particular, si el valor de un determinado rasgo A_i en un objeto o registro x es desconocido, se denota por $x_i = '?'$.

TABLA IV
Parámetros del CAGM

Parámetro	Descripción
w	Vector de pesos: indica la importancia de cada atributo.
θ_0	Es el valor que inicialmente tomará θ, y se utiliza en el cálculo de la similitud. Se considera θ=0.
ρ	Parámetro de paro; cuando $\theta = \rho$, el CAG dejará de iterar. Se considera el valor mínimo de los máximos valores de los atributos numéricos. Si todos los atributos fueran categóricos, se considera $\rho = m'$.
ρ_0	Parámetro de pausa; se utiliza para asignar clase desconocida. Se considera el valor máximo de los máximos valores de los atributos numéricos. Si todos los atributos fueran categóricos, se considera $\rho_0 = m'$.
u	Umbral para decidir si el patrón a clasificar pertenece a la clase desconocida; o bien, a alguna de las clases conocidas. Se considera u=0.

El entrenamiento del CAGM consiste en calcular primeramente los pesos de los atributos, utilizando para ello la metaheurística de Evolución Diferencial. Este proceso es detallado en la sección III.C. Posteriormente, de los atributos restantes, se calculan los valores de los parámetros de pausa y paro.

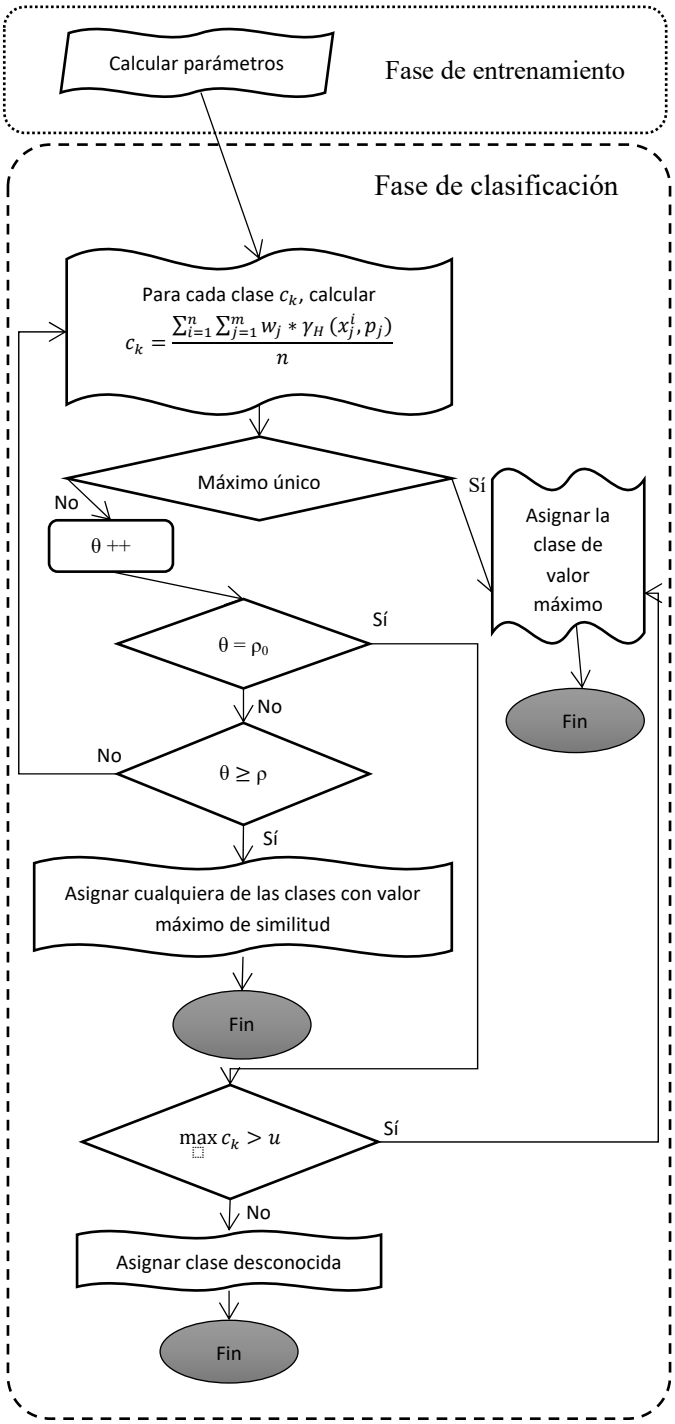

Fig. 1. Gráfico que representa el funcionamiento del clasificador Asociativo Gamma Modificado.

Para ello, los valores de los atributos numéricos son escalados, de forma tal que se conviertan de números reales a

números enteros. Luego, se seleccionan los máximos valores de cada uno de los atributos numéricos. Así, se define ρ como el valor mínimo de los máximos, y ρ_0 como el valor máximo de los máximos. En caso de no existir ningún atributo numérico, se considera el total de atributos restantes m' como el valor de ambos parámetros.

La fase de clasificación del CAGM comienza con escalar los valores numéricos del patrón a clasificar, al igual que se hizo en la fase de entrenamiento. Luego, comienza un proceso iterativo. Sea un patrón a clasificar $p \in P$ y sea p_j el valor correspondiente al j-ésimo atributo. Para analizar la similitud entre el patrón de prueba y los patrones de entrenamiento, se propone el **Operador Gamma Híbrido de Similitud** γ_H.

Sean dos patrones $x, y \in U$, con las correspondientes i-ésimas componentes denotadas por x_i y y_i, respectivamente. La definición del Operador Gamma Híbrido de Similitud γ_H consta de tres casos mutuamente exclusivos:

CASO I: Ambas componentes x_i y y_i son datos numéricos:

$$\gamma_H(x_i, y_i, \theta) = \begin{cases} 1 \text{ si } |x_i - y_i| \leq \theta \\ 0 \text{ si } |x_i - y_i| > \theta \end{cases}$$

CASO II: Ambas componentes x_i y y_i son datos categóricos:

$$\gamma_H(x_i, y_i) = \begin{cases} 1 \text{ si } x_i = y_i \\ 0 \text{ si } x_i \neq y_i \end{cases}$$

CASO III: Alguna de las dos componentes es un dato perdido, o ambas lo son $(x_i = '?') \vee (y_i = '?')$: $\gamma_H(x_i, y_i) = 1$.

Luego de obtenidas las similitudes, se calcula el promedio de la similitud generalizada de dicho patrón de prueba para cada clase c_k:

$$c_k = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} w_j * \gamma_H(x_j^i, p_j)}{n} \quad (1)$$

Se considera que el total de objetos de la clase c_k en el conjunto de entrenamiento está dado por n, que m es la cantidad de atributos, y que x_j^i representa el valor del j-ésimo atributo del i-ésimo objeto de la clase k, y que w_j representa el peso del j-ésimo atributo.

Si se encuentra un máximo único entre todos los valores de c_k, el proceso termina. Si esto no ocurre, se tienen en cuenta los valores de los parámetros de paro y pausa, así como el valor del parámetro θ, en un proceso iterativo. En la Fig. 1 se ilustra el proceso completo.

B. Selección de Rasgos Guiada por Evolución Diferencial

La metaheurística Evolución Diferencial (ED) [32] toma como inspiración el proceso de la evolución biológica. Así, se tiene una población de individuos (soluciones) generada inicialmente de forma aleatoria. Posteriormente, en cada una de las G generaciones del algoritmo, para cada elemento x de la población, se seleccionan tres individuos distintos entre sí, y denotados como $r1, r2, r3$.de forma aleatoria. A partir de dichos individuos, se genera una solución de prueba, como $x' = r3 + F(r1 - r2)$, donde F es un factor de escala, definido usualmente en el intervalo [0,2]. Posteriormente, cada elemento x'_j de esta solución de prueba es recombinado con el elemento correspondiente de la solución original x, como $x''_j = \begin{cases} x'_j & \text{si } rdm \leq CR \\ x_j & \text{en otro caso} \end{cases}$, donde rdm es un número aleatorio en el intervalo [0,1] y CR es una constante de cruza, definida en el intervalo [0,1]. Si el nuevo vector recombinado x'' es mejor que el vector original x, el vector original es sustituido por el vector recombinado.

Esta metaheurística se ha aplicado con éxito para estimar el peso de los rasgos en el Clasificador Asociativo Gamma [26]. Sin embargo, en el estado del arte no se reportan trabajos donde se haya explorado su uso en la selección de rasgos.

Para la estimación automática de pesos del CAG propuesta en [26], se utilizan valores de pesos en el intervalo [0,4]; a nuestro juicio, esto dificulta la interpretación de los resultados. Es por ello que en este artículo utilizamos una codificación de los individuos con vectores cuyos valores (pesos) se encuentran definidos en el intervalo [0,1].

Por otra parte, la propuesta de [26] ejecuta el algoritmo de ED con 10 individuos y 50 iteraciones, mientras que nuestra propuesta utiliza 25 individuos y 1000 iteraciones, con lo que se mejoran los resultados de [26].

Como función objetivo, se consideró la eficacia del CAGM con respecto al conjunto de entrenamiento. Como medida de eficacia se utilizó el promedio de sensibilidad por clase o promedio de aciertos por clase [28]. Eta medida es detallada en la sección IV. A. Como parámetros de ED se definió una constante de cruza $CR = 0.8$, un factor de escala $F = 1$, y una inicialización aleatoria de la población. Un umbral ε permite eliminar los rasgos irrelevantes. Sea w_i el peso asignado por ED al rasgo A_i; si $w_i \leq \varepsilon$, el rasgo A_i es eliminado.

IV. Diseño Experimental

Es pertinente hacer notar que, después de una amplia investigación documental, no se ha localizado en el estado del arte algún trabajo de investigación donde se utilicen datos de encuestas para predecir las intenciones de voto de ciudadanos mexicanos, o donde se aplique un clasificador asociativo a un problema de índole social. En esta sección se detalla el protocolo de experimentación, así como de las medidas de desempeño y tests estadísticos utilizados.

Considerando el desbalance de los datos (Tabla III), en la presente investigación se utilizó la validación cruzada estratificada en 5 hojas, que es la adecuada para el manejo de bancos de datos desbalanceados [33]; es decir, en cada iteración se utilizó el 20% del conjunto para prueba y el 80% para entrenamiento.

Se evaluó el desempeño del CAGM propuesto, con la selección de rasgos mediante ED, en la estimación de intenciones de voto de ciudadanos mexicanos. Se evaluó, además, el desempeño de seis clasificadores supervisados del estado del arte, que permiten el manejo de datos híbridos, en la estimación de intenciones de voto de ciudadanos mexicanos. Los clasificadores evaluados fueron: C4.5 [34], clasificador del vecino más cercano, utilizando 1 y 3 vecinos (1-NN, 3-

NN) [35], Perceptron Multicapa (MLP) con *Backpropagation* [36], Máquinas de Soporte Vectorial (SVM) [37] y Regresión Logística multinomial (LR) [38].

Todos estos algoritmos fueron aplicados mediante el uso del software KEEL [39], utilizando los parámetros por defecto de cada uno de ellos.

Se escogió este software debido a que permite una serialización de las particiones de entrenamiento y prueba, lo que permite garantizar la repetitividad de los experimentos; incluye métodos de validación especializados para el manejo de bancos de datos desbalanceados, como el *Distributed Optimally Balanced Stratified Cross Validation (DOB-SCV),* cuenta con la posibilidad de realizar test estadísticos, y permite el manejo directo de bancos de datos no balanceados. Finalmente, los resultados del CAGM fueron comparados con los obtenidos por estos clasificadores.

A. Medidas de Desempeño

Cuando se analizan conjuntos de datos desbalanceados, las medidas de desempeño estándar tales como la razón de instancias correctamente clasificadas, no se consideran adecuadas [28]. Esto se debe al sesgo de dichas medidas hacia la clase mayoritaria, puesto que no distinguen entre el número de clasificaciones correctas de las diferentes clases, lo cual puede conducir a conclusiones erróneas.

A fin de realizar la evaluación del desempeño en conjuntos de datos desbalanceados de múltiples clases, se ha propuesto el uso del promedio de sensibilidad por clase o promedio de aciertos por clase [28]. En un problema de dos clases, la sensibilidad (también conocida como tasa de verdaderos positivos o *Recall*), considera el total de instancias positivas correctamente clasificadas, respecto al total de instancias de la clase positiva.

En un problema de k clases, la sensibilidad considera el total de instancias correctamente clasificadas como de clase i, respecto al total de instancias de la clase i. Así, la sensibilidad de la clase i estima la probabilidad de clasificar correctamente a una instancia de clase i. Para el cálculo de la sensibilidad, es necesario contar con la matriz de confusión para cada una de las k clases del problema bajo estudio. En la Fig. 2 se muestra una matriz de confusión para k clases.

Sea n_{ii} el número de instancias correctamente clasificadas, en una matriz de confusión de k clases, y sea $t_i = \sum_{j=0}^{k} n_{ij}$ el total de instancias pertenecientes a la clase i. La sensibilidad (Recall o tasa de verdaderos positivos) de la clase i, denotada por S_i, se calcula como [40]:

$$S_i = Recall_i = TPR_i = n_{ii}/t_i \tag{2}$$

		Clase asignada		
		0	…	K
Clase real	0	n_{00}	…	n_{0k}
	…	…	…	…
	k	n_{k0}	…	n_{kk}

Fig. 2. Matriz de confusión para k clases.

En esta investigación se utiliza una medida de desempeño que da el mismo peso a cada una de las clases de problema, independientemente del número de ejemplos que tenga. Por lo tanto, se utiliza el promedio de sensibilidad por clase o promedio de aciertos por clase, que se define como [28]:

$$Avg\_S = \left(\sum_{I=1}^{k} S_i\right)/k \tag{3}$$

donde k es la cantidad de clases y S_i es la sensibilidad (tasa de verdaderos positivos o Recall) de la i-ésima clase. Esta medida permite evaluar el desempeño global de los algoritmos de clasificación en todas las clases del problema, no solo en la clase minoritaria.

B. Tests Estadísticos Utilizados

Para conocer qué algoritmos obtuvieron los mejores resultados en la determinación de las intenciones de voto, se utilizaron pruebas de hipótesis, las cuales permiten evaluar las diferencias en el desempeño alcanzado por los diferentes algoritmos en la solución de un determinado problema.

Se escogió el test no paramétrico de Friedman [41], recomendado para este tipo de estudios [42]. Esta prueba consiste en ordenar los datos, reemplazándolos por su respectivo rango. El mejor resultado corresponde al rango 1, el segundo mejor al rango 2, y así sucesivamente. Al ordenarlos, se considera la existencia de datos idénticos, en cuyo caso se asigna un rango promedio.

Si la hipótesis nula de igualdad de desempeños es rechazada por el test de Friedman, se hace necesario aplicar test post-hoc, para determinar entre qué algoritmos se encuentran las diferencias. Entre los test post-hoc recomendados para el análisis del desempeño de algoritmos en múltiples conjuntos de datos, se encuentra el test de Holm [42]. Existen herramientas automatizadas para el cálculo del test de Friedman, así como para el cálculo de los test post-hoc. En esta investigación se utilizó el software KEEL [39]. Aunque podría haberse utilizado el test de Wilcoxon para dos muestras relacionadas, escogimos utilizar los test de Friedman y de Holm, puesto que permiten establecer ranking entre los algoritmos comparados, a diferencia del test de Wilcoxon.

IV. Resultados y Discusión

En esta sección se aborda la influencia de la selección de rasgos mediante Evolución Diferencial en el Clasificador Asociativo Gamma Modificado. Finalmente, se evalúa el desempeño de la propuesta realizada, con respecto a otras del estado del arte, para la determinación de intenciones de voto de ciudadanos mexicanos.

A. Resultados de la Selección de Rasgos Basada en Evolución Diferencial

En esta investigación se exploró el uso de la Evolución Diferencial, no sólo para obtener los pesos de los rasgos para el CAGM, sino además para realizar un proceso de selección de éstos. Para ello, se consideraron umbrales en el intervalo

[0.0 – 0.9], y en cada caso, se seleccionaron los rasgos por encima del umbral deseado. Posteriormente, se utilizaron dichos rasgos en el CAGM, y se midió su desempeño, considerando el promedio de la sensibilidad por clase (3). Para establecer, o no, la existencia de diferencias significativas en el desempeño del CAGM, se utilizó el test de Friedman, con un nivel de significación de 0.05, para un 95% de confianza (Tabla V). El test de Friedman arrojó un valor de 0.064454, muy cercano al nivel de significación establecido.

TABLA V
RANKING DE FRIEDMAN AL COMPARAR
DIFERENTES UMBRALES PARA LA SELECCIÓN DE RASGOS

Ranking	3.5	3.6	4.4	5.0	5.2	5.3	5.8	6.0	7.2	9.0
$w_i \geq$	0.5	0.3	0.4	0.6	0.0	0.1	0.2	0.7	0.8	0.9

Luego de aplicado el test de Friedman, se procedió a buscar la existencia de diferencias en el desempeño de los algoritmos mediante el test de Holm.

La Tabla VI muestra los resultados del test de Holm al comparar la selección de rasgos con un umbral igual a 0.5 (mejor en el ranking) con respecto a la selección de rasgos considerando otros umbrales.

TABLA VI
TEST DE HOLM AL COMPARAR
DIFERENTES UMBRALES PARA LA SELECCIÓN DE RASGOS

I	9	8	7	6	5	4	3	2	1
$w_i \geq$	0.9	0.8	0.7	0.2	0.1	0.0	0.6	0.4	0.3
Holm	**0.006**	**0.006**	0.007	0.008	0.010	0.013	0.02	0.03	0.05

El test de Holm rechaza las hipótesis con un valor no ajustado $Holm \leq 0.00625$. Es decir, utilizar un umbral igual a 0.5 resulta en un desempeño del CAGM significativamente mejor que utilizar un umbral igual a 0.8 o 0.9.

Por otra parte, al utilizar un proceso de selección de rasgos, se disminuye la dimensionalidad de los datos, lo cual contribuye a facilitar su comprensión y análisis. En la medida en que el umbral aumenta, también lo hace la reducción obtenida. Sin embargo, para umbrales superiores a 0.8 se observa una disminución significativa en el desempeño del CAGM.

Es por ello que consideramos que un umbral de 0.5 ofrece el mejor compromiso en cuanto al desempeño del CAGM, dado que obtiene los mejores resultados en cuanto a eficacia (siendo el mejor en el ranking obtenido por el test de Friedman, Tabla V), con una reducción de la dimensionalidad de los datos del 54% al 57%, aproximadamente.

B. Resultados del Desempeño del Clasificador Asociativo Gamma Modificado

Luego de obtenidos los pesos del CAGM, y realizada la selección de los mismos, considerando un umbral de 0.5, se procedió a estudiar el desempeño de dicho clasificador. Para ello, se realizó una validación cruzada estratificada en 5 hojas y se promediaron los resultados. Posteriormente, se comparó el desempeño del CAGM con respecto a otros clasificadores del estado del arte, usando como medida de desempeño el el promedio de sensibilidad por clase (ecuación 3). La Tabla VII muestra los resultados en el desempeño de los clasificadores analizados. Se resaltan en negritas los mejores resultados en la estimación de la intención de voto, en cada caso.

TABLA VII
DESEMPEÑO DE LOS CLASIFICADORES EVALUADOS
MEDIDA DE DESEMPEÑO: PROMEDIO DE LA SENSIBILIDAD POR CLASE

Bancos de datos	CAGM	C45	1NN	3NN	MLP	SVM	LR
vn_diputados	0.75	**0.77**	0.65	0.67	0.25	0.59	0.70
vn_gobernadores	**0.91**	0.80	0.70	0.73	0.25	0.65	0.78
vn_presidente	**0.94**	0.75	0.70	0.90	0.33	0.64	0.86
vn_senadores	0.80	**0.85**	0.67	0.71	0.25	0.61	0.79
v_beneficio	**0.77**	0.54	0.60	0.63	0.50	0.75	0.76
v_programa	0.88	**0.94**	0.89	0.87	0.50	0.79	0.88

Como se aprecia, los mejores resultados en la determinación de la intención de voto fueron obtenidos por el CAGM con la selección de rasgos propuesta (mejor en 3 de los 6 bancos de datos), seguido del clasificador C4.5 (mejor en los 3 bancos de datos restantes).

Para establecer la existencia o no de diferencias significativas en el desempeño de los algoritmos comparados, se utilizó nuevamente el test de Friedman. El test de Friedman arrojó un valor de 0.00025, por lo que se rechaza la hipótesis nula. El ranking obtenido por el test se muestra en la tabla VIII. Como se aprecia, el CAGM obtuvo el primer lugar en el ranking, lo que muestra su excelente desempeño en la estimación de las intenciones de voto de ciudadanos mexicanos.

TABLA VIII
RANKING DE FRIEDMAN AL COMPARAR
EL DESEMPEÑO DE LOS CLASIFICADORES EVALUADOS

Ranking	1.75	2.50	2.91	3.83	4.50	5.50	7.00
Algoritmo	CAGM	C4.5	LR	3NN	1NN	SVM	MLP

Igualmente, para establecer la existencia, o no, de diferencias significativas con respecto al desempeño de los otros clasificadores analizados, se utilizó el test de Holm. En la tabla IX se muestran los resultados obtenidos.

TABLA IX
TEST DE HOLM AL COMPARAR
EL DESEMPEÑO DE LOS CLASIFICADORES EVALUADOS

i	6	5	4	3	2	1
Clasificador	MLP	SVM	1NN	3NN	LR	C4.5
Holm	**0.008**	**0.010**	0.012	0.017	0.025	0.050

El test de Holm rechaza las hipótesis con un valor no ajustado $Holm \leq 0.0125$. Es decir, el desempeño del CAGM en la estimación de las intenciones de voto de ciudadanos mexicanos es significativamente mejor que el de los clasificadores 1-NN, SVM y MLP. Además, con un 90% de confianza se aprecian diferencias significativas en el desempeño del CAGM con respecto al 3-NN, puesto que el test de Holm arroja un valor de probabilidad $p = 0.094844$, que es menor que el nivel de significación $\alpha = 0.1$. Sin embargo, no se apreciaron diferencias significativas entre el desempeño del CAGM y el del C4.5 y la regresión logística

(LR).

Analizando la reducción en cuanto a la dimensionalidad de los datos, podemos apreciar que el CAGM obtiene resultados de desempeño en cuanto al promedio de la sensibilidad por clase comparables a los del C4.5 y el LR, usando solamente el 50% de los rasgos. Cabe señalar, además, que el CAGM fue el primero en el ranking obtenido por el test de Friedman (Tabla VIII). Esto ilustra las bondades de nuestra propuesta.

Por otra parte, se realizó un análisis del tiempo de ejecución (en segundos) de los algoritmos comparados, que se muestra en las Tablas X y XI. Lamentablemente, KEEL no serializa los datos de tiempo de ejecución de los algoritmos C4.5, MLP, SVM ni LR. Es por ello que, para poder cumplir con el análisis de tiempo requerido por los revisores anónimos, utilizamos el software WEKA [43] y copiamos manualmente los resultados de tiempo de ejecución de estos clasificadores.

TABLA X
TIEMPO DE ENTRENAMIENTO DE LOS CLASIFICADORES EVALUADOS EN LA ESTIMACIÓN DE INTENCIONES DE VOTO MEXICANOS (EN SEGUNDOS)

Bancos de datos	CAGM	C45	1NN	3NN	MLP	SVM	LR
vn_diputados	55.2	0.2	3.8	3.8	111.6	17.1	0.9
vn_gobernadores	41.0	0.1	0.4	0.4	147.8	13.6	0.8
vn_presidente	38.9	0.1	3.9	4.0	143.6	13.6	0.8
vn_senadores	40.0	0.1	4.2	4.0	138.9	13.4	0.8
v_beneficio	82.3	0.3	22.3	22.1	182.1	98.0	2.3
v_programa	87.5	0.3	22.3	23.1	184.5	98.1	2.3

TABLA XI
TIEMPO DE CLASIFICACIÓN DE LOS CLASIFICADORES EVALUADOS EN LA ESTIMACIÓN DE INTENCIONES DE VOTO MEXICANOS (EN SEGUNDOS)

Bancos de datos	CAGM	C45	1NN	3NN	MLP	SVM	LR
vn_diputados	0.8	0.0	0.9	1.0	0.1	0.1	0.0
vn_gobernadores	0.1	0.0	0.1	0.1	0.1	0.1	0.0
vn_presidente	0.9	0.0	1.0	1.0	0.1	0.1	0.0
vn_senadores	0.9	0.0	1.0	1.0	0.1	0.1	0.0
v_beneficio	4.2	0.0	5.5	5.6	0.1	0.1	0.0
v_programa	4.2	0.0	5.7	6.0	0.1	0.1	0.0

Como se aprecia, el algoritmo más rápido en su entrenamiento es el C4.5, seguido del LR. Siguen los clasificadores del vecino más cercano (cuyos mayores tiempos de entrenamiento fueron menores a medio minuto), y el CAGM, cuyo mayor tiempo de entrenamiento fue de casi tres minutos. Tanto las Máquinas de Soporte Vectorial como el Perceptrón multicapa, tuvieron tiempos de entrenamiento máximos de más de tres y seis minutos, respectivamente.

En cuanto a los tiempos de clasificación, el C4.5, el LR, el SVM y el MLP tuvieron los mejores resultados, clasificando a todos los objetos en menos de una décima de segundo. Los tiempos de clasificación del CAGM son muy similares a los de los clasificadores 1-NN y 3-NN, sin embargo, su tiempo de entrenamiento fue mayor. Esto se debe a dos factores fundamentales: el primero, que el CAGM realiza un cálculo de los pesos de los atributos, mediante el algoritmo de Evolución Diferencial, y el segundo, que nuestra implementación no utiliza elementos de programación paralela y uso avanzado de recursos computacionales, como sí lo hacen las implementaciones de WEKA y KEEL.

Cabe señalar que los experimentos se realizaron en una computadora con procesador Intel(R) Core(TM) i7-8700, con 3.2 GHz, memoria RAM de 64GB y sistema operativo Windows 10.

V. CONCLUSIONES

Con los resultados de la presente investigación, se intenta revertir la ausencia, en México, de estudios donde se utilicen métodos computacionales para predecir las preferencias electorales de los votantes mexicanos. La primera tarea consistió en depurar los datos de las elecciones del año 2012, los cuales fueron proporcionados por la Secretaría de Gobernación. Con dichos datos, se crearon seis bancos de datos: cuatro para intenciones de voto (presidente, gobernadores, diputados y senadores) y dos para determinar la influencia en las intenciones de voto en los ciudadanos que manifestaron haber condicionado su voto a cambio de algún beneficio, o de algún programa de gobierno.

El modelo aquí presentado, el CAGM, se diseñó mediante una modificación realizada al Clasificador Asociativo Gamma original; además, se utilizó Evolución Diferencial para seleccionar rasgos relevantes. Los resultados indican que, al comparar el desempeño en los seis bancos de datos, nuestra propuesta exhibe los mejores resultados en tres de los bancos de datos, superando a algunos de los mejores modelos similares presentes en el estado del arte.

Como trabajo a futuro, creemos que es posible aplicar el CAGM al estudio de las intenciones de votos de ciudadanos de otros países, no solamente mexicanos.

AGRADECIMIENTOS

Los autores agradecen el apoyo de las siguientes instituciones mexicanas: Secretaría de Gobernación (SEGOB), Instituto Politécnico Nacional (Secretaría de Investigación y Posgrado, Secretaría Académica, COFAA, CIDETEC y CIC), CONACyT y Sistema Nacional de Investigadores (SNI).

REFERENCIAS

[1] I. P. Banai, B. Banai, and K. Bovan, "Vocal characteristics of presidential candidates can predict the outcome of actual elections," *Evolution and Human Behavior,* to be published. http://dx.doi.org/10.1016/j.evolhumbehav.2016.10.012.

[2] M. P. Cameron, P. Barrett, and B. Stewardson, "Can social media predict election results? evidence from new zealand," *Journal of Political Marketing,* vol. 15, pp. 416-432, 2016.

[3] D. Gayo-Avello, "No, you cannot predict elections with Twitter," *IEEE Internet Computing,* vol. 16, pp. 91-94, 2012.

[4] J. A. McCann and J. I. Domınguez, "Mexicans react to electoral fraud and political corruption: an assessment of public opinion and voting behavior," *Electoral Studies,* vol. 17, pp. 483-503, 1998.

[5] J. L. Klesner, "Adiós to the PRI? Changing Voter Turnout in Mexico's Political Transition," *Mexican Studies/Estudios Mexicanos,* vol. 17, pp. 17-39, 2001.

[6] J. A. McCann and C. H. Lawson, "An Electorate Adrift?: Public Opinion and the Quality of Democracy in Mexico," *Latin American Research Review,* vol. 38, pp. 60-81, 2003.

[7] J. T. Hiskey and S. Bowler, "Local context and democratization in Mexico," *American Journal of Political Science,* vol. 49, pp. 57-71, 2005.

[8] K. Bruhn and K. F. Greene, "Elite polarization meets mass moderation in Mexico's 2006 elections," *PS: Political Science and Politics,* vol. 40, pp. 33-38, 2007.

[9] C. Lawson and J. A. McCann, "Television news, Mexico's 2000 elections and media effects in emerging democracies," *British Journal of political science,* vol. 35, pp. 1-30, 2005.

[10] Y. A. Basallo, V. E. Senti, and N. M. Sanchez, "Artificial intelligence techniques for information security risk assessment," *IEEE Latin America Transactions,* vol. 16, pp. 897-901, 2018.

[11] E. M. Kakihata *et al.*, "Intrusion detection system based on flows using machine learning algorithms," *IEEE Latin America Transactions,* vol. 15, pp. 1988-1993, 2017.

[12] R. P. Bonidia, J. D. Brancher, and R. M. Busto, "Data Mining in Sports: A Systematic Review," *IEEE Latin America Transactions,* vol. 16, pp. 232-239, 2018.

[13] J. P. Castillo, C. D. Mafiolis, E. C. Escobar, A. G. Barrientos, and R. V. Segura, "Design, construction and implementation of a low cost solar-wind hybrid energy system," *IEEE Latin America Transactions,* vol. 13, pp. 3304-3309, 2015.

[14] S. M. R. Sanhueza and S. C. L. Freitas, "Overvoltage Forecast in a Urban Distribution Power Grid Considering PV Systems Conection," *IEEE Latin America Transactions,* vol. 16, pp. 2221-2227, 2018.

[15] F. A. R. Silva, "Analytical Intelligence in Processes: Data Science for Business," *IEEE Latin America Transactions,* vol. 16, pp. 2240-2247, 2018.

[16] A. B. de Sousa, A. S. Lima, J. N. de Souza, J. A. B. Moura, and A. C. B. Silva, "Business Risk-based Redundancy Points Identification in Synchronous Digital Hierarchy Optical Networks," *IEEE Latin America Transactions,* vol. 16, pp. 2254-2260, 2018.

[17] P. Pinto, I. Theodoro, M. Arrais, and J. Oliveira, "Data mining and social web semantics: a case study on the use of hashtags and memes in Online Social Networks," *IEEE Latin America Transactions,* vol. 15, pp. 2276-2281, 2017.

[18] D. Campus, G. Pasquino, and C. Vaccari, "Social networks, political discussion, and voting in Italy: A study of the 2006 election," *Political Communication,* vol. 25, pp. 423-444, 2008.

[19] P. Burnap, R. Gibson, L. Sloan, R. Southern, and M. Williams, "140 characters to victory?: Using Twitter to predict the UK 2015 General Election," *Electoral Studies,* vol. 41, pp. 230-233, 2016.

[20] R. Effing, J. van Hillegersberg, and T. Huibers, "Social media indicator and local elections in The Netherlands: towards a framework for evaluating the influence of Twitter, YouTube, and Facebook," in *Social Media and Local Governments. theory and Practice. Part V*, M. Z. Sobaci, Ed., ed: Springer, 2016, pp. 281-298.

[21] A. Jungherr, H. Schoen, and P. Jürgens, "The mediation of politics through Twitter: An analysis of messages posted during the campaign for the German federal election 2013," *Journal of Computer-Mediated Communication,* vol. 21, pp. 50-68, 2016.

[22] S. Muralidharan and Y. Sung, "Direct and Mediating Effects of Information Efficacy on Voting Behavior: Political Socialization of Young Adults in the 2012 US Presidential Election," *Communication Reports,* vol. 29, pp. 100-114, 2016.

[23] J. Wheatley, "Using VAAs to explore the dimensionality of the policy space: experiments from Brazil, Peru, Scotland and Cyprus," *International Journal of Electronic Governance,* vol. 5, pp. 318-348, 2012.

[24] S. Papagiannidis, C. K. Coursaris, and M. Bourlakis, "Do websites influence the nature of voting intentions? The case of two national elections in Greece," *Computers in human behavior,* vol. 28, pp. 300-307, 2012.

[25] J. DiGrazia, K. McKelvey, J. Bollen, and F. Rojas, "More tweets, more votes: Social media as a quantitative indicator of political behavior," *PloS one,* vol. 8, p. e79449, 2013.

[26] A. Ramirez, I. Lopez, Y. Villuendas, and C. Yanez, "Evolutive improvement of parameters in an associative classifier," *IEEE Latin America Transactions,* vol. 13, pp. 1550-1555, 2015.

[27] J. Ruiz-Shulcloper, "Pattern recognition with mixed and incomplete data," *Pattern Recognition and Image Analysis,* vol. 18, pp. 563-576, 2008.

[28] A. Fernández, V. López, M. Galar, M. J. De Jesus, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowledge-based systems,* vol. 42, pp. 97-110, 2013.

[29] C. López-Martın, I. López-Yánez, and C. Yánez-Márquez, "Application of Gamma classifier to development effort prediction of software projects," *Appl. Math,* vol. 6, pp. 411-418, 2012.

[30] I. López-Yáñez, A. J. Argüelles-Cruz, O. Camacho-Nieto, and C. Yáñez-Márquez, "Pollutants time-series prediction using the gamma classifier," *International Journal of Computational Intelligence Systems,* vol. 4, pp. 680-711, 2011.

[31] I. López-Yáñez, L. Sheremetov, and C. Yáñez-Márquez, "A novel associative model for time series data mining," *Pattern Recognition Letters,* vol. 41, pp. 23-33, 2014.

[32] R. Storn and K. Price, "Differential evolutionâ€“a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization,* vol. 11, pp. 341-359, 1997.

[33] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Information Sciences,* vol. 250, pp. 113-141, 2013.

[34] J. R. Quinlan, "C4.5: Programming for machine learning," *Morgan Kauffmann,* 1993.

[35] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory,* vol. 13, pp. 21-27, 1967.

[36] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, D. E. Rumelhart, M. J. L, and P. R. Group, Eds., ed: M.I.T. Press, 1986, pp. 318-362.

[37] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," *Neural Computation,* vol. 13, pp. 637-649, 2001.

[38] S. Le Cessie and J. C. Van Houwelingen, "Ridge estimators in logistic regression," *Applied statistics,* pp. 191-201, 1992.

[39] J. Alcalá *et al.*, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing,* vol. 17, pp. 255-287, 2010.

[40] F. Fernández-Navarro, C. Hervás-Martínez, and P. A. Gutiérrez, "A dynamic over-sampling procedure based on sensitivity for multi-class problems," *Pattern Recognition,* vol. 44, pp. 1821-1833, 2011.

[41] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics,* vol. 11, pp. 86-92, 1940.

[42] S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Information Sciences,* vol. 180, pp. 2044-2064, 2010.

[43] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA data mining software: an update", ACM SIGKDD explorations newsletter vol. 11(1), pp. 10-18, 2009.

Sonia Lizde Ortiz-Ángeles. Ingeniera en Telecomunicaciones por la UNAM, México en 2000. Se especializó en Gestión de las Telecomunicaciones (2002) por la Escuela de Organización Industrial en Madrid, España. Maestra en Tecnología de Cómputo (2018) por el CIDETEC-IPN, México. lizde00@yahoo.com.

Yenny Villuendas Rey. Licenciada y Maestra en Ciencias en Informática Aplicada por la Universidad de Ciego de Ávila; Doctora en Ciencias Técnicas por la Universidad de Las Villas, Cuba (2014). Profesora investigadora de tiempo completo en el CIDETEC IPN y miembro del SNI nivel C. yvilluendasr@ipn.mx.

Cornelio Yáñez Márquez. Licenciado en Física y Matemáticas de la ESFM del IPN en 1989, obtuvo el grado de Maestría en Ciencias de la Computación en 1995 y recibió el grado de Doctor en Ciencias de la Computación en 2002, ambos en el CIC IPN. Es profesor investigador de tiempo completo en el mismo centro y es miembro del SNI nivel 2. coryanez@gmail.com.

Itzamá López Yáñez. Ingeniero en Sistemas de Información por el ITSEM CSN. Obtuvo el grado de Maestría en Ciencias de la Computación en 2007 y recibió el grado de Doctor en Ciencias de la Computación en 2011, ambos en el CIC IPN. Es profesor investigador de tiempo completo en el CIDETEC IPN y miembro del SNI nivel 1. ilopezy@ipn.mx.

Oscar Camacho-Nieto. Ingeniero en Electrónica y comunicaciones por la ESIME-IPN. Graduado de Maestría y Doctorado en Ciencias de la Computación por el CINTEC-IPN y el CIC-IPN (1995, 2003). Se desempeña como director del CIDETEC IPN y es miembro del SNI nivel 1. ocamacho@ipn.mx.