Metodología de minería de datos para perfilamiento cuantitativo de la brecha digital de ciudades

Sergio R. Coria, Mónica Pérez-Meza, Rosibelda Mondragón-Becerra
y Darío Barragán-López

Instituto de Informática
Universidad de la Sierra Sur
Calle Guillermo Rojas Mijangos S/N, Esq. Av. Universidad Col. Ciudad Universitaria
Miahuatlán de Porfirio Díaz, Oax.,
México
{coria, mperez, rmondragon, dbarragan}@unsis.edu.mx
http://www.unsis.edu.mx

Resumen Se propone un enfoque novedoso para análisis y modelación del fenómeno de la brecha digital, concentrándose en el nivel de ciudad como unidad territorial. Se usan los algoritmos PART and J4.8 de Witten y Frank, implementados en el reconocido toolkit WEKA, sobre datos que describen diversos niveles de brecha digital en ciudades. Se implementan datasets del Censo 2010 de Población y Vivienda de México. Se selecciona la mayoría de sus variables y se transforman en porcentajes o en promedios por municipio. Se discretiza el porcentaje de hogares que tienen Internet para usarse como *target*. La mayoría de las otras variables transformadas se introducen como predictores a los algoritmos. Los resultados muestran que nuestro enfoque es altamente útil para producir perfiles de ciudades que describen interacciones entre variables demográficas, socio-económicas y educacionales asociadas a diversos niveles de presencia de Internet. Finalmente, se discuten las ventajas de estos algoritmos para este dominio.

Palabras clave: algoritmo PART, algoritmo J4.8, datos de censos, brecha digital, exclusión digital, inclusión digital, WEKA.

1. Introducción

La brecha digital es definida como *la distancia entre individuos, hogares, negocios y áreas geográficas en diferentes niveles socio-económicos respecto tanto a sus oportunidades para acceder a tecnologías de información y comunicación (TIC) como para su uso del Internet para una amplia variedad de actividades* [1]. Diversas teorías sobre el fenómeno de la brecha digital consideran que esta puede ser explicada por la interacción entre la presencia de bienes y servicios de TIC y un conjunto de características demográficas, socio-económicas y educacionales de la población en el territorio que sea estudiado.

Uno de los principales propósitos de la investigación de la brecha digital, tanto de países como de regiones o de ciudades, es descubrir patrones que ofrezcan conocimiento acerca de las interacciones entre la presencia de bienes y servicios de TIC en hogares y las características demográficas, socio-económicas y educacionales de sus habitantes. La investigación de la brecha digital requiere técnicas que faciliten el descubrimiento de relaciones causales entre variables. Por lo tanto, las técnicas de clasificación automática [2] pueden ser de gran utilidad.

La presencia de Internet en hogares es uno de los principales aspectos de las TIC involucrados en el fenómeno. Por ello, el problema de investigación específico en este trabajo consiste en evaluar cómo los algoritmos de clasificación automática PART [3] y J4.8 [4] (basado en C4.5 [5]) pueden contribuir a descubrir conocimiento en el campo de la investigación de la brecha digital, enfocándose en el nivel de ciudad como unidad para análisis y modelación y teniendo como *target* la presencia de Internet en hogares. El propósito principal no es explotar los modelos producidos para automatizar tareas de clasificación, sino ofrecer a los investigadores de la brecha digital una técnica que produzca modelos descriptivos capaces de guiar el descubrimiento de explicaciones para este fenómeno en ciudades de países específicos.

En esta investigación se eligió a J4.8 y PART porque, desde una perspectiva teórica, ambos forman parte de la familia de algoritmos de aprendizaje automático supervisado, lo cual permite clasificar instancias, descubriendo los patrones que caracterizan a sus clases. También, porque la precisión de ambos algoritmos es comparable a la de otros más complejos como el perceptrón multicapa (*multi-layer perceptron*, MLP). Desde una perspectiva práctica, fueron elegidos porque ofrecen una mayor facilidad explicativa y de interpretación a usuarios de dominio que no tienen conocimientos de aprendizaje automático y además porque facilitan la programación de sistemas de apoyo a la toma de decisiones (*decission support systems*, DSS).

El artículo está organizado en las siguientes secciones: la sección 2 presenta un breve panorama del estado del arte. La sección 3 describe los datos fuente que se usan para desarrollar la metodología y para realizar su evaluación. La sección 4 explica la metodología propuesta por esta investigación. La sección 5 presenta los resultados, que están organizados como un conjunto de siete modelos PART y siete modelos J4.8. La sección 6 discute los resultados. Finalmente, la sección 7 presenta algunas conclusiones y sugiere trabajo futuro en este campo.

2. Estado del arte

Uno de los principales propósitos de la investigación del fenómeno de la brecha digital consiste en identificar las variables que lo determinan en contextos territoriales determinados. En la mayoría de los trabajos relacionados con el tema, la unidad territorial más frecuente en análisis y modelación es el nivel de país. Por ejemplo, en [6] se analiza las determinantes de la brecha digital global realizando un análisis entre países acerca de la penetración de computadora e Internet.

Las técnicas más comunes para modelar y analizar la brecha digital aplican índices compuestos. Por ejemplo, [7] presenta el Índice de Acceso Digital (*Digital Access Index*, DAI), que es uno de los más utilizados actualmente. [8] propone un índice para análisis de infraestructura y acceso de TIC en países. Otro enfoque frecuente es el de análisis multivariable. Por ejemplo, [9] realiza un análisis crítico de una serie de índices de medición de brecha digital y propone como alternativa el enfoque multivariable y [10] hace un análisis empírico multivariable de brecha digital entre países. En [11] se describe una serie de retos metodológicos acerca de la medición de la brecha digital; entre estos, las deficiencias o errores frecuentes en el uso de índices compuestos, de los análisis multivariable y de los que incluyen la variable tiempo.

La ciudad como unidad territorial de análisis y modelación se ha vuelto relevante en la investigación de la brecha digital porque los gobiernos y las empresas definen políticas y estrategias dirigidas a ese ámbito. Muchos de los datos requeridos para ello están disponibles en censos nacionales. Al respecto, se han aplicado técnicas de minería de datos sobre censos, aunque en otras áreas distintas del fenómeno de la brecha digital. Algunos trabajos relacionados son, por ejemplo: [12], que hace un estudio de las clases sociales marginadas de Taiwán, y [13], que aplica el algoritmo CART [14] sobre datos del censo de China.

Pocos trabajos relacionados aplican aprendizaje automático al estudio de la brecha digital. Por ejemplo, [15] realiza una medición de la brecha digital internacional aplicando mapas auto-organizantes de Kohonen. Un trabajo similar a la presente investigación es [16], en el cual se muestran resultados sobre el uso de árboles clasificadores J4.8 para análisis y modelación de brecha digital de municipios mexicanos. Sin embargo, en ese trabajo no se usa el algoritmo PART, cuyo desempeño en diversos dominios es mejor que J4.8, según evaluaciones previas realizadas por [3].

J4.8 está inspirado en el reconocido algoritmo C4.5 [5]. A su vez, PART está basado en J4.8; por ello, en esta investigación se espera que algunos de los resultados producidos por PART sean similares a los producidos por J4.8. También, se espera que los modelos PART tengan un número de reglas menor que las producidas por J4.8 y que los valores de *accuracy* de PART sean mayores o iguales que los de J4.8.

3. Datos fuente

Los modelos se crean usando *datasets* experimentales producidos con la base de datos del Censo 2010 de Población y Vivienda de México [17]. Su diccionario de datos [18] es de gran utilidad para estos propósitos. La mayoría de sus ítems son frecuencias absolutas de variables demográficas, socio-económicas y educacionales de los habitantes o de las viviendas. Ofrece 200 atributos, organizados en 14 categorías, de 2,456 municipios. Las categorías, con sus respectivos números de atributos, son: identificación geográfica (9), población (47), fecundidad (1), migración (12), población indígena (13), discapacidad (9), educación (42), características económicas (12), servicios de salud (6), situación conyugal (3), religión (4), hogares censales (6), características de vivienda (35) y tamaño de localidad (1).

Una pequeña cantidad de estos atributos no son frecuencias absolutas, sino promedios; por ejemplo, de: hijos nacidos vivos, ocupantes por casa, años de educación, etc. Esta diferenciación entre frecuencias absolutas y promedios es relevante porque cada uno de estos dos tipos de atributos es procesado en forma distinta para propósitos de representación de la información. Una vez seleccionados los atributos, se crea una serie de *datasets* experimentales. No es solamente un *dataset* porque se usan diferentes tamaños de intervalo que definen diferentes conjuntos de etiquetas de clase para el atributo *target*.

4. Método

Para crear los modelos se realiza un proceso típico de minería de datos con base en los siguientes pasos generales: 1) selección de datos, 2) cálculo y discretización, 3) organización de *datasets*, y 4) creación y evaluación de modelos. Un aspecto destacable en este proceso es la discretización del porcentaje de hogares que tienen Internet en cada municipio, el cual se transforma en una etiqueta nominal que se usa como *target* en los modelos. A continuación, se describe cada paso del proceso.

4.1. Selección de datos

La gran mayoría de las variables del censo son seleccionadas para crear los *datasets* experimentales y solamente se descarta un pequeño conjunto de aquellas. Tres de los nueve atributos geográficos disponibles en la fuente se incorporan en los *datasets* experimentales: *LONGITUD*, *LATITUD* y *ALTITUD*. Los seis atributos descartados son dos identificadores de localidad, dos de municipio y dos de entidad federativa que no contribuyen al descubrimiento de patrones útiles en este caso.

Aunque algunos atributos están correlacionados estadísticamente en la fuente original (p. ej. población *total* entre 0 y 2 años de edad, con población *masculina* entre 0 y 2 años de edad, y población *femenina* entre 0 y 2 años de edad), no se realiza una reducción manual de características basada en análisis de correlación antes de generar los modelos. La razón es que PART y J4.8 omiten automáticamente de los modelos a los atributos que tienen una baja capacidad discriminativa y preservan a los que tienen mayor capacidad. Así, se adopta un enfoque exploratorio para descripción de los perfiles de ciudades considerando diversos niveles de presencia de Internet en hogares.

4.2. Cálculo y discretización de datos

De los atributos seleccionados, se calculan porcentajes por municipio a partir de las frecuencias absolutas, ya sea respecto al número total de hogares o al de habitantes, según corresponda a cada variable. Los atributos que son promedios en los datos fuente (p. ej. *GRAPROES*, promedio de años de educación de los habitantes del municipio) son incorporados en los *datasets* experimentales sin realizar ninguna transformación. El propósito es usar valores normalizados, ya sea porcentajes o

promedios. La excepción es un pequeño número de atributos que se usan como frecuencias absolutas o valores adimensionales; p. ej. las tres coordenadas geográficas. El atributo *POBTOT*, número total de habitantes de municipio, se incorpora directamente en los datasets sin aplicarle transformaciones aritméticas porque se supone que puede ser útil para distinguir perfiles de municipios respecto a la presencia de Internet.

Se realiza un proceso de discretización solamente al porcentaje de presencia de Internet en hogares (*VPH_INTER_%*). El propósito es producir un nuevo atributo, *VPH_INTER_presence*, de tipo nominal, usándolo como *target* para crear modelos PART y J4.8. Las interacciones entre este atributo y todos los demás tienen potencial para describir una parte considerable del fenómeno de la brecha digital.

La discretización de *VPH_INTER_%* se basa en un conjunto de siete tamaños de intervalo que se usan para crear siete datasets. La motivación para considerar diferentes tamaños de intervalo es evaluar su impacto en: 1) el número de clases mayoritarias generadas y sus porcentajes en cada dataset, y 2) accuracy de los modelos. El número de siete tamaños de intervalo es convencional; estos son: 2, 4, 5, 6, 10, 20 y 25 puntos porcentuales. Así, los valores de *VPH_INTER_%* son transformados en las etiquetas de clase de *VPH_INTER_presence*: c_1, c_2, c_3,…, c_n. La Tabla 1 presenta los datasets experimentales con sus respectivos tamaños de intervalo, número de clases y porcentajes de clases mayoritarias.

Table 1. Siete datasets experimentales que contienen a *VPH_INTER_presence* como *target* para modelos.

Dataset No.	Tamaño de intervalo (puntos porcentuales)	No. de clases	Clases mayoritarias
1	2	100/2=50	c_{50} (37.8%) + c_{49} (15.3%) + c_{48} (10.5%) + c_{47} (8.0%) + c_{46} (6.1%) + c_{45} (4.2%) = 82.0%
2	4	100/4=25	c_{25} (53.1%) + c_{24} (18.5%) + c_{23} (10.4%) = 82.0%
3	5	100/5=20	c_{20} (59.1%) + c_{19} (18.6%) + c_{18} (9.6%) = 87.3%
4	6 (de la fórmula de Sturges [18])	68.2/6= 11.3	c_{12} (63.6%) + c_{11} (18.4%) = 82.0%
5	10	100/10=10	c_{10} (77.7%) + c_9 (14.5%) = 92.2%
6	20	100/20=5	c_5 (92.2%)
7	25	100/25=4	c_4 (95.0%)

El tamaño de intervalo 6 es especial porque está determinado con una técnica [19] para calcular tamaño óptimo dependiendo del rango y del número de instancias de la variable en el dataset: $C = rango / (1 + 3.322 \text{ Log } N)$, donde C es el tamaño óptimo, *rango=máximo-mínimo* y N es el número de instancias. En los otros tamaños de

intervalo, la escala es de 0 a 100%; en cambio en este, es del valor mínimo (0.0%) al máximo (68.2%) de la variable *VPH_INTER_%*. Una vez generados los valores nominales de *VPH_INTER_presence*, el atributo *VPH_INTER_%* es eliminado de los datasets para evitar la generación de un patrón trivial que vincule a este valor porcentual con el nominal. La asignación de etiquetas de clase al atributo *target VPH_INTER_presence* está basada en las siguientes reglas: 1) la mejor clase es c_1 y corresponde a los porcentajes más altos de *VPH_INTER_%*; 2) la peor clase es c_n, donde n es igual al número de clases del *dataset* y corresponde a los porcentajes más bajos de la variable.

4.3. Organización de los *datasets*

Se producen siete *datasets* para el *target VPH_INTER_presence*. En cada uno, el número de atributos es 183 (incluyendo al *target*). La mayoría de los predictores son porcentajes que describen características de la población, de la vivienda y de los hogares. Un pequeño número de predictores son promedios municipales de años de educación, de hijos nacidos vivos, de ocupantes por casa, entre otros, así como valores absolutos de población total y de las tres coordenadas geográficas. Los análisis de Pareto de las clases muestran las mayoritarias y la existencia de un desbalance entre las clases de cada *dataset*: la mayoría de las instancias corresponde a las clases con menor presencia de Internet. Esto podría ser relevante porque el desbalance podría sesgar a los modelos, haciéndolos más capaces de reconocer a las clases mayoritarias. Sin embargo, como el principal propósito de estos modelos es descriptivo y no para clasificación automática, esta situación no es tan relevante.

4.4. Creación y evaluación de modelos

Los modelos PART y J4.8 en esta investigación se crean usando algoritmos implementados en el *toolkit* WEKA [4] con sus parámetros de configuración por omisión: para PART, -M 2 -C 0.25 -Q 1; para J4.8, -C 0.25 -M 2. En ambos tipos de modelos, el modo de prueba seleccionado es Validación Cruzada de 10 Subconjuntos (*10-fold cross validation*). Para cada dataset, se crea un modelo PART y un modelo J4.8. Los criterios de aceptación en la etapa de evaluación son: 1) el *accuracy* debe ser mayor o igual que 75%; 2) el estadístico Kappa [20] debe ser mayor o igual que 0.67; 3) el número de clases mayoritarias con base en análisis de Pareto en el correspondiente dataset experimental debe ser mayor que 1. Las razones para estos criterios son: los umbrales de aceptación de *accuracy* y Kappa están definidos de modo convencional con base en valores frecuentemente usados en la literatura. El valor de umbral para el número de clases mayoritarias está establecido considerando que una gran mayoría de los municipios mexicanos pertenece a la peor clase en cada *dataset* y, por lo tanto, necesita aplicarse una restricción al desbalance de las clases.

4.5. Creación de perfiles cuantitativos por clase de municipio

Considerando que el *target* representa clases de municipios con distintos porcentajes de presencia de Internet en hogares, al agrupar las reglas de un modelo PART (o las hojas de un árbol J4.8) con base en sus clases, se está describiendo el perfil de cada clase de municipio. De este modo, el perfil p de una clase específica c_i está constituido por la suma lógica de sus reglas: $p(c_i) = r_1(c_i)$ or $r_2(c_i)$ or $r_3(c_i)$… or $r_n(c_i)$.

5. Resultados

La Tabla 2 resume los resultados de los modelos generados con los siete *datasets*, presentando sus *accuracies*, Kappas y número de reglas y de hojas (los modelos completos pueden solicitarse a los autores). Los modelos PART y J4.8 que cumplen mejor los criterios de aceptación corresponden al *dataset* No. 4. Estos modelos son identificados como P_4 y J_4, respectivamente. De ellos, es preferible P_4 porque, aunque los valores de *accuracy* y de Kappa son muy similares a los de J_4, el número de reglas de P_4 (67) es considerablemente menor que el de hojas de J_4 (120). Esta diferencia implica que P_4 puede describir el fenómeno en forma más compacta que J_4.

Table 2. Resultados de modelos PART y J4.8 sobre siete datasets para el *target VPH_INTER_presence*.

Dataset No.	Modelos PART				Modelos J4.8			
	ID de modelo	Accu-racy %	Kappa	No. de reglas	ID de modelo	Accu-racy	Kappa	No. de hojas
1	P_1	53.1	0.4190	217	J_1	53.2	0.4199	345
2	P_2	70.9	0.5632	110	J_2	73.1	0.5691	205
3	P_3	78.9	0.6513	86	J_3	79.8	0.6645	150
4	P_4	82.2	0.6764	67	J_4	82.4	0.6797	120
5	P_5	91.7	0.7778	28	J_5	91.2	0.7639	67
6	P_6	98.2	0.8733	10	J_6	97.8	0.8428	17
7	P_7	99.2	0.9152	8	J_7	98.8	0.8786	9

Del modelo P_4 pueden obtenerse los perfiles de las ciudades (municipios), que se muestran a continuación:

$p(c_1) = r_{50}$ (la clase tiene solamente una regla).
$p(c_2) =$ (la clase no existe en el *dataset* y, por lo tanto, tampoco en los modelos).
$p(c_3) =$ (la clase tiene solo una instancia, por lo cual PART no la modela).
$p(c_4) = r_{67}$ (la clase tiene solamente una regla).
$p(c_5) = r_{60}$ (la clase tiene solamente una regla).
$p(c_6) = r_{59}$ or r_{63}.
$p(c_7) = r_{57}$ or r_{62} or r_{64}.
$p(c_8) = r_{53}$ or r_{61} or r_{66}.
$p(c_9) = r_{30}$ or r_{51} or r_{54} or r_{56}.

$p(c_{10}) = r_{14}$ or r_{23} or r_{31} or r_{36} or r_{41} or r_{43} or r_{48} or r_{52} or r_{55} or r_{58}.
$p(c_{11}) = r_7$ or r_{12} or r_{18} or r_{19} or r_{21} or r_{22} or r_{26} or r_{27} or r_{28} or r_{29} or r_{32} or r_{35} or r_{39} or r_{40} or r_{42} or r_{44} or r_{45} or r_{46} or r_{65}.
$p(c_{12}) = r_1$ or r_2 or r_3 or r_4 or r_5 or r_6 or r_8 or r_9 or r_{10} or r_{11} or r_{13} or r_{15} or r_{16} or r_{17} or r_{20} or r_{24} or r_{25} or r_{33} or r_{34} or r_{37} or r_{38} or r_{47} or r_{49}.

Como ejemplo de las reglas, se comentan tres de estas a continuación:

r_{50} de c_1: *P_3A5_M_porcent*<=1.9 AND *P3A5_NOA_M_porcent*<=0.5 AND *ALTITUD*<= 2249: c_1 (2).

Significa: Si la población entre 3 y 5 años de edad de sexo masculino es menor o igual al 1.9% y la población de 3 a 5 años de sexo masculino que no asiste a la escuela es menor o igual 0.5% y la altitud geográfica del municipio es menor o igual a 2,249 metros, entonces el municipio es de clase 1, teniendo entre 66% y 72% de presencia de Internet en hogares (ocurre en 2 de los municipios, sin excepción).

r_1 de c_{12}: *VPH_PC_porcent*<=11.9 AND *VPH_PC_porcent*<=8.5: c_{12} (1154).

Significa: si los hogares con PC son menores o iguales al 11.9% y además son menores al 8.5%, entonces el municipio es de clase 12, teniendo entre 0% y 6% de presencia de Internet en hogares (ocurre en 1,154 de los municipios, sin excepción).

r_{46} de c_{11}: *VPH_PC_porcent*<=31.6 AND *PNACOE_porcent*<=1.9: c_{11} (5/1).

Significa: Si los hogares con PC son menores o iguales a 31.6% y la población nacida en otra entidad es menor o igual a 1.9%, entonces el municipio pertenece a la clase 11, teniendo entre 6% y 12% de presencia de Internet en hogares (ocurre en 5 de los municipios, excepto en uno que tiene las condiciones descritas pero no pertenece a la clase 11).

En cada perfil, son más significativas las reglas que tienen los mayores valores de soporte (*support*, *coverage*) y de confianza (*confidence*). El soporte se calcula como: $(a/n)*100$ donde a es el número de instancias que tienen las condiciones descritas por la regla (pertenezcan, o no, a su clase) y n es el número total de instancias en el dataset (2,456). La confianza se calcula: $((a-b)/a)*100$, donde b es el número de instancias que no pertenecen a esa clase, es decir, las excepciones de la regla. Por ello, las reglas con mayor soporte describen a un mayor número de municipios y representan a los patrones más sólidos del fenómeno de la brecha digital en el país.

6. Discusión

El uso de la variable *VPH_INTER_presence* como *target* se justifica porque representa a las diversas clases de municipios y es el principal medio para descubrir los patrones socio-demográficos de cada clase. Los predictores útiles para describir los patrones son descubiertos automáticamente por los algoritmos evaluados. La Tabla 2 muestra que las *accuracies* y Kappas son altamente similares entre cada par de modelos PART y J4.8 para cada *dataset*. La diferencia más significativa entre cada

par es el número de reglas y de hojas, respectivamente. Las hojas en los árboles J4.8 constituyen los elementos terminales, es decir, los *consecuentes*, de reglas que son comparables a las reglas de los modelos PART. En ambos modelos, las reglas representan conocimiento descubierto no trivial. La medida de soporte de cada regla representa qué tan frecuente es ese patrón en el *dataset*. Por ello, al identificar las reglas con mayor soporte en P_4 se obtienen los patrones más significativos del fenómeno.

La regla con mayor soporte (1,154 / 2,456 * 100 = 47.0%) en P_4 es r_1; por ello, puede afirmarse que el patrón más significativo en el análisis de la brecha digital en municipios de México consiste en que: *los municipios con menor presencia de Internet (de 0 a 6% de los hogares) se caracterizan principalmente porque la presencia de PC es menor al 8.5%* (esta regla ha sido simplificada eliminando el valor 11.9%).

Contrastando los resultados con los de trabajos relacionados, la mayoría de estos últimos no son comparables porque aplican técnicas distintas para análisis y modelación y porque su unidad de análisis es el nivel de país, no el de ciudad. El único trabajo comparable es [15] porque en él se usan árboles clasificadores J4.8 con datos de municipios. Al comparar los dos trabajos, se observa que J4.8 y PART resultan útiles para este dominio porque ofrecen modelos descriptivos capaces de guiar el descubrimiento de explicaciones causales. Sin embargo, también se confirma que PART ofrece valores de *accuracy* y Kappa que son similares a los de J4.8, pero con PART los modelos tienen menos reglas y, en algunos casos, estas son más compactas. En ambos algoritmos, los perfiles de clases de municipios con distintos grados de presencia de Internet pueden generarse agrupando las reglas que corresponden a cada clase.

7. Conclusiones y trabajo futuro

Este artículo ha propuesto una metodología basada en la aplicación de los algoritmos PART y J4.8 para investigar la brecha digital de ciudades (particularmente, *municipios*) describiendo interacciones entre la presencia de Internet en hogares y una serie de aspectos demográficos, socio-económicos y educacionales de los habitantes. Con base en resultados empíricos, nuestra postura es que tanto PART como J4.8 son útiles para el análisis y la modelación del fenómeno de la brecha digital de ciudades porque: 1) las *accuracies* y Kappas de los modelos son satisfactorias sobre la base de criterios generalmente aceptados en el campo del aprendizaje automático, y 2) se descubren variables clave para la descripción y potencial explicación del fenómeno. Sin embargo, como era esperado con base en trabajos relacionados acerca de estos dos algoritmos, las reglas producidas por PART son menos y más cortas que las producidas por J4.8.

Nuestras principales contribuciones científicas en esta investigación son: 1) la propuesta de una técnica novedosa basada en PART y J4.8 para investigar el fenómeno de la brecha digital de ciudades, aplicable sobre datos de censos nacionales

de diversos países, y 2) el descubrimiento de patrones no triviales de este fenómeno en el contexto específico de los municipios de México.

En trabajo futuro se debería generar y evaluar otros modelos PART y J4.8 usando como *targets* otros datos respecto a presencia de bienes y servicios de TIC en hogares, tales como PC, teléfono fijo o teléfono celular. Los atributos de entidad federativa podrían incluirse dentro de los predictores. También, podría realizarse un análisis comparativo (*benchmarking*) entre estos dos algoritmos y uno o dos más usando los mismos datasets. Finalmente, deberían crearse otros modelos usando datos de censos de otros países.

Referencias

1. Organisation for Economic Co-operation and Development (OECD). Glossary of Statistical Terms. http://stats.oecd.org/glossary/index.htm. Accesado 01/Abr/2013.
2. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques (3rd ed.). The Morgan Kaufmann Series in Data Management Systems, Waltham (2011).
3. Frank, E., Witten, I.H.: Generating Accurate Rule Sets Without Global Optimization. In: 15th IMLS International Conference on Machine Learning, pp. 144–151. Morgan Kaufmann Publishers Inc., San Francisco (1998).
4. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Morgan Kaufmann Series in Data Management Systems, Burlington (2011).
5. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993).
6. Chinn, D.M., Fairlie, W.R. The Determinants of the Global Digital Divide: A Cross-country Analysis of Computer and Internet Penetration. Oxford Economic Papers 59(1), 16–44 (2007).
7. International Telecommunications Union (ITU). World Telecommunication Development Report. Access Indicators for the Information Society. Accesado 13/Ene/2013. http://www.itu.int/ITU-D/ict/publications/wtdr_03/material/DAI.pdf
8. Hanafizadeh, M.R., Saghaei, A., Hanafizadeh, P. An Index for Cross-country Analysis of ICT Infrastructure and Access. Telecommunications Policy 33, 385–405 (2009).
9. Bruno, G., Esposito, E., Genovese, A., Gwebu, K.L. A Critical Analysis of Current Indexes for Digital Divide Measurement. The Information Society 27, 16–28 (2011).
10. Iliadis, M.S., Paravantis, J.A. A Multivariate Cross-country Empirical Analysis of the Digital Divide. In: Procs. of ISCC '11 2011 IEEE Symposium on Computers and Communications, pp. 785–788. Washington, D.C., U.S.A. (2011).
11. Vehovar, V., Sicherl, P., Husing, T., Dolnicar, V.: Methodological Challenges of Digital Divide Measurements. The Information Society 22, 279–290 (2006).
12. Chang, C.J., Shyue, S.W. A Study on the Application of Data Mining to Disadvantaged Social Classes in Taiwan Population Census. Expert Systems with Applications 36, 510–518 (2009).
13. Sheng, B., Gengxin, S.: Data mining in census data with CART. In: Procs. of 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), pp. 260– 264. Coll. of Inf. Sci. & Eng., Qingdao Univ., Qingdao, China (2010).
14. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and regression trees. Wadsworth, Monterey (1984).

15. Deichmann, J.I., Eshghi, A., Haughton, D., Woolford, S.: Measuring the International Digital Divide: an Application of Kohonen Self-organising Maps. International Journal on Knowledge and Learning. 3(6), 552–575 (2007).
16. Coria, S.R., Mondragón-Becerra, R., Pérez-Meza, M., Ramírez-Vásquez, S.K., Martínez-Peláez, R., Barragán-López, D., Ávila-Barrón, O. CT4RDD: Classification Trees for Research on Digital Divide. Expert Systems with Applications (2013). DOI: 10.1016/j.eswa.2013.04.002. In press.
17. Instituto Nacional de Estadística y Geografía (INEGI). Base de Datos por Localidad del Censo Nacional de Población y Vivienda 2010. Accesado 01/Abr/2013. http://www.censo2010.org.mx
18. Instituto Nacional de Estadística y Geografía (INEGI). Conformación de la Base de Datos por Localidad del Censo Nacional de Población y Vivienda 2010. Accesado 01/Abr/2013. http://www.censo2010.org.mx
19. Sturges, H.A. The Choice of a Class Interval. Journal of the American Statistical Association 21(153), 65–66 (1926).
20. Cohen, J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 20, 37–46 (1960).