SIMULACIÓN DEL RAZONAMIENTO EN EL PROCESO DE IDENTIFICACIÓN BOTÁNICA BASADO EN REDES BAYESIANAS

SONIA I. MARIÑO Y PEDRO L. ALFONZO
Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura
Universidad Nacional del Nordeste
simarinio@yahoo.com-plalfonzo@hotmail.com

Fecha recepción: Agosto 2015 - Fecha aprobación: Abril 2016

RESUMEN

Se simula el proceso de identificación de especies vegetales de un dominio basado en técnicas de Minería de Datos, específicamente se optaron por modelos bayesianos. Se evalúan algunas soluciones inferenciales, se eligieron dos métodos comparando los resultados del aprendizaje y validación. Se exponen y justifican los resultados obtenidos en las simulaciones.

PALABRAS CLAVE: Inteligencia Artificial – Redes Bayesianas – Simulación – Minería de Datos – Dominios Botánicos.

ABSTRACT

The identification process of plant species is simulated using Data Mining techniques, using Bayesian models. It have been designed some solutions, using two methods in order to improve the learning and evaluation results of models. The simulation results are presented and justified.

KEYWORDS: Artificial Intelligence – Bayesian Networks – Simulation – Data Mining – Botanical Domains.

1. INTRODUCCIÓN

La Ciencia de la Computación (Red UNCI, 2006; ACM, 2013) se constituye con la convergencia de distintas aéreas del conocimiento, siendo una de ellas Inteligencia Artificial (IA). Fundamentalmente, los paradigmas que la constituyen son el simbólico y el conexionista.

Los paradigmas de la IA se desarrollan a través de métodos y herramientas en un intento de emular la mente de los sujetos cognoscentes, en la resolución de complejos problemas que requieren de la toma de decisiones.

En la IA, un tema de interés radica en la representación del conocimiento, es decir, aquellos mecanismos para representar y administrar la información de un dominio de conocimientos. (Castillo, Gutiérrez y Hadi, 1998; Rusell y Norving, 2004).

Siguiendo a (Castillo *et al*., 1998, p. 17), "los esquemas de representación resultantes deberán permitir una búsqueda o una operación eficiente de los mecanismos de inferencia".

El modelo descubrimiento del conocimiento (KDD o *Knowledge Discovery in Databases*, por sus siglas en inglés), surgió en el año 1996 para establecer las etapas principales de un proyecto de explotación de información (Frawley, Piatetsky-Shapiro y Matheus, 1992; Matheus, Chan y Piatetsky-Shapiro, 1993). Establece a la Minería de Datos (MD) como la etapa del proceso en la cual se realiza la extracción de patrones a partir de los datos (Moine, Haedo, Gordillo, 2011a, 2011b).

Su objetivo principal es proveer herramientas para automatizar el proceso de análisis de datos y el artesanal proceso estadístico de selección de hipótesis. El objetivo subyacente al concepto de KDD, es diferenciarse de la Minería de Datos (*Data Mining*), entendiendo a ésta como la actividad de aplicar distintos algoritmos en los datos para obtener patrones, del proceso necesario para generar conocimiento a partir de los datos. Es decir, entender a la Minería de Datos como un subproceso que integra a un proceso general destinado a obtener patrones de conocimiento (Martins, Pesado y Garcia-Martinez, 2014).

La Minería de Datos y el descubrimiento del conocimiento (KDD) contribuyen a la toma de decisiones tácticas y estratégicas, proporcionando un sentido automatizado para la generación de conocimiento y por ende a la toma acertada de decisiones y su aplicación es amplia en las diferentes ramas de la investigación (Valcárcel Asencios, 2004).

Moine *et al*. (2011a, 2011b,) indican que en la actualidad "el término KDD y Minería de Datos se utilizan indistintamente para hacer referencia al proceso completo de descubrimiento de conocimiento".

Existen diferentes métodos comprendidos en el KDD, entre ellos los cuantitativos, los probabilísticos y los estadísticos (Valcárcel Asencios, 2004). Entre los métodos de clasificación se mencionan a los Bayesianos.

Una Red Bayesiana se define como un grafo dirigido acíclico que consta de: i) Un conjunto de nodos, uno por cada variable aleatoria del "mundo"; ii) Un conjunto de arcos dirigidos que conectan los nodos; si hay un arco de X a Y, indica que X es un padre de Y (padres (X) denota el conjunto de variables aleatorias padres de X); iii) Cada nodo X_i contiene la distribución de probabilidad condicional P (X_i | padres (X_i)) (Ruiz Reina, 2005).

Por otra parte, una Red Bayesiana es una representación gráfica de dependencias del razonamiento probabilístico, en la cual los nodos representan variables aleatorias y los arcos representan relaciones de dependencia directa entre las variables. Es decir, modelan un fenómeno mediante un conjunto de variables y las relaciones de dependencia entre ellas. Entre los trabajos que establecen sus fundamentos se mencionan Pearl

(1988), Jensen (1996) y Jensen (2001). Definido un modelo, se puede hacer inferencia bayesiana; es decir, estimar la probabilidad posterior de las variables no conocidas, en base a las variables conocidas (Sucar, 2006). El aprendizaje de las redes bayesianas consiste en inferir, a partir de los datos, un modelo, estructuras y parámetros (Corso y Gibellini, 2011).

Un clasificador bayesiano se puede considerar un caso especial de una Red Bayesiana, que dispone de una variable especial representando la clase y las restantes variables son los atributos (Sucar, 2006). Para estimar estas probabilidades se han propuesto diversos clasificadores bayesianos (Corso y Gibellini, 2011).

Sucar (2006) sostiene que éste obtiene la probabilidad posterior de cada clase C_i, usando la regla de Bayes, como el producto de la probabilidad *a priori* de la clase por la probabilidad condicional de los atributos (E) dada la clase, dividido por la probabilidad de los atributos.

En Biología se recurre a diversos modelos computacionales para simular el proceso de identificación de entidades. En Mariño y Dematteis (2013) se expone un relevamiento aplicado al mencionado dominio de conocimiento.

En este trabajo, se presentan el método, los resultados y consideraciones preliminares referentes al diseño y evaluación de razonamientos que simulan la toma de decisiones de especialistas de un dominio de la Botánica utilizando modelos de Redes Bayesianas en un entorno de Minería de Datos. Específicamente se validan en la identificación de especies de Mirtáceas del Nordeste Argentino. Como se mencionó en Mariño (2001), estos taxones son un importante componente de las comunidades vegetales en el Microsistema Iberá (Corrientes) y en el Predio Guaraní (Misiones) y dado que muchas de ellas tienen valor económico, ya sea como frutales u ornamentales y en la medicina popular, se las seleccionó para el desarrollo del modelo que se expone a continuación. Por otra parte el herbario del Instituto de Botánica del Nordeste (CTES) posee una nutrida colección de Mirtáceas y un número importante de esos ejemplares han sido identificados por especialistas en la familia, lo que hace que las nuevas identificaciones puedan ser corroboradas por comparación con testigos fidedignos.

2. MÉTODO

A continuación se expone el método desarrollado para la simulación y validación del proceso de identificación, basado en CRISP-DM (Chapman, Clinton, Keber, Khabaza, Reinartz, Shearer y Wirth, 2000).

Fuente de información y conjunto de entrenamiento

El conjunto de datos utilizado en el entrenamiento y comprobación de los modelos analizados se obtuvo del herbario del Instituto de Botánica del

Nordeste (CTES). Éste posee una nutrida colección de Mirtáceas y un número importante de esos ejemplares han sido identificados por especialistas en la familia, lo que hace que las nuevas identificaciones puedan ser corroboradas por comparación con testigos fidedignos.

Se trató de incorporar la mayor cantidad de caracteres posibles para facilitar la identificación de ejemplares en los que, por no poseer frutos, se desconoce el tipo de embrión, carácter de gran importancia para la identificación en esta familia.

Se consideran como variables evidenciales los caracteres seleccionados por el especialista de conocimiento. La variable objetivo asume los distintos valores de 30 especies de Mirtáceas (Mariño, 2001). En el Anexo se detalla, en la TABLA 4 las variables evidenciales y sus posibles valores y en la TABLA 5 los valores que puede asumir la variable objetivo (nombre científico de la especie a identificar).

Análisis y comprensión del negocio

En esta fase se entendió el dominio de aplicación, específicamente consistió en simular con métodos inteligentes la identificación de especies vegetales pertenecientes a la familia Myrtaceae del NE Argentino. En Botánica el proceso de clasificación computacional se asocia al proceso de identificación de taxones.

Selección y preparación de los datos

Preparación de datos. Esta fase involucró aquellas actividades para construir el conjunto de datos final, el cual será utilizado como entrada a las herramientas de modelado. Las tareas se pueden aplicar múltiples veces y sin un orden pre-establecido. Incluyen extracción, transformación y carga, proceso conocido como ETL. Extraídos los datos de la fuente de la información se procedió a su transformación/conversión a un formato legible por la herramienta de MD.

Estimación de estadísticos sobre los atributos. Registrados los datos, desde la herramienta de MD se procedió a reconocer los atributos y computar algunas estadísticas básicas sobre cada atributo durante el análisis de los datos. Dado que el conjunto de datos seleccionados son atributos continuos/numéricos, se visualizan valores mínimo, máximo, media, desviación estándar, entre otros.

Modelado

Pueden seleccionarse y aplicarse distintas técnicas de modelado. Se optó por los clasificadores bayesianos, se basan en métodos inductivos –aplicados sobre los patrones de datos proporcionados por los expertos humanos-. Lo expuesto requiere disponer de numerosos ejemplos para lograr aprendizajes significativos. Es decir, pocos ejemplos dificultan la inferencia de los rasgos o evidencias que distinguen las distintas especies.

En esta fase se eligió el conjunto de datos sobre el cual se modelaría y simularía el problema. Se trabajó con un especialista del dominio, quien

seleccionó los patrones de datos y las variables evidenciales implicadas (Anexo). Se prepararon los datos según el formato requerido por Weka, software de modelado y simulación computacional.

Diseño de modelos, experimentación y evaluación

Esta fase consistió en definir la tarea de Minería de Datos. Se optó por la Clasificación basada en Redes Bayesianas, siendo el algoritmo de entrenamiento seleccionado el denominado BayesNet.

- Se seleccionó como algoritmos de búsqueda K2. Se ha propuesto para localizar redes Bayesianas de alta calidad. El algoritmo comienza su proceso con la definición de la red más simple posible, es decir, una red sin aristas, y supone que los nodos están ordenados. Para cada variable *Xi*, el algoritmo agrega a su conjunto de padres Π*i,* el nodo con número menor que *Xi,* que conduce a un máximo incremento en calidad correspondiente a la medida de calidad elegida para el proceso de búsqueda. Se repite el proceso hasta que no se incremente la calidad o se llega a una red completa (Castillo *et al.*, 1998).
- Se seleccionaron los siguientes estimadores:
  - BMAEstimator: Estima tablas de probabilidad condicional de una red de Bayes utilizando el Modelo de promedio Bayesiano (BMA).
  - SimpleEstimator: Estima directamente tablas de probabilidad condicional de una red de Bayes.
- Se diseñaron diversos experimentos, modificándose en cada uno de ellos, los estimadores mencionados.
- Se procedió a la simulación, interpretación y valoración de los resultados, confrontándose con aquellos valores meta esperados y determinados por el especialista del dominio.
- Se determinaron las métricas o medidas de calidad. Para evaluar la efectividad en la identificación de especies, se optaron por las siguientes métricas de precisión: Kappa *Statistic*, el Error Medio Absoluto (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el parámetro Alpha que indica el valor de la tasa de aprendizaje.
  - Kappa *Statistic* es una medida de concordancia entre las categorías pronosticadas por el clasificador y las categorías observadas, considera las posibles concordancias debidas al azar. La valoración del índice Kappa está dada:
    - Si el valor es 1: Concordancia perfecta.
    - Si el valor es 0: Concordancia debida al azar.
    - Si el valor es negativo: Concordancia menor que la que cabría esperar por azar.
  - RMSE o Root *Mean Squared Error* mide las diferencias entre los valores brindados por un modelo o un estimador y los valores realmente observados. Es una medida de

precisión, permite comparar diferentes errores de predicción de un mismo conjunto de datos, dado que es dependiente de la escala muestra.

- MAE o *Mean Absolute Error* se define como el Error Absoluto de una Media a la diferencia entre el valor medio obtenido y el hallado en esa media. El promedio de error absoluto es la suma de los errores absolutos de clasificación en cada uno de los sujetos llevados a promedio. El clasificador que arroje mayor cifra (mayor a 0.1) define un error de clasificación alto, por lo cual no se debe considerar sobre aquellos que arrojen una cifra menor.

- Parámetros de selección de los modelos. Se determinaron los siguientes valores de los indicadores para elegir los modelos: Clasificación correcta > 90 % instancias; Clasificación incorrecta < 10 % instancias; MAE < 0.1 ideal; Kappa *statistic* > 0.79, > 0.9 ideal; RMSE < 0.3, < 1 ideal. El clasificador que cumpla con estas especificaciones, se considera representativo para su integración en un sistema inteligente que emulará al especialista del dominio ante nuevos casos de identificación. Si los modelos proporcionan un mismo valor en la métrica MAE, se considerará aquel menor valor de RMSE para seleccionar el modelo más representativo.
- Comprobación. Entrenado el conjunto de datos, se procedió a su verificación utilizando la técnica denominada *Percent Split* en la herramienta de MD seleccionada. Esta opción divide los datos en dos grupos, el porcentaje especificado representa las instancias utilizadas para construir el modelo, y éste es evaluado respecto a las restantes. Cuando el número de instancias es suficientemente elevado, esta opción es suficiente para estimar con precisión las prestaciones del clasificador en el dominio.

Implementación.

Esta fase implica el despliegue de la solución tecnológica. Los modelos entrenados y validados pueden ser utilizados desde la herramienta. Sin embargo, para asegurar el empleo del simulador de identificación que emula el razonamiento de los especialistas del dominio se procederá al diseño de una interfaz de usuario.

3. DESARROLLO

A continuación se exponen los resultados distinguiendo la aplicación de un proceso de KDD utilizando una herramienta de Minería de Datos.

Se optó por un método de clasificación, dado que el objetivo es determinar la especie botánica a partir de un conjunto de variables evidenciales seleccionadas por el especialista en el dominio de conocimiento.

Además, se eligieron técnicas de aprendizaje supervisado para contrastar el aprendizaje en máquinas con respecto al valor esperado de los valores que puede asumir la variable objetivo.

Se procedió a la construcción de los modelos de clasificación siguiendo el método especificado en la fase Diseño de modelos, experimentación y evaluación de la sección anterior. El procedimiento descripto, se realizó tanto para el estimador BMAEstimator y SimpleEstimator, observándose los resultados obtenidos en la TABLA 1. Para fundamentar la elección de los modelos se toma como referencia una métrica que permite evaluar la calidad o nivel de confianza de un modelo entrenado como el estadístico Kappa. Éste mide el nivel de predicción respecto a la variable objetivo. Además, se consideraron las instancias clasificadas correctamente y los errores asociados al clasificador entre los que se mencionan las métricas MAE y RMSE.

TABLA 1. Fase de Entrenamiento del clasificador Bayes Network. Modelos diseñados

Estimador	Alpha	Corr	Kappa	MAE	RMSE
BMAEstimator	0.5	423	0.9974	0.0007	0.0156
SimpleEstimator	0.5	423	0.9974	0.0002	0.0123

Evaluados los modelos de la TABLA 1, se infiere que ambos estimadores detectan similares valores en las métricas seleccionadas para determinar el mejor comportamiento; siendo SimpleEstimator el más representativo al obtener una menor tasa de error MAE. Además, en ambos se detectó la identificación errónea de un caso correspondiente a la especie *Psidium kennedyanum* (cod. 60) como *Psidium guineense* (cod. 56). Se hipotetizaría la definición de otras variables que permita distinguir las especies del mismo género. En la FIGURA 1 se ilustra la predicción estimada en el proceso de entrenamiento.

En las TABLAS 2 y 3, las columnas refieren a los distintos modelos generados; % ICM: expresa el porcentaje de patrones sobre el total (424 en este caso) reservados para construir el modelo; I. Eval. indica el número de instancias sobre las que se aplica la evaluación del modelo aprendido.

La utilidad de un modelo inferencial se justifica si brinda resultados aceptables ante nuevos casos. Por lo expuesto, se procedió a la comprobación de los modelos –actividad comprendida en la fase diseño de modelos, experimentación y evaluación–, utilizando la técnica PercentSplit para su verificación. Se determinó un comportamiento válido ante nuevos casos, situación ilustrada al comparar los valores promedios del estadístico Kappa de las TABLAS 2 y 3, se observa que el estimador SimpleEstimator (TABLA 3), proporciona un mejor comportamiento en el proceso de aprendizaje y testeo. Otras métricas de error que permiten afirmar la elección del estimador son: MAE, RMSE, inversamente proporcional al estadístico Kappa, dado que al aumentar este último las métricas de error disminuyen considerablemente.

TABLA 2. Fases de entrenamiento y comprobación. BMAEstimator

Modelos	% ICM	I. Eval.	Clasificación correcta	Kappa	MAE	RMSE
Pro-1	10	382	73	0	0.0602	0.1729
Pro-2	20	339	68	0	0.0598	0.1722
Pro-3	30	297	58	0	0.0594	0.1720
Pro-4	40	254	49	0	0.0593	0.1720
Pro-5	50	212	40	0	0.0592	0.1721
Pro-6	60	170	34	0	0.0592	0.1721
Pro-7	70	127	26	0	0.0591	0.1717
Pro-8	80	85	15	0	0.0592	0.1721
Pro-9	90	42	7	0	0.0589	0.1713

TABLA 3. Fases de entrenamiento y comprobación. SimpleEstimator

Modelos	% ICM	I.Eval.	Clasificación correcta	Kappa	MAE	RMSE
Pro-1	10	382	333	0.8582	0.0085	0.0832
Pro-2	20	339	324	0.9512	0.0029	0.0477
Pro-3	30	297	282	0.9444	0.0033	0.0514
Pro-4	40	254	248	0.9741	0.0015	0.0368
Pro-5	50	212	209	0.9845	0.0009	0.0270
Pro-6	60	170	167	0.9807	0.0011	0.0297
Pro-7	70	127	124	0.974	0.0015	0.0327
Pro-8	80	85	83	0.9742	0.0015	0.0334
Pro-9	90	42	42	1	0.0055	0.0173

Se analizaron las matrices de confusión obtenidas en cada uno de los procesamientos sintetizados en las TABLAS 2 y 3. A modo de ejemplificar, una representación gráfica de las mismas se observa en la FIGURA 2, al utilizar la técnica PercentSplit con un valor igual a 70% se observan los errores predichos. En el caso ilustrado, la inferencia realizada estima incorrectamente un ejemplar identificado que correspondería a *Plinia rivularis* (cod. 52) como *Eugenia hyemalis var. marginata (cod.* 14) y dos ejemplares que corresponden a la especie *Eugenia moraviana (cod.* 18) como *Eugenia repanda* (cod. 26). Cabe aclarar que las FIGURAS 1 y 2 se corresponden con las estimaciones obtenidas en la matriz de confusión asociada al proceso de entrenamiento y testeo.

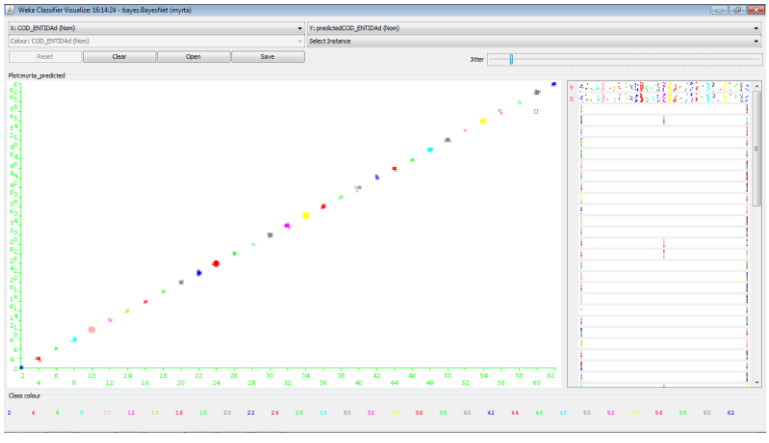

FIGURA 1. Predicciones del proceso de identificación utilizando el conjunto de entrenamiento.

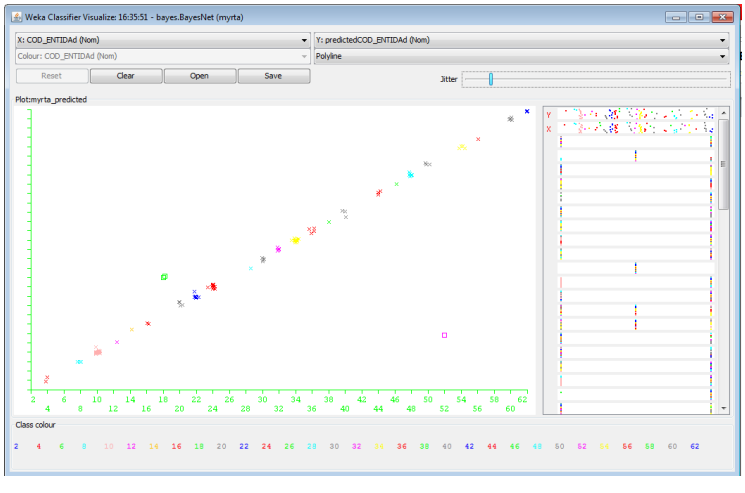

FIGURA 2. Predicciones del proceso de identificación aplicando la técnica PercentSplit.

A partir de la TABLA 3, en la FIGURA 3, se observa que la mayor instancia clasificada correctamente, respecto del total utilizadas en el entrenamiento es 333 y la más baja es 42, cuando el tamaño de los patrones evaluados se encuentra en 10% y 90% respectivamente. Estas instancias aumentan a medida que se incrementa el porcentaje de los patrones de evaluación y disminuyen cuando decrecen.

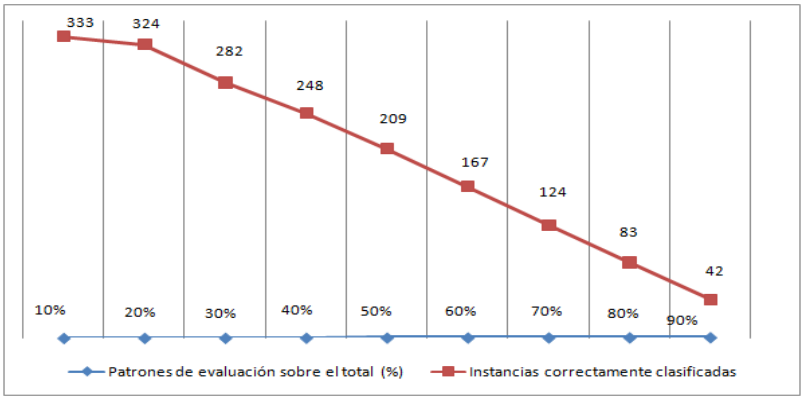

FIGURA 3. Instancias correctamente clasificadas.

A partir de la TABLA 3, en la FIGURA 4, se muestra el valor que asume Kappa de acuerdo a los diferentes patrones de evaluación. Se observa que la mayor precisión del estadístico es 1 y la más baja es 0.8582, cuando el entrenamiento es del 10%; y dependiendo de las instancias correctamente clasificadas, la precisión se incrementa o disminuye.

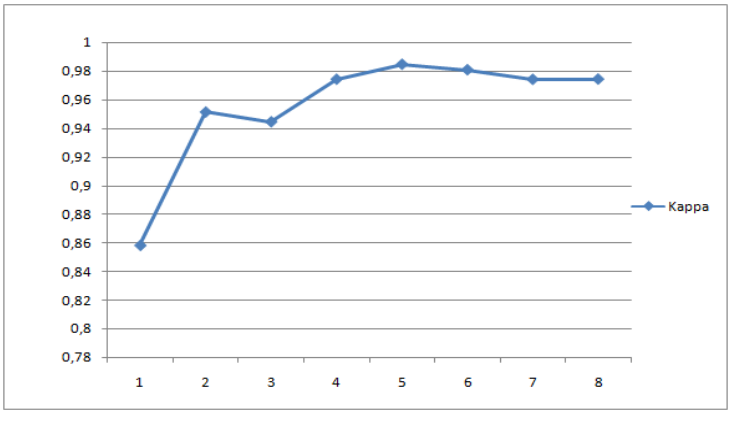

FIGURA 4. Variación de estadístico Kappa.

En este trabajo se ha enfocado el análisis de rendimiento de técnicas de inferencia basadas en Redes Bayesianas. Se han seleccionado dos estimadores: SimpleEstimator y BMAEstimator, que en el proceso de entrenamiento brindan métricas similares. La diferenciación se puede observar en procesos de validación utilizando la denominada técnica PercentSplit (TABLAS 2 y 3). Un análisis de los resultados obtenidos permitiría determinar que SimpleEstimator (TABLA 3) presenta mejores estimaciones en procesos de identificación de las especies seleccionadas.

4. CONCLUSIONES

Se ha expuesto un método de Minería de Datos enfocado a la resolución de un problema en Botánica empleando modelos de Redes Bayesianas.

Se han diseñado diversos modelos de aprendizaje y se han validado a fin de seleccionar aquel que mejor representaría al dominio de conocimiento elegido. Las simulaciones realizadas han permitido verificar el buen comportamiento de los métodos inferenciales seleccionados y experimentados en este trabajo.

Los distintos experimentos realizados permiten afirmar un mejor comportamiento del estimador SimpleEstimator. Lo expuesto se fundamenta en que aun aplicando distintos test/técnicas de prueba se obtiene una mejor estimación con el conjunto de datos desconocidos en el proceso de aprendizaje y utilizados para testear.

Como futuros trabajos se propone: analizar el grado de identificación de las especies botánicas utilizando otros clasificadores; estudiar individualmente los patrones erróneamente clasificados sobre la información proporcionada por la Matriz de Confusión y contrastar los resultados con un especialista en el dominio. Además, se avanzará en el diseño y desarrollo de una interfaz de usuario para asegurar la transferencia del modelo de razonamiento que emula a los especialistas del dominio seleccionado.

5. REFERENCIAS

ACM o Association for Computing Machinery (2013): “COMPUTER SCIENCE CURRICULA”. Documento en línea. Disponible en: http://ai.stanford.edu/users/sahami/CS2013/strawman-draft/cs2013-strawman.pdf.

CASTILLO, E.; GUTIÉRREZ, J.M.; HADI, A.S. (1998): “SISTEMAS EXPERTOS Y MODELOS DE REDES PROBABILÍSTICAS”. Ed. Academia Española de Ingeniería, España.

CHAPMAN, P., CLINTON, J., KEBER, R., KHABAZA, T., REINARTZ, T., SHEARER, C., WIRTH, R. (2000). CRISP-DM 1.0 Step by step BIguide. Edited by SPSS. Documento en línea. Disponible en: http://www-staff.it.uts.edu.au/~paulk/teaching/dmkdd/ass2/readings/methodology/CRISP WP-0800.pdf

CORSO, C. L.; GIBELLINI, F. (2011): “APLICACIÓN DE REDES BAYESIANAS USANDO WEKA”. CACIC 2011 - XVII CONGRESO Argentino de Ciencias de la Computación.

FRAWLEY, W. J., PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. (1992): "KNOWLEDGE DISCOVERY IN DATABASES: AN OVERVIEW". AI Magazine Vol. 13, num. 3, 57-70. Documento en línea. Disponible en: http://aaaipress.org/ojs/index.php/aimagazine/article/viewFile/1011/929.

GARCÍA MORATE, D. (2008): "MANUAL DE WEKA". Disponible en: http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf.

JENSEN, F. V. (1996): "AN INTRODUCTION TO BAYESIAN NETWORKS". London, UCL Press.

JENSEN, F. V. (2001): "BAYESIAN NETWORKS AND DECISION GRAPHS". New York, Springer.

MARIÑO, S. (2001): Tesis de Maestría en Informática y Computación. Universidad Nacional del Nordeste.

MARIÑO, S.; DEMATTEIS, M. (2013): "REVISIÓN DE SOLUCIONES DE TECNOLÓGICAS INTELIGENTES EN BIOLOGÍA". Telematique, Vol. 13, No 1, p. 30-50.

MARTINS, S.; PESADO, P.; GARCÍA-MARTÍNEZ, R. (2014): "PROPUESTA DE MODELO DE PROCESOS PARA UNA INGENIERÍA DE EXPLOTACIÓN DE INFORMACIÓN: MOPROPEI". Revista Latinoamericana de Ingeniería de Software, 2(5): 313-332, ISSN 2314-264.

MATHEUS J. C., CHAN, P. K. PIATETSKY-SHAPIRO, G. (1993): "SYSTEMS FOR KNOWLEDGE DISCOVERY IN DATABASE". IEEE, TKDE, special issue on Learning & Discovery in Knowledge-Based Databases, 1-16, Documento en línea. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.169&rep=rep1&type=pdf.

MOINE, J. M.; HAEDO, A.; N.; GORDILLO, S. (2011a): "ESTUDIO COMPARATIVO DE METODOLOGÍAS PARA MINERÍA DE DATOS". XIII Workshop de Investigadores en Ciencias de la Computación p. 278-281, ISBN: 978-950-673-892-1.

MOINE, J. M.; HAEDO, A.; N.; GORDILLO, S. (2011b): "ANÁLISIS COMPARATIVO DE METODOLOGÍAS PARA LA GESTIÓN DE PROYECTOS DE MINERÍA DE DATOS". VIII Workshop Bases de Datos y Minería de Datos (WBDDM), CACIC 2011 - XVII Congreso Argentino de Ciencias de la Computación, octubre 2011 : p. 931-938

PEARL, J. (1988): PROBABILISTIC REASONING IN INTELLIGENT SYSTEMS. Morgan-Kaufmann (San Mateo).

RED DE UNIVERSIDADES CON CARRERAS EN INFORMÁTICA (UNCI). (2006). Documento en línea. Disponible en: http://redunci.info.unlp.edu.ar/. Consulta: 14/10/2013.

RUSSELL, S.; NORVIG, P. (2004): "INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO". 2da edición, Prentice–Hall Hispanoamericana.

RUIZ REINA, J. L. (2005): "INTRODUCCIÓN A LAS REDES BAYESIANAS". Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Sevilla. Documento en línea. Disponible en: https://www.cs.us.es/cursos/ia2-2005/temas/tema-08.pdf.

SUCAR, L. E. (2006): "REDES BAYESIANAS. BS ARAUJO, APRENDIZAJE AUTOMÁTICO: CONCEPTOS BÁSICOS Y AVANZADOS". Pearson Educación.

VALCÁRCEL ASENCIOS, V. (2004): "DATA MINING Y EL DESCUBRIMIENTO DE CONOCIMIENTO". Revista de la Facultad de Ingeniería Industrial. Vol. (7) 2: pp. 83-86 (2004) UNMSM ISSN: 1560-9146. Diciembre de 2004.

WEKA. Disponible en: http://www.cs.waikato.ac.nz/ml/weka/.

Agradecimiento
Se agradece a la Lic. Sara G. Tressens quien ha proporcionado los datos para la elaboración del presente trabajo.

ANEXO

TABLA 4. Variables evidenciales y sus valores seleccionados (Fuente: EDC en Mariño, 2001).

Variables evidencia	Valores posibles	Identificador
Porte	1 - árbol 2 - arbusto 3 - árbol o arbusto	**1.** porte1a
Hojas	1 - hojas sésiles 2 - hojas pecioladas 3 - hojas sésiles o pecioladas	**2.** hoja1a
	1 - lámina cartácea 2 - lámina no cartácea	3. id_lamina1a (consistencia)
	1 - lámina coriácea 2 - lámina no coriácea	4. id_lamina1b
	1 - lámina subcartácea 2 - lámina no subcartácea	5. id_lamina1c
	1 - lámina subcoriácea 2 - lámina no subcoriácea	6. id_lamina1e
	1 - lámina ovada 2 - lámina no ovada	7. id_lamina2b
	1 - lámina obovada 2 - lámina no obovada	8. id_lamina2c
	1 - lámina oblonga 2 - lámina no oblonga	9. id_lamina2d
	1 - lámina linear 2 - lámina no linear	10. id_lamina2e
	1 - lámina con domacios 2 - lámina sin domacios	11. id_lamina3
	1 - lámina concolora 2 - lámina discolora	12. id_lamina4
	1 - epifilo glabro 2 - epifilo no glabro 3 - epifilo glabro o no glabro	13. id_lam_epifilo1
	1 - hipofilo glabro 2 - hipofilo no glabro 3 - hipofilo glabro o no glabro	14. id_lam_hipofilo1
	1 - ápiceagudo 2 - ápice no agudo	15. id_lam_apice1
	1 - ápiceobtuso 2 - ápice no obtuso	16. id_lam_apice2
	1 - ápice acuminado 2 - ápice no acuminado	17. id_lam_apice3
	1 - ápice redondeado 2 - ápice no redondeado	18. id_lam_apice5
	1 - ápice con apículo punzante 2 - ápice sin apículo punzante	19. id_lam_apice_foliar
	1 - base aguda 2 - base no aguda	20. id_lam_base1
	1 - base cordada 2 - base no cordada	21. id_lam_base2
	1 - base obtusa 2 - base no obtusa	22. id_lam_base3
	1 - base cuneada	23. id_lam_base4

Variables evidencia	Valores posibles	Identificador
	2 - base no cuneada	
	1 - base atenuada 2 - base no atenuada	24. id_lam_base5
	1 - base redondeada 2 - base no redondeada	25. id_lam_base6
	1 - flor sésil 2 - flor pedicelada 3 - flor sésil o pedicelada	26. id_flor2a
	1 - pedúnculo de inflorescencia uniflora glabro o casi glabro 2 - pedúnculo de inflorescencia uniflora esparcida a densamente pubescente	27. id_pedunculo_fs
	1 - bractéolas caducas 2 - bractéolas persistentes	28. id_f_brac1
	1 - bractéolasfoliaceas 2 - bractéolas no foliáceas	29. id_f_brac2
	1 - bractéolas connadas en la base 2 - bractéolas libres	30. id_f_brac3
	1 - cáliz tetrámero 2 - cáliz pentámero 3 - cáliz tetrámero o pentámero o hexámero	31. caliz1
	1 - cáliz rasgándose a la antesis 2 - cáliz no rasgándose	32. id_caliz2
	1- cáliz siempre no persistente en el fruto 2 - cáliz siempre persistente en el fruto 3 - cáliz siempre o no siempre persistente en el fruto	33. id_caliz4
	1- cáliz abriéndose por una calyptra 2 - cáliz no abriéndose por una calyptra	34. id_caliz5
	1 - sépalos siempre reflejos después de la antesis 2 - sépalos siempre no reflejos después de la antesis 3 - sépalos siempre o no siempre reflejos después de la antesis	35. id_sepalos1a
	1 - sépalos soldados en el botón floral 2 - sépalos no soldados en el botón floral	36. id_sepalos2
	1 - cara externa del sépalo siempre glabra 2 - cara externa del sépalo siempre no glabra 3 - cara externa del sépalo siempre o no siempre glabra	37. id_sepalos3
	1 - cara interna del sépalo siempre glabra 2 - cara interna del sépalo siempre no glabra	38. id_sepalos4

Variables evidencia	Valores posibles	Identificador
	2 - cara interna del sépalo siempre o no siempre glabra	
	1 - pétalos sólo cuatro 2 - pétalos sólo cinco 3 - pétalos cuatro, cinco o seis	39. nro_petals
	1 - pétalos rojos, rosados o purpúreos 2 - pétalos no rojos, rosados o purpúreos	40. col_petalos
	1- hasta 8 estambres 2- más de 20 estambres	41. cod_n_estambre
	1 - estambres rojos o purpúreos 2 - estambres no rojos o purpúreos	42. color_estambre
	1 - hipanto glabro 2 - hipanto no glabro 3 - hipanto glabro y no glabro	43. id-hipanto1a
	1 - hipanto prolongado sobre el ovario 2 - hipanto no prolongado	44. id-hipanto2
	1 - hipantocostillado 2 - hipanto no costillado	45. id-hipanto3
	1 - hipanto circunciso 2 - hipanto no circunciso	46. id-hipanto4a
	1 - estilo siempre glabro 2 - estilo siempre no glabro 3 - estilo glabro o no glabro	47. id_estilo2a
	1 - inflorescencia contraída 2 - inflorescencia laxa	48. id_inflor2
	1 - inflorescencia axilar 2 - inflorescencia terminal 3 - inflorescencia axilar o terminal	49. id_inflor3a
	1 - inflorescencia pluriflora (más de 3) 2 - inflorescencia no pluriflora	50. id_inflor4a
	1 - inflorescenciatriflora 2 - inflorescencia no triflora	51. id_inflor4b
	1 - inflorescencia uniflora 2 - inflorescencia no uniflora	52. id_inflor4c
	1 - inflorescencia biflora 2 - inflorescencia no biflora	53. id_inflor4d
	1 - racimo 2 - no racimo	54. id_inflor5c
	1 - panícula 2 - no panícula	55. id_inflor6a
	1 - tirso 2 - no tirso	56. id_inflor6b
	1 - dicasio 2 - no dicasio	57. id_inflor7a
	1 - inflorescencia fasciculada 2 - inflorescencia no fasciculada	58. id_inflor8
	1 - inflorescencia proliferante 2 - inflorescencia no proliferante	59. id_inflor9
	1 - inflorescencia cauliflora 2 - inflorescencia no cauliflora	60. id_caulifloros

Variables evidencia	Valores posibles	Identificador
Fruto	1 - fruto siempre globoso o subgloboso 2 - fruto siempre no globoso o subgloboso 3 - fruto no globoso o subgloboso	61. id_fruto1
	1 - fruto drupáceo (endocarpo leñoso) 2 - baya (endocarpo no leñoso)	62. id_frutos4
	1 - cubierta seminal (testa) ósea 2 - cubierta seminal (testa) no ósea	63. id_semilla3
	1 - cotiledones libres 2 - cotiledones soldados	64. id_cotiledon1
	1 - cotiledones foliáceos y contortuplicados 2 - cotiledones no foliáceos y contortuplicados	65. id_cotiledon2

TABLA 5. Posibles valores de la hipótesis o variables objetivos (Fuente: EDC en Mariño, 2001)

Hipótesis (Especies a identificar en el estudio)	Id
Blepharocalyxsalicifolius (B. tweediei)	*2*
Calyptranthesconcinna	*4*
Campomanesiaguaviroba	*6*
Campomanesiaguazumifolia	*8*
Campomanesiaxanthocarpavar. xanthocarpa	*10*
Eugenia burkartiana	*12*
Eugenia hyemalisvarmarginata	*14*
Eugenia involucrata	*16*
Eugenia moraviana	*18*
Eugenia pitanga	*20*
Eugenia pyriformisvar. pyriformis	*22*
Eugenia pyriformisvar. uvalha	*24*
Eugenia repanda	*26*
Eugenia sp.	*28*
Eugenia uniflora	*30*
Eugenia uruguayensis	*32*
Hexachlamysedulis	*34*
Hexachlamyshumilis	*36*
Myrcialaruotteanavar. australis	*38*
Myrciaselloi (M. ramulosa)	*40*
Myrciasp.	*42*
Myrcianthescisplatensis	*44*
Myrcianthespungens	*46*
Myrciariatenella	*48*
Myrrhiniumatropurpureumvar. octandrum	*50*
Pliniarivularis	*52*
Psidiumguajava	*54*
Psidiumguineense	*56*
Psidiumincanum	*58*
Psidiumkennedyanum	*60*
Psidiumnutans	*62*