MINERÍA DE DATOS APLICADA A LA CONSERVACIÓN *EX SITU* DE RECURSOS FITOGENÉTICOS DE SAN JUAN.

Karina Fernández[1], Carola Meglioli[2], Raúl Klenzi[1]
[1]Departamento de Informática. Facultad de Ciencias Exactas, Físicas y Naturales. Universidad Nacional de San Juan. Avda Ignacio de la Roza y Meglioli. Rivadavia.
[2]Área Conservación de Recursos Fitogenéticos. Instituto de Investigación y Desarrollo Agroindustrial Hortícola Semillero. Coll 3671 (oeste). Rivadavia (5407). San Juan.
karinaferh@gmail.com; caromeglioli@yahoo.com.ar; rauloscarklenzi@gmail.com

Resumen:
La digitalización y posterior análisis de datos biológicos es una actividad creciente en la actualidad. En la presente propuesta, se proyecta realizar tareas de agrupamiento y clasificación sobre los datos pertenecientes al área de conservación de recursos fitogenéticos del Instituto de Investigación y Desarrollo Agroindustrial Hortícola Semillero (INSEMI), de la provincia de San Juan. Se trabajará con una herramienta específica de Minería de Datos con conectividad al banco de datos sobre los recursos fitogenéticos conservados, donde se evaluarán diferentes algoritmos de segmentación, clasificación y visualización, que orienten la toma de decisiones con respecto a: a) la planificación de las campañas de recolección de recursos fitogenéticos, b) el manejo interno de las colecciones en cuanto al fortalecimiento de las mismas, y c) su manipulación posterior.

Contexto

La actividad de conservación *ex situ* de recursos fitogenéticos en la provincia de San Juan, está siendo fortalecida mediante la ejecución del Proyecto "Desarrollo competitivo del sector semillero de la provincia de San Juan", el cual es financiado por el Banco Mundial y con aportes locales en el marco del Ministerio de Producción y Desarrollo Económico del Gobierno de la Provincia, los cuales son administrados por el Programa de Servicios Agrícolas Provinciales (PROSAP/BIRF 7597-AR). Este proyecto se lleva a cabo con centro en el Instituto de Investigación y Desarrollo Agroindustrial Hortícola Semillero (INSEMI).

Por otro lado, en el ámbito de la Universidad Nacional de San Juan (UNSJ), se encuentra en ejecución el Proyecto "Minería de Datos en la Determinación de Patrones de Uso y Perfiles de Usuarios (21/E889)", financiado por CICITCA-UNSJ trienio 2011-2013, con el cual y atendiendo a la temática abordada, existe la posibilidad de un trabajo común de fuerte integración. En este contexto se cuenta con el aporte de la beca nacional TIC` s otorgada a la Alumna Karina Fernández, periodo 2012/2013 para el estudio de minería de datos en la segmentación y clasificación de un banco de germoplasma.

En el corto plazo se contará además con aportes del Ministerio de Ciencia, Tecnología e Innovación Productiva (MINCyT), para el fortalecimiento de las bases de datos y para la formación de

recursos humanos, ya que el INSEMI ha sido adherido al Sistema Nacional de Datos Biológicos (SNDB – Res 05/12)

Introducción

Se entiende por Recursos Genéticos, la variabilidad genética acumulada en todos los organismos vivos a lo largo de millones de años de evolución (REGENSUR, 2007). Entonces los Recursos Fitogenéticos se definen como cualquier material genético de origen vegetal de valor real o potencial para la alimentación y la agricultura, y que constituyen la materia prima utilizada por los agricultores, por consiguiente resulta fundamental su conservación y el mantenimiento para una producción agrícola sostenible (FAO, 1996; UICN, 1980).

La documentación de la información recopilada y generada es fundamental para la toma de decisiones sobre el manejo de las colecciones en un centro de conservación de recursos y sobre su valor de uso (Engels et al., 2007). Además, el valor del germoplasma conservado aumenta en la medida que se conoce, de ahí la Importancia de una adecuada documentación de los datos recopilados y generados (Rao et al., 2007). El sistema de documentación se utiliza para la recuperación, el almacenamiento, el mantenimiento o la actualización, procesamiento y análisis, e intercambio de datos con otros centros de conservación (Rivera-Gutiérrez et al., 2003).

El éxito de la revolución digital y el crecimiento de Internet aseguran que grandes volúmenes de datos de alta dimensión, están disponibles en todo lo que nos rodea. Esta información se mezcla a menudo, con la participación de diferentes tipos de datos tales como texto, imagen, audio, voz, hipertexto, gráficos y componentes de vídeo entremezcladas unas con otras. Sin embargo, a menudo la mayor parte de estos datos no son de mucho interés para la mayoría de los usuarios. Descubrimiento de Conocimiento en Bases de Datos (KDD) es un análisis automático y exploratorio que permite el modelado de depósitos de datos de gran tamaño. KDD es el proceso organizado de identificar patrones válidos, novedosos, útiles y comprensibles a partir de extensos y complejos conjuntos. Data Mining (DM) es el núcleo del proceso de KDD, que implica la inferencia de los algoritmos que analizan y modelan los datos, y permite descubrir patrones previamente desconocidos y de interés para los usuarios. Se trata de un área de investigación y desarrollo interdisciplinaria que abarca diversos dominios, y lejos de estar saturada, se amplía con nuevas técnicas y orientaciones (Mucherino A., y otros. 2009; O. Maimon O. Rokach L. 2010; Witten, Frank, Hall, 2011).

El procesamiento de información basada en minería de datos, está orientada a análisis de atributos pertenecientes al área de conservación *ex situ* de recursos fitogenéticos de San Juan. Se mantienen colecciones de herbario y de semillas tanto de especies nativas como también de cultivos tradicionales, importantes para la provincia. Este material genético responsable de las características de una planta que se transmite de una generación a la siguiente, para el futuro

beneficio de la humanidad y del ambiente.

En este contexto, contar con datos digitalizados y una base de datos actualizada, ágil y de fácil acceso, que contenga información georreferenciada y vinculante sobre los recursos fitogenéticos de San Juan, resulta además, una herramienta de alto valor, dado que no se han encontrado aún en la zona, instituciones que generen, mantengan y actualicen en forma continua datos biológicos.

Se plantea aplicar minería da datos a los diferentes registros colectados en el área de conservación de recursos fitogenéticos de San Juan. Las tareas principales radicarán en el uso de algoritmos específicos que permitan tareas de segmentación y clasificación de datos (Figuerola el al., 2005) provenientes de las colecciones de semillas de especies correspondientes a la geografía de San Juan. Para ello se hará uso de la herramienta de software Rapid Miner cuya última versión disponible es la 5.3.005. Se trata de un entorno de aplicación de algoritmos de aprendizaje de máquina y minería de datos, fácil de instalar y de ejecutar en cualquier plataforma y sistema operativo. Allí se pueden aplicar todos los pasos involucrados en la minería de datos desde el pre-procesamiento hasta la visualización de resultados al evaluar diferentes estrategias de segmentación, clasificación y reglas de asociación mediante una interfaz amigable (North 2012) y que se ofrece bajo una licencia AGPL versión 3.0 (software libre).

Se espera segmentar y clasificar datos biológicos provenientes de los Recursos Fitogenéticos conservados en San Juan para obtener agrupaciones requeridas para la planificación, manejo, regeneración y transferencia del germoplasma conservado.

Líneas de investigación y desarrollo

El equipo de trabajo desarrolla dos grandes líneas de trabajo: a) Digitalización y análisis de datos en el ámbito de la conservación de recursos fitogenéticos locales, b) Minería de Datos.

Resultados y Objetivos

Los datos sobre los recursos fitogenéticos conservados se encontraban 100% almacenados en planillas de recolección y cuadernos de campo, por lo que la primer tarea correspondió a la digitalización de los mismos para permitir posteriormente la migración de los mismos a una base de datos adecuada y poder realizar los análisis propuestos.

Hasta la fecha se ha logrado un elevado porcentaje de la sistematización y posterior digitalización de la información correspondiente a las colecciones de semillas de la flora nativa, como así también de las colecciones de semillas de los cultivos tradicionales. Estos registros son caracterizados por aproximadamente 80 atributos multivaluados. Estos datos han sido categorizados en planillas Excel, a partir de las cuales serán trasladados a la base de datos que se establecerá para el área de conservación de recursos fitogenéticos del INSEMI y desde donde la herramienta de aplicación RM tomará los datos, previa compatibilización en los

tipos de datos asociados a los atributos, que permita a su vez, viabilizar la aplicación de diferentes algoritmos de segmentación y clasificación. Esto último está iniciándose y los resultados que se obtengan de estos análisis definirán los próximos pasos a seguir dentro del ámbito de la minería de datos.

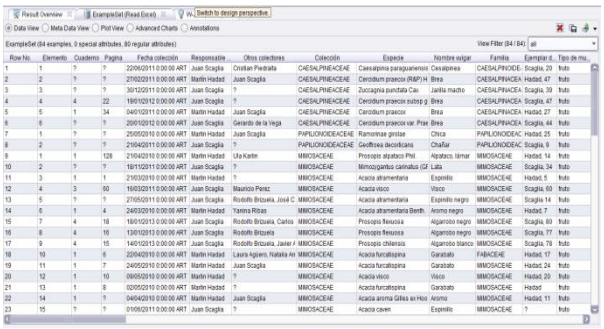

Figura 1

En la Figura1 se puede apreciar la lectura realizada desde RM de una porción de la fuente de datos original excell. En esta primer tanda de registros cargados se observa la existencia de 80 atributos entre los que figuran datos asociados al personal de la campaña de recolección, lugar físico de ubicación de los datos originales, que si bien pueden ser de suma importancia para la determinación de responsabilidades en el INSEMI puede tener escasa importancia a la hora comenzar con las tareas de segmentación y clasificación. Por lo expuesto anteriormente, el grupo de trabajo está en plena tarea de definición de atributos relevantes inherentes a la etapa de preprocesamiento de DM.

Se proseguirá con el relevamiento y el análisis de las herramientas de software y necesidades de hardware asociado, y con la generación de conocimiento relevante de los datos que favorezca la toma de decisiones para el manejo de las colecciones conservadas mediante la aplicación de herramientas específicas de minería de datos.

Formación de Recursos Humanos

El equipo de trabajo está formado por investigadores del Departamento de Informática de la Universidad Nacional de San Juan y del instituto de Investigación y Desarrollo Agroindustrial Hortícola Semillero (INSEMI), becarios y adscriptos a los proyectos. También se cuenta con apoyo de investigadores de INTA.

Tesis de posgrado en curso:

"Acondicionamiento de semillas de Poaceae nativas para su aprovechamiento en la recuperación de suelos degradados en zonas áridas". Doctorado en Agronomía. Facultad de Ciencias Agrarias, Universidad Nacional de Cuyo (Mendoza). Alumna: Lic. Carola Meglioli. Director: Dr. Carlos Parera.

Tesis de grado en curso:

Tras la carga de la totalidad de los registros existentes en papel a formato digital se llevará adelante el trabajo final de grado: "Minería de datos en la segmentación y clasificación de un banco de germoplasma". Licenciatura en Sistemas de Información. Departamento de Informática. Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan. Alumna: Karina Fernández. Director: Mg. Raúl Klenzi, Codirector: Lic. Carola Meglioli.

Se pretende además, y en el maro de un trabajo integrador de titulación intermedia de la carrera licenciatura en ciencias de la computación, desarrollar una

herramienta de software e interfaz adecuada, que permita la carga directa de la información durante la campaña de recolección, en dispositivos tipo tablets o celular avanzado a cargo del personal afectado, logrando una mayor agilidad al momento de carga y procesamiento de datos

Referencias

- Engels, J.M.M. y Visser, L. (eds.). 2007. Guía para el manejo eficaz de un banco de germoplasma. Manuales para Bancos de Germoplasma No. 6. Bioversity International, Roma, Italia. ISBN 978-92-9043-767-3
- FAO. 1996. Plan de Acción Mundial para la Conservación y la utilización sostenible de los Recursos Fitogenéticos para la Alimentación y la Agricultura y la Declaración de Leipzip. CuartaConferencia Técnica Internacional sobre Recursos Fitogenéticos, Leipzig, Alemania, 17-23 de junio, 64 pp.
- Maimon, O. Rokach L. 2010. “Data Mining and Knowledge Discovery Handbook” –Springer
- Mucherino A., Papajorgji P., Pardalos P. 2009. “Data Mining in Agriculture” (Springer Optimization and Its Applications Volume 34)
- North M. 2012. “Data Mining for the Masses” ISBN: 0615684378. A Global Text Project Book.
- Peña, D. 2002. Análisis de datos multivariantes. Ed. Mc. Graw Hill, España.
- Rao, N.K., J. Hanson, M.E. Dulloo, K. Ghosh, D. Novell y M. Larinde. 2007. Manual para el manejo de semillas en bancos de germoplasma. Manuales para Bancos de Germoplasma No. 8. Bioversity International, Roma, Italia. ISBN 978-92-9043-757-4.
- REGENSUR/PROCISUR. 2007. Acceso a los recursos genéticos: estado de situación en los países del cono sur.
- Rivera-Gutiérrez H. F., Suárez-Mayorga A. M., Varón-Londoño A. 2003. Estándar para la documentación de metadatos de conjuntos de datos relacionados con biodiversidad, versión (electrónica). Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, Bogotá, Colombia, 61 p.
- Unión Mundial para la Conservación de la Naturaleza (UICN). 1980. Estrategia Mundial para la Conservación.
- Witten I.; Frank E.,Hall M. 2011“Data Mining Practical Machine Learning Tools and Techniques”, Third Edition (The_Morgan Kaufmann Series in Data Management Systems) Morgan Kaufmann