Predicción de arribos de hacienda al Mercado de Liniers con algoritmos de aprendizaje supervisados

Daniel Negrotto
Instituto de Industria, Universidad Nacional de General Sarmiento
Departamento de Computación, FCEyN - Universidad de Buenos Aires
Email: dnegrotto@ungs.edu.ar, dnegrott@dc.uba.ar

***Resumen*—La administración de la hacienda en un mercado concentrador como el Mercado de Liniers requiere de una estimación diaria de la cantidad de cabezas disponibles para su ingreso. La confluencia de intereses de los actores participantes en este mercado determina que el ingreso total de hacienda sea una de las variables más consultadas diariamente. Predecir esta variable es importante tanto para los diferentes actores de la cadena de comercialización de hacienda, organismos gubernamentales de control, como también para otros actores de los negocios agropecuarios en general. En este trabajo se presenta un algoritmo de estimación del ingreso de hacienda para intentar pronosticar el ingreso total de animales al Mercado de Liniers en jornadas futuras. A partir del procesamiento de una dataset específico y su posterior enriquecimiento, se presenta un algoritmo supervisado de clasificación predictivo basado en árboles de decisión. Esta nueva herramienta intenta ser un aporte adicional a la transparencia y a la previsibilidad de la operatoria de compraventa de hacienda en el mercado formador de precios de referencia en el sector agropecuario.**

I. Introducción

El Mercado de Liniers [5] es el centro de operaciones pecuarias más importante de la República Argentina. Debido a características propias de su funcionamiento y operatoria, es considerado el mercado ganadero activo más importante del mundo.

Se encuentra ubicado en la Ciudad Autónoma de Buenos Aires y se extiende en un predio de 34 hectáreas.

El ganado que ingresa al Mercado de Liniers proviene de diversas provincias, siendo la provincia de Buenos Aires la que mayor hacienda aporta (entre el 75 y 85 % del total en condiciones normales). La Pampa, Sante Fe, Córdoba y Entre Ríos son las otras provincias con participación importante en los envíos de hacienda comercializados en Liniers.

Los productores agropecuarios (también llamados remitentes) envían su ganado para ser vendido en Liniers. Mediante el sistema de consignación, representado por los consignatarios de hacienda, se llevan a cabo los remates realizados en subasta pública. El comprador (o matarife) que ofrece el mejor precio, por último, es quien se lleva la hacienda para luego faenarla en un frigorífico faenador.

Los consignatarios de hacienda, los compradores, los productores agropecuarios, los medios de comunicación, el personal del Mercado involucrado en la operatoria y el resto de la comunidad del sector agropecuario están permanentemente atentos a las operaciones efectuadas en Liniers. Los precios del Mercado de Liniers son utilizados como precios de referencia en el sector ganadero y agropecuario en general y surgen como resultado de los remates públicos efectuados en su recinto de operaciones. La cantidad de hacienda ingresada es una de las variables importantes en la formación de los precios corrientes ya que la lógica de formación de precios se corresponde con las pujas existentes entre oferta y demanda. Los índices calculados diariamente en el Mercado son ampliamente utilizados como referencia para regir contratos de alquileres de campos, arrendamientos y negocios rurales en general.

El Mercado de Liniers intercambia información con diversos organismos públicos. Uno de estos organismos es el Ministerio de Asuntos Agrarios de la Provincia de Buenos Aires (MAA [4]). Debido a ello, este ministerio otorga diariamente al Mercado de Liniers la información de las guías emitidas en la provincia con destino al Mercado de Liniers. Esta información es recibida de manera automática mediante un mecanismo de importación de datos basados en el acceso a los sistemas del organismo mediante un Web-Service proporcionado por la entidad para tal fin. Los datos obtenidos son almacenados en las bases de datos del Mercado de Liniers y son utilizados para verificar la validez de la información ingresada al sistema una vez que la hacienda arriba al Mercado.

El día previo al ingreso de la hacienda se reciben en el centro de cómputos y en los atracaderos numerosas llamadas telefónicas para consultar el ingreso que se estima para el día siguiente. De esta manera, la confluencia de intereses de los actores participantes en el Mercado determina que el ingreso total de hacienda sea una de las variables más consultadas diariamente. Por lo tanto, se considera sumamente interesante la posibilidad de encarar el desarrollo de un algoritmo de estimación del ingreso de hacienda para intentar pronosticar el ingreso total de animales en la jornada laboral siguiente.

Diversos algoritmos de aprendizaje automático han sido utilizado con éxito en el sector ganadero para intentar predecir desde animales con enfermedades hasta comportamiento animal ej: [7] y [8]). En este trabajo se elaboró un algoritmo de clasificación de los lotes de hacienda basado en la información suministrada por el MAA. Para ello, se experimentó con diferentes técnicas de aprendizaje automático basadas en algoritmos de clasificación. El algoritmo de clasificación basado en árboles de decisión J48 (basado en el algoritmo C4.5 creado por Ross Quinlan en 1993 [1]) fue el que proporcionó mejores resultados.

La primera versión del algoritmo diseñado actualmente se

encuentra en etapa de evaluación por el Mercado de Liniers y ha mostrado hasta el momento excelentes resultados. Se planea implementar una nueva versión que incluya también información de otro dataset suministrado por el SENASA (Servicio Nacional de Sanidad y Calidad Agroalimentaria). Este dataset incluye la información de todos los documentos en tránsito (DTe [6]) hacia el Mercado de Liniers. Se trata de una información más completa que la proporcionada por el MAA ya que es nacional e incluye los envíos de hacienda al Mercado desde todas las provincias.

En la sección siguiente se presenta la metodología utilizada. Luego, se muestra la experimentación efectuada y los resultados computacionales obtenidos. Por último, se presentan las conclusiones del trabajo y futuras líneas de avance.

II. Metodología

En este trabajo se elaboró un algoritmo clasificador de lotes de hacienda basado en la información suministrada por el MAA. Para ello, se utilizaron técnicas de aprendizaje automático para implementar un algoritmo de clasificación. Se experimentó con diversos algoritmos de clasificación, siendo el algoritmo de clasificación basado en árboles de decisión J48 el que mejores resultados proporcionó.

La datos obtenidos del MAA son los siguientes:

- Número de guía.
- Fecha y hora de emisión de la guía.
- El código de partido de la provincia de Buenos Aires de dónde proviene la hacienda.
- Nombre del establecimiento productor.
- Número de CUIT del productor.
- Número de CUIT del consignatario de hacienda que venderá el lote.
- Los códigos de categoría de la hacienda que formará parte de la guía. Los códigos determinan que tipo de hacienda es trasladada (vacas, toros, novillos, terneros, etc).
- La cantidad de hacienda trasladada para cada una de las categorías consignadas.

La idea del presente trabajo es utilizar esta información y analizar las guías de traslado que todavía no ingresaron al Mercado de Liniers. Utilizando los datos disponibles de cada registro se pretende estimar la fecha de arribo al Mercado de Liniers y poder, de esa manera, calcular un valor estimado del ingreso total de hacienda para el próximo día de operaciones.

A. Algoritmo de clasificación

Para implementar el algoritmo predictivo se experimentó con diversos clasificadores basados en árboles de decisión, clasificadores probabilísticos de la familia de clasificadores Naive Bayes, clasificadores basados en vecindades (lazy learning) como k-NN y métodos de ensambles variados. Con los datasets utilizados, el clasificador que mejores resultados otorgó fue el clasificador basado en árboles de decisión J48. Un ensamble basado en varios clasificadores J48 ofreció también buenos resultados pero a un costo computacional mucho mayor. La ganancia final obtenida con el ensamble no fue lo suficientemente significativa como para justificar su uso frente al clasificador J48.

Para implementar el algoritmo de clasificación se utilizó el software weka [3] y su correspondiente algoritmo asociado de árboles de decisión J48 (weka.classifiers.trees.J48).

III. Experimentación y resultados computacionales

Se utilizó como dataset archivos arff generados a partir de los datos históricos de la base de datos del Mercado de Liniers. Se evaluaron diferentes datasets, variando el rango de fechas utilizado para efectuar la clasificación. Entre las opciones testeadas se usaron datasets con los movimientos ingresados a Liniers en los últimos 1, 2, 3, 6, 12, 18, 24 y 36 meses. Además, se testeó también datasets con los últimos 7 y 14 días de operaciones. Todos estos datasets fueron generados mediante un web-service creado para tal fin, y un correspondiente script bash para automatizar la tarea de preparación y generación del modelo en weka. Todos los clasificadores utilizaron Cross Validation 10-folds [2].

Según puede observarse en el gráfico 1, el dataset con mayor porcentaje de instancias correctamente clasificadas fue el dataset de 18 meses. En todas las pruebas se utilizó el

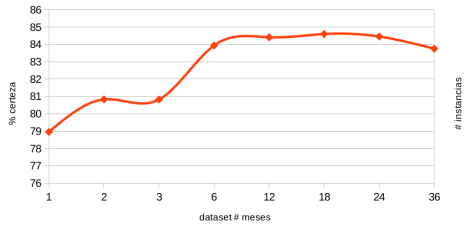

Figura 1. Análisis datasets

clasificador J48 con Confidence Factor 0.5 (CF - Intervalo de confianza), que fue el parámetro que mejores resultados proporcionó. Se testearon también los resultados con CF 0.25, 0.3, 0.4 y 0.6. En el gráfico 2 pueden observarse los resultados de tales comparaciones utilizando el dataset de 18 meses. Inicialmente se utilizó como dataset un subconjunto de los

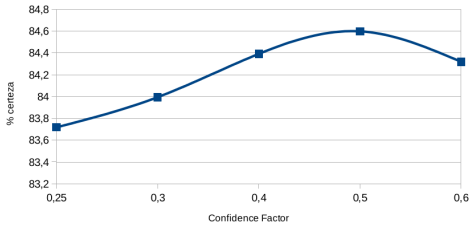

Figura 2. Confidence factor

campos mencionados anteriormente obtenidos de los datos del MAA. Estos son: código de partido (definido como numérico nominal en weka), el código de consignatario en el sistema de Liniers (numérico nominal), el número de CUIT del productor (numérico) y la cantidad de días entre la emisión de la guía y el arribo a Liniers (numérico nominal 0-8). Este último campo es el campo que se quiere estimar. Es importante aclarar que la

vigencia de las guías es de 72 horas hábiles. Sólo se alcanza en situaciones especiales valores de cantidad de días mayores a 5 días (feriados especiales), pero lo normal es 1-5 días de lapso. Los productores pueden tener más de un campo en diferentes localidades (por eso se incluye el código de partido como nominal) y pueden, a su vez, enviar la hacienda a través de distintos consignatarios de hacienda (por ello se incluye el código del Consignatario en el dataset). La idea es entrenar al clasificador para predecir la cantidad de días que demorará cada productor en enviar la hacienda al Mercado de Liniers después de haber gestionado la guía.

Si bien los envíos de hacienda realizados por productor son estacionales y pueden variar dependiendo del mes del año, se observó que dicha información no mostraba diferencias a los efectos de la predicción de la cantidad de días entre la gestión de la guía y el arribo a Liniers. Por ello, se asumió que la cantidad de días a predecir es una variable aleatoria independiente distribuida del período del año en el que se está considerando el envío.

Con los distintos datasets creados (variando el período de movimientos de acuerdo a los meses mencionados anteriormente) se testearon distintas clasificaciones que en todos los casos no superaron el 60 % de instancias correctamente clasificadas.

Para poder mejorar el modelo, se introdujeron nuevos campos al dataset. En primer lugar se agregó un campo numérico nominal con el número de día de la semana de la emisión de la guía (1-lunes, 2-martes, etc). Esta idea está basada en la observación de que la fecha de arribo está íntimamente relacionada con el día de la semana. En el Mercado de Liniers los días con mayores operaciones son los lunes, martes, miércoles y viernes. Los días jueves rara vez se superan los 2000 animales ingresados y los sábados y domingos no se registran ingresos. Por lo tanto, saber qué día de la semana se emitió la guía ofrece mucha más información para la clasificación. Con el agregado de ese campo se obtuvieron clasificaciones que rondaron el 75 % de instancias correctamente clasificadas.

Además, se vislumbró la posibilidad de agregar el dato distancia promedio en kilómetros. En el dataset inicial se tiene el campo código de partido, por lo tanto se codificó una nueva tabla con la distancia promedio desde las localidades del partido al Mercado y se agregó dicha información al dataset. Esta información también es relevante ya que la distancia a Liniers influye en el tiempo que la hacienda demorará en 'viajar' hasta el destino.

Para mejorar aún más el modelo se modificó el cálculo de la cantidad de días (campo a estimar). En el dataset utilizado como entrenamiento se definió la cantidad de días de manera secuencial eliminando los fines de semana. Es decir, en los casos donde la guía era emitida, por ejemplo, un viernes, la entrada estimada iba a ser recién el lunes o martes de la semana siguiente. Por lo tanto, el lapso de días en este caso debía ser 3 o 4. Analizando el dataset se verificó que esto generaba una cantidad de posibles estimaciones en el rango de 1 a 5 días. Cuantas mayores opciones se tenga que poder predecir, mayor es la posibilidad de error. Por ende, se modificó el cálculo de los días para eliminar los días sábados y domingos y calcular el campo sólo en función de los días hábiles. Con esta modificación se redujo la cantidad de estimaciones normales a 1, 2 o 3 días (72 horas es el tiempo máximo de vigencia de una guía).

Como última mejora se introdujeron 2 nuevos campos al dataset. Estos son el promedio de días de arribo histórico desde el partido, y el promedio de días histórico del productor. Se generó este promedio analizando los movimientos efectuados en el último año de operaciones y se obtuvieron estos 2 nuevos campos numéricos.

Con todas estas modificaciones se logró perfeccionar el modelo para lograr clasificadores con valores cercanos al 85 % de estimaciones correctamente clasificadas.

IV. Conclusiones

Como puede observarse en los resultados obtenidos, el clasificador implementado logra un porcentaje de certeza cercano al 85 %. En el presente trabajo se analizaron distintos clasificadores, diferentes conjuntos de datasets alternativos y los mejores parámetros de configuración para el clasificador elegido. Se inició el estudio con un dataset elemental y se propusieron diversas mejoras que lograron incrementar significativamente el porcentaje de clasificación correcta del algoritmo.

Si bien el total de hacienda ingresado es una variable bastante fluctuante y extremadamente sensible a numerosas variables externas, como ser huelgas, situación económica de coyuntura, problemas de sequía, inundaciones, etc.; se ha desarrollado un estimador que utiliza información alimentada en tiempo real para intentar predecir los posibles ingresos, utilizando para ello datos históricos y corrientes. La ventaja fundamental del algoritmo implementado, frente a otros enfoques basados en modelos estadísticos estáticos, es la elasticidad que el mismo ofrece para auto-adaptarse a los cambios.

V. Trabajo futuro

Actualmente el Mercado de Liniers ha implementado un convenio con el SENASA para recibir información diaria de los DTe con destino Mercado de Liniers. Este documento es de carácter obligatorio y tiene alcance nacional. Se proyecta la incorporación de esta información al algoritmo para mejorar la clasificación y obtener una estimación con mayor precisión.

Referencias

[1] T.Mitchell, *Machine Learning*, McGraw Hill, 1997.
[2] G.James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2014.
[3] The University of Waikato, Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/
[4] Ministerio de Asuntos Agrarios de la Provincia de Buenos Aires, http://www.maa.gba.gov.ar
[5] Mercado de Liniers S.A., http://www.mercadodeliniers.com.ar
[6] Servicio Nacional de Sanidad y Calidad Agroalimentaria, SENASA, DTe, http://www.senasa.gob.ar/sistemas-online/sigsa-dte
[7] D. E. Amrine, B. J. White, R. L. Larson, *Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease*, Computers and Electronics in Agriculture, Vol.105, pp 9-19, 2014.
[8] J. J. Valletta, C. Torney, M. Kings, A. Thornton, J. Madden, *Applications of machine learning in animal behaviour studies*, Animal Behaviour, Vol.124, pp 203-220, 2017.