Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Plan de Tesis de Maestría en Minería de Datos y Descubrimiento de conocimiento

Tesis a ser presentada para obtener al título de
Magíster en Explotación de Datos y Descubrimiento de Conocimiento

Walter Marcelo Lamagna

Directora de Tesis: María Elena Buemi

Buenos Aires, 2015

Índice

1. Tema de investigación

El reconocimiento de dígitos manuscritos es un tema que ocupa distintas áreas de conocimiento e investigación en aprendizaje de clasificación de patrones, resulta ser de interés por la posibilidad de su aplicación práctica y la necesidad de automatizar la comprensión del alto volumen de documentos impresos en las empresas, gobiernos y una sociedad en expansión permanente. Por ejemplo en la gestión automática de correo postal [9] [7], procesamiento de cheques bancarios [10], ingreso manual de datos en formularios [6], registros gubernamentales y tarjetas de crédito impresas. Los datos públicos a utilizarse contienen resultados electorales de la Ciudad de Buenos Aires, Argentina. El reconocimiento de los dígitos manuscritos en telegramas completados por las autoridades de cada mesa durante las elecciones de diputados y senadores en Argentina, Buenos Aires, Capital Federal datan del 27 de Octubre de 2013. Al final del presente trabajo se compara el tipeado manual realizado por data entries de la empresa contratada para dicha tarea, contra los números reconocidos en forma artificial.

Los datos abiertos, Open Data, en muchos casos de fácil acceso , y pese a que exista la posibilidad técnica de guardarlos en una máquina para luego disponer de ellos, la legalidad de su utilización no siempre es viable. En la legislación en Argentina se ha comenzado a incluir el término **Dato Abierto** ó **Dato Público** como una manera de que los ciudadanos puedan dispongan de ellos, e idealmente producir un valor agregado. El Gobierno de la Ciudad de Buenos Aires ha dispuesto un sitio web con datos públicos [1] reglamentado por la Ley de Acceso a la Información Pública Nro. 104 [2] y la Ley de Protección de Datos Personales Nro. 1845. El Gobierno de la Ciudad de Buenos Aires ha definido Datos Abiertos como: *Significa poner información del Estado en un catálogo al alcance de todos, en formatos digitales, estándar y abiertos, siguiendo una estructura clara que permita su fácil comprensión y reutilización por parte de la ciudadanía.*. Otras ciudades de la Argentina como Bahia Blanca [3] y Datos Abiertos Misiones [4], tienen su portal de Datos Públicos, lo que

[1] http://data.buenosaires.gob.ar/about

[2] http://www.cedom.gov.ar/es/legislacion/normas/leyes/ley104.html

[3] http://bahiablanca.opendata.junar.com/home/

[4] http://www.datos.misiones.gov.ar/

constituye una herramienta efectiva.

Los Datos Abiertos según el site Definicion Abierta [5] proyecto de la Fundación del Conocimiento Libre (Open Knowledge Foundation) [6], una organización sin fines de lucro, los define como:

- *Acceso y disponibilidad*: Los datos deben estar disponibles por completo, si existiese un costo asociado, este debe ser razonable, preferentemente accesible sin costo por Internet. Deben estar disponibles en un formato conveniente y modificable.
- `Reutilización y redistribución`: Los datos deben permitir su re-distribución y reutilización, incluyendo su comercialización ya sea individualmente o con entrecruzamiento y/o entremezclado con otras bases de datos.
- `Participación universal`: La licencia debe permitir que todos tengan la posibilidad de utilizar, reutilizar y redistribuir, sin ningún tipo de discriminación a personas o grupos.

Como puede observarse en el tercer punto se habla de licencias, toda base de datos abierta debería estar acompañada de su correspondiente licencia, la misma puede exigir una mención a quienes generaron los datos ó reservarse el derecho de que cualquier trabajo derivado de la base de datos abierta mantenga la misma licencia, etc. Es similar al `software libre`, donde libre no significa "gratis" y que la libertad se extiende hasta donde decide el o los autores originales del trabajo.

Los telegramas manuscritos disponibles por la Dirección Nacional Electoral [7], son resultados electorales y en virtud de la disposición 408/2013, lo dispuesto en el Anexo I del Decreto Nr. 682 del 14 de mayo de 2010, se aprueba la Directiva De Datos Públicos Abiertos Para La Administracion Electoral [8]. Esta disposición establece que los resultados electorales deberán adecuarse progresivamente a la directiva de permitir la interacción con los ciudadanos. A su vez, el acceso a la database de dígitos manuscritos MNIST tiene derechos de autor de Yann LeCun [9] y Corinna Cortes, bajo la licencia *Creative Commons Attribution-Share Alike 3.0*, son datos abiertos con la condición de mencionar a sus autores en trabajos derivados.
Esta tesis propone el reconocimiento automático de dígitos manuscritos utilizando técnicas de análisis de imágenes y minería de datos a partir de imágenes de planillas manuscritas provinientes de datos públicos.

[5] http://opendefinition.org/od/2.0/en/index.html
[6] https://okfn.org/
[7] http://www.resultados.gob.ar/inicio.htm
[8] http://www.infoleg.gob.ar/infolegInternet/anexos/220000-224999/221756/norma.htm
[9] http://yann.lecun.com/

2. Antecedentes

En 1929 Gustav Tauscheck [1] obtuvo una patente de su "Máquina que lee", en 1933 Paul W. Handel [2] patenta una "Máquina Estadísta" que aplica los primeros conceptos iniciales de reconocimiento de caracteres, que desde mediados de 1950 se ha convertido en un campo de investigación muy activo para investigación y desarrollo [12]. En 1951 con el advenimiento de la primera computadora disponible para su comercialización, la UNIVAC I, puesta en funcionamiento en el Centro de Estadística en Estados Unidos significó un avance tecnológico importante que permitió avanzar en el campo de lectura automática de carateres.

En 1992 el Instituto Nacional de Estándares y Tecnología de Estados Unidos (NIST), con el objetivo de motivar la creación de nueva tecnología, organizó una competencia en reconocimiento de dígitos manuscritos [3], se presentaron los conjuntos de datos "NIST *Special Data* 3 (SD3)" y "NIST *Test Data* 1 (TD1)". Estos datos consistían en dos discos compactos con caracteres manuscritos. El conjunto de datos SD3 es una de las mayores colecciones de dígitos y caracteres aislados disponible públicamente. En total, contiene mas de 300.000 imágenes extraídas de formularios que fueron completados por 2.100 individuos. Para su creación los empleados del departamento de censo fueron capacitados para completar los casilleros de los formularios. Se les pidió especial cuidado de que los dígitos y caracteres no se tocasen entre si ni tocasen los bordes de los casilleros. Luego, estos formularios fueron escaneados a una resolución de 300 ppi en formato binario y segmentados automáticamente. La competencia organizada por NIST evaluó con este conjunto de datos la performance de más de 20 algoritmos diferentes que reconociesen dígitos y caracteres manuscritos aislados. Las imágenes de SD3 fueron utilizadas como conjunto de entrenamiento, y con ellas las organizaciones que compitieron construyeron sus modelos. Un conjunto de datos diferente, llamado TD1 fue el utilizado como conjunto de test [5]. Los datos de TD1 fueron obtenidos utilizando un proceso de recolección diferente. Los sujetos eran estudiantes de escuela secundaria y la calidad de los datos en general podría describirse como dígitos más desprolijos en TD1 respecto a los datos en SD3. Esto causó una reducción en la performance de los clasificadores por que la mayoría de los sistemas entrenaron con dígitos más prolijos. Muchos de los que participaron en la competencia se desilusionaron al observar que obtenian muy buenos resultados durante el entrenamiento, pero una performance mucho pero sobre los datos de testing [4]. Esta experiencia muestra que una técnica puede dar muy buenos resultados para un conjunto de datos dado, sin embargo no significa que el problema de reconocimiento haya sido resuelto. Podría pensarse que se ha resuelto solamente el problema para ese conjunto de datos.

La base de datos MNIST (NIST Modificada) surge de combinar SD3 y TD1. Debido a las diferencias de distribución en SD3 y TD1, se generó un conjunto de entrenamiento y testing que no tuviese esa separación tan diversa entre entrenamiento y testing [4] [11]. MNIST puede describirse como un conjunto de datos de imágenes con dígitos manuscritos del 0 al 9 donde cada dígito tiene 28x28 pixels. El conjunto de datos para entrenamiento es de 60.000 muestras y 10.000 destinadas

para test. Durante el estudio del estado del arte acerca de la lectura artificiál de dígitos manuscritos se ha observado que MNIST es muy utilizada en los trabajos científicos para mostrar sus resultados. Se aplican redes neuronales convolucionales con éxito en la clasificación de caracteres manuscritos [8] [9]. En [4] se comparan algunos clasificadores utilizando MNIST como base de datos, y estas son las tasas de error sobre el conjunto de test.

A	B	C	D	E	F
8.4 %	2.4 %	1.6 %	1.7 %	1.1 %	0.9 %

- A) Clasificador Lineal: 8.4 %.
- B) Clasificador por vecino más cercano (k=3): 2.4 %.
- C) Red neuronal multicapa totalmente conectada: 1.6 %.
- D,E,F) LeNet 1, 4 , 5.

En esta tesis se utiliza la base de datos MNIST para entrenamiento y testing; una vez entrenado y probado el modelo se testea con los dígitos de telegramas de elecciones en Argentina y se observa que la performance se reduce aproximadamente un 30 %. Como un aporte, se incrementa la base de datos original MNIST con distorciones sobre los dígitos y se espera mejorar la capacidad de generalización del modelo predictivo superando una tasa de acierto del 80 %.

3. Aporte esperado

Se espera desarrollar un método automático de lectura artificiál que clasifique los dígitos manuscritos de los telegramas y los compare con los datos tipeados manualmente. Se utilizaron tres tipos de datos:

- Dígitos manuscritos que provienen de una imagen.
- El número en texto plano, resultado del clasificador sobre la imagen.
- El número en texto plano, tipeado por una persona que lee el telegrama.

La minería de datos es aplicada en diferentes etapas de este trabajo, se utilizan árboles de decisión en la clasificación de las regiones del telegrama y y redes neuronales convolucionales en la clasificación de los dígitos manuscritos.

4. Plan de trabajo

Comienzo: Septiembre de 2014 y se planifica finalizar en Diciembre 2015 - Marzo 2016.
Etapas del desarrollo de la tesis:

1. **Primera etapa: Estudio del estado del arte en lectura de dígitos manuscritos.** Viabilidad técnica y legal de la utilización de los datos.
   Tiempo estimado: 3 meses.

2. **Segunda etapa: Obtención de manera automatizada todos los telegramas en formato PDF.** Los datos son puestos a disposición por la Dirección Nacional Electoral en función de la iniciativa Open Data. Convertir los telegramas a un formato de imagen (jpg ó png) para aplicar algunas de las técnicas de procesamiento de imágenes y extraer las regiones de interés.
   Tiempo estimado: 3 meses.

3. **Tercera etapa: Iniciar el informe y la extracción de los datos.** Desarrollar un sistema que identifique las regiones de interés en los telegramas y extraer los números manuscritos. Segmentar los números para reconocer los dígitos que lo componen.
   Tiempo estimado: 2 meses.

4. **Cuarta etapa: Entrenar red neuronal**. Utilizar la base de datos MNIST de dígitos manuscritos y entrenar una red neuronal convolucional que generalice la clasificación sobre los números manuscritos en telegramas. Exposición en Jornadas de la Maestría. Avanzar en el informe escrito.
   Tiempo estimado: 3 meses.

5. **Quinta etapa: Refinar los parámetros de la red neuronal convolucional.** Presentar el trabajo para su evaluación, y corrección de los jurados la presentación final.
   Tiempo estimado: 4 meses.

6. **Sexta etapa: Preparar los resultados y entrega la final.**
   Tiempo estimado: 2 meses.

Referencias

[1] Gustav Tauschek (1935). Reading machine. US Patent 2.026.330.

[2] P. W. Handel (1933), Statistical machine. US. Patent 1.915.993.

[3] M. Costa, E. Filippi and E. Pasero (1994), "A Modular cyclic Neural Network for character recognition", Proceedings of the INNS World Congress on Neural Networks (WCNN '94), S. Diego (CA), Vol 3, June 5-9, pp 204-210

[4] LeCun, Y., Jackel, L. D., Bottou, L., Cortes, C., Denker, J. S., Drucker, H. & Vapnik, V. (1995). Learning algorithms for classification: A comparison on handwritten digit recognition. Neural networks: the statistical mechanics perspective, 261, 276.

[5] Guyon, I., Haralick, R. M., Hull, J. J., & Phillips, I. T. Data sets for OCR and document image understanding research. Handbook of character recognition and document image analysis, 779-799. 1997

[6] Van der Zwaag, B. J. (2001). Handwritten digit recognition: A neural network demo. In Computational Intelligence. Theory and Applications (pp. 762-771). Springer Berlin Heidelberg.

[7] Tay, Y. H., Lallican, P. M., Khalid, M., Viard-Gaudin, C., & Kneer, S. (2001). An offline cursive handwritten word recognition system. In TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology (Vol. 2, pp. 519-524). IEEE.

[8] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[9] Le Cun, B. B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In Advances in neural information processing systems.

[10] Palacios, R., & Gupta, A. (2008). A system for processing handwritten bank checks automatically. Image and Vision Computing, 26(10), 1297-1313.

[11] LeCun, Y., Cortes, C., & Burges, C. J. (1998). The MNIST database of handwritten digits.

[12] Mori, S., Suen, C. Y., & Yamamoto, K. (1992). Historical review of OCR research and development. Proceedings of the IEEE, 80(7), 1029-1058.