UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

TFG del Grado en Ingeniería Informática

título del TFG

Presentado por Nombre del alumno
en Universidad de Burgos — 25 de junio de 2019
Tutor: nombre tutor

UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

D. nombre tutor, profesor del departamento de nombre departamento, área de nombre área.

Expone:

Que el alumno D. Nombre del alumno, con DNI dni, ha realizado el Trabajo final de Grado en Ingeniería Informática titulado título de TFG.

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 25 de junio de 2019

Vº. Bº. del Tutor:	Vº. Bº. del co-tutor:
D. nombre tutor	D. nombre co-tutor

Resumen

La Minería de Datos es un campo de la informática que se ocupa de la obtención de información a partir de los datos dados. En los últimos años se ha encontrado con un enfoque más profundo debido a la implicación que tiene en los negocios, la economía o la política [3]. Los factores clave con los que Data Mining está tratando son la inferencia de la información de los datos mediante la búsqueda y el hallazgo de patrones en los datos. Teniendo esto en cuenta, se han desarrollado muchos algoritmos de reconocimiento de patrones para facilitar este proceso[13].

La gran cantidad de datos, por otro lado, trae consigo dificultades a la hora de inferir información de los mismos. Sobre todo porque una gran cantidad de ella, al ser multietiquetada, trae consigo problemas en términos de tiempo y espacio a la hora de procesarla. Por lo tanto, se han propuesto algunas soluciones para hacer frente a este problema. Entre ellos, se encuentran los algoritmos de selección de características en los que se selecciona un cierto número de etiquetas teniendo en cuenta la relevancia que tienen para las instancias de datos.

Por lo tanto, en este proyecto algunos de estos algoritmos han sido implementados en la biblioteca de Sklearn ml-sklearn utilizando el lenguaje python. Para ello se ha utilizado la guía python pep y el sklearn. Para la experiencia más práctica del usuario, se han construido algunos portátiles, en los que se han utilizado diferentes conjuntos de datos. Junto con estos, también se han construido algunos gráficos para que se vea la eficacia. Los algoritmos construidos son los siguientes: Binary Relevance, Label Powerset y Label Construction for Feature Selection[**?**].

Descriptores

Multi-label, Feature Selection, Binary Relevance, Label Powerset, Data Mining ...

Abstract

Data Mining is a field in computer science which deals with the gain of information out of the data given. In recent years has encountered a more deep approach because of the implication it has in various fields in which business, economy or politics can be included[3]. The key factors that Data Mining is dealing with are the inference of information from data by searching and finding patterns in the data. With this taken into account, a lot of pattern recognition algorithms have been developed in order to facilitate this process[13].

The big amount of data, on the other hand, brings difficulties when it comes to infer information out of data. Specifically, this problem arouses when the data to be mined is multi-labeled[2]. This brings issues in terms of time and space during the process. Therefore, some solutions have been proposed in order to handle this issue. Among which, is the Feature Selection methods in which a certain number of labels are selected taking into account the relevance they have for the data instances[1].

Therefore in this project some of these algorithms have been implemented over the Sklearn (Scikit-Learn)[7] and scikit-multilearn library[9], by using the Python language[11]. The Python style guide (PeP [5]) and the guide for Sklearn have been used in doing so. For the more hand on experience of the user, some notebooks have been build, in which different datasets have been used. Along with these, some graphics have been also plotted in order for the efficacy to be seen. The algorithms constructed are the following: Binary Relevance[14], Label Powerset[1], RAndom k-labELsets (RAKEL)[10] and Label Construction for feature Selection[8].

Keywords

Multi-label, Feature Selection, Binary Relevance, Label Powerset, Data Mining, RAndom k-labELsets,

Índice general

Índice de figuras

Índice de tablas

Introduction

Data Mining is a field in computer science which deals with the gain of information out of the data given. In recent years has encountered a more deep approach because of the implication it has in various fields in which business, economy or politics can be included[3]. Using different methods from the field of machine learning, statistics or database systems, through the process of data mining patterns in the data sets are discovered. The key factors that Data Mining is dealing with are the inference of information from data by searching and finding patterns in the data. After the inference or extraction of the information, the phase of making this information comprehensible for different uses takes place[**?**]. Data mining is also the analysis step in "knowledge discovery in databases"process (KDD)[4]. With this taken into account, a lot of pattern recognition algorithms have been developed in order to facilitate this process[13].

multi label, single label and feature selecting algorithms

the data given can sometimes be single label wich turn to be a single class clasification problem. for example, when it comes to classicifation the algorithm would predict if the instance to be predicted is or not of a specific class, which turns this problem into a binary clasification problem. Example, if its a dog or not in an image.

But, as for many real worl applications, usually the data is not single labeld which leads to the complication of the problem. Multi labeled data can have more labels that have to be classified for each instance. For instance, an image can have as label more things, sand, water, palm trees, sun, sky, chair, becnh. In order for an algorith to classifie an image from this kind of data set, it should take all the labels into account in order to do so.

This is when feature selection methods come in hand, because some

labels can be relevant to the classification process and others no, when it comes to the classification. In our case, the labels sand, water and palm trees cand be very relevant, as opossed to the ones that may not be so relevant, as 'bench', or chair. Methods to do this is by binary relevance or label powerset.

feature selection

Objetivos del proyecto

Este apartado explica de forma precisa y concisa cuales son los objetivos que se persiguen con la realización del proyecto. Se puede distinguir entre los objetivos marcados por los requisitos del software a construir y los objetivos de carácter técnico que plantea a la hora de llevar a la práctica el proyecto.

Conceptos teóricos

En aquellos proyectos que necesiten para su comprensión y desarrollo de unos conceptos teóricos de una determinada materia o de un determinado dominio de conocimiento, debe existir un apartado que sintetice dichos conceptos.

Algunos conceptos teóricos de LaTeX[1].

3.1. Secciones

Las secciones se incluyen con el comando section.

Subsecciones

Además de secciones tenemos subsecciones.

Subsubsecciones

Y subsecciones.

3.2. Referencias

Las referencias se incluyen [13] en el texto usando cite [12]. Para citar webs, artículos o libros [6].

[1] Créditos a los proyectos de Álvaro López Cantero: Configurador de Presupuestos y Roberto Izquierdo Amo: PLQuiz

3.3. Imágenes

Se pueden incluir imágenes con los comandos standard de LaTeX, pero esta plantilla dispone de comandos propios como por ejemplo el siguiente:

Figura 3.1: Autómata para una expresión vacía

3.4. Listas de items

Existen tres posibilidades:

- primer item.
- segundo item.

1. primer item.
2. segundo item.

Primer item más información sobre el primer item.

Segundo item más información sobre el segundo item.

-

3.5. Tablas

Igualmente se pueden usar los comandos específicos de LaTeXo bien usar alguno de los comandos de la plantilla.

Herramientas	App AngularJS	API REST	BD	Memoria
HTML5	X			
CSS3	X			
BOOTSTRAP	X			
JavaScript	X			
AngularJS	X			
Bower	X			
PHP		X		
Karma + Jasmine	X			
Slim framework		X		
Idiorm		X		
Composer		X		
JSON	X	X		
PhpStorm	X	X		
MySQL			X	
PhpMyAdmin			X	
Git + BitBucket	X	X	X	X
MikTEX				X
TEXMaker				X
Astah				X
Balsamiq Mockups	X			
VersionOne	X	X	X	X

Tabla 3.1: Herramientas y tecnologías utilizadas en cada parte del proyecto

Técnicas y herramientas

Esta parte de la memoria tiene como objetivo presentar las técnicas metodológicas y las herramientas de desarrollo que se han utilizado para llevar a cabo el proyecto. Si se han estudiado diferentes alternativas de metodologías, herramientas, bibliotecas se puede hacer un resumen de los aspectos más destacados de cada alternativa, incluyendo comparativas entre las distintas opciones y una justificación de las elecciones realizadas. No se pretende que este apartado se convierta en un capítulo de un libro dedicado a cada una de las alternativas, sino comentar los aspectos más destacados de cada opción, con un repaso somero a los fundamentos esenciales y referencias bibliográficas para que el lector pueda ampliar su conocimiento sobre el tema.

Aspectos relevantes del desarrollo del proyecto

Este apartado pretende recoger los aspectos más interesantes del desarrollo del proyecto, comentados por los autores del mismo. Debe incluir desde la exposición del ciclo de vida utilizado, hasta los detalles de mayor relevancia de las fases de análisis, diseño e implementación. Se busca que no sea una mera operación de copiar y pegar diagramas y extractos del código fuente, sino que realmente se justifiquen los caminos de solución que se han tomado, especialmente aquellos que no sean triviales. Puede ser el lugar más adecuado para documentar los aspectos más interesantes del diseño y de la implementación, con un mayor hincapié en aspectos tales como el tipo de arquitectura elegido, los índices de las tablas de la base de datos, normalización y desnormalización, distribución en ficheros3, reglas de negocio dentro de las bases de datos (EDVHV GH GDWRV DFWLYDV), aspectos de desarrollo relacionados con el WWW... Este apartado, debe convertirse en el resumen de la experiencia práctica del proyecto, y por sí mismo justifica que la memoria se convierta en un documento útil, fuente de referencia para los autores, los tutores y futuros alumnos.

Trabajos relacionados

Este apartado sería parecido a un estado del arte de una tesis o tesina. En un trabajo final grado no parece obligada su presencia, aunque se puede dejar a juicio del tutor el incluir un pequeño resumen comentado de los trabajos y proyectos ya realizados en el campo del proyecto en curso.

Conclusiones y Líneas de trabajo futuras

Todo proyecto debe incluir las conclusiones que se derivan de su desarrollo. Éstas pueden ser de diferente índole, dependiendo de la tipología del proyecto, pero normalmente van a estar presentes un conjunto de conclusiones relacionadas con los resultados del proyecto y un conjunto de conclusiones técnicas. Además, resulta muy útil realizar un informe crítico indicando cómo se puede mejorar el proyecto, o cómo se puede continuar trabajando en la línea del proyecto realizado.

Bibliografía

[1] Rafael B. Pereira, Alexandre Plastino, Bianca Zadrozny, and Luiz Merschmann. Categorizing feature selection methods for multi-label classification. *Artificial Intelligence Review*, 49, 09 2016.

[2] Alex de Carvalho, Andre ansd Freitas. *A Tutorial on Multi-label Classification Techniques*, volume 205, pages 177–195. 07 2009.

[3] The Economist. The worlds most valuable resource is no longer oil but data, 2017. [Internet; donwloaded 11-June-2019].

[4] Usama Fayyad, Gregory Piatetsky-shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996.

[5] Nick Coghlan Guido van Rossum, Barry Warsaw. Style guide for python code, 2013. https://www.python.org/dev/peps/pep-0008/.

[6] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, 1992.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[8] Newton Spolaôr, Maria Carolina Monard, Grigorios Tsoumakas, and Huei Diana Lee. A systematic review of multi-label feature selection and a new method based on label construction. *Neurocomput.*, 180(C):3–15, March 2016.

[9] P. Szymański and T. Kajdanowicz. A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*, February 2017.

[10] Grigorios Tsoumakas and Ioannis Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In Joost N. Kok, Jacek Koronacki, Raomon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron, editors, *Machine Learning: ECML 2007*, pages 406–417, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[11] Guido Van Rossum and Fred L Drake. *Python language reference manual*. Network Theory, 2003.

[12] Wikipedia. Latex — wikipedia, la enciclopedia libre, 2015. [Internet; descargado 30-septiembre-2015].

[13] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[14] Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202, Apr 2018.