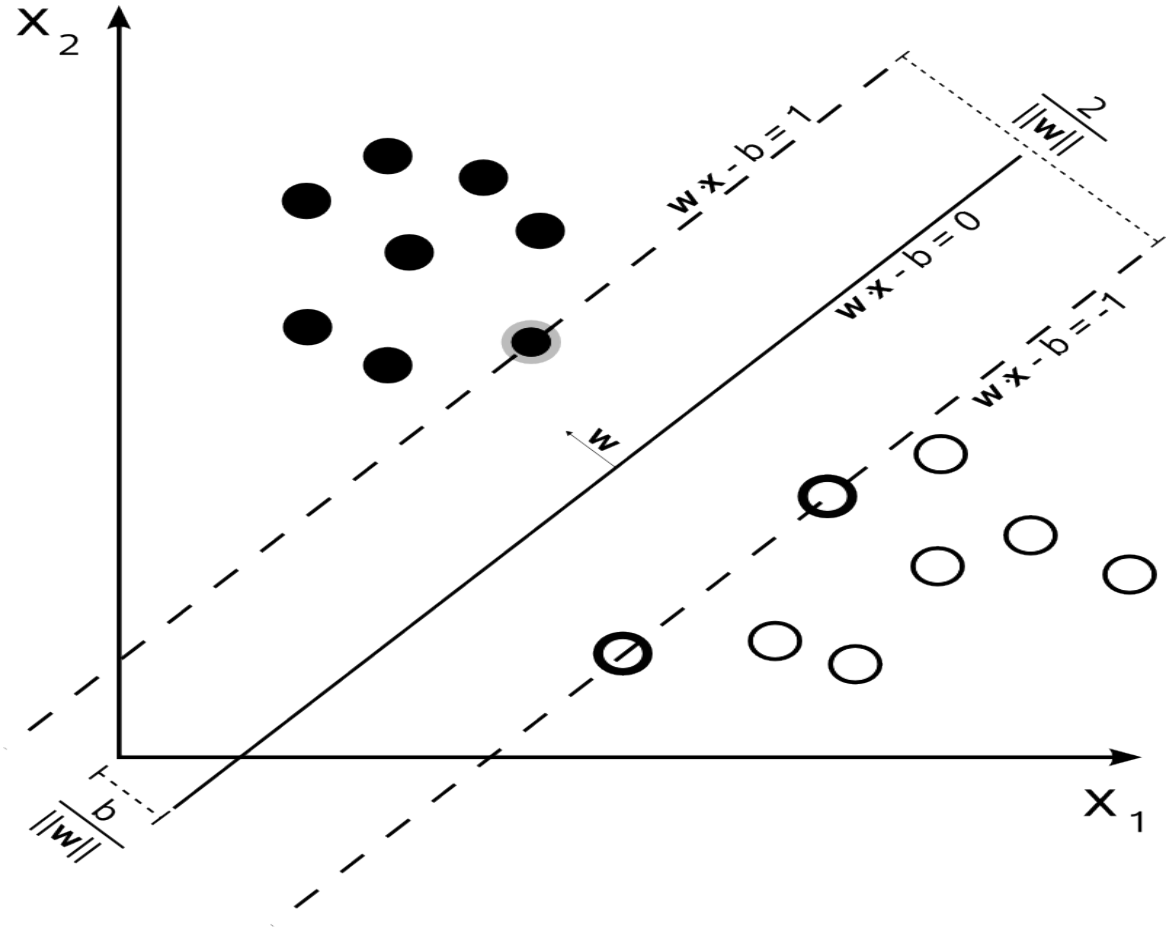
Machine Learning Workshop

Galicia 2016

Proceedings



Santiago de Compostela, 27 de octubre 2016

Comité Científico

Ricardo Cao	grupo MODES, CITIC e ITMATI, Universidade da Coruña
Milagros Fernández Gavilanes	Centro AtlantTIC, Universidade de Vigo
Andrés Gómez Tato	CESGA
David Mera Pérez	CiTIUS, Universidade de Santiago de Compostela

Comité Organizador

Andrés Gómez Tato	CESGA
Juan Touriño	Universidade da Coruña
Fernando Bouzas	CESGA
Raquel García	CESGA
Javier Cacheiro López	CESGA
José Carlos Mouriño	CESGA

Prefacio

Este volumen contiene los resúmenes de las presentaciones de **WGML2016: Workshop Machine Learning en Galicia 2016** celebrado el **27 de octubre de 2016 en Santiago de Compostela**. El objetivo de esta reunión fue presentar los proyectos y resultados de investigación de las tecnologías Machine Learning en las Universidades, centros de investigación, centros tecnológicos y empresas de Galicia, tanto a nivel de utilización como de desarrollos específicos de nuevos algoritmos, así como identificar las posibilidades de estas tecnologías en los sectores referentes de Galicia, las necesidades de infraestructuras y las posibles sinergias.

Se presentaron 25 trabajos de alto nivel que se resumen en este volumen. Su número y calidad indican la buena salud de este área de investigación e innovación en Galicia, tanto en aplicaciones industriales como en investigación básica.

Los organizadores quieren agradecer especialmente el apoyo de la Rede Galega de Tecnoloxías Cloud e Big Data para HPC, **R2014/041,** y la empresa TORUSWARE, sin cuyo apoyo económico no podría haberse realizado esta reunión. También al CSIC que prestó sus instalaciones para poder realizar la jornada, dado el elevado número de asistentes, más de 100, que desbordó las expectativas iniciales. Igualmente al CESGA, CITIC, CiTIUS y AtlantTIC, que colaboraron activamente en la organización de las jornadas. Finalmente, a los ponentes y asistentes, que son los verdaderos artífices del éxito de la reunión.

21 de octubre de 2016
Santiago de Compostela

El Comité Científico

Índice de presentaciones

Primera Sesión

Nonparametric Inference for big-but-biased data 1
Ricardo Cao and Laura Borrajo

Machine Learning Escalable con Spark ML en la plataforma BD—CESGA 2
Javier López Cacheiro

Aplicaciones del control estadístico de la calidad en eficiencia energética 3
Javier Tarrío Saavedra, Salvador Naya, Sonia Zaragoza, Miguel Flores, Manuel Febrero Bande and Manuel Oviedo

OTEA, el sistema experto que telegestiona instalaciones 4
Nerea Vilela Barreira, Anxo David Feijóo Lorenzo and Pedro Pérez Gabriel

Chequeando homogeneidad de dos muestras: Propuesta y Aplicaciones 5
Pablo Montero-Manso and José Vilar

Predicción puntual e intervalos de predicción en demanda y precio de la electricidad 6
Paula Raña, Juan Vilar and Germán Aneiros

Recommender Systems: machine learning vs. theoretical approaches 7
Paula Saavedra, Pablo Barreiro, Roi Durán, Ameed Almomani, Rosa Crujeiras, Maria Loureiro and Eduardo Sánchez Vila

Segunda Sesión

Distributed Embodied Evolution for Real-Time Optimization of Dynamic Engineering Problems 8
Abraham Prieto, Francisco Bellas, Pedro Trueba and Richard Duro

Sistema automatizado para la limpieza con láser de superficies no planas 9
Alberto Ramil, Javier Lamas and Ana J. López

Aplicación de Apache Spark y su librería MLlib para el desarrollo de sistemas recomendadores 10
Enrique Costa-Montenegro, Alexander Tsybanev, Héctor Cerezo-Costas, Francisco Javier González-Castaño, Felipe Gil-Castiñeira, Belén Barragáns-Martínez and Diego Almuiña-Troncoso

S-FRULER: aprendizaje automático escalable de reglas de predicción en Big Data 11
Ismael Rodríguez-Fernández, Manuel Mucientes and Alberto Bugarín

Using Deep Neural Networks for Discriminative Feature Localization 12
Javier Sánchez Rois and Daniel González Jiménez

Sistemas NLP para el análisis de sentimiento y detección de aspectos basados en machine learning 13
Milagros Fernández-Gavilanes, Jonathan Juncal-Martínez, Tamara Álvarez-López, Silvia García-Méndez, Enrique Costa-Montenegro and Francisco Javier González-Castaño

Tercera Sesión

Clasificación de Imágenes Hiperespectrales basada en Kernel ELM sobre GPU 14
Alberto S. Garea, Dora Blanco Heras and Francisco Argüello

Finis Terrae II como plataforma de Machine Learning 15
Andrés Gomez Tato, José Carlos Mouriño Gallego and Aurelio Rodríguez

Deep Learning para la detección de objetos en imágenes 16
Brais Bosquet, Manuel Mucientes and Victor Brea

Aplicación de técnicas de selección de características para la mejora de los sistemas automáticos de detección de vertidos de hidrocarburos 17
David Mera, Verónica Bolon-Canedo, Jose Manuel Cotos and Amparo Alonso-Betanzos

Retos en la Abstracción Semántica de Frases con Deep Learning 18
Héctor Cerezo-Costas

Machine learning for the management of agricultural soil data 19
M.S. Sirsat and M. Fernández Delgado

Cuarta Sesión

Predicción de la turbidez de un río con redes neuronales artificiales: aplicación al río Nalón 20
Carla Iglesias, Javier Martínez Torres and Javier Taboada Castro

BiGuardian: Sistema de detección proactiva y predictiva de amenazas de ciberseguridad .. 21
Diego Fustes Villadóniga, Eduardo San Miguel Martín and Juan Ramón González Hernández

Plataformas Big Data Eficientes y Escalables para Machine Learning 22
Guillermo L. Taboada

Desarrollo de un clasificador de placas de pizarra basado en técnicas de visión artificial y machine learning 23
Javier Martínez, Carla Iglesias and Javier Taboada

Detección de defectos en línea basado en Machine Learning 24
Jorge Rodríguez-Araújo and Antón García-Díaz

Robots que aprenden de ti y como tú. Aplicación en robots guía 25
Roberto Iglesias Rodríguez, Carlos Vázquez Regueiro, Xosé Manuel Parlo López and Miguel A. Rodríguez González

Nonparametric Inference for big-but-biased data

Ricardo Cao[1] *and Laura Borrajo*[1]
[1]Research Group MODES, Department of Mathematics, CITIC and ITMATI,
Campus de Elviña, Universidade da Coruña, 15071 A Coruña, Spain
e-mail: rcao@udc.es , laura.borrajo@udc.es

Crawford [1] has recently warned about the risks of the sentence "with enough data, the numbers speak for themselves". Some of the problems coming from ignoring sampling bias in big data statistical analysis has been recently reported by Cao [2]. The problem of nonparametric statistical inference in big data under the presence of sampling bias is considered in this work. The mean estimation problem is studied in this setup, in a nonparametric framework, when the biasing weight function is known (unrealistic) as well as for unknown weight functions (realistic). In the latter setup the problem is related to nonparametric density estimation. Asymptotic expressions for the mean squared error of the estimators proposed are considered. This leads to some asymptotic formula for the optimal smoothing parameter. The question of how big the sample size has to be to compensate the sampling bias in big data is considered. Some simulations illustrate the performance of the nonparametric methods proposed in this work.

[1] Crawford, K. The hidden biases in big data. *Harvard Business Review*, April 1st. (2013) Available at https://hbr.org/2013/04/the-hidden-biases-in-big-data
[2] Cao, R. Inferencia estadística con datos de gran volumen. *La Gaceta de la RSME*, **18** (2015) 393-417.

Machine Learning Escalable con Spark ML en la plataforma BD|CESGA

Javier López Cacheiro[1]
[1]CESGA, Avda de Vigo s/n, Campus Vida, Santiago de Compostela
e-mail: jlopez@cesga.es

Si tenemos en cuenta el aumento constante en el tamaño de los conjuntos de datos que son utilizados en los procesos de aprendizaje automático, nos damos cuenta que resulta cada vez más importante disponer de la capacidad de realizar los procesos de limpieza de datos, selección de características y aprendizaje de forma escalable, más allá de la capacidad de un sólo servidor.

En este sentido Spark ML, la nueva API de Machine Learning de Spark basada en DataFrames, nos proporciona algoritmos escalables que nos permitan realizar el proceso de aprendizaje de modo paralelo.

A través de un caso práctico mostraremos las posibilidades que nos ofrece Spark ML ejecutándose sobre la nueva plataforma Big Data del CESGA denominada BD|CESGA.

Aplicaciones del control estadístico de la calidad en eficiencia energética

Javier Tarrío Saavedra[1], *Salvador Naya*[1], *Sonia Zaragoza*[2], *Miguel Flores*[3], *Manuel Oviedo*[4], *Manuel Febrero*[4]

[1]Departamento de Matemáticas, Universidade da Coruña. Ferrol, España.

e-mail: javier.tarrio@udc.es

[2]Departamento de Ingeniería Industrial II, Universidade da Coruña. Ferrol, España.

[3]Escuela Politécnica Nacional. Quito, Ecuador.

[4]Departamento de Estadística e IO. Universidade de Santiago de Compostela. Santiago, España.

En este trabajo se presentan diversos casos de estudio en los que se han aplicado metodologías de aprendizaje estadístico adaptadas a la complejidad de los datos, dentro del marco del control estadístico de la calidad. Dentro de los denominados Building Management Systems (BMC), se han aplicado técnicas de gráficos de control univariantes, multivariantes y análisis de datos funcionales (FDA) para el control, análisis y mejora de instalaciones HVAC (heating, ventilating, and air conditioning), con datos obtenidos y gestionados a través de una Plataforma Web Energética, a partir del sistema Machine to Machine (M2M), mediante sensores de temperatura, humedad, CO_2 y consumo eléctrico. Los objetivos son establecer un sistema de control continuo de la calidad para detectar alarmas y situaciones atípicas, para identificar relaciones de dependencia entre variables críticas para la calidad del proceso y poder realizar predicciones. Todos estos objetivos están enfocados a incrementar la productividad, disminuir el consumo, cumplir con las especificaciones y facilitar la toma de decisiones de forma remota.

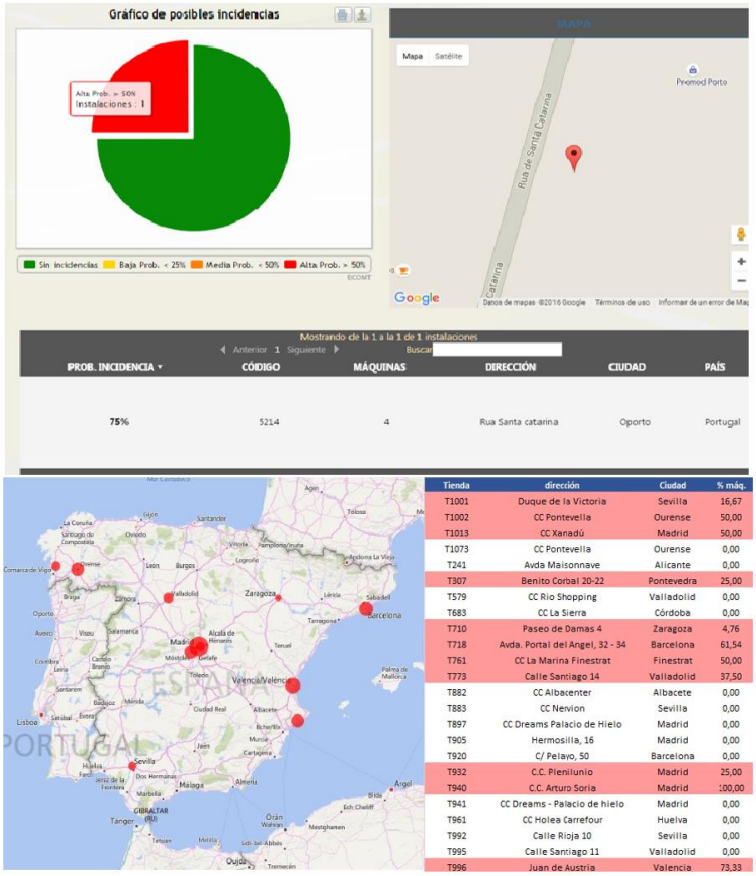
OTEA, EL SISTEMA EXPERTO QUE TELEGESTIONA INSTALACIONES

Anxo D. Feijóo Lorenzo[1], *Pedro Pérez Gabriel*[2] *y Nerea Vilela Barreira*[3]
[1] Director general, EcoMT. *afeijoo@ecomt.net*
[2] Consejero delegado, EcoMT. *pperez@ecomt.net*
[3] Responsable de I+D+i, EcoMT. *nvilela@ecomt.net*

Ecomanagement Technology, EcoMT, es una empresa TIC que analiza y gestiona la climatización e iluminación en más de 2.000 instalaciones con más de 600.000 variables y 3 billones desde hace más de 5 años. Para la telegestión se ha desarrollado un sistema software experto, OTEA, con el objetivo final de que sea capaz de tomar decisiones inteligentes para gestionar las instalaciones que controla sin intervención humana y con un bajo coste de implementación.

El comportamiento de los locales depende de variables con un marcado carácter probabilístico (temperatura exterior, temperatura ambiente, potencia de clima y general, ocupación, intervención de mantenimiento …), por lo que se está trabajando en la creación de un "oráculo" que pueda proponer soluciones a incidencias y genere reglas de funcionamiento para las cuestiones que el usuario no avanzado demande y no sea necesaria la consulta al experto. Algunos de los pasos que se han dado son el uso de redes neuronales, redes bayesianas y regresión, obteniendo algoritmos predictivos que generan mapas de riesgo que aportan información sobre que instalaciones tienen mayor probabilidad de incidencia provocando problemas de confort en determinadas circunstancias de ambiente y funcionamiento.

Probabilidad de incidencias en las instalaciones (Modelos de clasificación y regresión)



Tienda	dirección	Ciudad	% máq.
T1001	Duque de la Victoria	Sevilla	16,67
T1002	CC Pontevella	Ourense	50,00
T1013	CC Xanadú	Madrid	50,00
T1073	CC Pontevella	Ourense	0,00
T241	Avda Maisonnave	Alicante	0,00
T307	Benito Corbal 20-22	Pontevedra	25,00
T579	CC Río Shopping	Valladolid	0,00
T683	CC La Sierra	Córdoba	0,00
T710	Paseo de Damas 4	Zaragoza	4,76
T718	Avda. Portal del Angel, 32 - 34	Barcelona	61,54
T761	CC La Marina Finestrat	Finestrat	50,00
T773	Calle Santiago 14	Valladolid	37,50
T882	CC Albacenter	Albacete	0,00
T883	CC Nervión	Sevilla	0,00
T897	CC Dreams Palacio de Hielo	Madrid	0,00
T905	Hermosilla, 16	Madrid	0,00
T920	C/ Pelayo, 50	Barcelona	0,00
T932	C.C. Plenilunio	Madrid	25,00
T940	C.C. Arturo Soria	Madrid	100,00
T941	CC Dreams - Palacio de hielo	Madrid	0,00
T961	CC Holea Carrefour	Huelva	0,00
T992	Calle Rioja 10	Sevilla	0,00
T995	Calle Santiago 11	Valladolid	0,00
T996	Juan de Austria	Valencia	73,33

Referencias:
[1] Russell, S.J. y Norvig, P. Inteligencia artificial un enfoque moderno. Pearson (2004)
[2] Mayer-Schönberger, V. y Cukier, K. Big data. La revolución de los datos masivos. Turner (2013)

Chequeando homogeneidad de dos muestras: Propuesta y Aplicaciones

Pablo Montero Manso[1] *José A. Vilar Fernández*[1]
[1] Universidade da Coruña.
Grupo MODES. Departamento de Matemáticas. Facultad de Informática.
e-mail: p.montero.manso@udc.es jose.vilarf@udc.es

Chequear si dos conjuntos de datos han sido generados de la misma distribución de probabilidad es un tópico de gran interés que puede ser visto como un problema de aprendizaje supervisado, donde cada dato individual está etiquetado con el conjunto al que pertenece. Pruebas estadísticas con este fin han sido diseñadas para su aplicación a datos económicos [1], la comparación de series temporales de temperatura [2], evaluar similitud de campos en bases de datos [3] y en el contexto de problemas de aprendizaje automático como selección de variables [4], entre otras áreas. Para subsanar algunas de las limitaciones de los métodos clásicos, como puede ser el escenario de alta dimensionalidad y bajo tamaño muestral, se han propuesto desde la Estadística métodos basados en distancias [5,6] y desde Machine Learning métodos basados en kernels [3]. Distancias y kernels incorporan ventajas como la capacidad para comparar datos complejos, incluyendo datos categóricos [7] y grafos [3]. En este trabajo se presenta un nuevo método basado en distancias que, comparado a procedimientos similares propuestos recientemente, presenta mayor robustez al tipo de discrepancia entre las distribuciones generadoras y amplía el abanico de distancias que pueden ser utilizadas (propiedad deseable para tratar con datos de diferente naturaleza). Estas propiedades se ilustran mediante resultados de simulación y la aplicación a datos reales. Una discusión sobre posibilidades de aplicación exploradas recientemente enfatiza la utilidad del test propuesto.

.

Referencias

[1] Ma, Y., Wei L., and Hansheng W. A high dimensional two-sample test under a low dimensional factor structure. *Journal of Multivariate Analysis* **140** (2015): 162-170.

[2] Hall, P., and Nader T. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika* **89.2** (2002): 359-374.

[3] Gretton, A., et al. A kernel two-sample test. *Journal of Machine Learning Research* **13** (2012): 723-773.

[4] Landoni, E., et al. Parametric and nonparametric two-sample tests for feature screening in class comparison: a simulation study. *Epidemiology, Biostatistics and Public Health* **13** (2016).

[5] Henze, N. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics* (1988): 772-783.

[6] Székely, G., and Rizzo M L. Testing for equal distributions in high dimension. *InterStat* **5** (2004): 1-6.

[7] Cuadras, C. M. Distance analysis in discrimination and classification using both continuous and categorical variables. *Statistical data analysis and inference* (1989): 459-473.

Predicción puntual e intervalos de predicción en demanda y precio de la electricidad

Paula Raña[1], Juan Vilar[1] y Germán Aneiros[1].
[1]Departamento de Matemáticas, Universidade da Coruña.
e-mail: paula.rana@udc.es

Se aborda el problema de predicción puntual de demanda y precio de la electricidad mediante el uso de técnicas de análisis de datos funcionales. Los datos eléctricos componen una serie de tiempo funcional, en la que cada dato se corresponde con una curva diaria obtenida a partir de 24 observaciones horarias. Se propone el uso de modelos de regresión funcional para obtener predicciones en este contexto[1,2]. En primer lugar se considera un modelo de regresión funcional noparamétrico con respuesta escalar y explicativa funcional, en el que se predice la demanda o precio para una determinada hora de un día en función de la curva diaria anterior. En segundo lugar se propone un modelo semi-funcional parcialmente lineal en el que se añaden covariables escalares con efecto lineal sobre la respuesta. Dichas covariables incluyen información de temperatura, cuando se predice la demanda, e información de la propia demanda y de producción de energía eólica, cuando se predice el precio. Finalmente, mediante algoritmos bootstrap, se construyen intervalos de predicción que complementan a las predicciones puntuales obtenidas. Destacar que los métodos de predicción puntual e intervalos de predicción utilizados en este contexto se pueden aplicar en una amplia variedad de problemas.

Referencias

[1] Aneiros, G., Vilar, J and Raña, P. Short-term forecast of daily curves of electricity demand and price. *International Journal of Electrical Power and Energy Systems*, **80** (2016) 96-108.
[2] Raña, P., Aneiros, G., Vilar, J. and Vieu, P. Bootstrap confidence intervals in functional nonparametric regression under dependence. *Electronic Journal of Statistics*, **10(2)** (2016) 1973-1999.

Recommender Systems: machine learning vs. theoretical approaches

Paula Saavedra, Pablo Barreiro, Roi Durán, Ameed Almomani, Rosa Crujeiras, María Loureiro y Eduardo Sánchez Vila[1],
[1]CITIUS, University of Santiago de Compostela
e-mail: eduardo.sanchez.vila@usc.es

Recommender systems are personalization tools aimed at suggesting relevant products and items to end users. Mainstream companies like Google, Microsoft, Netflix, and Amazon, do apply these systems in a daily basis to continuously learn preferences, tastes and human behaviours. The generated profiles are then used to boost search engines, shopping baskets and the catalog of available products. Machine learning has played a big role in the development of the algorithms that work at the backend of these systems. Traditional content-based and user-based approaches rely on the popular k-Nearest Neighbours clustering technique to predict the utility/rating of an item. Nowadays, since the impact of the Netflix prize, ensemble models and learning-to-rank algorithms are the dominant concepts in the field. In spite of their current success, machine-learning algorithms show important limitations in order to satisfy the new demands of users. People are becoming reluctant to explicit recommendations as they feel it as another instrument of the advertising industry. As a result, there is an increasing need to explain why a website is recommending some items and not others. The complexity and opaque nature of machine-learning solutions make it hard to understand the reason why an algorithm delivers a certain solution. At this point, theoretical approaches come into rescue as they provide transparency and understanding about the calculations running behind the scenes. Following this line of thinking, we have developed choice-based models [1] that resort on decision-making principles to guide the recommendation process. We also discuss the future of recommender systems and how both theoretical and machine-learning approaches could work together to take advantage of the best of both approaches.

Referencias

[1] Saavedra P., Barreiro P., Durán R., Crujeiras R.,, Loureiro M., and Sánchez Vila E. Choice-based recommender systems. *Proceedings of RecSys'16*, Boston, 2016.

Distributed Embodied Evolution for Real-Time Optimization of Dynamic Engineering Problems

A. Prieto[1], *F. Bellas*[1], *P. Trueba*[1], *R.J. Duro*[1]
[1]Grupo Integrado de Ingeniería, Campus de Esteiro, Ferrol
Universidade da Coruña
e-mail: francisco.bellas@udc.es

There are several engineering optimization problems like routing, freight transportation, exploration, or layout design, which present a series of characteristics that make them very difficult to solve. Among these we find the absence of centralized updated information about all the variables, due to the spread out nature of the problems or lack of appropriate communications, or the dynamism of real-time operation. In this context, distributed population-based techniques have provided promising results by obtaining a solution through the concurrent behavior of several adequately constructed processing elements. The objective of our work is to study the application of a novel evolutionary paradigm, distributed Embodied Evolution (dEE), to obtain heterogeneous populations that solve this type of engineering optimization problems in real-time. dEE is inspired by natural evolution, and therefore, the individuals that make up the population are embodied and situated in an environment where they are forced to interact in a local, decentralized and asynchronous fashion. Hence, evolution in dEE is open-ended, leading to a paradigm that is intrinsically adaptive and highly suitable for real time learning in distributed dynamic problems, dEE interest has grown remarkably in the last decade, with several papers dealing successfully with different collective tasks, mainly in the multi-robot systems field.

This study is carried out applying a canonical version of dEE, which generalizes the three basic processes of evaluation, mating and replacement of a typical evolutionary algorithm. Moreover, in order to make it independent on the environment and specific task, the relevant evolutionary events have been replaced by stochastic variables, which follow specific probability functions.

- **Mating selection**: it has been modeled as an event that is triggered by a uniform probability function that depends on a single parameter, the probability of mating, that is $P_{mating} = \frac{S_{max}}{T_{max}}$, where S_{max} is the maximum window size of the tournament and T_{max} the maximum lifetime.
- **Selection policy**: the probability of being eligible as a candidate for mating ($P_{elegibility}$) is defined through a function that is based on the fitness value
- **Genotypic recombination**: a new intrinsic parameter is defined: the probability of using a local search strategy (P_{ls}), that is, a mutation operator. It is a measure of the exploration and exploitation balance through the ratio between crossover and mutation frequency.
- **Replacement**: the current canonical EE algorithm considers a fixed population size, therefore the replacement process in this case produces both, the removal of one current individual and the creation of a new one, and is modeled here as triggered by a replacement probability ($P_{replacement}$). This probability is defined based on a more intuitive and manageable parameter, which is the life expectancy (T_{exp}): $P_{replacement} = 1/T_{exp}$. T_{exp} is defined for each individual in each time step based on its current fitness, which depends on its genotype and the genotypes of the others.

We analyze in this study the canonical dEE response in two highly representative dynamic engineering problems: a Dynamic Fleet Size and Mix Vehicle Routing Problem with Time Windows (DFSMVRPTW) and a collective surveillance task with realistic location degradation, and we show the potential of this approach in such complex tasks.

References

[1] Trueba, P., Prieto, A., Bellas, F., Caamaño, P., Duro, R.J. Specialization analysis of embodied evolution for robotic collective tasks, *Robotics and Autonomous Systems*, Volume 61, Issue 7, 2013, pp. 682-693

[2] Prieto, A., Bellas, F., Trueba, P., Duro, R.J., Towards the standardization of distributed Embodied Evolution, *Information Sciences*, Elsevier, vol 312, 2015, pp. 55-77

[3] Prieto, A., Bellas, F., Trueba, P., Duro, R.J., Real-time optimization of dynamic problems through distributed Embodied Evolution, *Integrated Computer-Aided Engineering*, vol 23, 2016, pp 237-253

Sistema automatizado para la limpieza con láser de superficies no planas

Alberto Ramil[1], *Javier Lamas* [1], *Ana J. López* [1]
[1]Centro de Investigacións Tecnolóxicas. Escola Politécnica Superior. Universidade da Coruña.
Campus de Ferrol, 15471, Ferrol, Spain
e-mail: alberto.ramil@udc.es

Resumen:

Se presenta un sistema desarrollado para la limpieza automatizada con láser de superficies no planas. El sistema se basa en la adquisición del perfil de la superficie usando un escáner láser de línea acoplado a un sistema de control de tres ejes motorizados XYZ donde se sitúa la muestra. Utilizando el software desarrollado por nosotros los datos del perfil se emplean para generar un modelo de la superficie en forma de malla. Este modelo permite generar la trayectoria de los ejes motorizados de forma que el punto focal del haz del láser se mantenga a una distancia fija de cada punto de la superficie durante el proceso de limpieza. Para ello se generan automáticamente las instrucciones del controlador de los ejes que incluyen la activación del disparo del láser mediante una señal digital. Este sistema nos permite asegurar un tratamiento uniforme en toda la superficie. Hemos aplicado el sistema para eliminar costras y diferentes pátinas en rocas ornamentales pero puede ser aplicado en otros ámbitos.

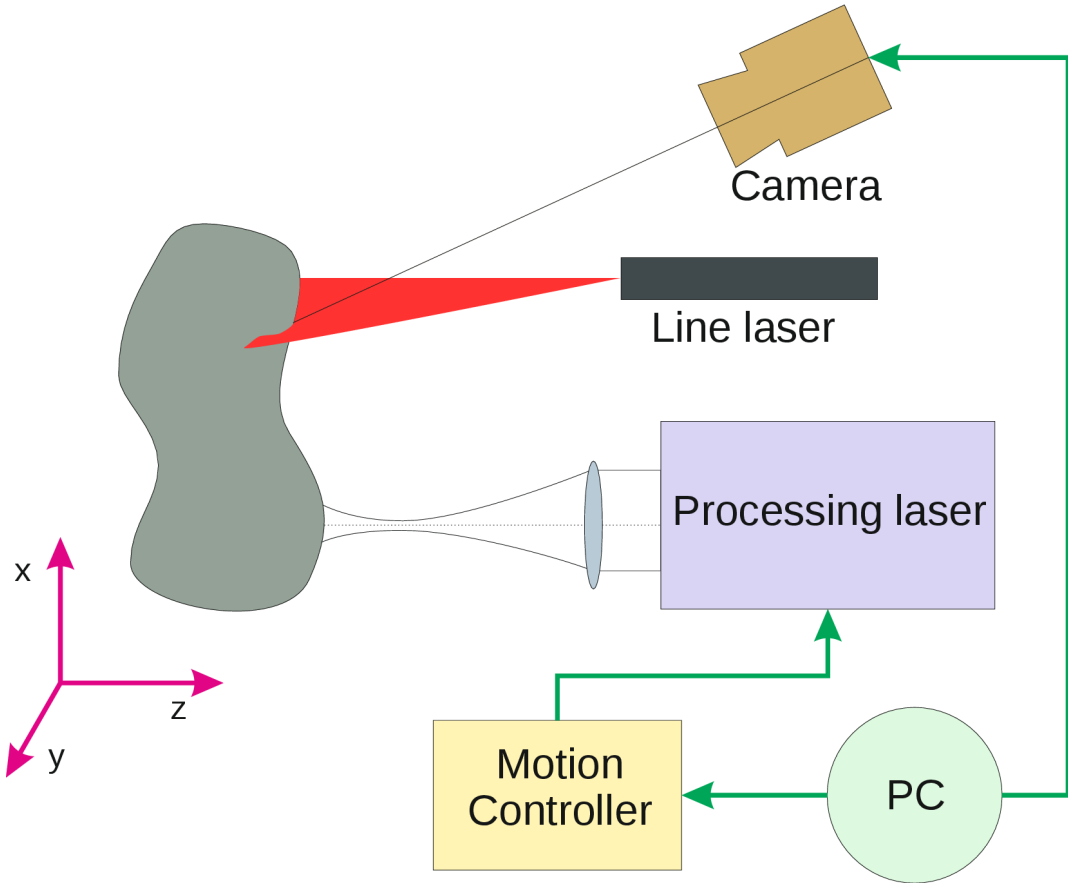


Figura 1. Esquema del sistema utilizado para procesar con láser materiales con superficies no planas.

Aplicación de Apache Spark y su librería MLlib para el desarrollo de sistemas recomendadores

Enrique Costa-Montenegro[1], *Alexander Tsybanev*[1], *Héctor Cerezo-Costas*[2], *Francisco Javier González-Castaño*[1], *Felipe Gil-Castiñeira*[1], *Belén Barragáns-Martínez*[3], *Diego Almuiña-Troncoso*[1]

[1]AtlantTIC, University of Vigo, Vigo, Pontevedra 36310, Spain

[2]Gradiant (Galician Research and Development Center in Advanced Telecomunications), Vigo, Pontevedra 36310, Spain

[3]Centro Universitario de la Defensa, Escuela Naval Militar, Marín, Pontevedra 36920, Spain

e-mail: kike@gti.uvigo.es, atsybanev@gti.uvigo.es, hcerezo@gradiant.org, javier@det.uvigo.es, xil@gti.uvigo.es, belen@cud.uvigo.es, dalmuina@gti.uvigo.es

Abstract:

La gran cantidad de información a la que pueden acceder fácilmente los usuarios en la actualidad hace necesario el uso de sistemas recomendadores para el procesado y filtrado de la misma. Para este fin se han propuesto diferentes tecnologías, entre las que destaca el reciente auge de las tecnologías Big-Data, cuyo uso nos planteamos en este trabajo.

En [1], presentamos de manera detallada la solución que hemos propuesto que hace uso de la tecnología Apache Spark junto con MLlib, su librería de Machine Learning, aplicada a un sistema recomendador de películas. Se demuestra en este trabajo que el uso de estas tecnologías facilita su implementación y mejora significativamente las prestaciones.

Para la evaluación se usaron dos datasets de películas (MovieLens y NetFlix) y se obtuvieron mejores valores de los parámetros empleados para evaluar la calidad de las recomendaciones con respecto a otros estudios en el estado del arte. También se desarrolló una interfaz web mediante la cual el usuario recibe las recomendaciones.

Esta contribución mostró las ventajas de la tecnología Apache Spark, presentando los resultados obtenidos cuando se ejecuta en un solo ordenador o cuando se hace uso de un cluster de ordenadores. Por último, se describió la influencia de diferentes parámetros (como *features*, *iterations*, *partitions*, *persistance*, etc.) en el funcionamiento de la tecnología Apache Spark.

Referencias

[1] Costa-Montenegro, Enrique; Tsybanev, Alexander; Cerezo-Costas, Héctor; González-Castaño, Francisco Javier; Gil-Castiñeira, Felipe; Barragáns-Martínez, Belén; Almuiña-Troncoso, Diego. *In-memory distributed software solution to improve the performance of recommender systems*. SOFTWARE—PRACTICE AND EXPERIENCE, En revisión (2016).

S-FRULER: aprendizaje automático escalable de reglas de predicción en Big Data

Ismael Rodríguez-Fdez, Manuel Mucientes, Alberto Bugarín[1]
[1]Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela
e-mail: manuel.mucientes@usc.es

S-FRULER es una versión distribuida y escalable de FRULER, que es un sistema genético-borroso que aprende automáticamente a partir de datos bases de conocimiento de tipo TSK-1 para problemas de regresión y predicción. S-FRULER obtiene modelos con una gran precisión y baja complejidad, y consigue reducciones significativas en el tiempo de cómputo, lo cual representa una gran ventaja en un ámbito como el aprendizaje automático, donde el tamaño de los datos influye de forma notable en la calidad de los modelos obtenidos. De forma transparente para el usuario, S-FRULER divide el problema en particiones más pequeñas e incorpora un proceso de selección de características para reducir el número de variables utilizadas en cada partición. Cada partición se resuelve a continuación de forma independiente utilizando el algoritmo FRULER. Después, una función de agregación obtiene la base de reglas final a partir de la información generada en cada partición. S-FRULER se ha aplicado en dos problemas reales de predicción: (i) en el ámbito de la bioinformática, la estimación del número de coordinación, utilizado para la predicción de la estructura 3D de cadenas de proteínas; (ii) en eficiencia energética de edificios, para la predicción de la temperatura del edificio y la temperatura de los depósitos de agua. La intervención del usuario en todo el proceso de aprendizaje y aplicación de los modelos obtenidos es mínima.

Using Deep Neural Networks for Discriminative Feature Localization

Javier Sánchez Rois, Daniel González Jiménez[1]
Àgata Lapedriza García[2]
[1]GRADIANT - Centro Tecnolóxico de Telecomunicacións de Galicia
[2]Universidad Oberta de Catalunya

Debido ós enormes logros dos sistemas de aprendizaxe visual baseados no uso de *convolutional neural networks* (CNN), o estado da arte no eido da visión por computador está a avanzar considerablemente. Comprender o proceso interno de aprendizaxe que se produce neste tipo de sistemas, así como as representacións de baixo nivel que son aprendidas nas capas internas das redes convolucionais é unha das claves da súa aplicación no problema da análise de escenas e a localización de obxectos. Recientes publicacións [1] plantexan a construcción de arquitecturas de rede orientadas á extracción da información máis discriminativa dentro dunha imaxe.

[1] Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. Learning deep features for scene recognition using places database. InAdvances in neural information processing systems (2014) 487-495.

Sistemas NLP para el análisis de sentimiento y detección de aspectos basados en machine learning

Milagros Fernández-Gavilanes[1], *Jonathan Juncal-Martínez*[1], *Tamara Álvarez-López*[1], *Silvia García-Méndez*[1], *Enrique Costa-Montenegro*[1], *Francisco Javier González-Castaño*[1]
[1]Grupo GTI, Departamento de Telemática, Escuela de Ingeniería de Telecomunicaciones.
Centro AtlantTIC, Campus Marcosende, Universidad de Vigo, 36310, Vigo
e-mail: {mfgavilanes, jonijm, talvarez, sgarcia, kike}@gti.uvigo.es, javier@det.uvigo.es

La investigación en el campo del análisis de sentimiento se ha incrementado considerablemente, en los últimos años debido al auge de contenidos generados por usuarios de diversas plataformas de Internet [1]. Conocer de antemano la opinión de estos usuarios hace que dichos contenidos sean considerados como valiososo en muchos sectores.

En este trabajo de investigación se ha desarrollado un sistema basado en técnicas de procesamiento del lenguaje natural (PLN) y machine learning, que permite automatizar el proceso de detección del sentimiento y detección de aspectos sobre los que trata. El sistema permite extraer una gran cantidad de información y peculiaridades lingüísticas expresadas por los usuarios en las redes. Teniendo en cuenta esa información, se ha desarrollado un algoritmo basado en dependencias sintácticas sin supervisión fácilmente adaptable a diversos entornos [2] sobre las que posteriormente se extraen características que son utilizadas por un sistema supervisado [3] basado en un clasificador Naive Bayes . Además, con el objetivo de determinar a qué hace referencia dicho sentimiento, se ha desarrollado un sistema de detección de aspectos usando CRF, en el que se enmarcan en función de categorías mediante máquinas de soporte vectorial (SVM) [4,5]. Todos estos sistemas se han diseñado para los idiomas inglés y español.

Referencias

[1] Liu, B. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies* (2012).
[2] Fernández-Gavilanes, M. ,Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., González-Castaño, F.J. Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications*, **58** (2016) 57-75.
[3] Juncal-Martínez, J., Álvarez-López, T., Fernández-Gavilanes, M., Costa-Montenegro, E., González-Castaño, F.J. GTI at SemEval-2016 Task 4: Training a Naive Bayes Classifier using Features of an Unsupervised System. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (2016) 115-119.
[4] Álvarez-López, T., Juncal-Martínez, J., Fernández-Gavilanes, M., Costa-Montenegro, E., González-Castaño, F.J. GTI at SemEval-2016 Task 5: SVM and CRF for Aspect Detection and Unsupervised Aspect-Based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (2016) 306-311.
[5] Álvarez-López, T., Fernández-Gavilanes, M., García-Méndez, S., Juncal-Martínez, J., González-Castaño, F.J. GTI at TASS 2016: Supervised Approach for Aspect Based Sentiment Analysis in Twitter. *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN* (2016) 53-57.

Clasificación de Imágenes Hiperespectrales basada en Kernel ELM sobre GPU

Alberto S. Garea , *Dora B. Heras, Francisco Argüello*
Centro Singular de Investigación en Tecnologías de la Información (CiTIUS)
Universidad de Santiago de Compostela
Rúa de Jenaro de la Fuente Domínguez, 15782 - Santiago de Compostela
e-mail: jorge.suarez.garea@usc.es; dora.blanco@usc.es; francisco.arguello@usc.es

El reciente desarrollo de sensores hiperespectrales que capturan información por pixel en un amplio rango del espectro electromagnético permite desarrollar aplicaciones en ámbitos que van desde la agricultura, el desarrollo urbano o la evolución de catástrofes naturales hasta ámbitos médicos [1]. Estas aplicaciones requieren realizar registrado, segmentación o clasificación entre otros procesos, con el objetivo de identificar elementos en las imágenes o de detectar cambios que se han producido a lo largo del tiempo. Este procesado es muy costoso computacionalmente y puede realizar utilizando diferentes técnicas muchas de las cuales se basan en algoritmos de *machine learning* [2]. Las técnicas deben además ser computadas eficientemente en plataformas computacionales de alto rendimiento.

Entre las diferentes técnicas que hemos aplicado para imágenes hiperespectrales podemos destacar las basadas en Support Vector Machines (SVMs) entre otras. Recientemente un algoritmo de aprendizaje basado en redes neuronales *feedforward* de capa simple llamado Extreme Learning Machine (ELM) [3] ha sido propuesto. Es eficiente términos de precisión de la clasificación, velocidad de aprendizaje y escalabilidad computacional. Presentamos resultados de aplicación de técnicas de clasificación basadas en la variante de ELM llamada kernel-ELM [4] y diseñadas para su ejecución eficiente en sistemas multicore y basados en GPU, que se aplican a imágenes hiperespectrales de teledetección de cobertura terrestre. Los resultados obtenidos superan en precisión a los disponibles en la bibliografía [5] y permiten la clasificación en tiempo real.

Referencias

[1] David Landgrebe, Hyperspectral image data analysis, *Signal Processing Magazine, IEEE*, **Número** 19:1 (2002) 17-28.

[2] Antonio Plaza, Jón Atli Benediktsson, Joseph W. Boardman, Jason Brazile, Lorenzo Bruzzone, Gustavo Camps-Valls, Jocelyn Chanussot, Mathieu Fauvel, Paolo Gamba, Anthony Gualtieri, et al., Recent advances in techniques for hyperspectral image processing, *Remote sensing of environment,* **Número** 113 (2009) 110-122.

[3] Guang-Bin Huang, An insight into extreme learning machines: random neurons, random features and kernels, *Cognitive Computation*, **Número** 6:3 (2014) 376-390.

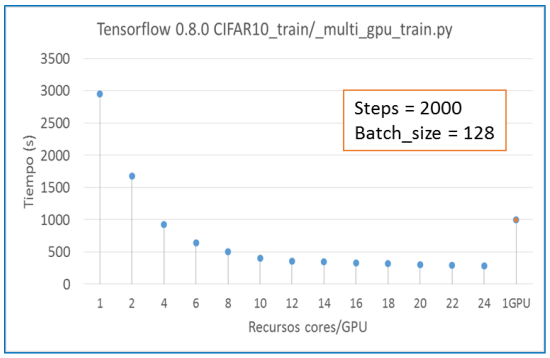
[4] Guang-Bin Huang, Hongming Zhou, Xiaojian Ding, and Rui Zhang, Extreme learning machine for regression and multiclass classification, *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **Número** 42:2 (2012) 513-529.

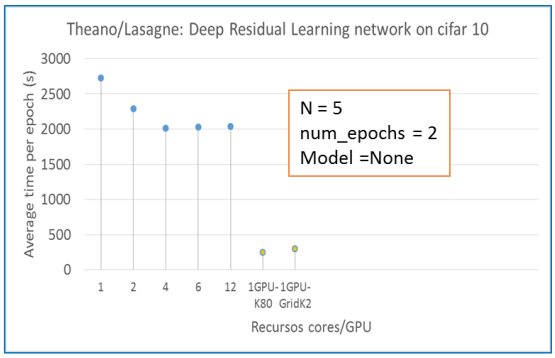
[5] Javier López-Fandiño, Pablo Quesada-Barriuso, Dora B. Heras, and Francisco Argüello, Efficient ELM-based techniques for the classification of hyperspectral remote sensing images on commodity GPUs, *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, **Número** 8:6 (2015), 2884–2893.

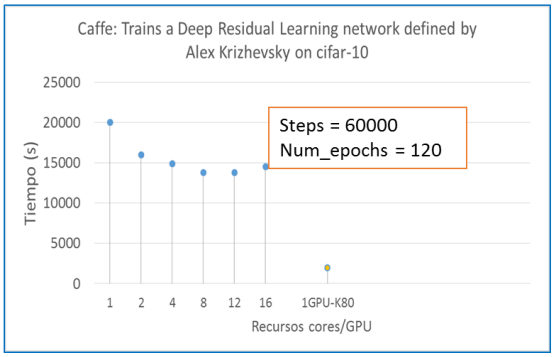
Finis Terrae II como plataforma de Machine Learning

A. Gómez[1], J.C. Mouriño, A. Rodríguez
[1]Fundación Pública Gallega Centro Tecnológico de Supercomputación de Galicia (CESGA)
e-mail: agomez@cesga.es

El entrenamiento de muchos de los algoritmos basados en tecnologías de Machine Learning requiere el acceso a grandes volúmenes de datos, abundante computación o ambos. Durante los últimos años han aparecido herramientas como TensorFlow[1], Caffe[2] o Theano[3] que simplifican el desarrollo de estos algoritmos, aprovechan las nuevas arquitecturas multi-núcleo, permiten la ejecución en distribuido y, en muchos casos, soportan la utilización de aceleradores como GPUs. El Finis Terrae, debido a su arquitectura y diseño, tiene posibilidades para acelerar el proceso de entrenamiento utilizando eficientemente esas capacidades, tanto usando la computación paralela en cada ejecución como ejecutando simultáneamente varios entrenamientos cuando se están realizando las búsquedas de parámetros óptimos. En esta presentación se mostrarán los primeros resultados obtenidos referentes a la utilización y escalabilidad de las tres herramientas citadas anteriormente obtenidos en el Finis Terrae. En algunos casos, los resultados preliminares muestran que el uso de varios núcleos del procesador Intel E2680 es más efectivo que su ejecución utilizando una GPU.







Referencias
[1] Martín Abadi, et al.. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
[2] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., … Darrell, T. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. ACM International Conference on Multimedia, 675–678. doi:10.1145/2647868.2654889 .
[3] The Theano Development Team, Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Zhang, Y. (2016). Theano: A Python framework for fast computation of mathematical expressions. arXiv E-Prints, abs/1605.0, 19. Retrieved from http://arxiv.org/abs/1605.02688

Deep Learning para la detección de objetos en imágenes

Brais Bosquet, Manuel Mucientes, Víctor Brea[1]
[1]Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela
e-mail: manuel.mucientes@usc.es

La utilización de redes neuronales profundas ha supuesto un importante avance en la detección de objetos en imágenes. Mediante técnicas de aprendizaje automático se determinan los pesos en redes neuronales con muchas capas (Deep Learning) con el fin de enmarcar y clasificar los diferentes objetos que aparecen en una imagen. En este trabajo se presentarán los últimos avances con redes neuronales convolucionales (CNNs) en dos líneas. En primer lugar, CNNs que generan regiones candidatas (en las que puede haber un objeto de interés), enmarcan de forma precisa el objeto de interés dentro de la región y lo clasifican, todo ello mediante una única CNN. En segundo lugar, CNNs que realizan la segmentación semántica de la imagen de una forma completamente convolucional, pudiendo recibir como entrada imágenes de cualquier tamaño y generando como salida una segmentación del tamaño correspondiente.

Aplicación de técnicas de selección de características para la mejora de los sistemas automáticos de detección de vertidos de hidrocarburos

David Mera[1], *Veronica Bolon-Canedo*[2], *J.M Cotos*[1], *Amparo Alonso-Betanzos*[2]
[1]Centro Singular de Investigación en Tecnoloxías da Información (CITIUS), Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez, 15782 - Santiago de Compostela, España.
[2]*Departamento de Computación, Universidade da Coruña, Campus de Elviña s/n, 15071 - A Coruña, España*

e-mail: david.mera@usc.es

Resumen:
Nuestras costas se ven afectadas habitualmente por vertidos de hidrocarburos generados en los pequeños accidentes y en las tareas de mantenimiento de los buques que las transitan. Un sistema de detección rápido y preciso es crucial para mitigar los efectos de los derrames y para perseguir a los infractores. Los sistemas de detección basados en el análisis de imagenes de radar de apertura sintético han demostrado ser muy eficientes en dicha tarea [1]. Basicamente estos sistemas segmentan todos los candidatos de una imagen, a continuación extraen un conjunto de características de cada uno de ellos y, finalmente, emplean algoritmos de machine learning para clasificarlos. La fase de caracterización acostumbra a ser computacionalmente costosa debido a la gran cantidad de candidatos que se suelen obtener durante la fase de segmentación. En general los sistemas de detección emplean conjuntos arbitrarios de características para caracterizar a los candidatos y, lamentablemente, escasean los trabajos que estudian su influencia en el resultado final. En este trabajo aplicamos y comparamos diferentes técnicas de selección de características sobre un conjunto formado por 141 elementos que representan a las características más habituales empleadas en los sistemas deteción de vertidos. El objetivo del estudio es eliminar las características irrelevantes y obtener un conjunto reducido que nos permita acelerar el proceso de caracterización y mejorar el rendimiento de los clasificadores. El resultado final del trabajo fue un vector de 6 características que fue validado a través de un clasificador SVM. Los experimentos revelaron que nuestro clasificador, empleando un menor número de características, obtenía un rendimiento similar o superior a los empleados en los trabajos previos analizados.

Referencias
[1] D. Mera, J. M. Cotos, J. Varela-Pet, P. G. Rodríguez, A. Caro, Automatic decision support system based on sar data for oil spill detection, Computers & Geosciences 72 (2014) 184–191

Palabras clave:
Machine Learning; Selección de características; Detección de vertidos; Clasificadores; SVM

Retos en la Abstracción Semántica de Frases con Deep Learning

Héctor Cerezo Costas[1]
[1]Gradiant
Edificio CITEXVI, local 14
Vigo, Pontevedra 36310, SPA
e-mail: hcerezo@gradiant.org

Los vectores de embeddings de palabras han supuesto un avance significativo en el procesado de lenguaje natural (PLN) con sistemas de aprendizaje máquina. En un vector n-dimensional son capaces de codificar información semántica de las palabras, obteniendo además un porcentaje alto de similitud con anotaciones humanas en varios tests [1]. Los vectores de embeddings de palabras se emplean junto con soluciones de *deep learning* para resolver problemas complejos en PLN a nivel de frase obteniendo los mejores resultados en muchas aplicaciones (análisis de sentimiento [2], traducción automática [3], etc).

Una gran ventaja de los embeddings de palabras es que se obtienen de forma no supervisada sobre millones de datos. Sin embargo, actualmente para hacer inferencia a nivel de frase hace falta entrenamiento específico, el cual es costoso de obtener para muchas aplicaciones. Por este motivo, es interesante obtener *embeddings* de frases sin supervisión humana que se puedan usar de forma genérica. Nuevos métodos que emplean recursos existentes como diccionarios o libros tratan de sistematizar la obtención de dichos embeddings (Skip-Thoughts [4], o DictRep [5]), siendo la principal baza de estos sistemas el gran volumen de información empleado para el entrenamiento. Estas soluciones no están exentas de retos, las implementaciones actuales de dichos métodos hacen un uso intensivo de recursos durante el aprendizaje de los modelos (p.e. el entrenamiento de Skip-Thoughts requiere de al menos dos semanas de computación sobre equipos con GPUs de última generación) y no existen métodos fiables que permitan asegurar que un método es superior a otro.

El objetivo de la presentación es ofrecer una visión de los métodos existentes para la obtención de ectores de frases y los retos que presentan en cuanto a precisión, velocidad de procesado, medida de su rendimiento y sus potenciales aplicaciones.

Referencias

[1] Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in Semantic Representation. Psychological review, 114(2), 211.
[2] Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved Semantic Representations from Tree-Structured Long Short-term Memory Networks. arXiv preprint arXiv:1503.00075.
[3] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence learning with Neural Networks. In Advances in neural information processing systems(pp. 3104-3112).
[4] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip- thought Vectors. In Advances in neural information processing systems (pp. 3294-3302).
[5] Hill, F., Cho, K., Korhonen, A., & Bengio, Y. (2015). Learning to Understand Phrases by Embedding the Dictionary. arXiv preprint arXiv:1504.00548.

Machine learning for the management of agricultural soil data

M.S. Sirsat, M. Fernández-Delgado[1]
[1]Centro Singular de Investigación en Tecnoloxías da Información da USC (CiTIUS)
e-mail: manuel.fernandez.delgado@usc.es

In this work, we use a variety of machine learning techniques for the classification of patterns composed by chemical measurements acquired from soils devoted to agriculture in the Indian state of Maharashtra. The India devotes 60.5% of its land to agriculture, which represents 11.3% of its Gross State Domestic Product. The machine learning methods are very interesting for the the design of a cultivation plan, recommendation of fertilizers and prediction of the soil fertility. Specifically, we classify several village-wise fertility indices (organic carbon, phosphorus pentoxide, manganese and iron), soil nutrients (nitrous oxide, phosphorus pentoxide and potassium oxide), preferable crop (bajra/soybean or irrigated/rainfed cotton), soil pH (slightly acidic/neutral/slightly alkaline/moderately alkaline) and soil type (light/medium). The classifiers, selected due to their good performances in the study [1], belong to several families including bagging (with several decision tree base classifiers), boosting (adaboost.M1), decision trees (J48, and decorate ensemble of J48 and recursive partitioning trees), K-nearest neighbors, neural networks (extreme learning machine with and without Gaussian kernel, multi-layer perceptron, probabilistic neural network and radial basis function neural network), random and rotation forests, rule-based (hybrid decision table-naive Bayes and ripper) classifiers and Gaussian kernel support vector machine. Globally, we apply 20 classifiers on 10 classification problems. The random forest achieves the best results for 6 of 10 problems, and adaboost.M1 is the best in 2 problems, while the support vector machine and decision table-naive Bayes are the best for 1 problem each one.

Referencias

[1] M. Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. Do we need hundreds of classifiers to solve real classification problems? J. Mach. Learn. Res., 15 (2015) 3133-3181.

Predicción de la turbidez de un río con redes neuronales artificiales: aplicación al río Nalón

Carla Iglesias[1], *Javier Martínez*[2], *Javier Taboada*[1]
[1]Departamento de Ingeniería de los Recursos Naturales y Medio Ambiente. ETSI de Minas. Universidad de Vigo.
e-mail: carlaiglesias@uvigo.es, jtaboada@uvigo.es
[2]Centro Universitario de la Defensa. Escuela Naval Militar. Marín, Pontevedra
e-mail: javier.martinez@cud.uvigo.es

Los controles de calidad del agua comprenden la medida de un buen número de parámetros químicos y físico-químicos [1], siendo frecuente su monitorización automática. En España, la red SAICA cubre el territorio nacional con unas 200 estaciones automáticas de alerta ubicadas en zonas con usos especialmente críticos, monitorizando en tiempo real el estado de las masas de agua. Sin embargo, la medida de todos los parámetros es costosa y precisa de tiempo.

En este trabajo de investigación [2] se han tomado los datos del año 2010 de una estación de la red SAICA en el río Nalón (Asturias), con medidas cada 5 minutos de turbidez, amonio, conductividad, oxígeno disuelto, pH y temperatura. Nuestro objetivo era predecir los valores de turbidez, un parámetro integrador de la calidad del agua cuyos equipos de medida son costosos y frágiles, a partir de los restantes parámetros citados, logrando así economizar recursos. Para ello, utilizamos redes neuronales artificiales con buenos resultados, demostrando la utilidad de las técnicas de machine learning en el análisis de la calidad de las aguas. Se trata de un trabajo inicial escalable al gran volumen de datos provenientes de la red SAICA, con una metodología aplicable a problemas similares (calidad del aire, p.e.).

Referencias

[1] Parlamento Europeo y el Consejo, Directiva 2000/60/CE del Parlamento Europeo y del Consejo, de 23 de octubre de 2000, por la que se establece un marco comunitario de actuación en el ámbito de la política de aguas, *D. Of. Las Comunidades Eur.* **L 327** (2000) 1–73.

[2] C. Iglesias, J. Martínez Torres, P.J. García Nieto, J.R. Alonso Fernández, C. Díaz Muñiz, J.I. Piñeiro, J. Taboada, Turbidity Prediction in a River Basin by Using Artificial Neural Networks: A Case Study in Northern Spain, *Water Resour. Manag.* **28** (2014) 319–331.

BiGuardian: Sistema de detección proactiva y predictiva de amenazas de ciberseguridad

Diego Fustes Villadóniga, Eduardo San Miguel Martín, Juan Ramón González Hernández
Oesía Networks S.L., Rúa Copérnico, 1, 15008 A Coruña
e-mail: dfustes@oesia.com, esanmiguel@oesia.com, jrgonzalezh@oesia.com

BiGuardian es un nuevo proyecto en desarrollo, destinado a combatir los distintos tipos de ciberataque mediante las más avanzadas y modernas tecnologías, incluyendo Big Data, Fast Data y Deep Learning. BiGuardian es capaz de adquirir datos en tiempo real, provenientes de diversos sensores instalados en la red interna (tráfico de red, logs de IDS, Firewall, aplicaciones, etc) y en la red externa (dorks de Google, redes sociales, foros, etc.). Los eventos adquiridos en tiempo real se integran, normalizan y enriquecen con información geográfica, listas negras, whois, etc. pasando a formar parte de un histórico infinito (gracias al Big Data) y a ser procesados con técnicas de Machine Learning en tiempo real (Fast Data). De esta forma, las técnicas de detección son capaces de ir más allá de lo que es posible con las tecnologías existentes, tanto en la cantidad de datos que pueden tratar, como en la velocidad a la que pueden ofrecer resultados, así como en la complejidad de los algoritmos utilizados. Se está realizando un estudio de algoritmos óptimos para el problema, con un enfoque centrado en la detección de anomalías semisupervisada, mediante técnicas de Deep Learning (autocodificadores apilados), análisis de grafos y árboles de decisión.

Plataformas Big Data Eficientes y Escalables para Machine Learning

Guillermo López Taboada[1,2]
[1] Grupo de Arquitectura de Computadores, Universidade da Coruña – A Coruña, España.
Web: http://gac.udc.es/~gltaboada
e-mail: taboada@udc.es
[2] *Torusware*
Web: http://www.torusware.com
e-mail: guillermo.lopez@torusware.com

Resumen:
En el contexto del Big Data, el rendimiento es clave, y al contrario de lo que pueda parecer, la infraestructura no es una commodity, siendo crítico diseñar y configurar las plataformas teniendo en cuenta su uso, en este caso el del Machine Learning (ML).
Para disponer de una plataforma eficiente, hay que tener en cuenta los diferentes motores de procesado de datos para Hadoop, como MapReduce, Spark, Flink, Storm y H 2 O. Los factores a tener en cuenta en su elección son latencia, ancho de banda, tolerancia a fallos, usabilidad, coste de recursos y escalabilidad.
Además, tenemos que tener en cuenta los kits de desarrollo de ML disponibles en cada uno de estos motores de procesado, entre otros Mahout, MLlib, Flink-ML, Samoa y H 2 O, los cuales deben ser evaluados en términos de escalabilidad, velocidad, cobertura, usabilidad y extensibilidad.
En esta charla se analizarán los diferentes factores a tener en cuenta dentro de una plataforma Big Data para ML, como los mencionados motores de procesado y kits de desarrollo para ML, con el objetivo de identificar los factores que determinan una plataforma eficiente, ilustrados con casos de éxito de Torusware en diseño, despliegue y operación de arquitecturas Big Data escalables y eficientes.

Desarrollo de un clasificador de placas de pizarra basado en técnicas de visión artificial y machine learning

Javier Martínez[1], *Carla Iglesias*[2], *Javier Taboada*[2]
[1]Centro Universitario de la Defensa. Escuela Naval Militar. Marín, Pontevedra.
javier.martinez@cud.uvigo.es
[2]Departamento de Explotación de Minas. ETSI de Minas. Universidad de Vigo.
carla.iglesias@uvigo.es, jtaboada@uvigo.es

En este trabajo de investigación, se ha desarrollado un sistema en base a técnicas de visión artificial y machine learning, que automatiza la fase de clasificación de las placas de pizarra que se lleva a cabo dentro del proceso de elaboración, realizada hasta el momento de forma manual por un experto en el área. Así, se configura un sistema híbrido 2D-3D láser escáner formado por una cámara lineal 2D y un láser escáner 3D, de tal modo que extrae la mayor cantidad posible de información de las placas de pizarra. En base a dicha información, se ha desarrollado un algoritmo basado en técnicas de visión artificial capaz de determinar un conjunto de variables que identifiquen cada placa de pizarra en función de los defectos contemplados en la normativa vigente [1]. A partir del conjunto de variables, se han implementado modelos de clasificación con técnicas de machine learning enfrentando los dos enfoques clásicos, el supervisado y el no supervisado [2]. Atendiendo a los resultados obtenidos en esta comparativa, se ha desarrollado un algoritmo novedoso de multiclasificación basado en SVM, clasificadores binarios 1vsAll y Direct Acyclic Graphs con el fin de mejorar los resultados obtenidos por los modelos más relevantes en este ámbito [3]. Los buenos resultados presentados demuestran la viabilidad técnica del prototipo desarrollado, tanto a nivel de detección de los defectos que condicionan la calidad final de la placa de pizarra, como de potencial de las técnicas de machine learning para la resolución del problema de clasificación de placas de pizarra para techar.
Una vez construido, testado y validado el prototipo anterior, se está desarrollando la evolución del mismo basando ahora la caracterización de las placas de pizarra en la información de una única cámara láser escáner 3D en color. Además, se han implementado nuevas mejoras de portabilidad al sistema conjunto y la implementación se está llevado a cabo bajo un único software de visión avanzado (HALCON).

Keywords: Pizarra, Clasificación, Machine Learning, Visión Artificial

Referencias

[1] López, M; Martínez, J; Matías, JM; Vilán, JA, Taboada, J. Application of a Hybrid 3D-2D Laser Scaninng System to the Characterization of Slate Slabs. *Sensors*, **10** (2010), 5949-5961.
[2] Martínez, J; López, M; Matías, JM; Taboada, J. Classifying Slate Tile quality using automated learning techniques. *Mathematical and Computer Modelling*, **57 (7-8)** (2013), 1716-1721.
[3] Martínez, J; Iglesias, C, Matías, JM; Taboada, J; Araújo, M. Solving the slate tile classification problemm using a DAGSVM multiclassification algorithm based on SVM binary Classifiers with a one-versus-all approach. *Applied Mathematics and Computation*. **230** (2014), 464-472.

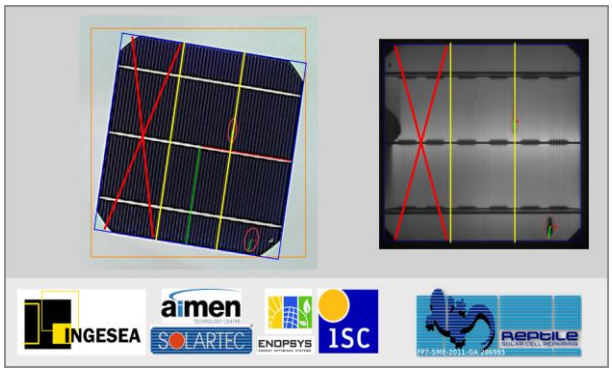
Detección de defectos en línea basado en *Machine Learning*

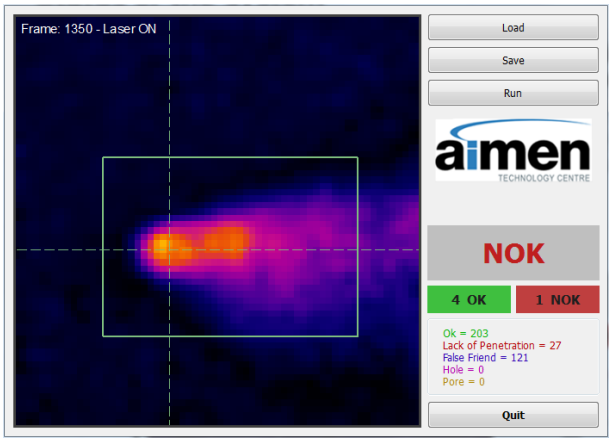
Jorge Rodríguez Araújo[1], *Antón García Díaz*
[1]Centro Tecnológico AIMEN, C/Relva, 27 A – Torneiros, 36410 Porriño – Pontevedra
e-mail: jorge.rodriguez@aimen.es

Los sistemas de producción demandan continuamente nuevos sistemas de control, predicción de fallos y detección de defectos que garanticen la calidad de los productos y mejoren la eficiencia de los procesos. Esta demanda, junto a la disponibilidad de entornos de computación más potentes está promoviendo el desarrollo de nuevas técnicas y sistemas basados en procesado de imagen y *machine learning* para la inspección y control de calidad en línea.

Así, basado en el análisis de imágenes de electroluminiscencia [1], se ha desarrollado una solución capaz de discriminar y localizar el tipo de defecto existente en una celda solar fotovoltaica usando máquinas de soporte vectorial (SVM), que además automatiza el proceso de reparación basado en láser. Lo que permite una significativa reducción de los desperdicios de producción mediante la utilización de celdas reparadas para la construcción de módulos a medida.

Por otro lado, mediante el análisis de imágenes térmicas de alta velocidad (obtenidas mediante sensores de imagen de PbSe no refrigerados en el rango MWIR) [2], se ha desarrollado una solución para la detección y clasificación de defectos en procesos de soldadura láser para automoción. La cual aplica el análisis de componentes principales (PCA) para la reducción dimensional de los datos del baño fundido, permitiendo el funcionamiento en línea (a una frecuencia de 1 kHz) y evitando posteriores inspecciones.





Referencias
[1] Rodríguez-Araújo, J., García-Díaz, A. Automated in-line defect classification and localization in solar cells for laser-based repair. *In 2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE)* (2014, June) pp. 1099-1104.
[2] Lapido, Y. L., Rodriguez-Araújo, J., García-Díaz, A., Castro, G., Vidal, F., Romero, P., Vergara, G. Cognitive high speed defect detection and classification in MWIR images of laser welding. *In Industrial Laser Applications Symposium 2015* (2015, July) pp. 96570B-96570B.

Robots que aprenden de ti y como tú. Aplicación en robots guía

R. Iglesias[1], *C.V. Regueiro*[2], *X. M. Pardo*[1], *M. A. Rodríguezs*[1]
[1]CiTIUS (Centro Singular de Investigación en Tecnoloxías da Información), Universidade de Santiago de Compostela
[2]Departamento de Electrónica y Sistemas, Facultad de Informática, Universidade da Coruña.
roberto.iglesias.rodriguez@usc.es

El aprendizaje máquina juega un papel imprescindible en al ámbito de la robótica, este trabajo recoge su importancia en los esfuerzos realizados por nuestro grupo de investigación con el fin de conseguir robots de servicio que aprendan de las personas y como las personas, y de forma más particular, en un robot guía en eventos o museos [1]. Por una parte han sido necesarios nuevos algoritmos basados en refuerzo que permitan que los robots aprendan incrementalmente a partir de su experiencia (interacción robot-entorno), incluso cuando existen muchos objetivos implícitos, o hay mucho ruido en la realimentación que el robot recibe del entorno [2]. Por otra parte, el robot debe construir su propia representación del "mundo", esto es, identificar eventos importantes en el flujo de información sensorial, memorizarlos temporalmente, y construir de forma adaptativa y dinámica *estados* que identifiquen el entorno en el que se encuentra. Con este fin hemos recurrimos a propuestas que crecen a partir de lo que se conoce como teoría de la resonancia-adaptativa. Estos elementos, junto con el reconocimiento de patrones temporales para el reconocimiento de gestos, han sido cruciales para el desarrollo del robot guía [3].

Finalmente, el aprendizaje máquina combinado con visión por computadora es también necesario en este tipo de robot, no solo para la identificación visual de las personas con las que el robot tiene que interaccionar, sino también para el reconocimiento de la escena: el comportamiento de los robots debe modularse no sólo con el tiempo sino también dependiendo de donde están. Los robots deben ser capaces de interpretar el entorno en el que se mueven, por lo que en este caso ha sido necesaria la construcción no supervisada de clasificadores capaces de "identificar" la escena a partir de imágenes adquiridas por el robot y que contienen suficiente información (imágenes canónicas) [4,5].

Referencias

[1] V. Alvarez-Santos, A. Canedo-Rodriguez, R. Iglesias, X.M. Pardo, C.V. Regueiro, M. Fernandez-Delgado, "Route learning and reproduction in a tour-guide robot". Robotics and autonomous systems, Vol. 63:206-213. 2015.
[2] J. García, Roberto Iglesias, Miguel A. Rodríguez, C. V. Regueiro, "Incremental Reinforcement Learning for multi-objective robotic tasks", Knowledge and Information Systems. 2016
[3] V. Alvarez-Santos, R. Iglesias, X.M. Pardo, C.V. Regueiro, A. Canedo-Rodriguez, "Gesture based interaction with voice feedback for a tour-guide robot", Journal of Visual Communication and Image Representation, Vol. 25(2):499–509. 2014.
[4] D.Santos-Saavedra, X.M. Pardo, R. Iglesias, "Canonical Views for Scene Recognition in Mobile Robotics", 7th Iberian Conference on Pattern Recognition and Image Analysis, IbPRIA 2015, Pattern Recognition and Image Analysis. LNCS 9117, 514-522, Springer. 2015
[5] David Santos-Saavedra, Roberto Iglesias and Xose M. Pardo, "Unsupervised Method to Remove Noisy and Redundant Images in Scene Recognition", Robot 2015: Second Iberian Robotics Conference. Advances in robotics, vol. 2". Advances in Intelligent Systems and Computing, Volume 418, págs.: 695-704, 2015

Índice de autores

Almomani, Ameed 7
Almuiña-Troncoso, Diego 10
Alonso-Betanzos, Amparo 17
Álvarez-López, Tamara 13
Aneiros, Germán 6
Argüello, Francisco 14

Barragáns-Martínez, Belén 10
Barreiro, Pablo 7
Bellas, Francisco 8
Blanco Heras, Dora 14
Bolón-Canedo, Verónica 17
Borrajo, Laura 1
Bosquet, Brais 16
Brea, Victor 16
Bugarín, Alberto 11

Cao, Ricardo 1
Cerezo-Costas, Héctor 10, 18
Costa-Montenegro, Enrique 10, 13
Cotos, José Manuel 17
Crujeiras, Rosa 7

Durán, Roi 7
Duro, Richard 8

Febrero Bande, Manuel 3
Feijóo Lorenzo, Anxo David 4
Fernández Delgado, M. 19
Fernández-Gavilanes, Milagros 13
Flores, Miguel 3
Fustes Villadóniga, Diego 21

García-Díaz, Antón 24
García-Méndez, Silvia 13
Gil-Castiñeira, Felipe 10
Gomez Tato, Andrés 15
González Hernández, Juan Ramón 21
González Jiménez, Daniel 12
González-Castaño, Francisco Javier 10, 13

Iglesias, Carla 20, 23
Iglesias Rodríguez, Roberto 25

Juncal-Martínez, Jonathan 13

L. Taboada, Guillermo 22
Lamas, Javier 9
López, Ana J. 9
López Cacheiro, Javier 2
Loureiro, Maria 7

Martínez, Javier 23
Martínez Torres, Javier 20
Mera, David 17
Montero-Manso, Pablo 5
Mouriño Gallego, José Carlos 15
Mucientes, Manuel 11, 16

Naya, Salvador 3

Oviedo, Manuel 3

Parlo López, Xosé Manuel 25
Pérez Gabriel, Pedro 4
Prieto, Abraham 8

Ramil, Alberto 9
Raña, Paula 6
Rodríguez, Aurelio 15
Rodríguez González, Miguel A. 25
Rodríguez-Araújo, Jorge 24
Rodríguez-Fernández, Ismael 11

S. Garea, Alberto 14
Saavedra, Paula 7
San Miguel Martín, Eduardo 21
Sánchez Rois, Javier 12
Sánchez Vila, Eduardo 7
Sirsat, M.S. 19

Taboada, Javier 23
Taboada Castro, Javier 20
Tarrío Saavedra, Javier 3
Trueba, Pedro 8
Tsybanev, Alexander 10

Vázquez Regueiro, Carlos 25
Vilar, José 5
Vilar, Juan 6
Vilela Barreira, Nerea 4

Índice de palabras clave

agricultura 19
análisis de datos funcionales 3
análisis de sentimiento 13
aprendizaje automático 2, 4
aprendizaje de reglas 11
aprendizaje por refuerzo en robótica 25
automatización 9

big data 1, 21, 22
bootstrap 6

Caffe 15
calidad de aguas 20
ciberseguridad 21
clasificación 19, 23
clasificadores 17
CNN 12
visión por computador 12
confort térmico. 3
crop recommendation 19

decision-making 7
deep learning 12, 16, 18, 21
detección de aspectos 13
detección de defectos 24
detección de objetos en imágenes 16
detección de vertidos 17
distancia 5
distribución 5

eficiencia energética 3, 4
embodied evolution 8
engineering optimization 8
entrenamiento 15
escalabilidad 22

fast data 21
Finis Terrae 15
forecasting 6
functional data 6

GPU 14
graph analytics 21
gráficos de control 3

Hadoop 2, 22
heterogeneous swarms 8
HVAC 3

imagen hiperespectral 14
informaciń imprecisa 1
incidencias 4
intervalos de predicción 6

kernel 5

laser 24

machine learning 2, 7, 10, 14, 15, 17, 21, 22, 23 24
métodos kernel 1
modelado superficial 9
monitorización automática 20

oráculo 4

PCA 24
performance 22
pizarra 23
PLN 13, 18
predicción en big data 11
procesado de materiales con láser 9
PySpark 2
Python 2

random forest 19
random utility models 7
reconocimiento de escenas 25
reconocimiento de patrones temporales 25
redes neuronales 12, 20

regresión en big data 11
robots guía 25
robótica 25

sistemas recomendadores 7, 10
segmentación semántica 16
selección de características 17
semántica 18
soil fertility 19
soil type 19
Spark 2, 10, 22

Spark ML 2
SVM 17, 24

teledetección 14
TensorFlow 15
test 5
Theano 15
tiempo real 14
two-sample 5

visión artificial 23

En colaboración con: Rede Galega de Tecnoloxías Cloud e Big Data para HPC y Torus Software Solutions S.L.