

UNIVERSIDAD DE CÓRDOBA



Departamento de Informática y Análisis Numérico

*Predicción ordinal utilizando metodologías de aprendizaje
automático: Aplicaciones*

Doctorado internacional

Programa de doctorado: Computación avanzada, energía y plasmas

Manuel Dorado Moreno

Directores

César Hervás Martínez

Pedro Antonio Gutiérrez Peña

Departamento de Informática y Análisis Numérico

Córdoba, septiembre de 2019

TITULO: *ORDINAL PREDICTION USING MACHINE LEARNING
METHODOLOGIES: APPLICATIONS*

AUTOR: *Manuel Dorado Moreno*

© Edita: UCOPress. 2019
Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

[https://www.uco.es/ucopress/index.php/es/
ucopress@uco.es](https://www.uco.es/ucopress/index.php/es/ucopress@uco.es)

UNIVERSITY OF CÓRDOBA



Department of Computer Science and Numerical Analysis

*Ordinal Prediction using machine learning methodologies:
Applications*

International Doctorate

Program: Advanced computing, energy and plasmas

Manuel Dorado Moreno

Supervisors

César Hervás Martínez

Pedro Antonio Gutiérrez Peña

Department of Computer Science and Numerical Analysis

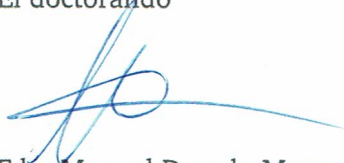
Córdoba, Sep 2019

La memoria titulada “Predicción ordinal utilizando metodologías de aprendizaje automático”, que presenta D. Manuel Dorado Moreno para optar al grado de Doctor, ha sido realizada dentro del programa de doctorado “Computación Avanzada, energía y plasmas” del Departamento de Informática y Análisis Numérico de la Universidad de Córdoba bajo la dirección del Doctor D. César Hervás Martínez y del Doctor D. Pedro Antonio Gutiérrez Peña.

El doctorando D. Manuel Dorado Moreno y los directores de la tesis D. César Hervás Martínez y D. Pedro Antonio Gutiérrez Peña garantizamos, al firmar esta Tesis Doctoral, que el trabajo ha sido realizado por el doctorando, bajo la dirección de los directores de la Tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

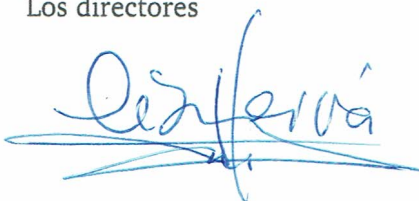
Córdoba, septiembre de 2019

El doctorando



Fdo: Manuel Dorado Moreno

Los directores



Fdo: César Hervás Martínez



Fdo: Pedro Antonio Gutiérrez Peña



TÍTULO DE LA TESIS:

Predicción ordinal utilizando metodologías de aprendizaje automático: Aplicaciones
Ordinal Prediction using machine learning methodologies: Applications

DOCTORANDO/A:

Manuel Dorado Moreno

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

En su Tesis, D. Manuel Dorado Moreno ha analizado diferentes aplicaciones del mundo real relacionadas con la clasificación ordinal, mediante el uso de técnicas de aprendizaje automático. La Tesis tiene cuatro partes principales: 1) en primer lugar, se proponen distintas técnicas de hibridación para obtener nuevos modelos de redes neuronales que permitan afrontar problemas de clasificación ordinal de forma más eficaz y eficiente (considerando, en esta parte, dos problemas reales de sociología); 2) posteriormente, se aborda un problema real de elevado interés, como es el de la asignación donante-receptor en trasplante de hígado, teniendo en cuenta características tanto del donante como del receptor y ofreciendo una alternativa más precisa a otros indicadores utilizados en medicina; 3) la Tesis continúa estudiando y ofreciendo distintas alternativas para otro problema real, el de la predicción de rampas de viento en parques eólicos, que tiene un gran interés debido al elevado coste que suponen este tipo de fenómenos para los gestores de las instalaciones; 4) el final de esta Tesis doctoral supone un nuevo enfoque metodológico para el problema ya mencionado asociado a la predicción de rampas de viento, de forma que se considera una combinación de modelos de aprendizaje profundo junto con aprendizaje multi-tarea, al considerar que la predicción de rampas de viento en puntos geográficos cercanos comparte mucha información común que puede ser aprovechada para mejorar la bondad de los modelos.

Estos resultados se ven avalados por la publicación de cinco artículos en revistas internacionales, un artículo de revista enviado y actualmente en segunda revisión y siete artículos presentados a congresos, de los cuáles seis son internacionales y uno es nacional. Queda patente de esta forma la calidad científica de las contribuciones de la Tesis. Esto nos lleva a presentar la Tesis como compendio de artículos.

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, 23 de septiembre de 2019

Firma del/de los director/es

Fdo.: César Hervás Martínez

Fdo.: Pedro Antonio Gutiérrez Peña

Esta Tesis Doctoral ha sido financiada en parte con cargo a los Proyectos **TIN2014-54583-C2-1-R** y **TIN2017-85887-C2-1-P** del Ministerio de Economía, Industria y Competitividad.

This work has been partially subsidized by the **TIN2014-54583-C2-1-R** and **TIN2017-85887-C2-1-P** projects from the Ministry of Economy, Industry and Competitiveness.



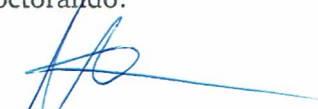
Mención de Doctorado Internacional

Esta tesis cumple los criterios para la obtención de la mención «Doctorado Internacional» concedida por la Universidad de Córdoba. Para ello se presentan los siguientes requisitos:

1. Estancia predoctoral realizada en otros países europeos:
 - **Università degli studi di Padova, Padua, Italia.** 3 meses de octubre a enero de 2018. Tutor de la estancia: **Dr. Alessandro Sperduti**, *Full professor* del Departamento de Matemáticas Aplicadas de la Universidad de Padua (Italia).
2. Esta tesis está avalada por los siguientes informes de idoneidad realizados por doctores de otros centros de investigación europeos:
 - **Dr. David Elizondo.** *Professor* de la Escuela de Ciencias de la Computación e Informática de la Universidad de De Montfort (Reino Unido) .
 - **Dr. Van Dinh Tran.** Investigador posdoctoral del departamento de Ciencias de la Computación en la Universidad de Freiburg (Alemania) .
3. La defensa de tesis y el texto se han realizado totalmente en inglés.
4. Entre los miembros del tribunal se encuentra un doctor procedente de un centro de educación superior europeo, tratándose del Dr. **Peter Tino**, *Full Professor* de la Escuela de Ciencias de la Computación de la Universidad de Birmingham (Reino Unido).

Córdoba, septiembre 2019

El doctorando:


Fdo.: Manuel Dorado Moreno

A César por todo el tiempo que me ha dedicado, por su paciencia infinita y sobretodo por la educación que me ha dado, no solo académica sino también profesional y personal que ha hecho que después de todos estos años trabajando con él me haya hecho mejor persona y profesional. Para mi, César no ha sido un jefe, sino un segundo padre, felicitándome cuando las cosas iban bien y enseñándome cómo hacerlas bien cuando iban mal.

A Pedro por ser mi guía durante estos años y enseñarme todo lo que he sé sobre Inteligencia Artificial, consiguiendo que sea algo que me apasione. También por enseñarme a enseñar durante los años que he tenido la suerte de dar clase junto a él, haciéndome mejor docente. Por último, por ser además de un buen director de tesis, un buen compañero de trabajo, dedicándome mucho de su tiempo y enseñándome a hacer las cosas bien.

A mis padres por apoyarme siempre, por haber sabido educarme y ponerme los pies en el suelo. Sin la dedicación que han invertido en mi durante toda mi vida, no defendería esta tesis, ni habría acabado mi carrera con las mejores calificaciones. Quien soy y adonde he conseguido llegar hasta ahora, en gran parte, es gracias a su educación, sus consejos y su apoyo incondicional.

A Elena por ser mi compañera de viaje desde que comencé la carrera, tanto en lo bueno como en lo malo. Por darme apoyo cuando estaba más agobiado y celebrar cualquier buena noticia. Por aguantar todas las veces que, estando de vacaciones o de descanso, ha estado a mi lado mientras preparaba la tesis y me animaba a terminar, en lugar de desentenderse. En resumen, por estar conmigo cada día desde el principio de mi carrera académica hasta el último haciéndome el camino mucho más fácil.

A los miembros de AYRNA, ya que de cada uno llevo una parte en mi. Por siempre, a pesar de tantos años y distintas situaciones, mantener un buen ambiente tanto en el laboratorio como fuera de él y por echarme una mano siempre que la he necesitado. Gracias en especial a Juan Carlos, Antonio Durán, David, Julio, Antonio Gómez, Javi y María por dejarme aprender de vosotros y por todos los buenos momentos.

Index

1	Introduction, motivation and objectives	1
1.1	Machine learning	2
1.1.1	Classification	5
1.1.2	Ordinal classification	6
1.1.3	Imbalanced classification	7
1.2	Artificial neural networks	10
1.2.1	Ordinal artificial neural networks	13
1.2.2	Echo State Networks	14
1.2.3	Deep multi-task learning	17
1.3	Areas of application	20
1.3.1	Liver allocation problem	20
1.3.2	Wind power ramp events	21
1.4	Motivation and challenges	23
1.5	Objectives	26
1.6	Summary of the thesis	27
1.7	Publications	29
2	Ordinal classification: hybridization of training algorithms and basis functions	33
2.1	Ordinal Classification Using Hybrid Artificial Neural Networks with Projection and Kernel Basis Functions	34
2.2	Rating the Rich: An Ordinal Classification to Determine Which Rich Countries are Helping Poorer Ones the Most	37
2.3	From outside to hyper-globalisation: an Artificial Neural Network ordinal classifier applied to measure the extent of globalisation	39
3	Ordinal prediction applications in medicine: the liver allocation problem	41
3.1	Ordinal Evolutionary Artificial Neural Networks for Solving an Imbalanced Liver Transplantation Problem	42
3.2	Dynamically weighted evolutionary ordinal neural network for solving an imbalanced liver transplantation problem	44
4	Ordinal prediction in time series: applications in wind energy	47
4.1	Robust estimation of wind power ramp events with reservoir computing	48
4.2	Combining Reservoir Computing and Over-Sampling for Ordinal Wind Power Ramp Prediction	50

4.3	Ordinal Multi-class Architecture for Predicting Wind Power Ramp Events Based on Reservoir Computing	52
4.4	Wind power ramp events ordinal prediction using minimum complexity echo state networks	54
5	Deep neural networks for multi-task learning	57
5.1	Multi-task Learning for the Prediction of Wind Power Ramp Events with Deep Neural Networks	58
6	Discussion and conclusions	61
6.1	Conclusions	62
6.1.1	Ordinal classification: hybridization of training algorithms and basis functions	62
6.1.2	Ordinal prediction applications in medicine: the liver allocation problem	63
6.1.3	Ordinal prediction in time series: applications in wind energy	64
6.1.4	Deep neural networks for multi-task learning	67
6.2	Generic discussion and future work	68

Figure index

1.1.1	Different examples of Machine learning applications.	3
1.1.2	Difference between classification and regression methodologies.	5
1.1.3	Example of an ordinal classification using thresholds.	7
1.1.4	Example of an imbalanced problem.	8
1.1.5	SMOTE applied to an imbalanced dataset.	9
1.2.1	Artificial Neural Network standard structure.	11
1.2.2	Ordinal Artificial Neural Network Structure.	13
1.2.3	Recurrent Neural Network structure.	15
1.2.4	Echo State Network reservoir and optional weights.	17
1.2.5	Multi-task learning applied to flower type and color classification.	18
1.2.6	Difference between standard and deep Neural Networks.	19
1.3.1	Graph showing the cumulative frequency of liver's graft loss.	21
1.3.2	Location of the three wind farms and the reanalysis nodes considered in the study.	23

1

Introduction, motivation and objectives

In the recent decades artificial intelligence has become part of our daily lives from the simplest applications, e.g., searching for information in a web browser, to more complex applications, such as automatically flying airplanes. There are millions of different uses for artificial intelligence and new applications are being developed every day. Most people relate it to humanoid robots that are capable of performing actions just as any human would do, but this is just part of science fiction, and even though research on humanoid robots is growing quite rapidly, the world will need some decades to see that transpire. The current artificial intelligence implementations focus on solving specific tasks, contributing with solutions that are correct for humans automatically, i.e., without the interaction of any person. Depending on the solution provided, it can be divided into many fields, such as: pattern recognition, data mining or machine learning.

For humans, classification is a simple task to perform, i.e., if someone is given a chair, he or she will instantly know that it is a chair, or if he or she sees a cat in the street, he or she will instantly know that it is a cat. We can envision in more complex tasks such as those involved in recognizing a person, and even without observing his or her face, someone who is known can be easily recognized just by his or her gait [49, 67]. This task, which is apparently a simple task for humans, is actually a very complex task, where thousands of variables need to be taken into account to be solved. People are constantly performing this task involuntarily, similar to breathing, but people become exhausted and can only recognize a few items in a short period of time. This is the point where machine learning

has an advantage over humans because once a machine learning model is prepared to perform a task, e.g., separate apples from pears, it can perform the task faster than humans and for longer periods of time, which is another benefit of developing machine learning algorithms in addition to providing solutions automatically.

In the beginning, artificial intelligence was successfully implemented in global companies, and slowly has made its way into international, national and local companies. This was mostly due to the considerable effort that researchers have made during the last decade, which improved all the classical methodologies to the extent that companies can implement them successfully, increasing their profitability. This is not the only important contribution that artificial intelligence has made; in areas such as the social sciences, medicine or renewable energies, it improves political decisions and investments in the social area; upgrades the medical processes, making them more successful and accessible to people and optimizes the production processes in renewable energy power plants.

Currently, online (and offline) available data are growing exponentially, e.g., every minute, more than a thousand Amazon packages are shipped, users perform nearly three million searches on Google and Americans use more than 3 million gigabytes of internet data. All these data are very valuable for the involved companies because they want to extract meaningful conclusions from them, and then, make optimal business decisions. Of course, that amount of information cannot be processed by a person, which is why machine learning is so influential. Additionally, banks, hospitals, industries and many other fields produce a huge amount of valuable data that need to be processed in one or another way to allow humans to make the most appropriate decision when solving a specific problem.

Most of the machine learning algorithms are structured in the same way: first, they need data to be trained; these data are produced by companies and users and will contain variables and objectives. Using the previous example, this training phase will relate the input variables (colour, shape, size, etc.) with an objective (differentiate apples from pears) in some way. After the training is finished, the model will be ready to be deployed in the machine that separates apples from pears and start working automatically. What does this mean? This means that the keys in artificial intelligence are the data, the model and the training algorithm, which are the main research fields in this area.

1.1. Machine learning

Machine learning is a branch of artificial intelligence whose goal is to develop techniques that allow computers to learn behaviours. This knowledge is acquired using training algorithms that build mathematical models and improve them to perform a specific task

correctly. This task is defined in the training data, which is a group of dependent and independent variables whose relation needs to be found with a mathematical model; that is why machine learning sometimes overlaps with statistics.

Machine learning, similar to artificial intelligence, is an interdisciplinary field as it can be applied in many different areas, as shown in Figure 1.1.1, which are not necessarily related to computer science, e.g., in health it is used for disease detection [19], in security it is used for tracking people with rare behaviours [1, 18], in agriculture it is used to predict the quantity of irrigation that a land will need [26] and finally, another field where machine learning has been applied widely is biometry [45]. While all of these applications need a computer to run the machine learning model, they are not computer science solutions, illustrating that machine learning is spreading fast and widely because it can be applied in any field that has a problem that can be solved automatically.



Figure 1.1.1: Different examples of Machine learning applications.

Machine learning methodologies can be classified according to some different criteria:

- Depending on the objective nature: whether the objective (label) is discrete (classification models) or continuous (regression models).
- Depending on the learning process: distinguishing between supervised (known objective) and unsupervised learning (unknown objective values).
- Depending on the reasoning process: inductive, if the model is trained using observed patterns and generalizes general rules, or transductive, if the model is trained using general rules to generalize unknown patterns.

- Depending on the model nature: it can be a probabilistic model or it can fit a deterministic one.
- Depending on the model itself: it would be discriminative if it simply focuses on discriminating two classes or generative if it states how the observations are assumed to have been generated.

While we consider the categories that define machine learning methods to be sufficient, there are, however, different criteria that can be applied to categorize machine learning models. In this thesis, the main categories that need to be clearly differentiated are regression/classification tasks and unsupervised/supervised learning; thus, a deeper definition will be needed.

Unsupervised learning [34] groups models whose objective is to find the clusters that group patterns into different classes. These patterns are initially only random data, as they are not labelled. The only counterpart in this branch is that there is no ground truth to compare the results with the actual cluster distribution.

However this thesis will focus on supervised learning [6], where the patterns are labelled before the training phase and these labels can be used to teach the model, so the main problem is to design a model capable of learning the underlying relation between the data and the labels. Here, we can differentiate two types of models, those that provide a clear intuition of how the result was obtained, i.e., given the input variables, a person can check the model and view clearly how they were transformed into the objective. The other types are known as black box models and are those that receive the input data, compute hundreds or thousands of mathematical operations and yield a result, and depending on the application, the data scientist will need to decide which model needs to be deployed.

Focusing on supervised learning and depending on the label nature, we can find regression techniques [32], where the labels to be learned are continuous, and classification techniques [41], where the labels are discrete. There are subdivisions of these two big branches; however, there is a learning paradigm that can be located between classification and regression, the one known both as ordinal regression or ordinal classification [29]. The reason why this paradigm lies between the other two is that the labels are discrete, such as those in classification tasks, but they follow a natural (or logical) order, such as continuous labels, where one label is higher or lower than the other label. Further explanation of classification and ordinal classification will be given in the following sections as they are the starting point of this thesis.

1.1.1. Classification

Together with regression, the classification task is one of the most common approaches that we encounter when tackling machine learning problems [66, 21, 62]. For instance, suppose that we want to predict if a person has a disease or not. For many years, we have been collecting information about millions of patients, including their characteristics and their symptoms, and then relating all that information to a disease. Using all of that information, we can focus on that specific disease and create a machine learning model that, in its training phase, will build the necessary mathematical structures to check, given the needed information, if that person has that disease. In this case, we have a binary classification task, as the objective has only two possibilities (yes and no); however, if the disease had multiple degrees (mild, moderate and severe), or if we wanted to predict which disease it is among a group of diseases (cold, flu or seasonal allergy), we would be defining a multi-class classification problem.

The difference in regression (see Figure 1.1.2) is mainly that the objective in regression is continuous instead of discrete, e.g., the *e immunoglobulin* that measures the level of allergy of a person ranges from 0 to 200, where 0 means no allergy. Therefore, predicting that *e immunoglobulin* level would have to be approached using regression techniques. In medicine, if that value is lower than 0.35 it is considered *class 0*, from 0.35 to 0.70 it is considered *class 1*, from 0.70 to 3.5 it is *class 2*, up to 17.5 it is *class 3*, up to 50 it is *class 4*, up to 100 it is *class 5* and finally, from 100 to 200 it would be *class 6*. . Now, the problem was discretized and the objective is not continuous anymore, but the classes follow a natural order (from *class 0* to *class 6*), such that problem should be approached using an ordinal classification model, which is the cornerstone of this thesis and will be detailed in the next section.

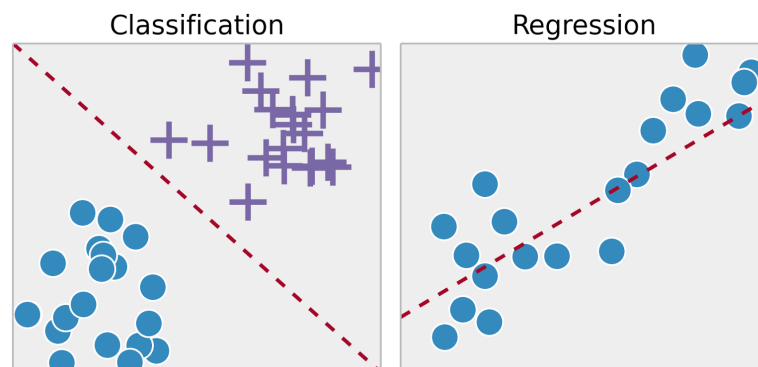


Figure 1.1.2: Difference between classification and regression methodologies.

1.1.2. Ordinal classification

Many fields of study need to classify patterns into naturally ordered classes [48, 17, 55], and in order to approach that classification, data scientists use what is called ordinal classification. That paradigm is located somewhere between regression and classification, so this type of classification has been traditionally handled by conventional methods for nominal classification (where the labels are not ordered). These kinds of supervised learning problems are also referred to as ordinal regression, where an ordinal scale ($Class_1 \prec Class_2 \prec \dots \prec Class_J$) is used to label the patterns. Therefore, similar to nominal classification, the goal is to learn how to classify those patterns in the correct class, but in ordinal classification, one should take into account that the greater the distance between the predicted and real class, the more the misclassification error committed [2].

Let us take the example from the previous section to describe the misclassification error in greater detail. After developing a machine learning model to predict the allergy degree of a person, it can predict that a person having an allergy of $Class_1$ has an allergy of $Class_2$, and the error will not be severe. However, if the machine learning model instead predicts that this person has an allergy of $Class_6$, then the misclassification error is much larger and it should be penalized more, which is the main difference from nominal classification, where the order or the class is not taken into account.

We can define ordinal classification more formally: consider an input vector \mathbf{x}_i and the objective label y_i , where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^k$ and $y_i \in \mathcal{Y} \in \{C_1, C_2, \dots, C_J\}$ (J is the number of classes) and $C_1 \prec C_2 \prec \dots \prec C_J$. The symbol \prec expresses that a label is before another in the ordinal scale. A classification rule or function has to be estimated, $f : \mathcal{X} \rightarrow \mathcal{Y}$, capable of predicting the categories of new patterns. As our setting is supervised learning, a training set of N patterns is given, $D = \{(\mathbf{x}_i, y_i), 1 \leq i \leq N\}$.

When solving ordinal problems, traditionally, nominal classifiers or regression methods were used [58, 9]. When using nominal classifiers, the ordinal information is lost; moreover, when using regression methods, one has to assume that the distances between the categories are known and equal as the labels are replaced by numbers, which is not true in most of the cases. In both of the cases, the information that is provided by the labels has to be figured up; however, this can be avoided using ordinal classification techniques, as they have proved to lead to better performance when dealing with ordinal problems. Researchers have approached these problems in some different ways; some of them approached the ordinal problem by decomposing it into a set of binary classification tasks or by formulating the problem as a larger augmented binary classification. However, the most used techniques are those based on thresholds, where one assumes that there is an unobservable latent variable that represents the output labels in a continuous function.

This means that, in order to solve ordinal problems using threshold models, we need to estimate:

1. The function $f(x)$ that predicts the latent variable trying to imitate its behaviour.
2. A set of $J - 1$ thresholds to define the intervals for every class, taking into account that, $b_1 \leq b_2 \leq \dots \leq b_{J-1}$

An example of ordinal classification following this threshold approach is shown in Figure 1.1.3. The idea underlying these two estimations is to give the model enough freedom to find the function $f(x)$ and the set of thresholds b that keeps the classes ordered and, at the same time, provide a greater separation of the classes in the following way:

$$C(x) = \begin{cases} C_1, & \text{if } f(x, \theta) \leq b_0^1, \\ C_2, & \text{if } f(x, \theta) \leq b_0^2, \\ \dots & \\ C_J, & \text{if } f(x, \theta) > b_0^{J-1}. \end{cases} \quad (1.1)$$

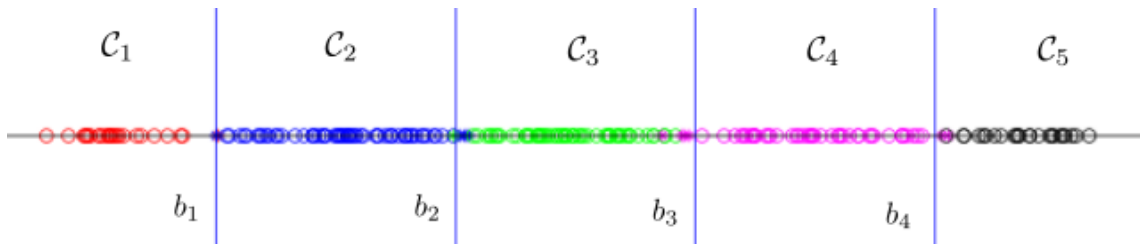


Figure 1.1.3: Example of an ordinal classification using thresholds.

1.1.3. Imbalanced classification

One of the most common issues when dealing with ordinal classification tasks is that the data are not distributed homogeneously among the classes, as shown in Figure 1.1.4 where one or more of the classes clearly have fewer patterns than the others. If we go back to the allergy degree prediction (see Section 1.1.1, most of the patterns will belong to $Class_0$ and $Class_1$ as they are the most frequent ones; in contrast, only a small percentage of the population is located in $Class_6$. One could conjecture that, as they are a minority, their misclassification is not important; however, the reality is that in most of the cases, these minority classes are the most important to be correctly predicted, e.g., in cancer detection, most people will not have this illness, but what is important is to correctly detect the people who have cancer.

This is a vast disadvantage for machine learning algorithms, as they need patterns to be trained properly. Briefly, when training a model, it tries to classify a pattern with

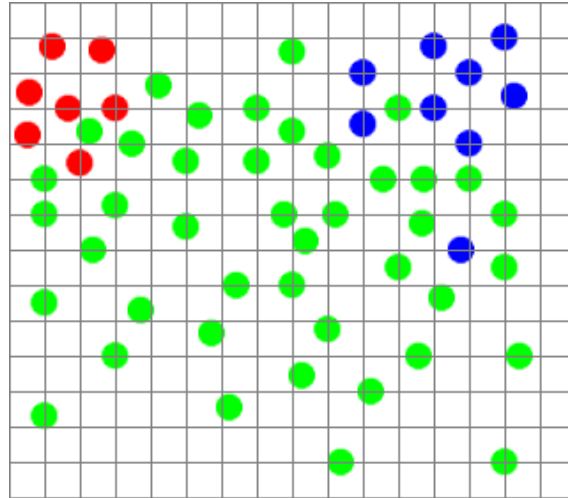


Figure 1.1.4: Example of an imbalanced problem.

its current state; if it classifies the pattern into the wrong class, it is penalized and tries to rebuild its structure to classify it properly. That means that the fewer the patterns in a class, the fewer the times the algorithm will rebuild in order to correctly classify the patterns belonging to that class.

Because of that, imbalanced classification is acquiring more attention in the data science community [31, 42, 8]. However, from a formal perspective, imbalanced datasets are harder to train not only because of the unbalanced distribution of the patterns but also because of the complexity of the data (commonly, noise, non-representative patterns and class overlapping) and the size of the training set (high dimensionality or a small number of patterns). Data scientists have approached this problem in two different ways (which are not exclusive):

1. Augmenting the training set data quality: based on sampling methods, it includes over-sampling the minority classes, under-sampling the majority ones, or a combination of both.
2. Improving the algorithm: forcing the model to focus on the minority classes.

The algorithmic approach is very sensitive, since data scientists have usually applied it to modify the cost function, which measures the model performance [47], exacting a greater penalty on the misclassification of the minority classes by multiplying them by a defined coefficient. Some more accurate approaches try a set of coefficients and validate which of them fits the data better, but generally, both of these techniques have been proven to lead to overfitting [31], which is not desirable. There are more complex approaches that are known as dynamic costs that modify the coefficients of the cost function for one

or more classes every time the performance is checked [15]. These techniques that are detailed later smooth the overfitting, leading to better results at the expense of increasing the computational cost (time) to train the model.

Augmenting the training set data quality is not the philosopher's stone that makes the model work perfectly; it avoids overfitting, but applying it correctly is complex, as using it is not as simple as creating synthetic patterns in the minority classes or removing patterns from the majority ones randomly. When creating synthetic patterns, which is known as over-sampling [61], data scientists use the existing patterns to create new ones, and if the modification of the variables of the pattern is not large enough, it can lead to non-representative patterns; or if the modification is too large, it can lead to patterns that do not belong to these classes, resulting in a model that misclassifies patterns from other classes into the minority ones. On the other hand, removing patterns from the majority classes [38, 23] can destroy the structure of the class when removing the representative patterns, losing important information and making the model unable to fit the data properly. More importantly, the pattern removal does not provide extra information to the minority classes, which most of the time are the ones that the model should focus on.

However, over-sampling methodologies have been improving during the last years and some techniques such as SMOTE (synthetic minority over-sampling technique) [11], when applied correctly, provide very good synthetic data to the training algorithm, as shown in Figure 1.1.5. SMOTE focuses on creating synthetic patterns in the minority classes by combining the already existing ones, with some variants developed to be applied in ordinal classification, and they are chosen to be applied in the different works presented in Chapters 3 and 4, and so will be described in more detail. However, SMOTE has three shortcomings as it may cause:

1. an over-generalization problem due to over-sampling of noisy samples,
2. an over-sampling of uninformative samples,
3. an increase of the overlaps between different classes around the class boundaries.

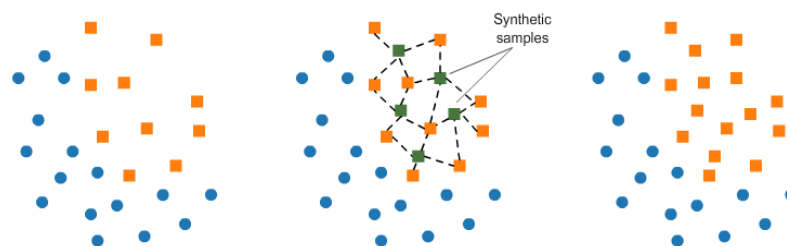


Figure 1.1.5: SMOTE applied to an imbalanced dataset.

There are numerous SMOTE-based techniques, and some of them are focused on solving the data imbalance in ordinal classification problems, as is the case with this thesis. Specifically, in this thesis, these SMOTE-based techniques used are the ones proposed [59].

1.2. Artificial neural networks

Among all the different model types that we can find in the machine learning literature, we selected artificial neural networks (ANNs) [5] to develop most of the work in this thesis. They were created in the 50s and are still under study due to their high computational power, which is developed using an adaptive learning process. They mainly have two components: neurons and links, and they are generally structured in a group of neurons called layers that are ordered, and whose neurons are only connected (through links) to the ones in the next layer. Each neuron will perform a defined mathematical operation depending on their type (sigmoid, radial basis function, product unit, etc.) [28] and each link will have an associated weight that multiplies the output of the neuron. There are three types of layers:

1. Input layer: this layer will contain as many neurons as the number of independent variables. This layer is only used to help define the network structure, as most of the time, no operations are carried out in this layer and the inputs are directly sent to the next layer.
2. Hidden layer(s): these layers are where the computational power of the neural networks resides. They contain neurons that receive an input, transform it (depending on the neuron type) and multiply the result by the link weight to the next layer. Usually, only one hidden layer is used, but recently, with arrival of deep learning [44], many hidden layers are used and each of them performs different transformations to the input data (usually images).
3. Output layer: this layer is where all the inputs that were transformed through the hidden layers and rescaled by the weights of the links are transformed to the result sought by the data scientist. In the unidimensional regression case, it is composed of a single neuron whose output will be the corresponding value of a pattern input into the solution space. In the classification case, it will have as many neurons as classes defined in the problem (or the number of classes minus one, in the case of using a probabilistic output), and their output will be a probability value whose meaning is the probability of that pattern belonging to that class, and the neuron with the higher output value will be the one that indicates the class that the pattern belongs to.

This layer distribution can be observed in Figure 1.2.1. Hornik et al. [36] demonstrated that artificial neural networks with sigmoid basis function are universal approximators of vector-valued functions, provided sufficiently hidden units are available. All these properties have made them an attractive tool for data scientists to successfully solve diverse classification and regression problems [69, 33, 3].

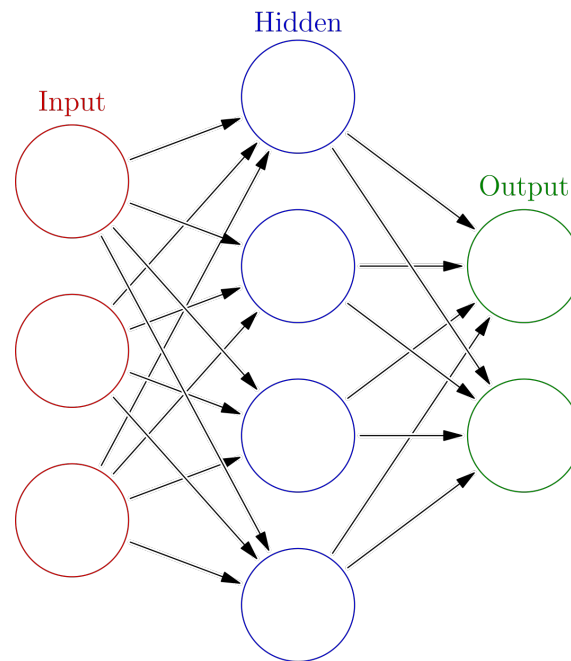


Figure 1.2.1: Artificial Neural Network standard structure.

ANNs, as mentioned before, develop their computational power through an adaptive learning process for which the objective is to adjust the network weights (links) to fit the training data. Generally, in order to train them, gradient descent algorithms [22] such as backpropagation are used. In some cases, gradient descent algorithms can get stuck in a local minimum of the solution space, resulting in a model with poor performance; hence, metaheuristics such as evolutionary algorithms [70], ant colonies [16] or particle swarm optimization [40] are a good alternative to train ANNs. Moreover, even though their computational cost is higher, they are more flexible than gradient descent approaches, as they can, for example, adjust the network structure (number of neurons) or set any objective (fitness) function that fits the problem more specifically even though it is not continuous, which gradient descent algorithms are not capable of performing. If the problem is complex enough to use metaheuristics, one of the best solutions would be to use an evolutionary algorithm for some generations and then, once the structure is optimized, apply a gradient descent algorithm to the best individual in the population. In this way, we can reduce the computational cost of evolutionary algorithms, as the gradient descent will optimize the best solution much faster and will keep the flexibility of the network struc-

ture optimization. This technique has been researched in detail in the works presented in Chapter 2 and is called hybridization of evolutionary algorithms with gradient descent approaches.

Hidden layer neurons can be classified depending on their activation function, which is responsible for determining whether to activate a neuron (send the transformed information through its links) or not. In this thesis, we will focus on three of them.

- Sigmoidal units: employ an activation function defined by:

$$\sigma_i(\mathbf{x}, \mathbf{w}_i) = \frac{1}{1 + e^{-\sum_{j=1}^k x_j w_{ij}}},$$

where e is the natural logarithm, x_j is the input neuron j and w_{ij} is the input weight from the input neuron j to the hidden neuron i .

- Radial basis function units: employ an activation function based on a radius and is defined by:

$$\sigma_i(\mathbf{x}, (\mu_i, r_i)) = e^{-\frac{\|\mathbf{x} - \mu_i\|^2}{2r_i^2}},$$

where \mathbf{x} is the vector of input neurons, μ_i is the centre of the kernel i and r_i defines its radius.

- Product units: employ an activation function defined by:

$$\sigma_i(\mathbf{x}, \mathbf{w}_i) = \prod_{j=1}^k x_j^{w_{ij}},$$

where x_j is the input from the neuron j and w_{ij} is the weight from the input neuron j to the hidden neuron i .

Each type of neuron has an objective, i.e., sigmoidal and product units are called projection-based functions as they simply rely on transforming their input through a non-linear projection to a point, while radial basis function units are known as kernel functions, as they depend on the distance from the input to the centre of the kernel. Projection-based functions contribute to a global recognition model [52] as they give different outputs for every input when they are activated, while radial basis functions contribute to a local recognition one [4], as they only provide outputs to the inputs that are next to their kernel centre. Some of the works in this thesis use a mixture of projection and kernel-based neurons in order to take advantage of the properties of both activation functions.

1.2.1. Ordinal artificial neural networks

Standard neural networks were not developed to carry out ordinal classifications, but some researchers have adapted artificial neural networks to ordinal classification [30, 14]. In this thesis, we will use an ordinal artificial neural network that was developed based on the proportional odds model (POM) [54]. This model is an extension of binary logistic regression for dealing with ordered categories. It was the first statistical model developed to deal with this kind of problem, and its adaptation to artificial neural networks is simple: add a second hidden layer with a single linear neuron whose objective is to project the outputs of the first hidden layer into a one-dimensional space (line). Then, the output layer will have one bias for each class, and the objective of this bias is to set the optimum thresholds to split the one-dimensional space (where the patterns were projected) and obtain the different intervals for each class. The structure of this neural network can be observed in Figure 1.2.2.

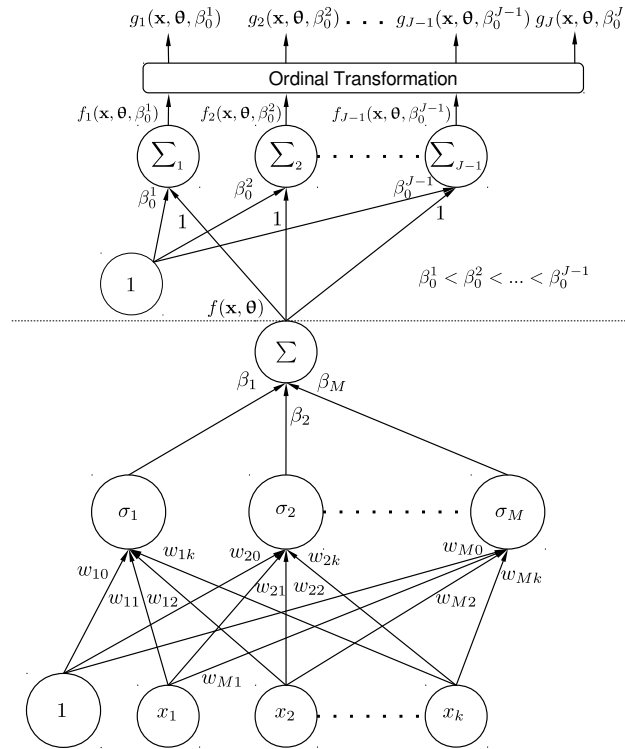


Figure 1.2.2: Ordinal Artificial Neural Network Structure.

The goal in ordinal classification is to assign an input vector \mathbf{x} to one of J discrete classes C_j , $j \in \{1, \dots, J\}$, where there exists a given ordering between the labels $C_1 \prec C_2 \prec \dots \prec C_J$. Hence, the objective is to find a prediction rule $C : \mathcal{X} \rightarrow \mathcal{Y}$ by using an i.i.d. training sample $X = \{x_i, y_i\}_{i=1}^N$ where N is the number of training patterns, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $X \subset \mathbb{R}_k$ is the k -dimensional input space and $\mathcal{Y} = \{C_1, C_2, \dots, C_J\}$ is the

label space.

An adaptation of the POM to artificial neural networks is used in this work. This adaptation is based on two elements: the first one is a second hidden linear layer with only one node whose inputs are the non-linear transformations of the first hidden layer. The task of this node is to project the values into a line, to impose an order. Apart from the single-node linear layer, an output layer is included with one bias for each class, whose objective is to set the optimum thresholds to classify the patterns into the class they belong to.

The structure of our model is presented in Figure 1.2.2 and has two main parts. The part at the bottom is formed by two layers of neurons, where $\mathbf{x} = (x_1, \dots, x_k)$ is the vector of input variables, and k is the number of variables in the database. $\mathbf{W} = \{\mathbf{w}_1 \dots \mathbf{w}_M\}$ is the matrix of weights of the connections from the input nodes to the hidden layer nodes (for each neuron, $\mathbf{w}_i = \{w_{i0}, w_{i1}, \dots, w_{ik}\}$, w_{i0} being the bias of the neuron).

The upper part of the figure shows a single node in the second hidden layer of the model, which is the one that performs the transformation to one dimension of the POM model. Its result, $f(\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = \{\mathbf{W}, \beta_1, \dots, \beta_M\}$, is connected, together with a second bias, to the output layer, where J is the number of classes, and $\beta_0^0, \dots, \beta_0^{J-1}$ are the thresholds for the different classes. These $J - 1$ thresholds are able to separate the J classes, but they have to fulfil the order constraint shown in the figure. Finally, the output layer obtains the outputs of the model, $f_j(\mathbf{x}, \boldsymbol{\theta}, \beta_0^j)$, for $j \in \{1, \dots, J - 1\}$. These outputs are transformed into a probability ($g_j(\mathbf{x}, \boldsymbol{\theta}, \beta_0^j)$), using the POM model structure. $g_j(\mathbf{x}, \boldsymbol{\theta}, \beta_0^j)$ is the probability that each pattern has of belonging to the different classes, and the class with the greatest probability is the one selected by the neural network.

1.2.2. Echo State Networks

There are special types of ANNs that allow cycles, i.e., the connections with neurons in previous layers, or even between neurons in the same layer, and they are known as recurrent neural networks [13] (see Figure 1.2.3). These networks were developed to deal with data ordered in time. In these databases, called time series, the problem to solve is usually to perform predictions or forecasts. When dealing with time series, e.g., meteorological or economics databases [27, 64, 51], we know that current events depend direct or indirectly on what happened previously. These temporal relations to previous events contribute a large amount of information to machine learning models, and there have been many different proposals to deal with time series.

For some of these proposals that are focused on modifying the data by adding information from previous patterns to the current one [43, 63], the following question arises: how many past patterns should be included? There were many different proposals, most

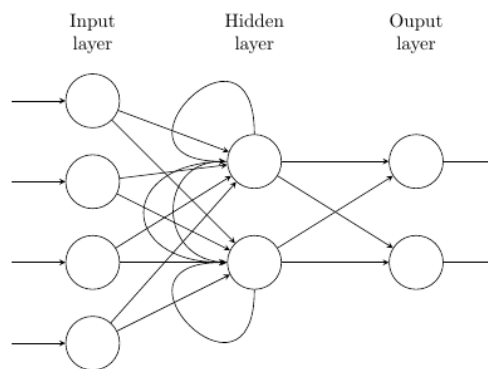


Figure 1.2.3: Recurrent Neural Network structure.

of them using a delay window:

- Fixed size: when the data scientist fixes the size of the window, e.g., if the window had size 2, it would include information about the two previous patterns in the current one. The size is usually cross-validated or set according to the statistical properties of the time series.
- Dynamic size: it assumes that all the patterns are not equally complex to predict, i.e., a rare event (storm) will need to be studied deeper in the past than the most common event in the time series (sunny weather). There are many different approaches to set this dynamic size, but the most used approaches take into account how many patterns of the same event happened before turning into a different event to set the size.

On the other hand, some proposals modify the models to take into account the temporal information [13]. In the case of artificial neural networks, recurrent neural networks (previously mentioned) were developed. The idea is that the links to the previous or the current layer (cycles) will make an input go back (transformed) to the same layer after a number of iterations (one, if it is connected to itself; two, if it is connected to the previous layer, and so on), mixing with newer pattern inputs. In this way, a hidden layer will have information about the current pattern and the transformed information about the previous ones (memory). Although the idea is good, some important difficulties emerged [35]:

- Vanishing gradient: when trying to optimize the network, gradient descent algorithms, where the gradient of the error is computed from the output to the input layers, are used. The problem is that the cycles make the gradient go slowly to zero, and then, it does not provide any information to the algorithm; consequently it will not be able to optimize the model. The more layers there are, the faster the gradient transforms to zero.

- **Model initialization:** the model needs some time before becoming stable; this happens because of the cycles that make the inputs flow slowly through the network. This means that if the maximum cycles a network goes through are five neurons, the model will need five patterns to fill all of its weights, and then the network will be stabilized. This is not a big issue if the dataset has a large number of patterns, as the patterns used to stabilize the network will just be discarded after this process.

However, some variants of recurrent neural networks were developed by researchers in order to tackle these well-known difficulties. This new paradigm is known as reservoir computing [50] and it mainly contains two types of models, echo state networks [37] and liquid state machines [25]; we will focus on echo state networks. The main difference between reservoir computing approaches and recurrent neural networks is that they have a large hidden layer containing randomly generated connections (including cycles) that are not trained in order to avoid the abovementioned gradient vanishing. This hidden layer is known as the reservoir, and it needs to fulfil some properties in order to ensure the convergence of the training algorithm. In contrast, the only weights that are trained are those connecting the reservoir to the output layer, and optionally, the weights that directly connect the input to the output layer (skipping the reservoir). This training methodology has proven to be sufficient to obtain good and competitive results in time series datasets, as the temporal relations of the data are latent in the reservoir. The standard echo state networks were developed to solve regression problems.

Echo state networks consist of a layer of input neurons that are connected by random weights to a large recurrent layer of neurons randomly interconnected among them, and finally a linear readout layer that will be the trainable part of the network while the random weights will remain unchanged. Mathematically, the states of the reservoir neurons are updated as follows:

$$\mathbf{x}_{t+1}^S = \tanh(\mathbf{W}^{\text{in}}[1; S_t] + \mathbf{W}\mathbf{x}_t^S),$$

where \tanh denotes the hyperbolic tangent function, t represents the time variable, $S_t \in \mathbb{R}$ is the output function, $\mathbf{x}_t^S \in \mathbb{R}^N$ is the vector containing the outputs of the neurons in the reservoir, and \mathbf{x}_{t+1}^S represents the updated outputs of the neurons in the next instant. Moreover, \mathbf{W}^{in} is the reservoir input weight matrix, \mathbf{W} is the random weight matrix of the reservoir, and, finally, $[\cdot; \cdot]$ indicates a vertical concatenation of vectors. Then, the standard output of these networks is defined as:

$$\hat{S}_{t+1} = \mathbf{W}^{\text{out}}[1; \mathbf{x}_{t+1}^S],$$

where \mathbf{W}^{out} is the output weight matrix, which is obtained using an optimization algo-

rithm known as Ridge Regression [50].

The standard echo state network architecture is defined in Figure 1.2.4, where there is an input layer, a reservoir and an output layer. In the reservoir, there are N neurons, which means that the maximum memory of the model, i.e., the previous data it can store in its weights, is N (instances). The easiest way to extend this model to a classification one would be to discretize its output by setting up some thresholds; however in this thesis, different echo state network architectures and extensions to classification and ordinal classification will be detailed.

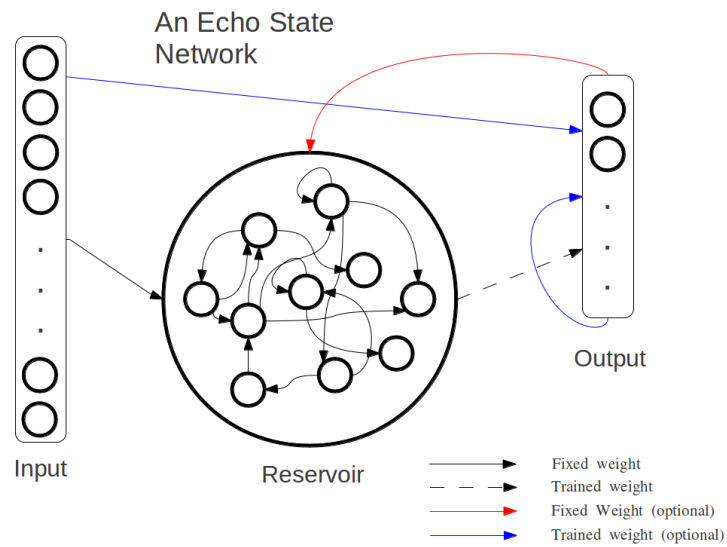


Figure 1.2.4: Echo State Network reservoir and optional weights.

1.2.3. Deep multi-task learning

Multi-task learning is a branch of Multi-output learning, which maps each input (instance) to multiple outputs. Assume $X = \mathbb{R}^d$ is the d -dimensional input space and $Y = \mathbb{R}^m$ is the m -dimensional output label space. The goal of multi-output learning is to learn a function $f : X \rightarrow Y$ from the training $D = \{(\mathbf{x}_i, \mathbf{y}_i) | 1 \leq i \leq n\}$, where each training sample $\mathbf{x}_i \in X$ is a d -dimensional feature vector and $\mathbf{y}_i \in Y$ is its corresponding output vector. Specifically, the multi-task learning approach [10] is based on the idea of sharing information from different tasks (each of them with a single label or output), which have some logical relation, to learn both of them. It has shown how modelling multiple related tasks simultaneously by sharing their information achieved better results than learning all of them independently. Therefore, it can be subsumed under the multi-output learning paradigm since learning multiple tasks is similar to learning multiple outputs. The major

difference is that in the multi-task scenario, each task could be learned with a different training set while in the multi-output scenario, all of the multiple outputs would share the same training data.

For instance, we can develop two independent models, one to classify the colour of a flower and another one to classify its type; the first model will receive a photo of the flower, while the second will only receive some characteristics, e.g., its petal width and height. It is likely their performance will be good, but if these models receive both image and flower characteristics, then their performance can be increased in both of them, as they would have more relevant information to perform their classification task. The idea underlying multi-task learning is that a single model is able to learn both of the tasks at the same time, as seen in the example shown in Figure 1.2.5. This approach was developed in the early 90s, but due to its high computational cost, it was discarded for several years. With the recent arrival of deep learning and its optimized algorithms and structures, this technique has resurfaced, showing again its high potential in dealing with multiple related tasks [72, 73, 71].

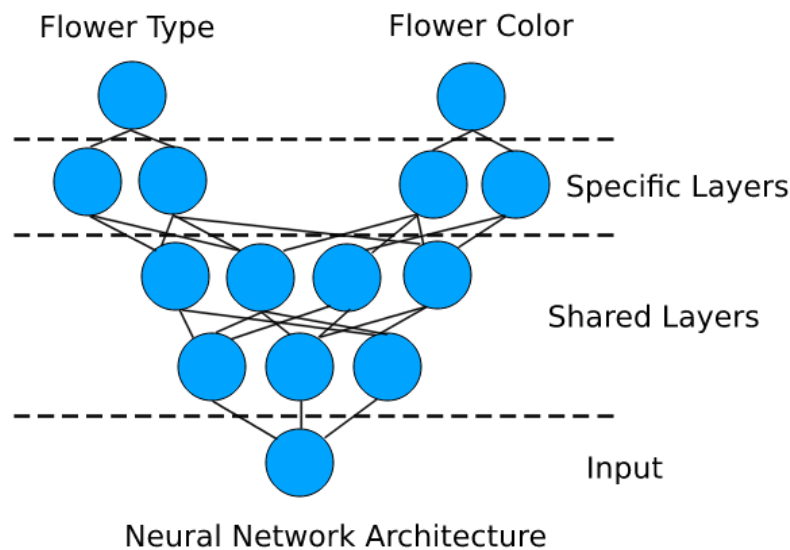


Figure 1.2.5: Multi-task learning applied to flower type and color classification.

Deep learning arose as a tool for solving image-related problems [46, 68, 12], mainly using convolutional layers, which were capable of extracting the most important information from the images; they became relevant in the machine learning area. The idea was to apply filters and transformations to the pixels of an image, similar to classical image processing algorithms, to process them and obtain the most relevant information. Again,

the most flexible model, the ANN, was used to structure the image processing in a machine learning framework and renamed as the deep neural network. In this way, each layer would perform a single processing task on the image. Among all the different transformations that can be performed on an image, the most common ones in deep neural networks are convolution and pooling [7]. In addition, autoencoders [65] are very useful to filter the image information.

Deep neural networks are built placing one layer on another layer as many times as needed to process the image properly, and finally a fully connected networks is used to obtain the output. This leads to neural networks with a large number of hidden layers in contrast to standard neural networks as shown in Figure 1.2.6 (this is why they are called deep), making their processing and computation of their gradient very costly with the traditional neuron activation functions. For this reason, the rectified linear units [56] were chosen as the activation function for the neurons, as their gradient is easily computable.

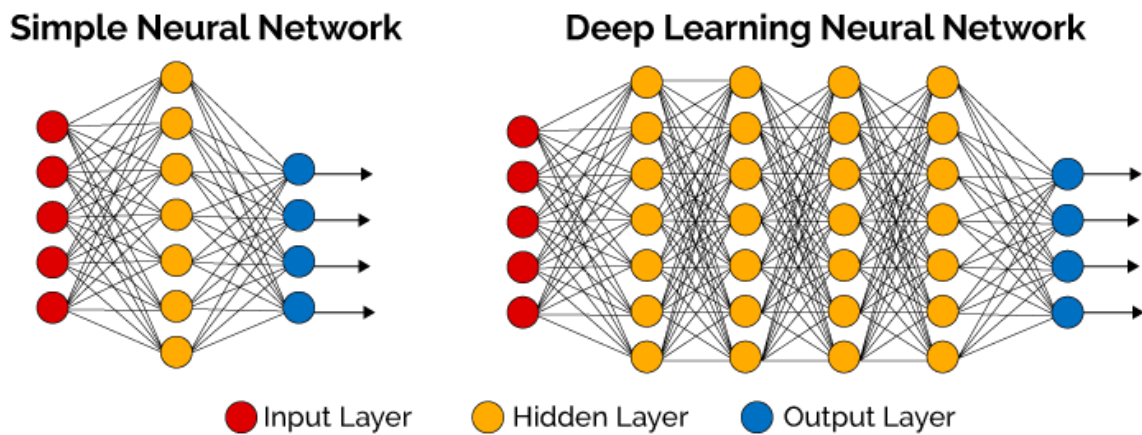


Figure 1.2.6: Difference between standard and deep Neural Networks.

However, the processing was still costly for a single CPU because their architecture does not allow them to compute a large number of processes at the same time, as they have a low number of processing units but a huge computational power. Installing a huge cluster of CPUs is very expensive, but graphical processing units (GPUs) were the best solution to solve this problem as they have a huge number of processing units with low computational power, which is not a problem since the processing task for a single neuron is very simple. At this point, there was a massive trend among deep learning researchers to migrate the different algorithms from CPU architectures to GPU, which made deep learning achieve its current relevance in the machine learning field.

Moreover, deep learning is not only important because of image processing, as other types of hidden layers and activation functions for deep neural networks were developed, e.g., to process signals, time series, etc. The work presented in Chapter 5, which concludes the thesis, will explain deep multi-task learning, which is the application of multi-task

learning techniques in deep neural networks, in more in detail.

1.3. Areas of application

This thesis was focused mainly on three application areas, of which one of them is a collaboration with the Loyola University. These works focused on the development of countries, beginning with a study about the investment that rich countries made to the poorer ones to conclude with a ranking of the countries depending on their globalisation level.

Then, a second application area came through a collaboration with the Reina Sofia Hospital in the city of Córdoba, which is one of the hospitals where more organ transplantations are performed every year in Spain. The problem was to develop an organ allocation system based in machine learning to select the most appropriate candidates for an organ.

The second application focuses on solving a wind energy production problem known as wind power ramp events, which is one of the most harmful events for wind farms. This problem was proposed in a collaboration with the University of Alcalá de Henares and Iberdrola, which is one of the main energy producer companies in Spain.

1.3.1. Liver allocation problem

Liver transplantation is an accepted treatment for patients presenting an end-stage liver disease [53]. However, not all donors are compatible with all patients, and there are not enough donors for all patients; thus, there needs to be a liver allocation system. The first approach of a liver allocation system was called Donor Risk Index [20], which establishes the quantitative risk associated with the surgery considering the donor information. However, the system being used currently is actually the opposite; it takes into account the severity of the recipient and is called the Model for End-stage Liver Disease (MELD) [39].

Nonetheless, none of these methods can be considered good predictors of graft failure, as they only consider the characteristics from one of the parts in the surgery, but not both. Furthermore, MELD is able to provide a reasonable graft survival rate but there are still some cases where the graft fails before the first week (the fewer cases), which directly means that the liver should have been given to another person. Now, imagine that a recipient rejects the graft after three months; it is likely that organ would have been a better fit to another patient on the list, and the graft would have survived for 5 or more years. The current systems provide reasonable results, but they can be potentially optimized to make the most of the available donors.

The papers included in this thesis consider a transplantation dataset with data from 1406 surgeries performed in seven Spanish transplant units and the King's College Hospital (UK). This dataset will include data not only from the donor and the recipient but also some characteristics of the surgery process, such as the cold ischaemia time. Since MELD provides reasonable results (a high percentage of grafts survive longer than three months), intuition tells us that the data will be unbalanced, and nothing is closer to reality, as usual in medicine problems. The most important classes to predict, which are graft failures before the first year, only represent 15 % of the dataset, as seen in Figure 1.3.1.

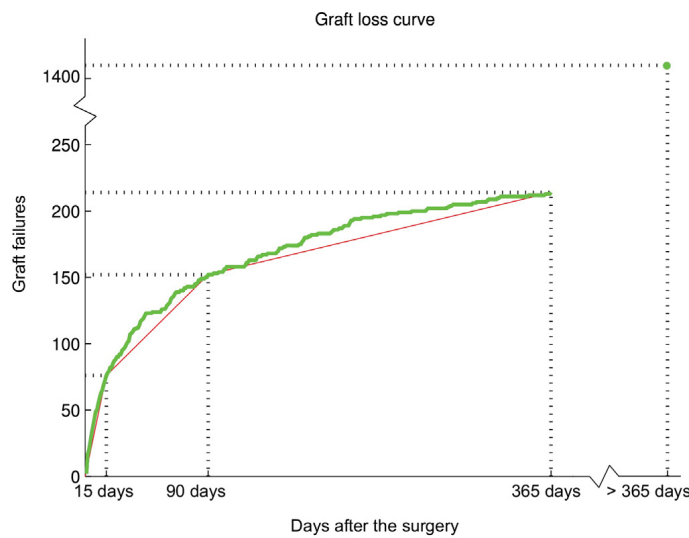


Figure 1.3.1: Graph showing the cumulative frequency of liver's graft loss.

In order to solve this problem, the dataset was split in four different classes according to the time before the graft loss:

- Class 1: graft loss before first 15 days.
- Class 2: graft loss between 15 days and 3 months.
- Class 3: graft loss between 3 months and 1 year.
- Class 4: graft loss after 1 year or no graft loss.

In the works presented in Chapter 3, this liver transplantation problem will be approached using ordinal ANNs, cost-based evolutionary algorithms and SMOTE techniques.

1.3.2. Wind power ramp events

Wind power ramps events (WPRES) are one of the most harmful natural events for wind farms [24, 57] and their prediction is currently among the hot topics in wind energy

research. WPREs can increase maintenance costs as they are able to damage the wind turbines when there is a high increase in the wind speed and, in contrast, they can also produce losses in the energy production because in a short period of time, the wind speed can be drastically reduced.

As can be intuited, WPREs consist of important fluctuations of wind speed in a short period of time, leading to the problems mentioned above. The origin of WPREs is in meteorological processes, usually crossing fronts that derive from drastic changes in the wind speed. Currently, the most effective way of dealing with them is to perform a correct prediction with enough time to allow the operators to take the necessary actions to prevent profit losses.

To solve this prediction problem, different classification techniques and databases were used. The different datasets included the following:

1. Information collected from three different wind farms located in the Spanish geography from 2002 to 2012.
2. Data from the ERA-Interim reanalysis project with wind and temperature predictors at different heights that are computed every 0.75 kilometres all around the globe.

The first dataset contained information about the wind speed every hour, and knowing the power curve of every wind farm, we could transform the wind speed into produced energy. Then, the dataset was discretized into the different classes according to the defined ramp function [24] which decides whether a ramp event occurred or not, according to previous and current power productions. Mathematically, the ramp function is defined as:

$$S_t = \max_{i \in [t-\Delta, t]}(P_i) - \min_{i \in [t-\Delta, t]}(P_i), \quad (1.2)$$

where Δ is the time interval considered. For the binary prediction, the classification problem is stated by defining a threshold power value S_0 that splits the different patterns into different classes. Depending on the S_0 definition, different problems were stated, from a binary classification in the first work, to an ordinal multi-class classification problem.

The second dataset computed all the information every 6 hours; that time restriction set our Δ and our prediction horizon to 6 hours, meaning that we are able to predict WPREs that would occur after 6 or more hours. As mentioned previously, these predictors are computed every 0.75 kilometres around the globe, which means that there is a grid of points where these predictors are available, and the probability that one of these grid points is on the same geographical location of a wind farm is very low. As wind can come in any direction, picking only the point that is nearest to the wind farm is not sufficient,

so we build the dataset with the four nearest points that include the wind farm inside the square drawn joining them. This can be clearly seen in Figure 1.3.2.

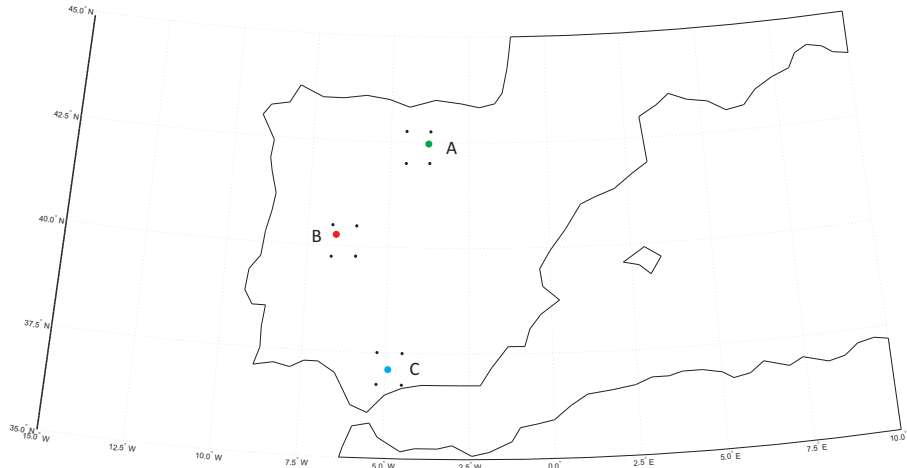


Figure 1.3.2: Location of the three wind farms and the reanalysis nodes considered in the study.

We selected 10 predictors from every point, which means we had 40 predictors for each wind farm. To avoid dealing with so many variables, which in many cases can lead to correlation problems and low performance of the classification models, we perform a weighted average of every predictor considering their distance from the reanalysis node to the wind farm centre, thus obtaining again a database with 10 variables, but with their values now extrapolated to the wind farm centre.

Finally, this problem was solved using recurrent neural network architectures, in particular, echo states networks with different structures and architectures. The work concluding this problem solution proposes a deep multi-task neural network that includes information of all the wind farms and is able to predict whether a WPRE is going to occur or not in each of them at the same time with a single model.

1.4. Motivation and challenges

From the previous sections, multiple issues related to ordinal classification performance and real life problems can be noted. Ordinal classification problems appeared when ANNs and their corresponding algorithms obtained average quality results even though they are computationally costly. When discussing the performance of the models, most of the problems appear because the models are not flexible enough to fit the shape of the data. On the other hand, the low performance of the training algorithms is usually related to the problem: low performance appears when the algorithm is not able to reach a good minimum in the solution space. One of the main reasons is because the algorithm is not able to reach a good solution in the defined time usually as a consequence of its

high computational cost; in addition, this can occur because the problem has particular characteristics such as class imbalance, which most of the algorithms are unable to handle.

On the other hand, due to collaborations with different companies and institutions, problems related to medicine and renewable energy areas came to our attention. First, the medicine issues were related to liver allocation in transplants, where the current solution was to transplant the liver according to the MELD, a metric that only takes into account the health status of the recipient, such that the sicker he or she is, the higher his or her MELD score is. This seemed insufficient, as more valuable information is available that should be taken into account when making these kinds of decisions, for example, donor characteristics or characteristics directly related to the liver and the surgery process.

Then, the issues we faced in the renewable energy area were related to wind farms, where one of the most known problems concerns the WPREs that highly decrease the companies' profit due to the increase of maintenance costs and the decrease of the overall production of the wind farms. This occurs because turbines can be damaged with WPREs, and consequently, they need to be repaired and will not produce any energy while they are damaged; therefore, the best solution to this issue is to correctly predict WPREs and turn off the turbines before they occur. The low prediction rates of the current methodologies do not allow the wind farms to properly optimize their production, and thus, their profit.

Considering the mentioned issues, we can synthesize the following challenges that constitute the objectives of this thesis

Artificial Neural Networks. The performance of ANNs is usually competitive with state-of-the-art models, given that they are very powerful models that usually obtain good results by just applying a standard training algorithm. However, when other state-of-the-art models are fine-tuned, they can easily overcome an ANN's results. Thus, it is necessary to provide strategies to fine-tune ANNs to improve their performance. This thesis analyses different activation function combinations and output layer proposals in order to provide ANNs with better flexibility.

Echo State Networks. Echo state networks have been used in the literature to solve regression problems when data have temporal relations. Nevertheless, scant effort has been devoted to developing networks and training algorithms based on classification, and even less on ordinal classification; hence, this thesis will provide new structures and algorithms based on this type of recurrent neural network.

Training algorithms. This could be decomposed in two sub-challenges: improving training algorithms and training algorithms for imbalanced classification. The first one will focus on the balance between the exploration and exploitation characte-

istics of the algorithms; in this way, the probabilities of finding a good solution to the problem will increase. This thesis will develop a combination of evolutionary and gradient descent algorithms to achieve this target. The second one consists of improving the training algorithm capacity to distinguish minority classes from majority ones. Usually, training algorithms focus on increasing the global performance considering all classes as equals, which makes the algorithm include the patterns belonging to the minority class in the majority one, resulting in a high global performance as most of the patterns belong to the majority class. This thesis will study different ways to implement cost-based approaches, avoiding well-known issues such as overfitting of the training data.

Multi-task Learning. Multi-task learning algorithms, from our point of view, are a very powerful tool when a problem can be divided into subproblems; thus, learning all the subproblems at the same time will provide much more useful information. In this thesis, this approach will be the baseline to develop a prediction problem in multiple locations at the same time, with the objective of obtaining a global prediction model in a defined geography.

Liver Allocation. Currently, the methodology to allocate a liver is to check the MELD score, which only depends on recipient information. This does not consider compatibility with the donor and does not try to achieve the largest survival time. If a recipient rejects a liver and the graft dies, he or she will be re-transplanted in the next days because his or her MELD score will be even higher; thus two organs would have been wasted in a single person, and it is likely with a deeper study this would not have happened. The challenge is to provide the medical experts with novel tools to help them make the best decision for liver allocation in hepatic transplants. To do so, this thesis will develop a machine learning methodology including specific algorithms and models to optimize the survival time of the graft, considering information from the donor, recipient and surgery.

Wind Energy. Renewable energy technologies are under constant research and upgrade, and the main goal is to obtain as much profit as possible, i.e., increase the amount power production by reducing the direct and derived production costs. One of the biggest issues in optimizing the profit in wind farms are WPREs, and the best way of dealing with these events is to correctly predict them well in advance. This thesis will research and develop several data mining and machine learning techniques optimized for this particular problem that will be easily extended to other time series prediction problems.

1.5. Objectives

The present thesis addresses the aforementioned challenges. All these challenges result in the following objectives.

- Artificial Neural Networks:

ANN1 *To propose different artificial neural network architectures.* The first part of this thesis will research approaches to hybridize activation functions with different characteristics in the hidden layers and search for a good balance to achieve a model that takes advantage of them.

ANN2 *To enhance state-of-the-art output layers.* After achieving a good balance for activation functions in the hidden layers, the next step is to enhance output layers, and since one of the most employed output layers for ordinal classification is the ordinal logistic regression, which has some well known shortcomings this thesis will search for a way to overcome them, and thus improve its performance.

- Training Algorithms:

TA1 *To improve evolutionary algorithms' performance.* . The objective is to find different ways to adapt the costly evolutionary algorithms (EAs) to new artificial neural networks (ANNs) and problems. EAs for ANNs were left apart in the last years because the increment on the performance was not worth the time that an ANN took to be trained when compared to gradient descent performance. The counterpart is that gradient descent algorithms are only able to train the weights of the ANN but not its structure, which provides plenty of room to optimize its performance. This thesis will work on a mixture of EAs and gradient descent algorithms to decrease the time of the training phase while keeping the capabilities to optimize the ANNs' architecture.

TA2 *To prepare algorithms for handling imbalanced data.* Usually, when working with time series or medical data, the most important classes to predict are those with fewer patterns. Most of the algorithms are not prepared to deal with imbalanced databases; thus, this thesis will study how to adapt current algorithms to train ANNs with imbalanced databases.

- Liver allocation:

LA1 *Improve the performance of previous works when predicting the most suitable recipient for a liver.* The predictive model that takes information from the donor,

the recipient and the surgery process, would be improved through the use of SMOTE techniques to palliate the imbalance degree of the dataset.

LA2 *To develop an algorithm* capable of dealing with the data imbalance of this problem, the baseline would be cost-based algorithms, but the goal is to achieve an algorithm that evolves its cost function at the same time that it trains the model.

- Wind Energy:

WE1 *Extract a database to predict WPREs.* Considering the raw data gathered from all the wind farms, the objective is to transform the database and find other data sources that are compatible with the original data and provide extra information in order to improve the prediction performance of the developed models. After finding a suitable second source of information, these works will focus on finding a function to determine whether a WPRE occurred or not and finally perform a deeper study to remove all the unnecessary data.

WE2 *Develop models to predict WPREs.* This objective consists of finding the most appropriate model to perform successful predictions of WPREs. Knowing that recurrent neural networks are a suitable model to perform predictions in time series and considering their shortcomings, we will search for alternative models that share recurrent neural networks' capabilities to compute time series, such as echo state networks.

WE3 *Prediction of WPREs in multiple farms.* After developing an appropriate model to compute the time series database that was developed in the first objective, the last objective is to develop a new database and model, starting from the already developed ones as a baseline. This new definition of the problem consists of predicting WPREs in all the wind farms at the same time combining all their information, in order to provide the necessary tools to predict WPREs in a global area, instead of in a single wind farm. To do so, we will establish multi-task learning and deep neural networks as the starting point.

1.6. Summary of the thesis

Artificial Intelligence is part of our everyday life, not only as consumers but also in most of the productive areas since companies can optimize most of their processes with all the different tools that it can provide. There is one topic that has been especially useful in the artificial intelligence implementation process which is machine learning, as it can be used in most of the practical applications that appear in real-life problems. Machine learning is the part of artificial intelligence that focuses on developing models that are

able to learn a function that transforms input data into a desired output. One of the most important parts in machine learning is the model, and one of the most successful models in the state-of-the-art approaches is the artificial neural network. This is why the current thesis, for its first challenge, will study how to improve them to be able to learn more complex problems without needing to apply computationally costly training algorithms. The next important step to improve the model's performance is to optimize the algorithms that are used to let them learn how to transform the inputs into the desired outputs, and the second challenge of this thesis is to optimize the computational cost of evolutionary algorithms, which are one of the best options to optimize ANNs due to their flexibility when training them.

Ordinal classification (also known as ordinal regression) is an area of machine learning that can be applied to many real-life problems since it takes into account the order of the classes, which is an important fact in many real-life problems. In the area of social sciences, we will study how potential countries are helping the poorer ones the most, and then we will perform a deeper study to classify the level of globalisation of a country. These studies will be performed by applying the models and algorithms that were developed in the first stage of the thesis.

After these first works, continuing with the ordinal classification approaches, we focused on the area of medicine, where there are many examples of applications of these techniques, e.g., any disease that may have progression is usually classified in different stages depending on its severity from low to high. In our case, this thesis will study how a treatment (liver transplantation) can affect different patients (survival time of the graft), and therefore decide which patient is the most appropriate for that specific treatment.

The last chapter of the thesis will delve in ordinal classification to achieve ordinal prediction of time series. Time series have been usually processed with classical statistical techniques since machine learning models that focused on time series were too costly. However, currently, with the arrival of powerful computation machines together with the evolution of models such as recurrent neural networks, classic statistical techniques can hardly be competitive versus machine learning. In areas such as economics, social sciences, meteorology or medicine, time series are the main source of information, and they need to be correctly processed to be useful. The most common consideration when dealing with time series is to learn from past values to predict future ones, and the works in this last chapter will focus on performing ordinal predictions of WPREs in wind farms, creating novel models and methodologies.

The thesis will conclude with a work that implements a deep neural network to predict WPREs in multiple wind farms at the same time; therefore, this model would allow predicting WPREs in a global area instead of in a specific geographical point.

1.7. Publications

The work performed in this thesis is reflected in seven papers in international conferences and seven papers in international journals.

The following international journal papers include some of the ideas of this thesis:

- J1 A .Sianes, M. Dorado-Moreno y C. Hervás-Martínez. Rating the Rich: An Ordinal Classification to Determine Which Rich Countries are Helping Poorer Ones the Most. *Social Indicators Research*, 116:47–65, 2014, Impact Factor (2014): 1.395 (Q1).
- J2 M. Dorado-Moreno, A. Sianes y C. Hervás-Martínez. From outside to hyper-globalisation: an Artificial Neural Network ordinal classifier to measure the extent of globalisation. *Quality & Quantity*, 50(2):549–576, 2016, Impact Factor (2016): 1.094 (Q2).
- J3 M. Dorado-Moreno, M. Pérez-Ortiz, P. A. Gutiérrez, R. Ciria, J. Briceño y C. Hervás-Martínez. Dynamically weighted Evolutionary Ordinal Neural Network for solving an Imbalanced Liver Transplantation Problem. *Artificial Intelligence in Medicine*, 77:1–11, 2014, Impact Factor (2017): 2.879 (Q1).
- J4 M. Dorado-Moreno, L. Cornejo-Bueno, P. A. Gutiérrez, L. Prieto, C. Hervás-Martínez y S. Salcedo-Sanz. Robust estimation of wind power ramp events with reservoir computing. *Renewable Energy*, 111:428–437, 2017, Impact Factor (2017): 4.900 (Q1).
- J5 M. Dorado-Moreno, P. A. Gutiérrez, L. Cornejo-Bueno, L. Prieto, S. Salcedo-Sanz y C. Hervás-Martínez. Ordinal multi-class architecture for predicting wind power ramp events based on reservoir computing. *Neural Processing Letters*, In press, 2018, Impact Factor (2017): 1.787 (Q2).

Some ideas have been submitted to different journals and are currently under review process:

- J6 M. Dorado-Moreno, N. Navarin, P.A. Gutiérrez, L. Prieto, A. Sperduti, S. Salcedo-Sanz and C. Hervás-Martínez. Multi-task Learning for the Prediction of Wind Power Ramp Events with Deep Neural Networks. *Neural Networks* (Under Revision), 2018, Impact Factor (2017): 7.197 (Q1).

Moreover, some works have also been published in national and international conferences:

- C1 M. Dorado-Moreno, P. A. Gutiérrez y C. Hervás-Martínez. Ordinal Classification Using Hybrid Artificial Neural Networks with Projection and Kernel Basis Functions.

In *7th International Conference on Hybrid Artificial Intelligence Systems (HAIS)*, pages 319–330, 2012, Salamanca, Spain.

- C2 M. Dorado-Moreno, P. A. Gutiérrez, J. Sánchez-Monedero y C. Hervás-Martínez. Overcoming the linearity of Ordinal Logistic Regression adding non-linear covariates from Evolutionary Hybrid Neural Network models. In *16th Conference of the Spanish Association for Artificial Intelligence*, pages 301–311, 2015, Albacete, Spain.
- C3 M. Dorado-Moreno, P. A. Gutiérrez, J. Sánchez-Monedero y C. Hervás-Martínez. Nonlinear Ordinal Logistic Regression using covariates obtained by Radial Basis Function neural networks models. In *13th International Work-Conference on Artificial Neural Networks (IWANN)*, pages 80–91, 2015, Palma de Mallorca, Spain.
- C4 M. Dorado-Moreno, A.M. Durán-Rosal, D. Guijo-Rubio, P.A. Gutiérrez, L. Prieto, S. Salcedo-Sanz y C. Hervás-Martínez. Multiclass Prediction of Wind Power Ramp Events Combining Reservoir Computing and Support Vector Machines. In *17th Conference of the Spanish Association for Artificial Intelligence*, pages 300–309, 2016, Salamanca, Spain.
- C5 M. Dorado-Moreno, M. Pérez-Ortiz, M. D. Ayllón-Terán, P. A. Gutiérrez y C. Hervás-Martínez. Ordinal Evolutionary Artificial Neural Networks for Solving an Imbalanced Liver Transplantation Problem. In *11th International Conference on Hybrid Artificial Intelligent Systems (HAIS)*, pages 451–462, 2016, Sevilla, Spain.
- C6 M. Dorado-Moreno, L. Cornejo-Bueno, P. A. Gutiérrez, P. A. Gutiérrez, S. Salcedo-Sanz y C. Hervás-Martínez. Combining Reservoir Computing and Over-Sampling for Ordinal Wind Power Ramp Prediction. In *International Work-Conference on Artificial and Natural Neural Networks (IWANN)*, pages 708–719, 2017, Cádiz, Spain.
- C7 M. Dorado-Moreno, P. A. Gutiérrez, S. Salcedo-Sanz, L. Prieto y C. Hervás-Martínez. Wind power ramp events ordinal prediction using minimum complexity echo state networks. In *2018 International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*, pages 180–187, 2018, Madrid, Spain.

Other publications done during the PhD.:

- E. Rodero, A. González, M. Dorado-Moreno, M. Luque y C. Hervás-Martínez. Classification of goat genetic resources using morphological traits. Comparison of machine learning techniques with linear discriminant analysis, *Livestock Science*, 180:14–21, 2015 Impact Factor (2015): 1.293 (Q2).

- A. M. Durán-Rosal, M. Dorado-Moreno, P. A. Gutiérrez y C. Hervás-Martínez. Identification of extreme wave heights with an evolutionary algorithm in combination with a likelihood-based segmentation. *Progress in Artificial Intelligence*, 6:59–66, 2017.
- A. M. Durán-Rosal, M. Dorado-Moreno, P. A. Gutiérrez y C. Hervás-Martínez. On the Use of the Beta Distribution for a Hybrid Time Series Segmentation Algorithm. In 17th Conference of the Spanish Association for Artificial Intelligence, pages 418–427, 2016, Salamanca, Spain.

2

Ordinal classification: hybridization of training algorithms and basis functions

This chapter presents different research works concerning ordinal classification models and algorithms, including new proposals and a thorough set of experiments. This chapter also includes two applications in real problems in the social sciences.

Main publications associated with this chapter:

- M. Dorado-Moreno, P. A. Gutiérrez y C. Hervás-Martínez. Ordinal Classification Using Hybrid Artificial Neural Networks with Projection and Kernel Basis Functions. In *7th International Conference on Hybrid Artificial Intelligence Systems (HAIS)*, pages 319–330, 2012, Salamanca, Spain.
- A .Sianes, M. Dorado-Moreno y C. Hervás-Martínez. Rating the Rich: An Ordinal Classification to Determine Which Rich Countries are Helping Poorer Ones the Most. *Social Indicators Research*, 116:47–65, 2014, Impact Factor (2014): 1.395 (Q1).
- M. Dorado-Moreno, A. Sianes y C. Hervás-Martínez. From outside to hyper-globalisation: an Artificial Neural Network ordinal classifier to measure the extent of globalisation. *Quality & Quantity*, 50(2):549–576, 2016, Impact Factor (2016): 1.094 (Q2).

Other publications associated with this chapter:

- M. Dorado-Moreno, P. A. Gutiérrez, J. Sánchez-Monedero y C. Hervás-Martínez. Nonlinear Ordinal Logistic Regression using covariates obtained by Radial Basis Function neural networks models. In *13th International Work-Conference on Artificial Neural Networks (IWANN)*, pages 80–91, 2015, Palma de Mallorca, Spain.
- M. Dorado-Moreno, P. A. Gutiérrez, J. Sánchez-Monedero y C. Hervás-Martínez. Overcoming the linearity of Ordinal Logistic Regression adding non-linear covariates from Evolutionary Hybrid Neural Network models. In *16th Conference of the Spanish Association for Artificial Intelligence*, pages 301–311, 2015, Albacete, Spain.

The three main publications are now presented in the different sections of this chapter.

2.1. Ordinal Classification Using Hybrid Artificial Neural Networks with Projection and Kernel Basis Functions

Artificial neural networks (ANNs) have been widely used in the machine learning area, as they have been demonstrated to be a flexible and successful tool when solving classification and regression problems. This paper proposes an optimization of ANNs whose architectures have been modified (based in the proportional odds model) to perform an ordinal classification, which consists of combining two different basis functions in the same hidden layer.

The objective of combining two different basis functions is that by considering basis functions with different characteristics, the model can take advantage of both of their characteristics when solving a machine learning problem. In this case, we selected product unit neurons, which are a projection function to be combined with radial basis function neurons, which are kernel functions, since it has been demonstrated that any continuous function can be decomposed in two mutually exclusive functions, such as radial (RBF) and crest ones (PU).

This hybrid ANN will be trained using an evolutionary algorithm, whose cost function has been modified in order to take into account the class order of each problem because we are tackling an ordinal classification. In this way, if a problem with six classes has a pattern that belongs to class one and is misclassified in class five, the error will be penalized to a higher degree than if the same pattern is misclassified in class two, as it is the adjacent class to the real class.

The model was tested on six different datasets from the UCI repository and compared to the same model but using only RBF and only PU neurons and further compared

against two state-of-the-art models, concluding that the proposal obtains better performance than the non-hybrid models and is able to achieve competitive results versus state-of-the-art models.

Ordinal Classification Using Hybrid Artificial Neural Networks with Projection and Kernel Basis Functions

M. Dorado-Moreno, P.A. Gutiérrez, and C. Hervás-Martínez

Department of Computer Science and Numerical Analysis, University of Córdoba,
Campus de Rabanales, 14071, Córdoba, Spain
{i92domom, pagutierrez, chervas}@uco.es
<http://www.uco.es/ayrna>

Abstract. Many real life problems require the classification of items into naturally ordered classes. These problems are traditionally handled by conventional methods intended for the classification of nominal classes, where the order relation is ignored. This paper proposes a hybrid neural network model applied to ordinal classification using a possible combination of projection functions (product unit, PU) and kernel functions (radial basis function, RBF) in the hidden layer of a feed-forward neural network. A combination of an evolutionary and a gradient-descent algorithms is adapted to this model and applied to obtain an optimal architecture, weights and node typology of the model. This combined basis function model is compared to the corresponding pure models: PU neural network, and the RBF neural network. Combined functions using projection and kernel functions are found to be better than pure basis functions for the task of ordinal classification in several datasets.

Keywords: Ordinal Classification, Projection basis functions, Kernel basis functions, Evolutionary neural networks, Evolutionary algorithm, Gradient-descent algorithm.

1 Introduction

In the real world, there are many supervised learning problems referred to as ordinal classification, where examples are labeled by an ordinal scale [24]. For instance, a teacher who rates his students using labels (A,B,C,D) that have a natural order among them ($A > B > C > D$). In this paper, we are selecting artificial neural networks to face this kind of problems. They are a very flexible modeling technique, whose computing power is developed using an adaptive learning process. Properties of artificial neural networks made them a common tool when successfully solving classification problems [15,17].

The objective of this paper is to adapt the hybrid model previously proposed in [14] to ordinal regression, adding a local search algorithm to result in a hybrid training method with both evolutionary and gradient-directed algorithms.

2.2. Rating the Rich: An Ordinal Classification to Determine Which Rich Countries are Helping Poorer Ones the Most

In 2004, the Organization for Economic Cooperation and Development (OECD) accepted the "Policy Coherence for Development", which states that economic help is not a sufficient indicator to assess the contribution to poorer countries' development; moreover, the OECD outlined that it would work to ensure that the countries' development policies are feasible, and they will be determined by other policies that affect countries under development.

Currently, the most acknowledged way of measuring the Policy Coherence for Development is the Commitment to the Development Index (CDI), but it has been subject to multiple criticisms (explained in detail in the paper), which this paper will try to overcome with an ordinal artificial neural network.

The paper considers 22 rich countries belonging to the OECD and analyses seven different policy areas that are scored in terms of their help in the development of poorer countries. In total there are 33 macroeconomic indicators that will provide them with a rank in five different classes, and these have been defined as follows: $AAA > AA > A > B > C$ which represent higher to lower help in the development. This is clearly the definition of an ordinal classification, problem and provides the rationale of this paper's proposal of an ordinal classification ANN. The model used is the one proposed in the previous section, and it will be compared against ordinal SVM models

Our model obtained the best results in the comparison over 30 executions, and the best obtained model only misclassified 3 countries when comparing to the original CDI score. It is important to note that the misclassification appeared in countries whose classification score was in the lowest part of their class, and the model classified them in the previous class instead of the CDI score one.

Rating the Rich: An Ordinal Classification to Determine Which Rich Countries are Helping Poorer Ones the Most

Antonio Sianes · Manuel Dorado-Moreno · César Hervás-Martínez

Accepted: 4 February 2013 / Published online: 26 February 2013
© Springer Science+Business Media Dordrecht 2013

Abstract When talking about poverty, a lot of energy is expended by academics and sociologists in the identification and classification of the poor. Less attention is paid to classifying the rich. The Center for Global Development created the Commitment to Development Index in 2003, which ranks countries according to their contribution to the reduction of poverty in developing countries. Since its first report, “Ranking the rich, the Index has been quite successful. However, it has also been subject to multiple criticisms. This paper proposes the use of an ordinal classification to rate, not rank, the performance of rich countries. An ordinal classification, where an ordinal scale labels the examples, can help discovering the level of each country’s commitment to development, automatically and independently from others’ performances. It could stimulate both advocacy from civil society and the determination of more coherent public policies in rich countries for poorer ones. The methodology used is Artificial Neural Networks, a common machine learning tool for successfully solving classification problems. Experiments yield robust results, showing better outcomes than other alternative ordinal classifiers, opening the possibility of developing a classification technique which could overcome the limitations of the current ranking technique.

Keywords Fight against poverty · Policy coherence for development · Commitment to development index · Ordinal classification · Artificial neural networks

A. Sianes
ETEA Foundation for Development and Cooperation, Universidad Loyola Andalucía,
Escritor Castilla Aguayo 4, 14004 Córdoba, Spain
e-mail: antonio.sianes@fundacionetea.org

M. Dorado-Moreno (✉) · C. Hervás-Martínez
Department of Computer Science and Numerical Analysis, University of Córdoba,
Campus de Rabanales, C2 building, 14004 Córdoba, Spain
e-mail: i92domom@uco.es

C. Hervás-Martínez
e-mail: chervas@uco.es

2.3. From outside to hyper-globalisation: an Artificial Neural Network ordinal classifier applied to measure the extent of globalisation

The term globalisation is a concept used to understand the acceleration of changes that have occurred in the society in the last decades. Globalisation has many different aspects that can be economic, social or political. There are many different indexes used to measure the extent of globalisation of a country; we selected the KOF (Konjunkturforschungsstelle) developed by the Swiss Economic Institute as a baseline to perform this research.

The KOF Index combines 23 variables to obtain six indicators and then combines these six indicators linearly to obtain the three main branches of globalisation, i.e., economic, political and social. According to this index, economic and social globalisation are equally important, with an influence of 37% of the final score, while the political branch is not that important, with an influence of 26%.

This index has some problems that this research will propose to overcome by making use of a hybrid ANN ordinal classifier, as presented in Section 2.1. The KOF Index measures the globalisation from 0 to 100; therefore, we first split the problem into 6 different classes: $A+ > A > B+ > B > C+ > C$. There will be 147 countries included in the study, and 80 of them belong to classes B and $C+$. The methodology consists of training the hybrid ANN using an evolutionary algorithm optimized for ordinal classification. Then, because the ordinal ANN is an adaptation of a threshold methodology, we will set the thresholds for the different classes according to those of the model, and the score of each country will be the projection of that country in one dimension after all the ANN processing is completed.

After comparing our model proposal with other state-of-the-art ordinal classifiers, we can conclude that in this paper, a new score based on KOF was created with a high degree of accuracy (more than 98%), which only misclassified two countries in the adjacent class compared with the original KOF Index. In addition, we proposed a reordering of the countries in terms of globalisation according to our model scoring, where some countries were able to win up to 9 positions in the ranking.

From outside to hyper-globalisation: an Artificial Neural Network ordinal classifier applied to measure the extent of globalisation

Manuel Dorado-Moreno · Antonio Sianes ·
César Hervás-Martínez

Published online: 28 January 2015
© Springer Science+Business Media Dordrecht 2015

Abstract Globalisation has become a key concept in the social sciences to understand the accelerating changes occurred in modern societies during recent decades. As a consequence, measuring the influence of globalisation on the economic, social and political aspects of nations has been a requirement. There are many indices at present to calculate the extent of globalisation reached by each country. However, most of the methods used to build those indices suffer certain methodological limitations that hinder the wider dissemination and usefulness of their results. As an alternative, in this paper, we propose a methodology for ordinal ranking of countries associated with their globalisation level, which gives us an easier and more useful information about the different levels where countries are regarding to this criteria. Among Computational Intelligence techniques, Artificial Neural Networks (ANNs) have become dominant modelling paradigm. We have built a novel non-linear ordinal classifier by combining the Proportional Odd Models (POM) with ANNs that is able to classify countries according to their level of globalisation in six classes, which range from hyperglobalised countries to countries that remain outside the process of globalisation. The results could not be more encouraging. Our experiments yield robust results and show better outcomes than alternative linear and non-linear ordinal classifiers, which raises the possibility of developing a model of classification that might overcome some of the limitations of the indices currently employed to measure globalisation.

M. Dorado-Moreno (✉) · C. Hervás-Martínez
Department of Computer Science and Numerical Analysis, University of Cordoba,
Campus de Rabanales, C2 building, 14071 Cordoba, Spain
e-mail: i92domom@uco.es

C. Hervás-Martínez
e-mail: chervas@uco.es

A. Sianes
ETEFA Foundation for Development and Cooperation, Loyola University Andalusia,
Escritor Castilla Aguayo 4, 14004 Cordoba, Spain
e-mail: antonio.sianes@fundacionetea.org

3

Ordinal prediction applications in medicine: the liver allocation problem

This chapter presents some articles regarding the liver allocation problem. They propose a recommendation system to help medical experts in the selection of a suitable recipient for a given organ in a liver transplantation. Using artificial neural networks and evolutionary algorithms, these models will identify in which recipient the organ will be able to survive longer.

Main publications associated with this chapter:

- M. Dorado-Moreno, M. Pérez-Ortiz, M. D. Ayllón-Terán, P. A. Gutiérrez y C. Hervás-Martínez. Ordinal Evolutionary Artificial Neural Networks for Solving an Imbalanced Liver Transplantation Problem. In *11th International Conference on Hybrid Artificial Intelligent Systems (HAIS)*, pages 451–462, 2016, Sevilla, Spain.
- M. Dorado-Moreno, M. Pérez-Ortiz, P. A. Gutiérrez, R. Ciria, J. Briceño y C. Hervás-Martínez. Dynamically weighted Evolutionary Ordinal Neural Network for solving an Imbalanced Liver Transplantation Problem. *Artificial Intelligence in Medicine*, 77:1–11, 2014, Impact Factor (2017): 2.879 (Q1).

These publications are now presented in the different sections of this chapter.

3.1. Ordinal Evolutionary Artificial Neural Networks for Solving an Imbalanced Liver Transplantation Problem

The following paper proposes a model to predict in which recipient a liver would survive longer. In contrast to the current recipient scoring methodology known as MELD, which only considers the recipient's health status, this model combines information from the donor, the recipient and the surgery process in order to select the most suitable recipient for a liver.

The data collected contains the survival time of the graft and information from the donor, the recipient, and the surgery process. This data was collected from 7 Spanish liver transplant units and the King's College Hospital (London), achieving a total of 1406 surgeries to perform the study. The patterns were classified considering the survival days of the liver after the surgery, establishing five classes: less than 15 days, from 15 days to 3 months, from 3 months to 1 year and more than one year. The issue in this problem is that the model must learn how to distinguish early failures of the graft and there are only a few patterns where the graft survived for less than one year; thus, the model can hardly learn how to differentiate the patterns from this class to the rest of the classes.

In this paper, different perspectives have been considered for this topic: in the first instance, an ordinal SMOTE technique is implemented to over-sample groups of interesting examples taking into account the class order information. Then, the algorithmic approach in which the cost function of the evolutionary algorithm is modified is employed in order to penalize the misclassified patterns of the interesting minority classes that are harder than the rest of the classes.

The experiments compare different ordinal classifiers that worked successfully in similar problems with the state-of-the-art approaches versus our evolutionary artificial neural network proposal with the modified cost function. We also compared the results among three different ordinal over-sampling techniques. In conclusion, over-sampling was needed in order to achieve non-trivial classifiers, and the state-of-the-art models that were able to achieve non-trivial classifiers were not competitive with the performance of our proposal.

Ordinal Evolutionary Artificial Neural Networks for Solving an Imbalanced Liver Transplantation Problem

Manuel Dorado-Moreno^{1(✉)}, María Pérez-Ortiz^{2,3},
María Dolores Ayllón-Terán³, Pedro Antonio Gutiérrez¹,
and Cesar Hervás-Martínez¹

¹ Department of Computer Science and Numerical Analysis, University of Córdoba,
Rabanales Campus, Albert Einstein Building, 14071 Córdoba, Spain

{i92domom,pagutierrez,chervas}@uco.es

² Department of Mathematics and Engineering,
Universidad Loyola Andalucía, Córdoba, Spain

i82perom@uco.es

³ Liver Transplantation Unit, Reina Sofía Hospital, Córdoba, Spain
lolesat83@hotmail.com

Abstract. Ordinal regression considers classification problems where there exists a natural ordering among the categories. In this learning setting, thresholds models are one of the most used and successful techniques. On the other hand, liver transplantation is a widely-used treatment for patients with a terminal liver disease. This paper considers the survival time of the recipient to perform an appropriate donor-recipient matching, which is a highly imbalanced classification problem. An artificial neural network model applied to ordinal classification is used, combining evolutionary and gradient-descent algorithms to optimize its parameters, together with an ordinal over-sampling technique. The evolutionary algorithm applies a modified fitness function able to deal with the ordinal imbalanced nature of the dataset. The results show that the proposed model leads to competitive performance for this problem.

Keywords: Ordinal regression · Artificial neural networks · Imbalanced classification · Liver transplantation · Donor-recipient matching

1 Introduction

Liver transplantation is an accepted treatment for patients who present end-stage liver disease. However, transplantation is restricted by the lack of suitable

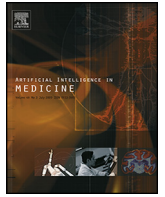
This work has been subsidized by the TIN2014-54583-C2-1-R project of the Spanish Ministerial Commission of Science and Technology (MICYT), FEDER funds and the P2011-TIC-7508 project of the “Junta de Andalucía” (Spain). The authors M. Dorado-Moreno, M. Pérez-Ortiz and M.D. Ayllón-Terán have contributed equally to the preparation of this paper.

3.2. Dynamically weighted evolutionary ordinal neural network for solving an imbalanced liver transplantation problem

This paper considers the same dataset employed in our previous work, which consists of 1406 surgeries with information from the donor, the recipient and the surgery. To solve this problem, a dynamic cost function is developed to be used in the evolutionary algorithm. This means that after some iterations of the algorithm, the cost function that penalizes the misclassification of patterns belonging to the minority classes is recalculated considering the current classification performance of the algorithm in each class.

This paper performed a comparison of the results considering only recipient information because this approach is currently being applied in the hospitals with the MELD score. Then, it considered only donor information, obtaining similar results. Observing that only one source of information was insufficient, the combined donor and surgery information and the combined recipient and surgery information were tested, resulting in performances with poor quality. In conclusion, all the information from the donor, recipient and surgery is needed in order to achieve non-trivial classifiers that attain competitive performance.

The conclusion from the results is that the dynamic cost function added to the evolutionary algorithm is able to double the classification performance of the static cost function proposed in the previous work. Furthermore, there are no state-of-the-art models that are able to compete with our proposal's results, as they are not directly prepared to handle the imbalanced data.



Dynamically weighted evolutionary ordinal neural network for solving an imbalanced liver transplantation problem



Manuel Dorado-Moreno^{a,*}, María Pérez-Ortiz^b, Pedro A. Gutiérrez^a, Rubén Ciria^c, Javier Briceño^c, César Hervás-Martínez^a

^a Department of Computer Science and Numerical Analysis, University of Córdoba, Campus Universitario de Rabanales, "Albert Einstein Building", Third Floor, 14071 Córdoba, Spain

^b Department of Quantitative Methods, Universidad Loyola Andalucía, Escritor Castilla Aguayo 4, 14004 Córdoba, Spain

^c Liver Transplantation Unit, Reina Sofía Hospital, Av. Menéndez Pidal, 14004 Córdoba, Spain

ARTICLE INFO

Article history:

Received 21 July 2016

Received in revised form 17 January 2017

Accepted 5 February 2017

Keywords:

Artificial neural networks

Ordinal classification

Imbalanced classification

Survival analysis

Liver transplantation

ABSTRACT

Objective: Create an efficient decision-support model to assist medical experts in the process of organ allocation in liver transplantation. The mathematical model proposed here uses different sources of information to predict the probability of organ survival at different thresholds for each donor–recipient pair considered. Currently, this decision is mainly based on the Model for End-stage Liver Disease, which depends only on the severity of the recipient and obviates donor–recipient compatibility. We therefore propose to use information concerning the donor, the recipient and the surgery, with the objective of allocating the organ correctly.

Methods and materials: The database consists of information concerning transplants conducted in 7 different Spanish hospitals and the King's College Hospital (United Kingdom). The state of the patients is followed up for 12 months. We propose to treat the problem as an ordinal classification one, where we predict the organ survival at different thresholds: less than 15 days, between 15 and 90 days, between 90 and 365 days and more than 365 days. This discretization is intended to produce finer-grain survival information (compared with the common binary approach). However, it results in a highly imbalanced dataset in which more than 85% of cases belong to the last class. To solve this, we combine two approaches, a cost-sensitive evolutionary ordinal artificial neural network (ANN) (in which we propose to incorporate dynamic weights to make more emphasis on the worst classified classes) and an ordinal over-sampling technique (which adds virtual patterns to the minority classes and thus alleviates the imbalanced nature of the dataset).

Results: The results obtained by our proposal are promising and satisfactory, considering the overall accuracy, the ordering of the classes and the sensitivity of minority classes. In this sense, both the dynamic costs and the over-sampling technique improve the base results of the considered ANN-based method. Comparing our model with other state-of-the-art techniques in ordinal classification, competitive results can also be appreciated. The results achieved with this proposal improve the ones obtained by other state-of-the-art models: we were able to correctly predict more than 73% of the transplantation results, with a geometric mean of the sensitivities of 31.46%, which is much higher than the one obtained by other models.

Conclusions: The combination of the proposed cost-sensitive evolutionary algorithm together with the application of an over-sampling technique improves the predictive capability of our model in a significant way (especially for minority classes), which can help the surgeons make more informed decisions about the most appropriate recipient for an specific donor organ, in order to maximize the probability of survival after the transplantation and therefore the fairness principle.

© 2017 Elsevier B.V. All rights reserved.

* Corresponding author.

E-mail address: i92domom@uco.es (M. Dorado-Moreno).

4

Ordinal prediction in time series: applications in wind energy

This chapter encompasses the contributions of the thesis related to ordinal prediction, specifically the prediction of the wind power ramp events in wind farms. These articles include different methodology proposals based on the standard echo state networks, which have been enhanced for this specific problem through different modifications. The first works include nominal prediction as a first step to solve this problem, which were then extended to ordinal prediction.

Main publications associated with this chapter:

- M. Dorado-Moreno, L. Cornejo-Bueno, P. A. Gutiérrez, P. A. Gutiérrez, S. Salcedo-Sanz y C. Hervás-Martínez. Combining Reservoir Computing and Over-Sampling for Ordinal Wind Power Ramp Prediction. In *International Work-Conference on Artificial and Natural Neural Networks (IWANN)*, pages 708–719, 2017, Cádiz, Spain.
- M. Dorado-Moreno, L. Cornejo-Bueno, P. A. Gutiérrez, L. Prieto, C. Hervás-Martínez y S. Salcedo-Sanz. Robust estimation of wind power ramp events with reservoir computing. *Renewable Energy*, 111:428–437, 2017, Impact Factor (2017): 4.900 (Q1).
- M. Dorado-Moreno, P. A. Gutiérrez, L. Cornejo-Bueno, L. Prieto, S. Salcedo-Sanz y C. Hervás-Martínez. Ordinal multi-class architecture for predicting wind power

ramp events based on reservoir computing. *Neural Processing Letters*, In press, 2018, Impact Factor (2017): 1.787 (Q2).

- M. Dorado-Moreno, P. A. Gutiérrez, S. Salcedo-Sanz, L. Prieto y C. Hervás-Martínez. Wind power ramp events ordinal prediction using minimum complexity echo state networks. In the *2018 International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*, pages 180–187, 2018, Madrid, Spain.

Other publications associated with this chapter:

- M. Dorado-Moreno, A.M. Durán-Rosal, D. Guijo-Rubio, P.A. Gutiérrez, L. Prieto, S. Salcedo-Sanz y C. Hervás-Martínez. Multiclass Prediction of Wind Power Ramp Events Combining Reservoir Computing and Support Vector Machines. In *17th Conference of the Spanish Association for Artificial Intelligence*, pages 300–309, 2016, Salamanca, Spain.

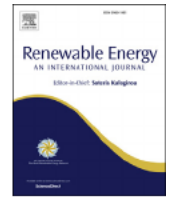
The four main publications are now presented in the different sections of this chapter.

4.1. Robust estimation of wind power ramp events with reservoir computing

This section contributes to the study of different echo state network (ESN) architectures when applied to WPREs prediction, and in this case, it only distinguishes between ramp and non-ramp classes. The different architectures under research were combinations of the inputs of both the reservoir and the output layer. As a baseline, the standard ESN with a discretization function after its ridge regression output layer was compared with the different proposals that made use of logistic regression to perform the classification.

In the paper, two different problems were solved: the first proposal tries to predict WPREs in the next 6 hours, while the second one predicts whether a WPRE will occur or not in any of the next 24 hours. The different proposals were compared in both of the problem proposals using different evaluation metrics for binary classification problems.

The results showed which architecture was the most suitable for this problem; thus, it would be the one used in the next research work on this topic. There were no papers that solved the prediction of WPREs in the same manner as our proposal, so the state-of-the-art results left us to conjecture whether our results were competitive or not. Consequently, we implemented the standard ESN as a baseline of our results and showed that the performed modifications greatly increased its performance.



Robust estimation of wind power ramp events with reservoir computing



M. Dorado-Moreno ^{a, *}, L. Cornejo-Bueno ^b, P.A. Gutiérrez ^a, L. Prieto ^c,
C. Hervás-Martínez ^a, S. Salcedo-Sanz ^b

^a Department of Computer Science and Numerical Analysis, Universidad de Córdoba, Córdoba, Spain

^b Department of Signal Processing and Communications, Universidad de Alcalá, Alcalá de Henares, Spain

^c Department of Energy Resource, Iberdrola, Madrid, Spain

ARTICLE INFO

Article history:

Received 3 March 2016

Received in revised form

9 March 2017

Accepted 9 April 2017

Available online 12 April 2017

Keywords:

Wind power ramp events prediction

Recurrent neural networks

Reservoir computing

Echo state networks

Reanalysis data

Time series

ABSTRACT

Wind power ramp events are sudden increases or decreases of wind speed within a short period of time. Their prediction is nowadays one of the most important research trends in wind energy production because they can potentially damage wind turbines, causing an increase in wind farms management costs. In this paper, 6-h and 24-h binary (ramp/non-ramp) prediction based on reservoir computing methodology is proposed. This forecasting may be used to avoid damages in the turbines. Reservoir computing models are used because they are able to exploit the temporal structure of data. We focus on echo state networks, which are one of the most successfully applied reservoir computing models. The variables considered include past values of the ramp function and a set of meteorological variables, obtained from reanalysis data. Simulations of the system are performed in data from three wind farms located in Spain. The results show that our algorithm proposal is able to correctly predict about 60% of ramp events in both 6-h and 24-h prediction cases and 75% of the non-ramp events in the next 24-h case. These results are compared against state of the art models, obtaining in all cases significant improvements in favour of the proposed algorithm.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

The prediction of Wind Power Ramps Events (WPRES) in wind farms is one of the current hot topics in wind energy research. WPRES deeply affect local wind speed prediction, also increasing the management costs of wind farms, because of their potential damage effect in wind turbines [13,25]. WPRES consist of important fluctuations of wind power in a short period of time (within a few hours), leading to a significant increase or decrease of the power produced in the wind farm. The origin of WPRES are specific meteorological processes (usually crossing fronts, local fast changes in the wind, etc.). Currently, the most effective way of dealing with WPRES in wind farms is the correct prediction of these events, as has been recently reported [6,10].

Different previous works have dealt with the problem of WPRES prediction, many of them applying computational intelligence techniques. In Ref. [30], different time series prediction models

have been evaluated in a problem of WPRES prediction, with a short-term prediction horizon between 10 min and 1 h. Experiments in a large wind farm with 100 wind turbines reports good performance of this data-mining approach. However, the short prediction time-horizon of this study makes difficult to apply it in real cases within wind farms. In Ref. [1], a hybrid Autoregressive Moving Average (ARMA) – Hidden Markov model approach has been proposed to predict wind ramp events. Experiments in two with farms in the USA show a good performance of the methodology proposed. In Ref. [27], a dynamic programming approach is proposed for detecting WPRES in time series of wind power. In Ref. [12], a neural network approach for switching between three different regimes of WPRES (ramp-up, ramp-down and no-ramp) is proposed, in which depending on the WPRES type, a different neural network is trained with specific structure and training algorithm. In Ref. [6], a neural network is used as a surrogate model of the wind power generation at a wind farm, then the neural network is used to simulate different possible future scenarios of wind power generation and WPRES. In Ref. [29], a Support Vector Machine (SVM) for classification is used to forecast WPRES, after grouping the ramp events in

* Corresponding author.

E-mail address: manuel.dorado@uco.es (M. Dorado-Moreno).

4.2. Combining Reservoir Computing and Over-Sampling for Ordinal Wind Power Ramp Prediction

WPREs are one of the most harmful issues in wind farms and the most effective way of dealing with them is to predict their occurrence. Thus, this paper focuses on an ordinal prediction of WPREs in three classes (negative ramp, non-ramp, positive ramp), which are naturally ordered as they depend on wind speed.

The model used to solve this problem was a modified echo state network whose input is the wind speed in the previous time instant, and the network output is connected to a kernel mapping layer together with reanalysis data that contains information of wind speed and temperature at different heights. Finally, its output is connected to an ordinal logistic regression layer that will be trained to predict whether a WPRE will occur or not in the next 6 hours. On the other hand, the database is imbalanced since WPREs are not common events, and most of the time, there is a non-ramp state, meaning that the wind speed is stable. To solve this, SMOTE was applied to the reservoir activations instead of the original data because it would damage the temporal information by including synthetic patterns in the dataset that would not have a temporal relation to the previous nor the following pattern.

The experiments show that the model is able to achieve good results, improving those attained by the state-of-the-art techniques, due to the ordinal classification approach that helped the model to distinguish between negative ramps and positive ramps correctly.

Combining Reservoir Computing and Over-Sampling for Ordinal Wind Power Ramp Prediction

Manuel Dorado-Moreno¹(✉), Laura Cornejo-Bueno²,
Pedro Antonio Gutiérrez¹, Luis Prieto³, Sancho Salcedo-Sanz²,
and César Hervás-Martínez¹

¹ Department of Computer Science and Numerical Analysis,
Universidad de Córdoba, Córdoba, Spain

`manuel.dorado@uco.es`

² Department of Signal Processing and Communications,
Universidad de Alcalá, Alcalá de Henares, Spain

³ Department of Energy Resource, Iberdrola, Madrid, Spain

Abstract. Wind power ramp events (WPREs) are strong increases or decreases of wind speed in a short period of time. Predicting WPREs is of vital importance given that they can damage the turbines in a wind farm. In contrast to previous binary approaches (ramp versus non-ramp), a three-class prediction is proposed in this paper by considering: negative ramp, non-ramp and positive ramp, where the natural order of the events is clear. The independent variables used for prediction include past ramp function values and meteorological data obtained from physical models (reanalysis data). The proposed methodology is based on reservoir computing and an over-sampling process for alleviating the high degree of unbalance of the dataset (non-ramp events are much more frequent than ramps). The reservoir computing model is a modified echo state network composed by: a recurrent neural network layer, a nonlinear kernel mapping and an ordinal logistic regression, in such a way that the order of the classes can be exploited. The standard synthetic minority oversampling technique (SMOTE) is applied to the reservoir activations, given that the direct application over the input variables would damage its temporal structure. The performance of this proposal is compared to the original dataset (without over-sampling) and to nominal logistic regression, and the results obtained with the oversampled dataset and ordinal logistic regression are found to be more robust.

Keywords: Wind power ramp events · Reservoir computing · SMOTE · Over-sampling · Ordinal prediction · Kernel mapping

This work has been subsidized by the TIN2014-54583-C2-1-R, TIN2014-54583-C2-2-R and TIN2015-70308-REDT projects of the Spanish Ministerial Commission of Science and Technology (MINECO, Spain) and FEDER funds (EU). Manuel Dorado-Moreno's research has been subsidized by the FPU Predoctoral Program of the Spanish Ministry of Education, Culture and Sport, grant reference FPU15/00647.

4.3. Ordinal Multi-class Architecture for Predicting Wind Power Ramp Events Based on Reservoir Computing


This work is an extension of the one presented in Section 4.2. Furthermore, this work first redefined the ramp function used in the previous paper, as it was unable to detect some of the WPREs, thus resulting in a dataset with a higher degree of imbalance and patterns belonging to the majority class that should have been classified in the minority classes. This modification consisted of calculating the ramp function, i.e., the function evaluated to decide whether there is a WPRE or not, using the maximum and the minimum values of the wind speed in the time interval (6 hours) instead of using only the wind speed value at the first and the last hour.

Another addition to the previous paper is a function to impute all the missing data through the use of a regression algorithm and the reanalysis data. Previously, if there were more than 6 hours lost in a row, the data were just discarded from the final dataset. With these two additions to the raw data treatment, the amount of both positive and negative ramps was increased by 30 %.

The results were compared with different state-of-the-art methods that were used for nominal and ordinal regression of time series to determine whether the ordinal output layer is useful. In addition, different percentages of the over-sampling ratio were used. The results illustrated that the novel data treatment increased the prediction performance and that except for one of the three wind farms, the proposed architecture was still the one obtaining the best performance.



Ordinal Multi-class Architecture for Predicting Wind Power Ramp Events Based on Reservoir Computing

M. Dorado-Moreno¹  · P. A. Gutiérrez¹ · L. Cornejo-Bueno² · L. Prieto³ · S. Salcedo-Sanz² · C. Hervás-Martínez¹

© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

Wind power ramp events (WPRES) are strong increases or decreases of wind speed in a short period of time. Predicting WPRES in wind farms is of vital importance given that they can produce damages in the turbines, and, in any case, they suddenly affect the wind farm production. In contrast to previous binary definitions of the prediction problem (ramp vs non-ramp), a three-class prediction model is used in this paper, proposing a novel discretization function, able to detect the nature of WPRES: negative ramp, non-ramp and positive ramp events. Moreover, the natural order of these labels is exploited to obtain better results in the prediction of these events. The independent variables used for prediction include, in this case, past wind speed values and meteorological data obtained from physical models (reanalysis data). Reanalysis will be also used for recovering missing data from the measuring stations in the wind farm. The proposed prediction methodology is based on Reservoir Computing and an over-sampling process for alleviating the high degree of unbalance in the dataset (non-ramp events are much more frequent than ramps). Three elements are combined in the prediction method: a recurrent neural network layer, a nonlinear kernel mapping and an ordinal logistic regression, to exploit the information provided by the order of the classes). Preprocessing is based on a variation of the standard synthetic minority over-sampling technique, which is applied to the reservoir activations (since the direct application over the input variables would damage its temporal structure). The performance of the method is analysed by comparing it against other state-of-the-art classifiers, such as Support Vector Machines, nominal logistic regression, an autoregressive ordinal neural network, or the use of leaky integrator neurons instead of the standard sigmoidal units. From the results obtained, the benefits of the kernel mapping and the ordinal model are clear, and, in general, the performance obtained with the Reservoir Computing approach is shown to be very robust in the detection of ramps.

✉ M. Dorado-Moreno
manuel.dorado@uco.es

¹ Department of Computer Science and Numerical Analysis, Universidad de Córdoba, Córdoba, Spain

² Department of Signal Processing and Communications, Universidad de Alcalá, Alcalá de Henares, Spain

³ Department of Energy Resource, Iberdrola, Madrid, Spain

4.4. Wind power ramp events ordinal prediction using minimum complexity echo state networks

Previous works proposed the use of ESNs in order to predict WPREs. The nature of ESNs is purely stochastic, as the reservoir is generated randomly by following a Gaussian distribution; the only control over the reservoir is that when generated, it needs to fulfil the echo state property. There is a proposal called minimum complexity echo state networks, which proposes three different ESN architectures whose reservoir links are generated following a reasonable pattern, and in this way, they can be initialized without yielding a stochastic result because in this scenario, the ESN's link connections have been determined by the user.

In this work, the three different minimum complexity ESN structures in combination with our three architecture proposals are first tested to check if the minimum complexity ESNs obtain competitive results because they drastically reduce the number of links in the reservoir, and in the second instance, to determine the best structure/architecture combination to predict WPREs.

The results showed that minimum complexity ESNs, in addition to reducing the computational time of the training, were able to achieve competitive results when compared to the results obtained in previous works.



Wind Power Ramp Events Ordinal Prediction Using Minimum Complexity Echo State Networks

M. Dorado-Moreno^{1(✉)}, P. A. Gutiérrez¹, S. Salcedo-Sanz², L. Prieto³,
and C. Hervás-Martínez¹

¹ Department of Computer Science and Numerical Analysis, University of Cordoba,
Córdoba, Spain

`manuel.dorado@uco.es`

² Department of Signal Processing and Communications, University of Alcalá,
Alcalá de Henares, Spain

³ Department of Energy Resource, Iberdrola, Madrid, Spain

Abstract. Renewable energy is the fastest growing source of energy in the last years. In Europe, wind energy is currently the energy source with the highest growing rate and the second largest production capacity, after gas energy. There are some problems that difficult the integration of wind energy into the electric network. These include wind power ramp events, which are sudden differences (increases or decreases) of wind speed in short periods of times. These wind ramps can damage the turbines in the wind farm, increasing the maintenance costs. Currently, the best way to deal with this problem is to predict wind ramps beforehand, in such way that the turbines can be stopped before their occurrence, avoiding any possible damages. In order to perform this prediction, models that take advantage of the temporal information are often used. One of the most well-known models in this sense are recurrent neural networks. In this work, we consider a type of recurrent neural networks which is known as Echo State Networks (ESNs) and has demonstrated good performance when predicting time series. Specifically, we propose to use the Minimum Complexity ESNs in order to approach a wind ramp prediction problem at three wind farms located in the Spanish geography. We compare three different network architectures, depending on how we arrange the connections of the input layer, the reservoir and the output layer. From the results, a single reservoir for wind speed with delay line reservoir and feedback connections is shown to provide the best performance.

This work has been subsidized by the projects with references TIN2017-85887-C2-1-P, TIN2017-85887-C2-2-P and TIN2017-90567-REDT of the Spanish Ministry of Economy and Competitiveness (MINECO) and FEDER funds. Manuel Dorado-Moreno's research has been subsidised by the FPU Predoctoral Program (Spanish Ministry of Education and Science), grant reference FPU15/00647. The authors acknowledge *NVIDIA Corporation* for the grant of computational resources through the *GPU Grant Program*.

5

Deep neural networks for multi-task learning

This chapter presents the work that concludes the thesis, including most of the knowledge that was acquired through the development of this thesis. It includes the prediction of time series, techniques to ease imbalanced data issues and a new definition for the wind power ramp events (WPREs) prediction problem, which is now addressed from a global perspective instead of predicting wind power ramp events in a specific location, i.e., a single wind farm. To evaluate this work, multi-task and deep learning methodologies were studied deeply and combined to achieve competitive results.

Publication associated with this chapter:

- M. Dorado-Moreno, N. Navarin, P.A. Gutiérrez, L. Prieto, A. Sperduti, S. Salcedo-Sanz and C. Hervás-Martínez. Multi-task Learning for the Prediction of Wind Power Ramp Events with Deep Neural Networks. *Neural Networks* (Under Revision), 2018, Impact Factor (2017): 7.197 (Q1).

This publication is presented in the next section.

5.1. Multi-task Learning for the Prediction of Wind Power Ramp Events with Deep Neural Networks

Multi-task learning is now experiencing a resurgence due to the good results obtained when combined with deep neural networks. This paper proposes to solve the WPREs prediction problem from a different perspective than that proposed in the previous works, which was to predict these events in a specific wind farm. Now, this paper proposes a multi-task problem where each task is the prediction of these events in a wind farm. Thus, a single model is able to not only predict WPREs but also to enhance the results due to the exploitation of multi-task learning characteristics, which allow the model to be enhanced through the use of information of different related tasks included in the model. The model selected to apply the multi-task learning was a deep neural network composed of two main parts: some shared layers that will compute the input information from the different tasks together, and some specific layers that will take the last shared layer output and use it to predict each of the specific tasks.

To handle the imbalanced nature of the database, over-sampling techniques could not be applied because of multi-task learning, and the time series of each input were synchronized. Specifically, over-sampling WPREs in one of the tasks could mean creating a new pattern from the majority class in the other classes, and this would lead to a dataset with the same degree of imbalance but with synthetic patterns. To solve this, we developed a training algorithm based on dynamic costs that are modified according to the misclassification rate of each task.

The different experiments show that the proposed model and algorithm are able to predict multiple WPREs in different wind farms at the same time, with the results that reach the performance of state-of-the-art methods that focus on a single wind farm.

Multi-task Learning for the Prediction of Wind Power Ramp Events with Deep Neural Networks

M. Dorado-Moreno^{a,*}, N. Navarin^{b,c}, P.A. Gutiérrez^a, L. Prieto^e, A. Sperduti^c,
S. Salcedo-Sanz^d, C. Hervás-Martínez^a

^a*Department of Computer Science and Numerical Analysis, University of Cordoba,
Córdoba, Spain*

^b*Department of Computer Science, University of Nottingham, Nottingham, United Kingdom*

^c*Department of Mathematics, University of Padova, Padova, Italy*

^d*Department of Signal Processing and Communications, University of Alcalá, Alcalá de
Henares, Spain*

^e*Department of Energy Resource, Iberdrola, Madrid, Spain*

Abstract

In Machine Learning, the most common way to address a given problem is to optimize an error measure by training a single model to solve the desired task. However, sometimes it is possible to exploit latent information from other related tasks to improve the performance of the main one, resulting in a learning paradigm known as Multi-Task Learning (MTL). In this context, the high computational capacity of deep neural networks (DNN) can be combined with the improved generalization performance of MTL, by designing independent output layers for every task and including a shared representation for them. In this paper we exploit this theoretical framework on a problem related to Wind Power Ramps Events (WPREs) prediction in wind farms. Wind energy is one of the fastest growing industries in the world, with potential global spreading and deep penetration in developed and developing countries. One of the main issues with the majority of renewable energy resources is their intrinsic intermittency, which makes difficult to increase the penetration of these technologies into the energetic mix. In this case, we focus on the specific problem of WPREs prediction, which deeply affect the wind speed and power prediction, and they are also related to different turbines damages. Specifically, we exploit the fact that WPREs are spatially-related events, in such a way that predicting the occurrence of WPREs in different wind farms can be taken as related tasks, even when the wind farms are far away from each other. We propose a DNN-MTL architecture, receiving inputs from all the wind farms at the same time to predict WPREs simultaneously in each of the farms locations. The architecture includes some *shared layers* to learn a common representation for the information from all the wind farms, and it also includes some *dedicated layers*, which refine the representation to match the specific characteristics of each location. Finally we modified the Adam optimization algorithm for dealing with imbalanced data,

*Corresponding author

Email addresses: manuel.dorado@uco.es (M. Dorado-Moreno),
nnavarin@math.unipd.it (N. Navarin), pagutierrez@uco.es (P.A. Gutiérrez),
sperduti@math.unipd.it (A. Sperduti), sancho.salcedo@uah.es (S. Salcedo-Sanz),
chervas@uco.es (C. Hervás-Martínez)

6

Discussion and conclusions

This chapter concludes the thesis, including the main conclusions obtained in the previous chapters and highlighting some future research derived from this work.

This thesis has focused on solving different real-life problems related not only to time series through the use and improvement of ordinal classification models and algorithms. As stated in the introduction, several objectives have been established, including proposing different neural network architectures and enhancing the state-of-the-art output layers. Subsequently, improving the performance of evolutionary algorithms when applied to artificial neural networks, as well as preparing training algorithms for handling imbalanced data followed as related objectives. As the most substantive aspect of this thesis concentrates on applications of these techniques, the goals referring to that aspect includes classifying rich countries considering their help to poorer to predict the globalisation of countries, predicting the most suitable recipient for a given liver in an imbalanced dataset, to extract a database to predict WPREs, develop models to predict WPREs and achieve the prediction of WPREs in multiple wind farms with a single model.

In our opinion, all these objectives have been achieved and proven in different applications. To support this statement, this chapter briefs the contributions together with some conclusions and will end with some future research lines. Please note that this section summarises the conclusions; further details are provided in the corresponding chapters.

6.1. Conclusions

This section presents the research performed on the topic of ordinal classification with respect to the four main lines of work: hybridization of training algorithms and basis functions; applications in medicine, exemplified by the liver allocation problem; time series prediction, illustrated by applications in wind energy; and finally, deep neural networks for multi-task learning. In this section, we summarise the thesis contributions grouped by topics.

6.1.1. Ordinal classification: hybridization of training algorithms and basis functions

This thesis contribution started with Chapter 2, which proposes different improvements for ANNs and evolutionary algorithms, both based in hybridizations, and proposes how to obtain a ranking of patterns from an ordinal ANN model.

We first worked on the development of an ANN model that outperformed the state-of-the-art models in ordinal classification problems. To do so, a new ANN based on POM was developed, and then its hidden layer was tested with a pure and hybrid hidden layer. To train this model, an evolutionary algorithm was used, as it can optimize the number of hidden neurons in the hidden layer. Moreover, an ordinal fitness function was set and the algorithm was combined with a local search algorithm, for which the objective was to optimize the best solution of the evolutionary algorithm. The model performance was tested using three different metrics on six ordinal datasets, showing that its performance was statistically significantly better than both state-of-the-art models and the pure (no hybridization applied) hidden layers of ANNs.

In the next work, this model was tested in a real application where the contribution of rich countries in the third world was measured. Considering 22 rich countries (from 2003 to 2010) classified in 5 different classes and ordered from the lowest to the highest contribution, our model was able to classify 19 of the 22 countries correctly when compared to the original score, while ordinal SVM versions could only classify a maximum of 15 countries correctly. It is important to note that this misclassification was determined for the 3 countries closest to the threshold between the original and the misclassified class.

To close the ordinal classification section, we used ANNs to measure the extent of globalisation of 147 countries included in the KOF Index, which is one of the most accepted metrics to analyse the globalisation of countries. However, this index is subject to limitations, e.g., it has no mathematical form for combining the different variables and must acquire the information from all the countries in the study before being able to obtain a ranking. Therefore, we proposed a hybrid ordinal ANN to perform the ranking and

compared it with two different metrics versus other ordinal state-of-the-art models. First, our proposal was able to achieve better performance in terms of accuracy and MAE than the rest of the models, only misclassifying 2 countries out of the 147 in the class previous to the original. Then, we performed a sensitivity analysis of every KOF variable (group of indicators) to check how every variable affected the final result, concluding that the "Political Globalisation" is the most important variable in this study, while the KOF Index gives it the lowest importance. Finally, we were able to achieve a new ranking of the countries using our ANN model, where some of the countries won/lost up to 9 positions when compared to the KOF Index.

6.1.2. Ordinal prediction applications in medicine: the liver allocation problem

Chapter 3 of this thesis focuses on ordinal classification to tackle a liver allocation problem. This problem consists of selecting the most suitable recipient for a new liver considering variables related to the donor, the recipient and surgery. This problem, as most of those in medicine, is highly imbalanced and the most important classes are those having fewer numbers of patterns. Therefore, in order to solve this, we selected ANNs, over-sampling techniques and a training algorithm to handle the dataset imbalance. The dataset was built from information of 7 Spanish liver transplant units together with information from the King's College Hospital (London), achieving a total of 1406 surgeries to perform the study. The patterns were classified considering the survival days of the liver after the surgery, establishing five classes: less than 15 days, from 15 days to 3 months, from 3 months to 1 year and more than one year.

First, we used SMOTE techniques to over-sample the groups of interesting patterns and thus to palliate the skewness of the distribution of the patterns belonging to each class. Together with the SMOTE techniques, in order to train an ordinal ANN, we developed an evolutionary algorithm with a modified function to penalize the misclassification of the patterns belonging to the interesting minority classes. The results of this proposal were compared with different state-of-the-art models, and additionally, three different ordinal SMOTE methodologies were used. Three metrics were considered in this comparison, where the most important was the GMS (geometric mean of the sensitivities), which instantly discards all trivial models (those that classify all the patterns in the majority class), obtaining 0 points in this metric. The best results were obtained with the ordinal ANN that used the cost function to handle the imbalance together with the OGO-SP (ordinal graph-based over-sampling via shortest paths using a probability function for the intra-class edges).

The next step was to improve the algorithm capabilities to handle the data imbalan-

ce; to do so, the proposal was to implement a dynamic cost function, which was evolved together with the algorithm results considering the worst classified class. The results showed that this improved cost function was able to double the GMS result obtained with the static version of the cost function, which also means that the results obtained with the state-of-the-art models are even more far from this model than they were in the previous work. This work also performs an analysis to decide whether all groups of variables (donor, recipient and surgery) are necessary or not, concluding that all are necessary since the model is only able to obtain a non-trivial classifier when all of them are used. Finally, even though ANNs are usually treated as black box models, this work interprets the ANN model in order to extract the importance of each variable in the final result. Considering the interpretability of the model, it is also important to note that this is a threshold-based model; thus, each pattern result can be compared with the class border to check if it is located close to it (it could belong to the next or previous class) or if it is far from the threshold, which would mean a robust result.

6.1.3. Ordinal prediction in time series: applications in wind energy

Wind power ramp events (WPREs) are one of the most harmful events in wind farms, and currently, the most effective way of dealing with them is to make an accurate prediction with enough time to take the necessary actions to prevent any damage or loss.

The first problem with this prediction is that there is not a selected formula to decide whether a wind event can be considered a WPRE or not; the second and more important problem is that less than 10% of a dataset with nearly 100,000 patterns (for each wind farm) are WPREs; therefore, there is a high degree of imbalance to address. The dataset is composed of three wind farms located in the Spanish geography that contain information of wind speed recorded by sensors every hour from 2002 to 2013. In addition, we decided to use reanalysis data from the ERA-Interim Reanalysis Project, which calculates information from meteorological processes every six hours, so that we could enhance the data quality by adding this information, even though a loss of the temporal resolution of the data to some degree was incurred, which decreased from 1 to 6 hours.

In our first work, we focused on solving a binary classification problem where an event could be a WPRE or not. To handle the temporal relation of the patterns, i.e., the time series characteristics, we decided to use echo state networks, a type of recurrent neural network generated stochastically that has achieved good results in the last years and can process the temporal order of the data. We proposed four different architectures of ESN; the first one was a discretized version of the standard ESN including only the wind power from the previous time that we wanted to predict, and the rest directly used a logistic regression as the output layer. The second proposal also included the reanalysis data and

was connected directly to the output layer; the third additionally connected the reanalysis data to the hidden layer (reservoir); and finally, the last proposal had independent reservoirs, one for the wind power and another for the reanalysis data.

With this established, we performed two different predictions, one of them to predict if a WPRE would occur or not in 6 hours, and the other one to predict whether a WPRE would occur in any of the next 24 hours. We compared our proposals versus a persistence model and logistic regression using four metrics (GMS, AUC, specificity and sensitivity), concluding that the best results in the 6-hour prediction were obtained by the model including both the wind power and the reanalysis data in the same hidden layer (CRC), while in the next 24-hour prediction, the best results were obtained by the model that connected the wind power with the hidden layer and the reanalysis data directly to the output layer (RCX). After performing a ranking of the models considering all the test performances, RCX was the best in GMS and AUC and the second best in SP, while CRC was only the best in ST.

These results were derived in different works, and the first of them was subsequently considered as an ordinal classification problem, where now, instead of distinguishing between WPRE and non-WPRE events, we differentiated negative WPREs (when the wind speed falls), non-WPREs and positive WPREs (when the wind increases). To handle the imbalanced nature of the dataset, we decided to apply SMOTE techniques to it, but we found that it was not that trivial, as applying any standard over-sampling technique to time series would damage the temporal relation of the data. Moreover, over-sampling techniques consider single patterns but not the adjacent ones in time (the previous and following patterns). The solution was to feed the ESN reservoir with all the data to extract a new dataset in which the temporal relations of the data are not reflected, which was non-linearly transformed by the reservoir considering the temporal relations, and then apply the SMOTE over that dataset.

We finally added a kernel mapping layer prior to the ordinal logistic regression (output layer) in order to overcome its well-known limitations, due to its linear nature. In contrast to the previous work, where solving the binary problem did not need over-sampling to achieve acceptable results, for the three-class problem, the imbalance ratio of the dataset was much higher, and models that did not apply over-sampling to the dataset obtained trivial classifiers. Using six different metrics, we compared the proposal versus three models, where one of them did not consider the kernel mapping, and the other two were the nominal versions of the former. The results justify the use of both the kernel mapping and the ordinal output layer because it was the only model able to achieve the best or the second best result for most of the metrics in all the wind farms.

After achieving a methodology that could handle WPREs prediction with a good

performance, we then performed a more fine-grained management of the data. First, we modified the discretization function that had been used by the expert, as we found that it was not able to detect all the possible WPREs. The amount of negative and positive WPREs increased 30 % in this work, slightly reducing the imbalance ratio of the dataset. Then, we decided to recover the missing data instead of ignoring it, which caused the model training to underperform in the cases where the interval from one pattern to the next was not 6 hours. To do so, we used the reanalysis data (which is calculated and has no missing data) together with the non-missing data to train a gradient boosted regression tree ensemble to recover the wind speed missing data from the wind farms sensors. Finally, for this work, we decided to use the wind speed instead of the wind power to train the model, as it would provide us with information that would be more relevant than the power that the wind farm produced in order to predict a WPRE. For this task, we compared our proposal with different state-of-the-art-models that were used to predict time series, e.g., neural networks using input windows and cost-based models, and we also compared the results with the proposal using leaky neurons, which are a special type of reservoir unit that can handle the leak rate with a hyper-parameter. We finally compared the model versus some nominal classifiers and compared two different versions of the ordinal logistic classifier for our proposal. Ultimately, we tested different over-sampling ratios (0 %, 50 %, and 80 %). The best results were obtained using an over-sampling ratio of 50 %, and our proposal using the immediate thresholds version of the ordinal logistic regression was able to achieve results close to 80 % in both the accuracy and the GMS, while in our first work (binary classification), we could only achieve 65 % in the GMS metric.

Finally, the last work of this chapter proposes something different: it tries to reduce the complexity of the reservoir models and at the same time remove their stochastic nature, so that their performance cannot be attributed to chance. At the same time, we wanted to re-validate the conclusions of the best model architecture obtained in our first work.

To achieve the first goal, we used minimum complexity echo state networks proposed in [60], where they establish a fixed structure for the reservoir so that it contains the minimum cycles and is able to process the temporal information of the data. To re-validate our architecture, we again proposed three different architectures for the reservoir of the network. This work concludes that using minimum complexity ESNs does not drastically affect the models' performance (losing approximately 5 % of classification performance) and that the use of a single reservoir to handle the wind speed and connect the reanalysis variables directly to the output layer is still the architecture that performs better after all the transformations are performed on the dataset.

6.1.4. Deep neural networks for multi-task learning

In the last chapter, we conclude the thesis with a completely different approach to solve the WPREs prediction problem. Instead of solving the prediction in each of the wind farms, this work proposes a single model to predict WPREs in all the wind farms, which could be easily extended to any new wind farm located in the Spanish geography. To solve this, we applied the multi-task learning approach, which allows learning different tasks (outputs) in parallel with a single model by assuming that the tasks are related somehow, being able to enhance the results of single-task learning through the exploitation of the information transfer from one task to another. In our case, the outputs are clearly related, as the meteorological processes that are responsible for WPREs are correlated through time and space, and the information of a WPRE occurring in one wind farm could be beneficial to predict the next WPRE in a wind farm next to it.

The model selected for this work is a deep neural network, so that we are capable of processing the vast amount of information from the three wind farms and achieve good results. The model receives the inputs of the three wind farms at the same time, and these inputs are processed together by a number of hidden layers and then spread in three different groups of hidden layers (one for each wind farm), which will process that information specifically to achieve each of the outputs (prediction of WPRE in a wind farm). The information from the sensors in the wind farms now includes the wind direction, as it is a critical factor to decide the positive or negative influence of these meteorological processes in other wind farms depending on their location. Over-sampling was not possible in this work, as the combination of different output possibilities is too high, and over-sampling a minority pattern in one farm would likely lead to creating one belonging to the majority class in the other two farms. To solve this, we applied a dynamic weight function that is modified depending on the worst classified class.

The experiments included a cross-validation of different hyper-parameters, such as: the number of shared layers (those processing all the inputs together), the number of specification layers (those processing the information for each of the wind farms), the number of hidden neurons in each of these layers, the window size of the inputs, the learning rate of the algorithm, the batch size, the dropout probability and the imbalance weight update factor. The results are compared versus the reservoir model from previous works (using the same dataset as the MTL model), and versus the three independent models (without the shared layers) using five different metrics. We can conclude that multi-task learning is a suitable methodology to handle the problem of predicting WPREs in different locations, as the results obtained are robust, and better in most of the metrics than those obtained by independent models and the old reservoir proposal (using this dataset). This means that MTL is able to extract and exploit the relations of the different

tasks, and the deep neural network model using a time window is able to handle all the data together. Finally, another important finding to note is that the cost-based algorithm proposed was sufficient to achieve competitive results without using SMOTE techniques, an outcome that did not happen in the single-task approaches.

6.2. Generic discussion and future work

Future prediction is currently and has always been one of the most trending topics in statistics and artificial intelligence research, not only because it is an interesting challenge to solve but also because of the intrinsic need of human beings to know the future. Currently, it is widely used in meteorology, renewable energies, business processes, industry, and many other fields, which means that there are tons of different applications that need to be solved, and those who need to solve them are the data scientists.

Ordinal classification is a good approach to solve most of these problems, as usually, the predictions are performed on an ordinal scale, from low to high levels of the output variable, i.e., what needs to be predicted. If the output variable is already discrete, it can be directly addressed by ordinal classification algorithms, and if the data are a time series, we can use models that are able to handle the temporal relations of the patterns such as recurrent neural networks with the ordinal classifiers. If, in contrast, the output is continuous, it can be discretized into ordinal classes (simplified), so that the order information of the patterns is still present in the classes, and then ordinal techniques can be applied to learn how to predict it.

One of the fields where predictions are very useful is medicine, where medical decisions could be supported by evidence of future results, helping the experts in the decision-making. In our case, we developed an ordinal model that was able to predict which recipient was the best match (longest survival time) for a liver, which could help the medical expert to make the best possible decision in seconds. However, while machine learning (ML) is being introduced into the medical area little by little, the challenge will be to make the patients (not the doctors) believe in the ML decisions, as doctors already understand them, but there are strong social and ethical barriers to surpass before ML decisions can truly take an important role in this field.

In contrast, renewable energy is a field where ML importance is growing exponentially, as it can optimize the energy generation depending on many different factors not only taking into account the meteorology but also predicting the price of the energy and selecting when to produce or when to stop producing it. Additionally, predictive maintenance is truly helpful, as renewable energy generators are generally not monitored exhaustively; this helps to plan maintenance and avoid the unplanned urgent maintenance,

which are costly. In our case, we focused on wind power generation, where one of the most harmful and unpredictable events is known as WPREs, causing important economic losses to companies. The model we developed is able to predict with nearly an 80% of accuracy these WPREs 6 hours in advance, which means there is enough time for the operators to take all the necessary actions to prevent economic losses.

Changing topics, it is important to note that there are many prediction problems whose data form a time series, e.g., any data recorded by sensors. To achieve good results in addressing time series prediction, one needs to account for the temporal relation of the data, which is truly valuable in obtaining the best prediction performance. This can be achieved with models that are specially designed for time series, such as recurrent neural networks, or adapting them by using techniques such as time windows to process the inputs. Finally, it is important to take into account that some processes that are usually applied in ML are not suitable for time series, so they need to be adapted in order to maintain the temporal relation of the data, which is truly valuable.

To conclude the discussion, ML is applicable to nearly any knowledge area, and with the immense amount of data that all the companies are currently collecting, it is necessary to use ML in order to predict and make the best decisions in all the problems that a professional can face. Additionally, the ordering of the classes is absolutely necessary, as most of the prediction problems work with continuous variables or with the severity of a variable from low to high (ordinal scale). Finally, professionals in ML tend to find a single task (the main task in the dataset) in a dataset and solve it without considering that in the same dataset, there are other tasks (secondary) that are highly related to that one and their learning process could help in solving the main task. In our case, we were able to achieve better results when learning the three prediction tasks in the same model than by learning them independently. Therefore, we would like to encourage researchers and data scientists to use the MTL approach.

As future work, several applications and methodologies will be developed. An imminent research line would include the prediction of optimal energy production in solar plants, including important business information, such as the future energy price. This will allow companies to optimize their energy production and reduce the operating time of the generators, also increasing their lifetime. Another source of future work could be found in searching for new information sources in order to optimize the prediction of WPREs.

Moreover, with the arrival of Industry 4.0 and the internet of things, factories are starting to apply ML techniques to optimize their production processes, e.g., applying predictive maintenance to the machines and robots of the most critical processes is a crucial task to solve if medium and large companies want to be competitive with worldwide companies in the near future. Here, some works can be developed under the area of ordinal

deep learning in order to early detect damages in robot parts or to enhance the building process of pieces reducing the movements performed by the machines.

In the area of medicine, achieving a decision support model that performs well to assist medical experts to choose the best donor-recipient match worldwide would be an interesting future study. This can be achieved using our current datasets together with datasets such as UNOS. This can also be helpful to achieve more robust models, as we would have more diverse data to train the models.

Finally, from the MTL approach, a promising line of research would be to try to generalize the model so that more wind farms could be included in the model with minimal refactorization of the model. A good starting point would be to study the transfer learning approach in order to keep the model knowledge and apply a minimum learning process to the model when adding a new wind farm to achieve a good performance.

References

- [1] R. Arroyo, J. J. Yebes, L. M. Bergasa, I. G. Daza, and J. Almazán. Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls. *Expert Systems with Applications*, 42(21):7991–8005, nov 2015.
- [2] S. Baccianella, A. Esuli, and F. Sebastiani. Evaluation measures for ordinal regression. In *Proc. of the Ninth Int. Conf. on Intelligent Systems Design and App. (ISDA)*, Pisa, Italy, 2009.
- [3] H.-B. Bao and J.-D. Cao. Projective synchronization of fractional-order memristor-based neural networks. *Neural Networks*, 63:1–9, mar 2015.
- [4] C. M. Bishop. Improving the generalization properties of radial basis function neural networks. *Neural Computation*, 3(4):579–581, 1991.
- [5] C. M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, Oxford, UK, 1996.
- [6] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Inf. Science and Statistics. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1st edition, 2006.
- [7] Y.-L. Boureau, J. Ponce, and Y. LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10*, pages 111–118, USA, 2010. Omnipress.
- [8] P. Branco, L. Torgo, and R. P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2):1–50, aug 2016.
- [9] A. Camurri, I. Lagerlöf, and G. Volpe. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1-2):213–225, jul 2003.
- [10] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

- [11] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.
- [12] D. Cireşan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2012.
- [13] J. Connor, R. Martin, and L. Atlas. Recurrent neural networks and robust time series prediction. *IEEE Transactions on Neural Networks*, 5(2):240–254, mar 1994.
- [14] M. Dorado-Moreno, P. A. Gutiérrez, and C. Hervás-Martínez. Ordinal classification using hybrid artificial neural networks with projection and kernel basis functions. In *7th International Conference on Hybrid Artificial Intelligence Systems (HAIS2012)*, page 319–330, 2012.
- [15] M. Dorado-Moreno, M. Pérez-Ortiz, P. A. Gutiérrez, R. Ciria, J. Briceño, and C. Hervás-Martínez. Dynamically weighted evolutionary ordinal neural network for solving an imbalanced liver transplantation problem. *Artificial Intelligence in Medicine*, 77:1–11, mar 2017.
- [16] M. Dorigo, V. Maniezzo, and A. Colorni. Ant system: optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 26(1):29–41, 1996.
- [17] O. M. Doyle, E. Westman, A. F. Marquand, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, S. Lovestone, S. C. R. Williams, and A. Simmons. Predicting progression of alzheimer’s disease using ordinal regression. *PLoS ONE*, 9(8):e105542, aug 2014.
- [18] M. Elhamod and M. D. Levine. Automated real-time detection of potentially suspicious behavior in public transport areas. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):688–699, jun 2013.
- [19] Y. Fan. Ordinal ranking for detecting mild cognitive impairment and alzheimer’s disease based on multimodal neuroimages and csf biomarkers. In T. Liu, D. Shen, L. Ibanez, and X. Tao, editors, *Proceedings of the First International Workshop on Multimodal Brain Image Analysis (MBIA2011)*, volume 7012 of *Lecture Notes in Computer Science*, pages 44–51. Springer Berlin Heidelberg, 2011.
- [20] S. Feng, N. Goodrich, J. Bragg-Gresham, D. Dykstra, J. Punch, M. DebRoy, S. Greenstein, and R. Merion. Characteristics associated with liver graft failure: The concept of a donor risk index. *American Journal of Transplantation*, 6(4):783–790, feb 2006.

- [21] I. Fette, N. Sadeh, and A. Tomasic. Learning to detect phishing emails. In *Proceedings of the 16th international conference on World Wide Web*. ACM Press, 2007.
- [22] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *Computer Journal*, 7:149–154, 1964.
- [23] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 42(4):463–484, 2012.
- [24] C. Gallego-Castillo, A. Cuerva-Tejero, and O. Lopez-Garcia. A review on the recent history of wind power ramp forecasting. *Renewable and Sustainable Energy Reviews*, 52:1148–1157, dec 2015.
- [25] D. Goldenholz. Liquid computing: A real effect, 2002.
- [26] C. Goumopoulos, B. O’Flynn, and A. Kameas. Automated zone-specific irrigation with wireless sensor/actuator network and adaptable decision support. *Computers and Electronics in Agriculture*, 105:20–33, jul 2014.
- [27] E. Guresen, G. Kayakutlu, and T. U. Daim. Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, 38(8):10389–10397, aug 2011.
- [28] P. Gutiérrez, C. Hervás, M. Carbonero, and J. Fernández. Combined projection and kernel basis functions for classification in evolutionary neural networks. *Neurocomputing*, 72(13-15):2731–2742, aug 2009.
- [29] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernandez-Navarro, and C. Hervás-Martínez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, January 2016. JCR(2016): 3.438 Position: 21/146 (Q1) Category: COMPUTER SCIENCE, INFORMATION SYSTEMS.
- [30] P. A. Gutiérrez, P. Tiño, and C. Hervás-Martínez. Ordinal regression neural networks based on concentric hyperspheres. *Neural Networks*, 59:51–60, nov 2014.
- [31] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, may 2017.
- [32] F. E. Harrell. *Regression Modeling Strategies*. Springer New York, 2001.

- [33] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, jan 2017.
- [34] G. Hinton and T. Sejnowski. *Unsupervised Learning: Foundations of Neural Computation*. Computational Neuroscience. Mit Press, 1999.
- [35] S. Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 06(02):107–116, apr 1998.
- [36] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.
- [37] H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks. *GMD Report*, 148:1–43, 2001.
- [38] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, Oct. 2002.
- [39] P. Kamath and W. Kim. The Model for End-stage Liver Disease (MELD). *Hepatology*, 45(3):797–805, 2007.
- [40] J. Kennedy and R. Eberhart. Particle swarm optimization. In D. Touretzky, editor, *Proceedings of the IEEE International Conference on Neural Networks*, volume 4, pages 1942–1948, Perth, WA, Australia, 1995.
- [41] S. Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica (Ljubljana)*, 31(3):249–268, 2007. cited By 874.
- [42] B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, apr 2016.
- [43] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2008.
- [44] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015.
- [45] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, jun 2015.

- [46] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, dec 2017.
- [47] X. Liu, J. Wu, and Z. Zhou. Exploratory under-sampling for class-imbalance learning. In *Proceedings of the Sixth International Conference on Data Mining*, 2006.
- [48] Y. Liu, Y. Liu, S. Zhong, and K. C. Chan. Semi-supervised manifold ordinal regression for image ranking. In *Proceedings of the 19th ACM international conference on Multimedia 11*. ACM Press, 2011.
- [49] D. López-Fernández. Contributions to gait recognition using multiple-views. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, 15(2):22, nov 2016.
- [50] M. Lukoševičius and H. Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, aug 2009.
- [51] I. Maqsood, M. Khan, and A. Abraham. An ensemble of neural networks for weather forecasting. *Neural Computing and Applications*, 13(2), may 2004.
- [52] A. C. Martínez-Estudillo, F. J. Martínez-Estudillo, C. Hervás-Martínez, and N. García. Evolutionary product unit based neural networks for regression. *Neural Networks*, 19(4):477–486, 2006.
- [53] V. Mazzaferro, E. Regalia, R. Doci, S. Andreola, A. Pulvirenti, F. Bozzetti, F. Montalto, M. Ammatuna, A. Morabito, and L. Gennari. Liver transplantation for the treatment of small hepatocellular carcinomas in patients with cirrhosis. *New England Journal of Medicine*, 334(11):693–700, mar 1996.
- [54] P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society*, 42(2):109–142, 1980.
- [55] O. A. Montesinos-López, A. Montesinos-López, J. Crossa, J. Burgueño, and K. Es-kridge. Genomic-enabled prediction of ordinal data with bayesian logistic ordinal regression.
- [56] V. Nair and G. Hinton. Rectified linear units improve restricted boltzmann machines. pages 807–814, 2010. cited By 2748.
- [57] T. Ouyang, X. Zha, and L. Qin. A survey of wind power ramp forecasting. *Energy and Power Engineering*, 05(04):368–372, 2013.
- [58] C. Pena-Reyes and M. Sipper. Designing breast cancer diagnostic systems via a hybrid fuzzy-genetic methodology. In *FUZZ-IEEE. 1999 IEEE International Fuzzy Systems. Conference Proceedings (Cat. No.99CH36315)*. IEEE, 1999.

- [59] M. Pérez-Ortiz, P. A. Gutiérrez, C. Hervás-Martínez, and X. Yao. Graph-based approaches for over-sampling in the context of ordinal regression. *IEEE Transactions on Knowledge and Data Engineering*, 27(5):1233–1245, May 2015.
- [60] A. Rodan and P. Tiño. Minimum complexity echo state network. *Neural Networks, IEEE Transactionss on*, 22(1):131–144, jan. 2011.
- [61] N. Rout, D. Mishra, and M. K. Mallick. Handling imbalanced data: A survey. In *Advances in Intelligent Systems and Computing*, pages 431–443. Springer Singapore, dec 2017.
- [62] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multcategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, sep 2004.
- [63] S. K. Tanbeer, C. F. Ahmed, B.-S. Jeong, and Y.-K. Lee. Sliding window-based frequent pattern mining over data streams. *Information Sciences*, 179(22):3843–3865, nov 2009.
- [64] J. van Lint, S. Hoogendoorn, and H. van Zuylen. Accurate freeway travel time prediction with state-space neural networks under missing data. *Transportation Research Part C: Emerging Technologies*, 13(5-6):347–369, oct 2005.
- [65] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010. cited By 1948.
- [66] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias. Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach. *Transportation Research Part C: Emerging Technologies*, 13(3):211–234, jun 2005.
- [67] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518, dec 2003.
- [68] N. Wang and D.-Y. Yeung. Learning a deep compact image representation for visual tracking. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 809–817. Curran Associates, Inc., 2013.
- [69] C. Yang, Z. Li, R. Cui, and B. Xu. Neural network-based motion control of an underactuated wheeled inverted pendulum model. *IEEE Transactions on Neural Networks and Learning Systems*, 25(11):2004–2016, nov 2014.

- [70] X. Yao. A review of evolutionary artificial neural networks. *International Journal of Intelligent Systems*, 4:203–222, 1993.
- [71] D. Zhang and D. Shen. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer’s disease. *NeuroImage*, 59(2):895–907, jan 2012.
- [72] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via multi-task sparse learning. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2012.
- [73] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Computer Vision – ECCV 2014*, pages 94–108. Springer International Publishing, 2014.

