Graphical representation of quantitative datasets for medical visualisation

Luis Gonzalez-Abril[1], Mariano González[2], Juan Antonio Ortega[3], Cecilio Angulo[4]

Abstract. A methodology to show understable graphical information to health professionals is introduced in this paper. Datasets containing continuous and discrete features are considered. Our algorithmic approach allows the representation of knowledge where each instance (usually, patient information) is translated to an image in order to provide a quickly informative view of its main features. An example on a dataset of blood tests is considered. Moreover, a discussion about the implications of this proposal on health knowledge representation is analysed and future improvements are displayed.

1 Introduction

The generation of information and knowledge valid for users and her/his environment is a key point in the field of health. Generated knowledge will allow immediate advantage, bringing explainable information to the patient and the rest of the actors involved around: family, health professionals, administration and society. At this point, it is worth to note that during the last few years progress has been experienced in applying new information analysis technologies and their introduction in different aspects of society, including, of course, medicine and health. Most of these technologies are coming from the social media domain due to the huge volume of information to deal. Medical information from a patient can imply also a large volume of data (not ever), however the nature and purpose of this information seems more aligned with dealing critical industrial processes and devices that social media [1]. How data should be structured, stored and analysed the information, with the aim of extracting knowledge form it, is highly correlated with industrial issues about data protection, safety, regulations, expert knowledge, and so on [9, 2].

In recent years, there has been a huge increase in the number of studies using deep learning techniques in computer vision applied to medical images analysis, stories such as MRI, X-ray, or ultrasound. So, it is estimated that there were over 400 papers published in 2016 and 2017 in major medical imaging related conference venues and journals [7]. Deep learning is a novel computer vision technique starting in 2012 [5] with a very explosive adoption in all the vision areas as in medical imaging community due to its demonstrated potential to complement image interpretation and augment image representation and classification. Furthermore, it is well-known that deep learning techniques are able to find out complicated structures in high-dimensional data, which eventually reaps benefits in many areas of society.

In visual field, the records of image classification have been broken in the ImageNet Challenge 2012 by using deep convolutional neural network (CNN) [5]. Additionally, deep learning has a significant impact on other visual problems, such as face detection, image segmentation, general object detection, and optical character recognition. Deep learning can also be used for speech recognition, natural language understanding, and many other domains, such as recommendation systems, web content filtering, disease prediction, drug discovery, and genomics [6]. Accuracy in image classification attained by the convolutional neural networks is highly influenced by the way in that these images are provided to the CNN. Thus, accuracy can be improved depending on the format of the provided images. In a similar way, it should be also taken into account the image format when researchers provide images to the medical staff. Hence, the focus in our work is about determine a graphical format in the presentation of the medical information such that the medical staff can better understand it, implicitly considering that our format are images.

There exist a large number of information visualisation techniques which have been developed over the last decade to support the exploration of large data sets [4]. Among them, in [10] is studied the problem of visualising in a low-dimensional (2D or 3D) space. In this paper, the initial goal is to visualise medical images. However, the main aim of the research is more ambitious. It is sought, given some data, to transcribe these graphically in order to be able to make use of the powerful tools provided by deep learning of this type of information in order to carry out a problem of classification and / or diagnosis of diseases.

Let us also indicate that there are data visualisation platforms as D3 (http://d3js.org/), VisPy (http://vispy.org/) and Kibana (https://www.elastic.co/fr/kibana). Nevertheless, the aim of the presented approach is to benefit the medical knowledge in order to obtain a graphical representation more cognitive by the health personal.

The approach presented in this paper was initially defined in [8], where the UCI repository's database "thyroid" is considered. This dataset contains user features that are relevant for thyroid illnesses detection. This database from KEEL has been released to identify if a given patient is normal, suffers from hyperthyroidism or hypothyroidism. Data contain both, continuous and binary features, and it is displayed in the form of an image (see for instance Figure1). Nevertheless, this graphical procedure transforming data into images was only defined as an intermediate step to use on them a Generative Adversarial Network (GAN) [3].

The main aim in this paper, however, is to consider from a dataset, with continuous and discrete features, a consistent way to show the

[1] Applied Economics I, FCEE, Universidad de Sevilla, Av. Ramón y Cajal 1, 41016 Sevilla, Spain. luisgon@us.es
[2] Computer Systems and Languages, Universidad de Sevilla, Av. Reina Mercedes, 41012 Seville, Spain. mariano@us.es
[3] Computer Systems and Languages, Universidad de Sevilla, Av. Reina Mercedes, 41012 Seville, Spain. jortega@us.es
[4] IDEAI-UPC, Universitat Politècnica de Catalunya, Jordi Girona 29, 08034 Barcelona, Spain. cecilio.angulo@upc.edu

Figure 1. Image representing an instance from the UCI repository's database Thyroid [8]. Dark purple colour is represent null values.

information as an image in order to facilitate the work to the medical staff. The rest of the paper is structured as follows. Section 2 introduces the methodology used. Section 3 describes an implementation with a dataset on analytical results of blood to several patients. Finally, conclusions and future work are drawn.

2 Methodology

A laboratory analysis performed on a blood sample that is usually extracted from a vein in the arm using a hypodermic needle, or via finger prick. Multiple tests for specific blood components often grouped together into one test panel called a blood panel or blood work. Blood tests are often used in healthcare to determine physiological and biochemical states, such as disease, mineral content, pharmaceutical drug effectiveness, and organ function.

Blood tests results should always be interpreted using the ranges provided by the laboratory that performed the test. Example ranges used in this study are shown in Table 1. As it can be observed, al-

Table 1. Examples of ranges of several test

Test	Low	High	Unit
Red blood cells	4.30	5.90	10^6 cells$/mm^3$
Haemoglobin	13.00	17.00	g/dL
Hematocrit	39.00	50.00	%
Platelets	150.00	450.00	$10^3/microL$
White blood cells	3800.00	9800.00	cells$/mm^3$
Eosinophils	0.05	0.50	$10^9/L$
Basophils	0.01	0.10	$10^9/L$
Lymphocytes	1.50	4.00	$10^9/L$
Monocytes	0.10	0.50	$10^9/L$
Neutrophils	3.00	5.00	$10^9/L$

though ranges among different laboratories are similar, medical staff should check them in order to analyse whether the values are normal or not. Hence, a normalisation process would be a good idea to overcome this problem.

By taking into account that the main aim is to obtain a graphical representation from blood test, the range of the normalisation for each test instance is from 1 to 255, since the colour map of rainbow will be used (see Figure 2). Furthermore, value 0 is representing NaN's in the tests, that is, those tests that have not been realised but are in a template previously considered for the health professionals.

The rainbow colour map is used since it is a perceptual uniform colour map, which is very suitable for the representation of scientific

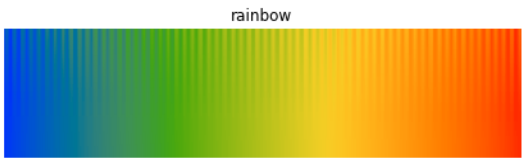

Figure 2. Colour map of rainbow.

data[5]. This is what the rainbow colour map looks like, with a central part in the green-yellow range and the extremes in blue and red, thus distinguishing the lower and upper ends, which is interesting when it is important to distinguish when a value is below or above the desired values, and not simply away from them.

The normalisation process carried out is as follows: The range $[1\ ,\ 255]$ of the RGB scale splits in three ranges: $[1\ ,\ 64]$, $[64\ ,\ 191]$ and $[191\ ,\ 255]$ because these ranges approximately contain to 25%, 50%, and 25% of the RGB scale values. Hence, in each test, the value 64 in RGB scale is assigned to the Low value and the value 191 in RGB scale is assigned to the High value, respectively.

Given a x value of a test with low and high values, then the normalised value of x, denoted by $vn(x)$, is obtained from

$$vns(x) = E\left[\frac{x - Loss}{High - Low} \times (191 - 64) + 64\right]$$

where $E[\cdot]$ denoted integer part, as follows,

$$vn(x) = \begin{cases} 1 & \text{if } vns(x) <= 1 \\ 255 & \text{if } vns(x) >= 255 \\ vns(x) & \text{otherwise} \end{cases}$$

Let us indicate that the values must be integers in order to be used in the RGB scale. The continuous version of $vn(x)$ is depicted in Figure 3. Moreover, an intuitive interpretation of the normalisation process can be seen in Figure 4.

When a Blood Test is carried out, different features of the blood are analyzed. Let $\mathbf{x}$ a vector of size n with the normalised values of each one of the n test. Hence, an array $N \times N$ is considered such that $N^2 \geq n$ in order to show an image of the blood test. Elements in the array that exceed n are encoded as a value of 0.

However, this value 0, which denotes the absence of data, should have an associated colour in the palette, like the other values. This colour should be distinguished from the rest of the values and to happen the white colour has been assigned to it.

Furthermore, many arrangements can be done in a $N \times N$ array from a vector of size n. In this work, the arrangement used is the next one,

$$(x_1, x_2, \cdots, x_n) \Rightarrow \begin{pmatrix} x_1 & x_2 & \cdots & \cdot & \cdots & x_N \\ x_{N+1} & x_{N+2} & \cdots & \cdot & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdots & x_n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 \end{pmatrix}$$

[5] https://colorcet.holoviz.org/

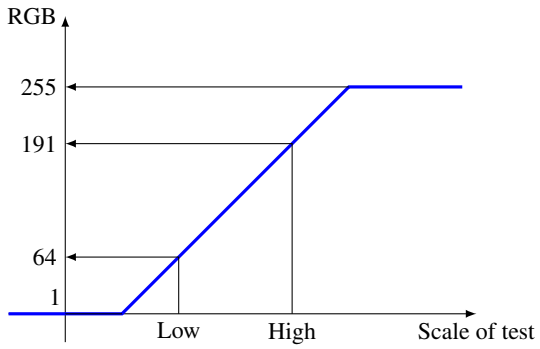

Figure 3. Graphical representation of the continuous version of $vn(x)$.

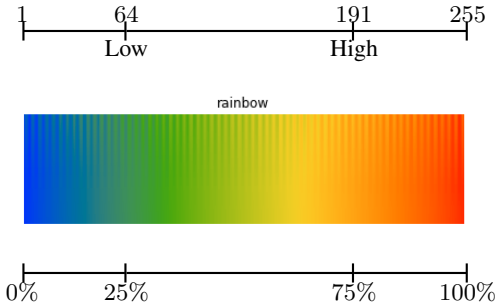

Figure 4. Graphical representation of the normalisation.

3 Implementation

Let us consider the blood test given in Table 2.

In this case, $n = 20$ and $N = 5$ is considered since it is the smallest integer such that $n < N^2$ ($4^2 = 16 < 20 < 25 = 5^2$). The normalisation and the arrangement on the blood test gives the next array,

$$\begin{pmatrix} 119 & 88 & 95 & 100 & 116 \\ 132 & 178 & 80 & 116 & 76 \\ 92 & 75 & 107 & 82 & 189 \\ 106 & 120 & 82 & 86 & 77 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

which provides Figure 5 (left).

A medical doctor trained in this methodology can see that all test are ok, that is, the values are between the low and high values of each feature.

Let us consider a blood test such that its graphical representation is shown in Figure 5 (right).

In this case the doctor can seen that the analytic is not ok because there are six features with anomalous values. Four are very high (features 8, 13, 17 and 18, that is, Platelets, Basophils, Monocytes and Monocytes2, respectively) since the colour is red. Two are very low (features 19 and 20, that is, Neutrophils and Neutrophils2, respectively) since the colour is purple.

An advantage of this approach for the doctors is that it is very easy to study the evolution of a patient. Figure 6 shows five blood

[ht]

Table 2. Examples of a blood test (Patient 1).

Test	Value	Low	High
Red blood cells (in 10^5)	49.0	42.0	58.0
Haemoglobin	14.3	13.5	17.5
Hematocrit	42.3	39.0	52.0
Vol. Erythrocyte Medium	86.33	81.0	99.0
Hemoglobin Erythroc.media	29.18	25.0	35.0
Conc.Hemoglobin Erythroc.media	33.81	30.0	37.0
C.V. Volum. Erythrocyte	15.1	11.5	15.5
Platelets (in 10^4)	16.3	12.5	40.0
Platelet Volume MPW Medium	9.1	7.0	12.0
Leukocytes (in 10^2)	43.0	35.	110.0
Eosinophils	1.4	0.0	6.0
Eosinophils2	60.2	0.0	600.0
Basophils	0.7	0.0	, 2.0
Basophils2	30.1	0.0	200.0
Lymphocytes	41.8	, 17.0	42.0
Lymphocytes2	1797.4	600.0	4100.0
Monocytes	5.5	1.0	11.0
Monocytes2	236.5	100.0	1000.0
Neutrophils	50.6	45.0	75.0
Neutrophils2	2175.8	1500.0	7500.0

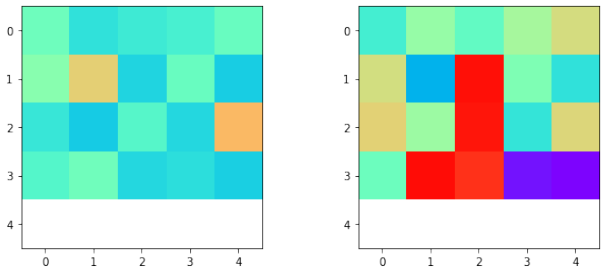

Figure 5. Proposed Graphical to the Patient 1 (left) and Patient X (right).

tests performed on the same patient over time. An aspect relevant in these blood tests is relative to the 'C.V. Volum. Erythrocyte'. The value is high in the first four tests and in the fifth it seems that it has been corrected. Furthermore, it can be seen that in the last blood test, the Neutrophils2 is null.

4 Conclusion and Future works

A way to graphically represent a blood test has been proposed in this paper. This representation uses the RGB colour scale and it is represented in a square matrix of a certain order depending on the blood characteristics analysed.

This is an initial proposal and there are still many details to analyse. For example: What is the best way to put, according to medical personnel, the analytical values on a graph? There are many types of analytic, therefore, it would be necessary to analyse, together with the medical personnel, which would be the most appropriate to carry out this graphic conversion? Will this approach really help doctors?

Furthermore, from the figures, we think that qualitative reasoning based on colours could be carried out to provide support for the doctor's decision.

Figure 6. Blood test evolution of a patient.

Acknowledgements

This research has been partially supported by the EDITH Research Project (PGC2018-102145-B-C21,C22 (AEI/FEDER, UE)), funded by the Spanish Ministry of Science, Innovation and Universities.

REFERENCES

[1] Cecilio Angulo, Luis González Abril, Cristóbal Raya, and Juan Antonio Ortega, 'A proposal to evolving towards digital twins in healthcare', in *Bioinformatics and Biomedical Engineering - 8th International Work-Conference, IWBBIO 2020, Granada, Spain, May 6-8, 2020, Proceedings*, eds., Ignacio Rojas, Olga Valenzuela, Fernando Rojas, Luis Javier Herrera, and Francisco M. Ortuño Guzman, volume 12108 of *Lecture Notes in Computer Science*, pp. 418–426. Springer, (2020).

[2] Cecilio Angulo, Juan Antonio Ortega, and Luis González Abril, 'Towards a healthcare digital twin', in *Artificial Intelligence Research and Development - Proceedings of the 22nd International Conference of the Catalan Association for Artificial Intelligence, CCIA 2019, Mallorca, Spain, 23-25 October 2019*, eds., Jordi Sabater-Mir, Vicenç Torra, Isabel Aguiló, and Manuel González Hidalgo, volume 319 of *Frontiers in Artificial Intelligence and Applications*, pp. 312–315. IOS Press, (2019).

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, 'Generative adversarial nets', in *Advances in Neural Information Processing Systems 27*, eds., Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, 2672–2680, Curran Associates, Inc., (2014).

[4] D. A. Keim, 'Information visualization and visual data mining', in *IEEE Transactions on Visualization and Computer Graphics*, , vol. 8, no. 1, pp. 1-8, (2002).

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, 'Imagenet classification with deep convolutional neural networks', in *Advances in Neural Information Processing Systems 25*, eds., F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 1097–1105, Curran Associates, Inc., (2012).

[6] Hinton G. LeCun Y., Bengio Y., 'Review: Deep learning', *Nature*, (2015).

[7] Geert J. S. Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez, 'A survey on deep learning in medical image analysis', *CoRR*, **abs/1702.05747**, (2017).

[8] Esteban Piacentino, 'Generative adversarial network based machine for fake data generation', Technical report, Master Thesis. Universitat Polièc nica de Catalunya, (09 2019).

[9] Raghupathi V. Raghupathi, W., 'Big data analytics in healthcare: Promise and potential.', *Health Information Science and Systems*, (2014).

[10] Jian Tang, Jingzhou Liu, Ming Zhang, Qiaozhu Mei, 'Visualizing large-scale and high-dimensional data', in *Proceedings of the 25th International Conference on World Wide Web*, pp. 287–297. (2016).