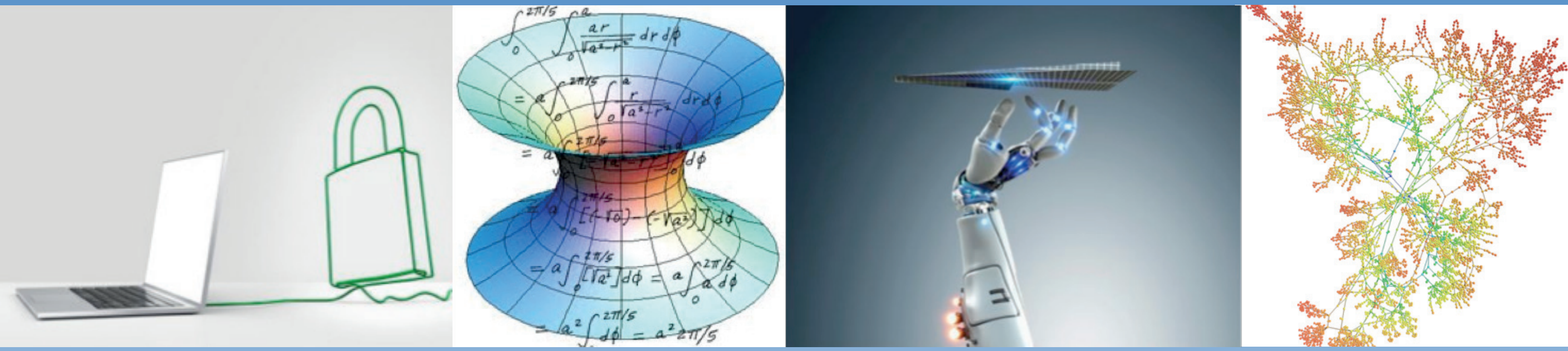
4TH URV DOCTORAL WORKSHOP IN COMPUTER SCIENCE AND MATHEMATICS

Edited by Alexandre Viejo & David Sánchez



UNIVERSITAT ROVIRA i VIRGILI

[DΣIM] Departament d'Enginyeria
Informàtica i Matemàtiques

ETSE Escola Tècnica Superior
d'Enginyeria
Universitat Rovira i Virgili

Title: 4th URV Doctoral Workshop in Computer Science and Mathematics
Editors: Alexandre Viejo & David Sánchez
November 2017

Universitat Rovira i Virgili
C/ de l'Escorxador, s/n
43003 – Tarragona
Catalunya (Spain)

ISBN: 978-84-8424-631-2

Preface

This book of proceedings gathers the contributions presented at the *4th URV Doctoral Workshop in Computer Science and Mathematics.* After the successful previous editions in 2014, 2015 and 2016 the fourth edition was held in Tarragona (Catalonia, Spain) on November 16th, 2017. It was jointly organized by the Security and Privacy research group (CRISES) and the Doctoral Program on Computer Science and Mathematics of Security of Universitat Rovira i Virgili (URV), which celebrates its 10th anniversary this year. The main aim of this workshop is to promote the dissemination of the ideas, methods and results that are developed in the Doctoral Thesis of the students of this doctorate program, and to promote the knowledge sharing, collaboration and discussion between their respective research groups.

The workshop had two invited talks and fourteen oral presentations. The first invited talk was given Prof. Josep Domingo-Ferrer, who is the Academic Director of the Serra Húnter Programme (SHP). The SHP is part of the new academic staff model that the Government of Catalonia is promoting to reinforce the internationalization of the Catalan universities, with the ultimate goal of consolidating Catalonia as the knowledge hub of Southern Europe; the talk described the selection process and the career expectations of Serra Húnter faculty members. The second invited talk was given by Dr. Sara Hajian, a former Ph.D. student of the Doctoral Program on Computer Science and Mathematics of Security of the URV, who is currently a research scientist at the Eurecat Technology Center in Catalonia. Her talk discussed the algorithmic bias that may happen in decision making based on Big Data, and the technical solutions to detect and prevent algorithmic discrimination.

In this book, the reader will find the contributions of the fourteen Ph.D. students that presented their works in the Workshop. Each chapter presents the research topic of each student, the goals of the Doctoral Thesis and some preliminary results. Contributions were framed in a variety of research lines, which include security and privacy in computer systems, artificial intelligence, medical informatics, hardware architectures and mathematics. All contributions present innovative proposals, methods or applications, with the aim of opening new and strategic research lines. The editors and organizers invite you to contact the authors for more detailed explanations and encourage you to send them your suggestions and comments, which will certainly help them in their PhD theses.

The members of the organizing committee were Dr. Alexandre Viejo, Dr. David Sánchez, Dr. Aïda Valls (Coordinator of the Ph.D. program), Mr. Jesús Manjón and Mrs. Olga Segú.

We could not finish without first thanking the invited speakers for such interesting talks. Second, we thank all the participants and, especially, the students that presented their work in this DCSM workshop. Finally, we also want to thank Universitat Rovira i Virgili (URV), the Departament d'Enginyeria Informàtica i Matemàtiques (DEIM), and the Escola Tècnica Superior d'Enginyeria (ETSE) for their support.

Alexandre Viejo and David Sánchez (Editors)

Contents

Universal measures of disclosure risk and information loss in individual data anonymization

Nicolas Ruiz [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
nicolas.ruiz@oecd.org

Data on individual subjects are increasingly collected and exchanged. By their nature, they provide a rich amount of information that can inform statistical and policy analysis in a meaningful way. However, due to the legal obligations surrounding these data, this wealth of information is often not fully exploited in order to protect the confidentiality of respondents. In fact, such requirements shape the dissemination policy of micro data at national and international levels. The issue is how to ensure a sufficient level of data protection to meet releasers' concerns in terms of legal and ethical requirements, while offering to users a reasonable richness of information. Moreover, over the last decade the role of micro data has changed from being the preserve of National Statistical Offices and government departments to being a vital tool for a wide range of analysts trying to understand both social and economic phenomena. As a result, more parties, often very heterogeneous in their privacy and information requirements, are now involved in micro data transactions. This has opened a new range of questions and pressing needs about the privacy/information trade-off and the quest for best practices that can be both useful to users but also respectful of respondents' privacy.

Statistical disclosure control (SDC) research has a rich history in addressing those issues, by providing the analytical apparatus through which the privacy/information trade-off can be assessed and implemented. Over the years, it has burgeoned in many directions. But streaming from the large variety of practical cases that can occur in micro data exchange is the diversity of techniques available for data anonymization. Such diversity is undoubtedly useful but has however one major drawback: a lack of agreement and clarity on the appropriate choice of tools in a given context, and as a consequence a lack of general view (or at best an incomplete one) across the relative performances of the techniques available.

Moreover, a variety of parties is involved in micro data exchange. Indeed, it is natural to assert that across each party different sensitivities to privacy

[*] PhD advisor: Josep Domingo-Ferrer

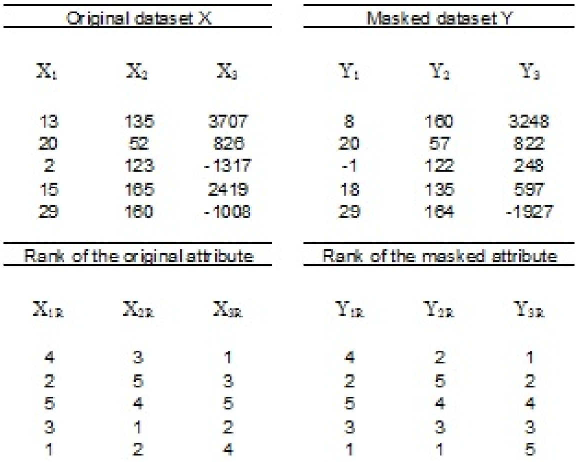
and information prevail. Some may place greater emphasis on the preservation of privacy, e.g. typically the data releasers, while others are relatively more concerned by the extent to which information is preserved, e.g. typically the researchers. Additionally, these sensitivities can differ also within groups, e.g. one researcher can have a low sensitivity to information loss and consider a release better than no release at all, while another could simply disregard the data above a certain threshold of loss set according to his intended use of the data.

A step toward the resolution of such limitations has been recently proposed ([1]), by establishing that any micro data masking method can be viewed as functionally equivalent to a permutation of the original data plus eventually a small noise addition. This insight, called the permutation paradigm, unambiguously establishes a common ground upon which any masking method can be gauged. It is independent of the underlying parameters of the masking mechanism and the characteristics of the data. Moreover, it presents the advantage of being meaningful and easy to grasp and implement, as the only necessary and sufficient information for the comparative evaluation of some methods, being under different parametrizations and/or different data sets, is a distribution of permutation distances. Thus, the permutation paradigm is also a tremendous simplifier for data anonymization.

While this paradigm is not considered by its author as a new anonymization method per se ([1]), it offers the potential to re-interpret all the techniques available through the same lens. It remains however to develop a set of appropriate measures of disclosure risk and information loss based on permutation distances. This is the objective of this contribution.

To recall the permutation paradigm, we use a simple toy example which consists (without loss of generality) of five records and three attributes $X = (X_1, X_2, X_3)$ generated by sampling $N(10, 102)$, $N(100, 402)$ and $N(1000, 20002)$ distributions, respectively. Noise is then added to obtain $Y = (Y_1, Y_2, Y_3)$, the three masked version of the attributes, from $N(0, 52)$, $N(0, 202)$ and $N(0, 10002)$ distributions, respectively. One can see that the masking procedure generates a permutation of the records of the original data (Figure 1).

Now, as long as the attributes' values of a dataset can be ranked, which is obvious in the case of numerical and categorical ordinal attributes, but also feasible in the case of nominal ones ([1]), it is always possible to derive a dataset Z that contains the attributes X_1, X_2 and X_3, but ordered according to the ranks of Y_1, Y_2 and Y_3, respectively, i.e. in Figure 1 re-ordering (X_1, X_2, X_3) according to (Y_{1R}, Y_{2R}, Y_{3R}). This can be done following a post-masking reverse procedure. Finally, the masked data Y can be fully reconstituted by adding small noises (E_1, E_2, E_3) (small in the sense that they cannot re-rank Z while they can still be large in absolute values) to each observation in each attribute (Figure 2).



Original dataset X			Masked dataset Y		
X_1	X_2	X_3	Y_1	Y_2	Y_3
13	135	3707	8	160	3248
20	52	826	20	57	822
2	123	-1317	-1	122	248
15	165	2419	18	135	597
29	160	-1008	29	164	-1927
Rank of the original attribute			**Rank of the masked attribute**		
X_{1R}	X_{2R}	X_{3R}	Y_{1R}	Y_{2R}	Y_{3R}
4	3	1	4	2	1
2	5	3	2	5	2
5	4	5	5	4	4
3	1	2	3	3	3
1	2	4	1	1	5

Fig. 1: An illustration of the permutation paradigm

By construction, Z has the same marginal distribution as X, which is an appealing property. Moreover, under a maximum-knowledge intruder model of disclosure risk evaluation, the small noise addition turns out to be irrelevant: re-identification via record linkage can only come from permutation, as by construction noise addition cannot alter ranks. Reverse mapping thus establishes permutation as the overarching principle of data anonymization, allowing the functioning of any method to be viewed as the outcome of a permutation of the original data, independently of how the method operates. This functional equivalence leads to the following proposition:

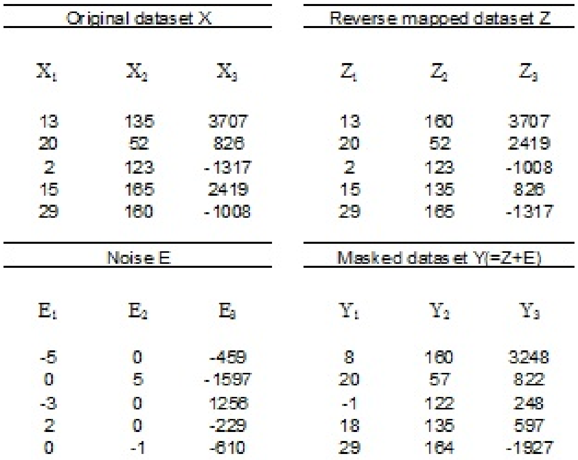
Proposition 1 *For a dataset $X_{(n,p)}$ with n records and p attributes $(X_1, \ldots, X_p)$, its anonymized version $Y_{(n,p)}$ can always be written, regardless of the anonymization methods used, as:*

$$Y_{(n,p)} = (P_1 X_1, \ldots, P_P X_p)_{(n,p)} + E_{(n,p)} \tag{1}$$

where $P_1, .., P_p$ is a set of p permutation matrices and $E_{(n,p)}$ is a matrix of small noises.

Proposition 1 is simply a restatement of the permutation paradigm. It has however several implications. The first is that it characterises permutation matrix as an encompassing tool for data anonymization: the analytical framework of anonymization mechanisms can in fact be viewed as functionally equivalent to a set of permutation matrices. Permutation matrices are meaningful, readable and practical in comparison to the sometimes quite complex analytical apparatus of some masking methods.

From each underlying permutation matrices, one can count, columns by columns of Pj, how many times the 1s have been moved, using the identity



Original dataset X			Reverse mapped dataset Z		
X_1	X_2	X_3	Z_1	Z_2	Z_3
13	135	3707	13	160	3707
20	52	826	20	52	2419
2	123	-1317	2	123	-1008
15	165	2419	15	135	826
29	160	-1008	29	165	-1317

Noise E			Masked dataset Y(=Z+E)		
E_1	E_2	E_3	Y_1	Y_2	Y_3
-5	0	-459	8	160	3248
0	5	-1597	20	57	822
-3	0	1256	-1	122	248
2	0	-229	18	135	597
0	-1	-610	29	164	-1927

Fig. 2: Equivalence in anonymization: postmasking reverse mapping plus noise addition

matrix as a starting point (which is a particular case of a permutation matrix with no permutation applied), then assigning a negative (resp. positive) sign if the 1 has been moved up (resp. down) and compute ranks displacement vectors (where each zero is replaced by an epsilon value). From the running example above, one gets:

$$r_1 = \begin{pmatrix} \epsilon \\ \epsilon \\ \epsilon \\ \epsilon \\ \epsilon \end{pmatrix} r_2 = \begin{pmatrix} \epsilon \\ \epsilon \\ \epsilon \\ 1 \\ -4 \end{pmatrix} r_3 = \begin{pmatrix} \epsilon \\ 2 \\ 2 \\ -2 \\ -2 \end{pmatrix}$$

Now, r_j has to be evaluated in some way for assessing disclosure risk based on permutation distances. Bearing in mind that different user can have different views about disclosure risk (and thus about permutation distances), the following proposition establishes a measure of disclosure risk sensitive to different aversions, with an adjustable degree of focus on permutation distances:

Proposition 2 *For any attribute $j = 1, \ldots, p$ of $Y_{(n,p)}$, a quantitative measure of disclosure risk in the permutation paradigm is given by:*

$$D_j(\alpha) = \left[\frac{1}{n} \sum_{i=1}^{n} |r_{j(i)}|^{\alpha} \right]^{1/\alpha} \textit{ for } \alpha \leq 1 \wedge \alpha \neq 0$$
$$\textit{and } D_j(\alpha) = \prod_{i=1}^{n} |r_{j(i)}|^{1/n} \textit{ for } \alpha = 0, \tag{2}$$

where $r_{j(i)}$ denotes the elements of r_j and α the parameter of aversion to disclosure risk.

$D_j(\alpha)$ makes use of a power mean[2]([3]) for the aggregation of the components of r_j, with the parameter α substantiating the notion of aversion to disclosure risk. The arithmetic mean becomes a special case ($\alpha = 1$ of $D_j(\alpha)$, which forms a natural starting point by computing the average level of permutation distances. In that case, all distances are given the same weight and there is a one-to-one substitution between them, e.g. two records permuted two ranks are equivalent to one record permuted four ranks. From this benchmark, the more α decreases, the more weight is given to the smallest permutation distances[3]. In fact, the more α approaches $-\infty$, the more $D_j(\alpha)$ converges towards the smallest permutation distance in r_j[4]. As a result, for a given r_j and, we have $D_j(\alpha') \leq D_j(\alpha)$: the lower is α, the stronger is the aversion to disclosure risk.

Now, information loss can be assessed through a similar but also general approach, by considering the degree of similarity between the permutations that took place for the two attributes and allowing different weights for different *relative* distances. To do so, it can be observed that a vector $\Delta(r_k)$ of differences between the vectors r_j and $r_{j'}$ is a vector of dissimilarity between the anonymization procedures that have been applied to the couple of attributes $k = (j, j')$ (with $j \neq j'$). When each of the components of $\Delta(r_k)$ are equal to zero, j and j' having been permuted the same way; the permutation matrices applied to them are identical, despite the fact that the anonymization methods used can be different in practice. There is no loss of information as the joint distribution of j and j' is preserved. But when $\Delta(r_k)$ has some non-zero elements information has been modified. This leads to the following proposition:

Proposition 3 *For two attribute j and j' of $Y_(n, p)$, a quantitative measure of information loss in the permutation paradigm is given by:*

$$I_j(\theta) = \left[\frac{1}{n}\sum_{i=1}^{n} |\Delta r_{j(i)}|^{\theta}\right]^{1/\theta} \quad \textit{for } \theta \geq 1, \tag{3}$$

where $\Delta r_{j(i)}$ denotes the elements of $\Delta(r_k)$ and θ the parameter of aversion to information loss.

The measure $I_j(\theta)$ bears strong analytical similarities with $D_j(\alpha)$, but while the latter is concerned about average or small permutation distances across records for a given attribute, the former considers average or large relative permutation distances between two attributes across records.

[2] In linear algebra power mean is also the formula for the computation of p-norms.

[3] $D_j(0)$ is the geometric mean and $D_j(-1)$ the harmonic mean.

[4] The limit case $D_j(\infty)$ is strictly equal to the shortest permutation distance in r_j.

The measures of Proposition 2 and 3 establish some universal measures of disclosure risk and information losses. They are universal in the sense that they can be applied on any dataset and any method, but also because they can account for the variety of preferences that occur in micro data transaction.

Acknowledgement. The author wants to thank Josep Domingo-Ferrer and Krish Muralidhar for useful discussions on a preliminary version of this paper.

References

[1] J. Domingo-Ferrer and K. Muralidhar. New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. *Information Sciences*, Vol. 337, pp. 11-24, Apr 2016.

[2] J. Domingo-Ferrer, D. Sánchez and G. Rufian-Torrell. Anonymization of nominal data based on semantic marginality. *Information Sciences*, Vol. 242, pp. 35-48, May 2013.

[3] G. H. Hardy, J. E. Littlewood and G. Polya. Inequalities Cambridge University Press, 1988.

Effects on blood flow of inferior vena cava filters

Josep M. López Besora [⋆]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
josep.m.lopez@urv.cat

1 Introduction

The use of inferior vena cava (IVC) filters has become frequent among thrombotic patients suffering from additional complications. The IVC filters, which are surgically placed within the cava vein, are designed to prevent pulmonar embolisms by holding back any circulating blood clot until it can be dissolved by medication. There exist several different filters with different shapes, whose different effects and actual effectiveness are to date known only in a heuristic way or through statistical analysis performed on pulmonary embolism patients; detailed dynamical studies are available only for one specific filter model [1]. The purpose of the current research is to perform a preliminary comparative study to assess how the presence of different filters affects the blood flow in the inferior vena cava.

2 Methods

We have obtained a real accurate 3D model of a portion of a patient's cava vein; also, we have generated 3D models for some different representative filter geometries (see 1). We have used these models to simulate how these filters affect the blood flow within the vein. In particular, computer fluid dynamics studies have been performed for the inferior vena cava model either with or without filters within it.

3 Results

In all of the cases investigated it was found that the presence of a filter produces a significant increase of both wall shear stress levels and pressure drop. Significant differences among calculated values of blood velocity and

[⋆] PhD advisor: Joan Herrero and Dolors Puigjaner

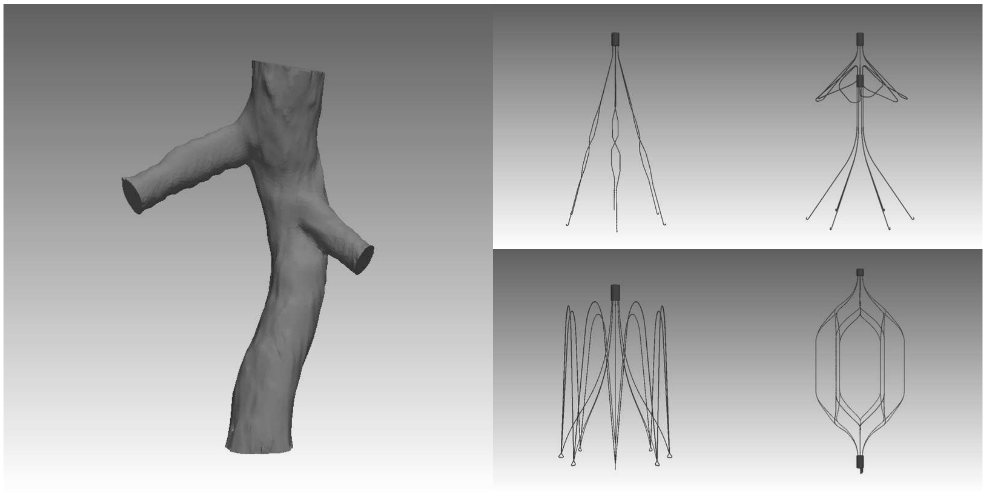


Fig. 1: Geometrical models of the portion of cava vein (left) and four different filter geometries (right). For the sake of clarity a different scale was used for vein and filters.

viscosity were found for different cases. Notwithstanding, different placements of the filter have minor effects on the blood dynamics.

4 Discussion

The present results suggest, in the context of a real patient, that strong WSS levels might provoke the detachment of the filter from the vein wall, which would result in fatal consequences.

Acknowledgement. This work was supported by grant 2016PFR-URV-B2-66 from Universitat Rovira i Virgili.

References

[1] Kenneth, I et al. *J Biomech Eng.* 136(8),081003, 2014.

GABLE: GAmification for a Better LifE

Hamed H. Aghdam [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
hamed.habibi@urv.cat

1 Introduction

Cerebral Palsy (CP) is one of the most frequent causes of disability in childhood, with an incidence of 2 per 1000 live births. In the EU, there are 1.3 million out of 15 million people with CP in the world. This neurological disorder affects body movement, balance and posture and often is accompanied by cognitive or sensory impairments like mental retardation, deafness and vision problems. The severity of these problems varies widely, from very mild and subtle to very profound. But what most is affected by this disease, from the youngest age, is the ability to play.

Play is probably the main activity for any child. Through play, children start to explore their world, and put the basis of their own system of values, which will be the cornerstone of their adult life. Play in children with CP becomes difficult due to the disability, and this in turn can affect child's self-esteem. In addition, the sensory and motor problems experienced by children with CP affect how they interact with their surroundings, including the environment and other people. Youth affected by CP have fewer opportunities to participate in traditional games and exercises such as playing basketball, riding a bike or playing ball with their friends.

In this context, video games, and in particular exergames, represent a very promising way to enable youths with CP to perform the exercise they need to break the cycle of deconditioning, while allowing them to socialize with others in fun ways from the comfort of their homes. Exergames are a combination of exercise (or exertion) and video games. In particular, we refer to digital games that require actions of large body parts or the whole body to control gameplay. Reviews of exergames indicate that they have positive effects both on motivation for active participation in rehabilitation and on impaired functions. However, the design of these games can be challenging, if our goal is to help them to socialize with others. First of all, limitations in physical abilities of youth with CP make it difficult for them to play many of

[*] PhD advisor: Domènec Puig. Additional author: Julián Cristiano.

the existing exergames. Second of all, there are challenges to social play such as establishing player groups and playing with players with different abilities that need special consideration.

2 GAmification for a Better LifE

GABLE[2] aims to create the first online video games service for youths with CP, which will be the hosting platform of games focused on improving motor skills and visual-motor coordination for youths with CP. These games will leverage the latest advances in Computer Vision techniques, in order to improve accessibility. The platform will be constructed with social networking in mind, which will allow parents, caregivers and patients to socialize in a common environment, to share experiences, in an effort to provide the best care for CP patients. Caregivers would also be able to share best practices and lessons learned among themselves, which will help them to provide a better service across Europe and beyond.

GABLE was born out of the idea that there is little or no help for youths with CP to play games specifically suited for their disability, while motivating them to play more and helping to rehabilitate their motor and motor-visual skills, and at the same time, introducing them to multiplayer/online gaming that would improve their social skills, and by extension, their social inclusion among their peers. Looking at the research done into these problems, we have seen that only baby steps have been made in an attempt to solve them. Several isolated demonstrations, in a research environment, have shown that there is great potential in exergames to be used as rehabilitation tools, while others have shown that multiplayer can bring about social interaction leading to an improvement of social skills for disabled patients.

What GABLE wants to do is to create the first online platform of games for patients with CP, where they, and their caregivers and parents, can have instant access to games, join a community of peers, and share their knowledge in order to provide the best possible help for all youth with CP. The platform will be built around the idea that online multiplayer games take advantage of the motivational aspects of group activity, and can provide additional motivation compared to single player games. These games are inherently promising for people with disabilities that are confined to their homes or care centres, so they have the potential to reach the widest audience.

3 Machine Learning in GABLE

One of the important aspects of GABLE is to harvest useful information from raw data in order to help caregivers and patients. To this end, we are

[2] www.projectgable.eu

designing a machine learning module, integrated in the backbone of the development platform and game-authoring tools, which mainly performs three different tasks including *Statistical analysis*, *Predictive analysis* and *Recommender system.*

3.1 Statistical Analysis

This module will provide basic analysis to the caregivers and parents, who will be able to track the progress of the patients using score of the games. The module will analyse all data collected, and will display statistical data on the types of games a patient has played, the progress she/he has made while playing a certain game repeatedly, etc. It will give an overall view of all the activity of the player, including game time, preference to a certain type of game, and will assist the caregiver in the active monitoring of a patient.

From a statistical perspective, social inclusion and education are latent variables of the data that are related to the game score. For example, a game that is mainly played using the body motions teaches the patients to improve their motor functions. Consequently, the game score of a patient is improved during time by playing this game, we can imply that the motor function of the patient is improved as well. The second set of tools will conduct more advanced discovery techniques using data collected from all users of GABLE platform. In the above example, it is possible to create a multivariate regression model for predicting the score of a user in a particular game if she/he plays the game more often. The regression model in this example may consider "age", "times played", "gender", "genre of game" in order to predict the game score. We will create various regression models by automatically selecting an appropriate set of independent variables to predict a dependent variable. Using this model, caregivers can predict the score of their patients in order to see if playing a particular game will be useful for their patients.

3.2 Predictive Analysis

This module will use the data gathered on the games played by the patients and their progress within a game in order to provide the caregivers recommendations on changing the difficulty level of a certain game, or recommend more challenging games. By analysing the data generated by the patients playing the game, the system can learn a regression or a classification model to predict the score of the patient in a certain scenario, and suggest to the caregiver in which configuration of a scenario the score of a particular patient will increase or decrease. With this information at hand, caregivers/parents will be able to change the configuration of the games in order to make them easier/more challenging. We will utilize a probabilistic graphical model[1] to achieve this goal. To be more specific, we will create a Bayesian network to model the dependencies between different variables and factorize the joint probability density function into smaller functions. There are two major advantageous with this model. First, it can deal with missing values by marginalizing the probabilistic model over them. Second, given values of some of the parameters of a game, it can suggest the other values of the parameters such that the probability of winning the game is reduced/increased. This can be done by computing the most probable explanation of the unknown variables. It should be noted that each game might have a different scenario. That said, we must create a unique graphical model for each game taking into account the interaction between their parameters and characteristics of the patients.

3.3 Recommender System

This module will provide rating for each game, as a measure of how many patients are playing this game, or how high it was scored by the patients and other caregivers and parents. Using this rating system, the module can then recommend a certain game for a patient, or for the caregiver/parent. We will follow the collaborative filtering [2][3] approach for this purpose. Taking into account the fact that collaborative filtering mainly works with user preferences, we will explicitly collect information about taste of each user by asking them to rate the games and analysing their search queries. We may also implicitly estimate taste of each user by computing the amount of time that they have played a particular game. In addition, we may also consider a hybrid recommendation system by storing some information about each game such as genre, type of game, graphical information (2D/3D).

Acknowledgement. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.732363.

References

[1] D. Koller and N. Friedman and L. Getoor and B. Taskar, "Graphical Models in a Nutshell", Introduction to Statistical Relational Learning, MIT Press, 2007

[2] Xiaoyuan Su and Taghi M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," Advances in Artificial Intelligence, vol. 2009, Article ID 421425, 19 pages, Hindawi, 2009. doi:10.1155/2009/421425

[3] Christian Desrosiers and George Karypis, "A Comprehensive Survey of Neighborhood-based Recommendation Methods", Recommender Systems Handbook, pp. 107-144, Springer, 2010, doi:10.1007/978-0-387-85820-3_4

Analysis of the Spiral Structure in Disc galaxies using the FFT Transform

Carlos Barberà *

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
carlos.barbera@urv.cat

1 Introduction

Our own interest in deprojection angles stems from the fact that we started a quantitative study of the properties of spiral structure in near-by disc galaxies, and for this we first need to deproject all our images. Indeed the list of all studies for which it is necessary to know the spatial orientation of the galaxy is too long to include here.

We will study the spiral structure, in disc galaxies, decomposing each image by means of bidimensional Fourier transforms. The first step is to deproject the galaxy image. It is thus necessary to determine the two deprojection angles, namely the position angle (hereafter PA) and the inclination angle (hereafter IA). The PA is the angle between the line of nodes of the projected image and the north, mesured towards the east, while the IA is the angle between the perpendicular to the plane of the galaxy and the line of sight.

Several methods have been proposed so far to obtain these angles. All of them suffer from some kind of systematic errors. The two methods we present can obtain very accurate values of the deprojection angles and also perform well for very low resolutions. This indicates that our methods can be used for very distant galaxies.

2 Deprojection methods

We introduce two methods, based on the Fourier transforms and which are closely linked to the two methods used by Garcia-Gomez and Athanassoula in [10] for HII region distribution in 1991. Let $I(u, \theta)$ be the image of the galaxy written in polar coordenates (r, θ), and $u = ln(r)$. We define the Fourier transform of this image as:

* PhD advisor: Carlos García Gómez

$$A(p,m)=\int_{u_{min}}^{u_{max}}\int_{0}^{2\pi} I(u,\theta)e^{i(pu+m\theta)}\,d\theta du \qquad (1)$$

In this equation, p corresponds to the radial frequency and m to the azimuthal frequency. Thus the $m=1$ values correspond to one-armed components, the $m=2$ values to two-armed components and so on. The values of $u_{min}=ln(r_{min})$ and $u_{max}=\ln(r_{max})$ are set by the inner and outer radius of the part of the image that we will analyze. Fixing the value of m, we can calculate the power associated to this compoment simply as:

$$P_m=|A(p,m)|=\left|\int_{-p_{max}}^{p_{max}} A(p,m)dp\right| \qquad (2)$$

The p_{max} value is related to the resolution in Fourier space through $p_{max}=\frac{1}{2\Delta u}=\frac{N-1}{2(u_{max}-u_{min})}$ where N is the number of points used in the Fourier transform in the radial dimension, usually $N=256$ or 512. In our first method we try to minimize the effect of the spiral structure by minimizing the ratio:

$$BAG1=\frac{P_1+P_2+\cdots+P_6}{P_0+P_1+\cdots+P_6}$$

This is equivalent to maximizing the contribution of the axisymmetric component. Since a badly deprojected galaxy will look oval, and thus contribute to the $m=2$ components as a bar, for our second method we simply minimize the ratio

$$BAG2=\frac{P_2}{P_0}$$

We tested them using artificially generated, yet realistic, galaxies described by an exponential disk with spiral components

$$I(r\theta)=e^{(-r/\alpha)}+Ae^{(-r_0/\alpha)}e^{-(\frac{r-r_0}{\sigma})^2}cos(p\ln(r)+2\theta)$$

We obtain very accurate values of the deprojection angles for a variety of situations; from inclinations greater of 80° to face on galaxies. One goal of the test is also the good performance of our methods in the case of very low resolutions, this indicates that our methods can be applied to very distant galaxies for cosmological interest.

3 Deprojection of the galaxies

We applied the two methods to the Frei sample galaxies as follows. First we constructed a grid covering all the possible range of values of PA and IA in increments of 2°. For each pair of angles (PA,IA) we deproject the galaxy image and we compute the Fourier transform (1) with the help of a polar grid. Using Eq. (2) we then calculate the power in each component and then the value of the ratios BAG1 and BAG2. We repeat this for every (PA,IA) pair in the grid. The optimum values are those for which we have a minimum. We illustrate the use of our methods with the help of galaxy NGC4501.

Finally, we perform correlations for comparing with other methods in the literature. Our two methods give mean correlation coefficients with the rest of the methods of 0.89 for BAG1 and 0.9 for BAG2 in the case of PA and of 0.87 and 0.88 for the BAG1 and BAG2 method respectively in the case of IA. This indicates that our methods are wel suited for the derivation of the deprojecton angles. In general, we can conclude that all the methods for deriving the deprojection angles are well suited from a statistical point of view.

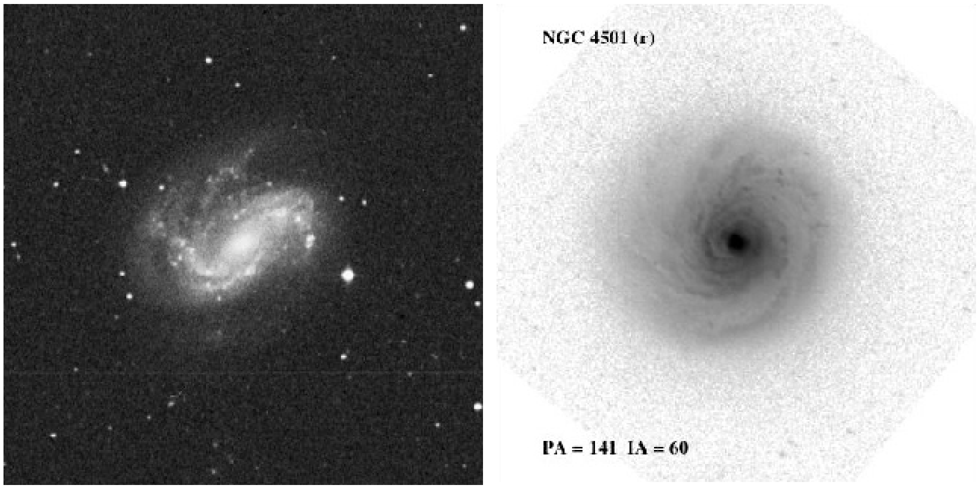


Fig. 1: Image of NGC4501 before and after deprojection

References

[1] Athanassoula, E., Misiriotis, A. Morphology, photometry and kinematics of N -body bars - I. Three models with different halo central concentrations *Monthly Notices of the Royal Astronomical Society*, 330(1):35–52, 2002.

[2] Buta, R. The structure and dynamics of ringed galaxies. V - The kinematics of NGC 1512, NGC 3351, NGC 4725, and NGC 4736 *Astrophysical Journal Supplement Series*, 66:233–259, 1988

[3] Considère, S., Athanassoula, E. The distribution of H II regions in external galaxies. I *Astronomy and Astrophysics* 111(1):28–42, 1982

[4] Considère, S., Athanassoula, E. Analysis of spiral components in 16 galaxies *Astronomy and Astrophysics Supplement Series* 76(3):365–404, 1988

[5] de Vaucouleurs, G. Revised Classification of 1500 Bright Galaxies. *Astrophysical Journal Supplement* 8:31, 1963

[6] de Vaucouleurs, G., de Vaucouleurs, A., Corwin, H.G., Buta, R.J., Paturel, G., Fouqué, P. Third Reference Catalogue of Bright Galaxies *Sky and Telescope* 82(6):621, 1991

[7] Danver, C.G. A morphological investigation of some near galaxies *Annals of the Observatory of Lund* 10:7–8, 1942

[8] Elmegreen, B.G., Elmegreen, D.M. Arm classifications for spiral galaxies *Astrophysical Journal* 314:3–9, 1987

[9] Frei, Z., Guhathakurta, P., Gunn, J., Tyson,J.A. A Catalog of Digital Images of 113 Nearby Galaxies *Astronomical Journal* 111:174, 1996

[10] García-Gómez, C., Athanassoula, E. Analysis of the distribution of HII regions in external galaxies. I - Position and inclination angles *Astronomy and Astrophysics Supplement Series*, 89(1):159–184, 1991

[11] García-Gómez, Barberà, C., C., Athanassoula, E., Bosma, A., Whyte, L. Deprojecting spiral galaxies using Fourier analysis. Application to the Ohio sample *Astronomy and Astrophysics*, 421:595–601, 2004

[12] Laurikainen, E., Salo, H. BVRI imaging of M 51-type pairs. II. Bulge and disk parameters *Astronomy and Astrophysics Supplement*, 141:103–111, 2000

[13] Whyte, L.F., Abraham, R.G., Merrifield, M.R., Eskridge, P.B., Frogel, J.A., Pogge, R.W. Morphological classification of the OSU Bright Spiral Galaxy Survey *Monthly Notices of the Royal Astronomical Society*, 336(4):1281–1286, 2002

Cryptographic protocols for Low Emission Zones access control

Carles Anglés Tafalla [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
carles.angles@urv.cat

1 Introduction

Reducing environment pollution and achieving greater sustainability into urban mobility are two of the major challenges that big cities confront in the 21st century. Promoting a rational use of vehicles, such as vehicle sharing incentives or electric vehicles use, are just some of the current strategies. On this basis, many cities have started establishing the so-called Low Emission Zones (*LEZ*), which are zones where a number of restrictions and penalties are applied to their users. These measures are aimed at reducing the traffic of combustion engine vehicles and encouraging the use of less polluting and low emission ones, preferably electric vehicles.

Although these strategies have proven to be effective in large cities, on a practical level, their implementation is neither simple nor economical. One of the main technological challenges regarding the *LEZ* scheme is to design a secure and reliable system which automatically controls the access of vehicles to these areas. Privacy also arises important challenges to the field and reveals that alternative user detection systems should be proposed instead of the use of video cameras that record all the vehicles plates that access the *LEZs*.

Our general objective is to provide secure protocols that automatically control the vehicle accesses to *LEZ*, but preserving the privacy of the drivers as long as they behave honestly.

2 Related Work

In recent years, several *LEZ* access control approaches, known as Electronic Road Pricing systems (*ERP*), on the basis of privacy by design have been proposed [1, 2, 3, 4, 5, 6, 7, 8]. All these systems require the use of an On-Board Unit (*OBU*) fitted with a *GPS* and a wireless communication system. The price of the fare is calculated according to the route the vehicle has

[*] PhD advisors: Jordi Castellà Roca and Alexandre Viejo

traveled. On the one hand, in [1] and [2], the information related to the external server is sent by the *OBU* to the external server, owned by the Service Provider (*SP*), which is in charge of setting the prices in each billing period. On the other hand, in [3, 4, 5, 6] it is the *OBU* which calculates the fees and sends them to the *SP* server in each billing period. In that way, the revealed information relating to the location of the vehicle is minimal. These systems use cryptographic evidences along with physical random-located checkpoints to demonstrate that the *OBU* has been honest when calculating the amounts corresponding to the traveled routes. The work in [7] presents a user privacy preserving protocol based on a time approach which, unlike the aforementioned works, offers a non-probabilistic fraud control. A further improvement of this protocol has been published in [8]. This proposal enhances the pricing system to dynamically adapt fares to the traffic changing conditions aiming at a better traffic distribution. Even when these protocols tackle the most important drawbacks of the systems proposed to date, due to their particularities, specific *OBU*s and full access to some of its functionalities are required for their feasibility. Nevertheless, *OBU*s integration in nowadays vehicles is not widespread and, as proprietary devices, most of their capabilities can be restricted to third parties.

3 Model of the system

Our general objective consists of encouraging the smartphone integration to the *LEZ* access control systems. The current anonymous approaches to control access to *LEZs* rely on the vehicles' On Board Units (*OBUs*), nevertheless, their integration in nowadays vehicles is not widespread. The adoption of the drivers' smartphone for this purpose may ease the rollout and acceptance of these zones. In any case, privacy is a mandatory issue and should be preserved as long as the drivers do not try to commit fraud. Only when a user accesses the *LEZ* without the proper authorization she should be identified and her anonymity revoked.

The scheme we propose in [9] presents a lightweight ERP solution that controls the access to a *LEZ* in a secure and reliable way, while providing privacy to honest users. In contrast to other systems, our approach uses the drivers' smartphone to validate their access instead of relying on an *OBU*. Those users who access the *LEZ* without proper authorization are automatically identified for their subsequent sanction. Accordingly, all anti-fraud measures do not affect the privacy of honest drivers.

The lifecycle of our system is divided into eight phases: i) Registration; ii) Installation; iii) Vehicle Registration; iv) Access; v) Exit; vi) Payment; vii) Fraud Control and; viii) Privacy Configuration.

Before a user could start using the proposed access model, she should complete the Registration (i), Installation (ii) and Vehicle Registration phases (iii);

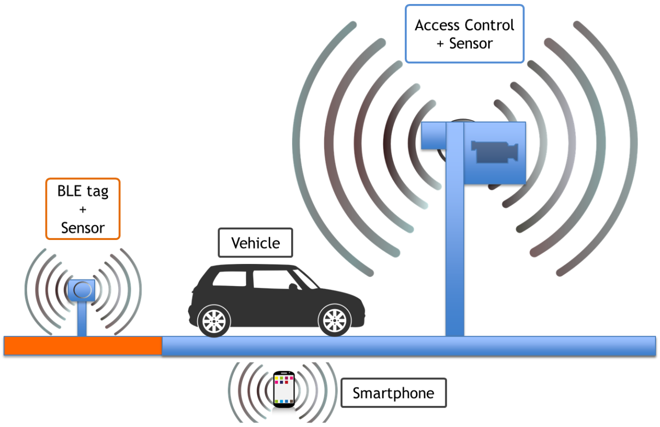


Fig. 1: Access infrastructure.

where she registers her personal data to the Competent Administration of the LEZ ($CALEZ$), installs the mobile application into her smartphone and registers the vehicles she will use to access the LEZ, respectively. Figure 1 shows a general scheme of the LEZ Access (iv) and Exit (v) phases. Both phases are presented together as they perform the same operations. When the vehicle approaches the LEZ access area, a Bluetooth Low Energy (BLE) tag awakens the application on the user's smartphone. This process is automatically done without the intervention of the user. For its part, the Input Sensor notifies to the Access Control (AC) entity that a vehicle has entered the LEZ access area. The mobile phone application establishes a secure communication with the AC entity through a cryptographic protocol and proves that it is a valid user. During the process, the user's anonymity is preserved through the use of a pseudonym. Then, the AC verifies whether the user's access permissions are correct or not. Moreover, the access and exit points are equipped with several sensors to obtain the vehicle's profile (height and length). If the user's credentials are valid, the access is registered and the user can privately access the LEZ. This access information will be used during the Payment phase (vi) to calculate the fee the user has to pay. Conversely, if the user does not have valid access permissions, the AC will take a photo of the vehicle license plate. With this photo the system will be able to identify the offending user. Additional anti-fraud measures are performed in Fraud Control phase (vii), where an independent entity looks for inconsistent patterns in the registered accesses and exits. Finally, to avoid that all the registered accesses of a user could be bind together though her pseudonym, a user can ask for a new one by running the protocol defined on the Privacy configuration phase (viii).

Acknowledgement. This work was partially supported by the DGT under the project SPARK & GO SPIP2015-01783 and the Spanish Government under AccessTur TIN2014-54945-R AEI/FEDER UE, SmartGlacis TIN2014-57364-C2-R and Red de excelencia Consolider ARES TIN2015-70054-REDC projects. Some of the authors are members of the UNESCO Chair in Data Privacy, yet the views expressed in this paper neither necessarily reflect the position of the UNESCO nor commit with that organization.

References

[1] R. A. Popa, H. Balakrishnan, A. J. Blumberg, Vpriv: Protecting privacy in location-based vehicular services, in: USENIX Security Symposium, USENIX 2009.

[2] X. Chen, G. Lenzini, S. Mauw, J. Pang, A group signature based electronic toll pricing system, in: ARES, IEEE Computer Society, 2012, pp. 85–93.

[3] J. Balasch, A. Rial, C. Troncoso, B. Preneel, I. Verbauwhede, C. Geuens, Pretp: Privacy-preserving electronic toll pricing, USENIX Security Symposium, 2010.

[4] S. Meiklejohn, K. Mowery, S. Checkoway, H. Shacham, The phantom tollbooth: Privacy-preserving electronic toll collection in the presence of driver collusion, in: USENIX Security Symposium, 2011, pp. 32–32.

[5] J. Day, Y. Huang, E. Knapp, I. Goldberg, Spectre: spot-checked private ecash tolling at roadside, in: WPES, ACM, 2011, pp. 61–68.

[6] F. D. Garcia, E. R. Verheul, B. Jacobs, Cell-based privacy-friendly roadpricing, Computers & Mathematics with Applications 65 (5) (2013) 774–785.

[7] R. Jardí-Cedó, M. Mut Puigserver, M. Payeras-Capellà, J. Castellà-Roca, A. Viejo, Time-based low emission zones preserving drivers' privacy, Future Generation Computer Systems, Available online 27 June 2016.

[8] R. Jardí-Cedó, J. Castellà-Roca, A. Viejo, Privacy-preserving electronic road pricing system for low emission zones with dynamic pricing, Security and Communication Networks, 2016, vol. 9, no 16, p. 3197-3218.

[9] J. Castellà-Roca, M. Mut Puigserver, M. Payeras-Capellà, A. Viejo, C. Anglès-Tafalla, Secure and Anonymous Vehicle Access Control System to Traffic-Restricted Urban Areas, 3rd International Workshop on Vehicular Networking and Intelligent Transportation Systems (VENITS 2017), August 2017.

Deep learning techniques for prediction of diabetic retinopathy severity

Jordi de la Torre [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
jordi.delatorre@gmail.com

1 Introduction

Diabetic Retinopathy (DR) is a leading disabling chronic disease and one of the main causes of blindness and visual impairment in developed countries for diabetic patients. Studies reported that 90% of the cases can be prevented through early detection and treatment. Eye screening through retinal images is used by physicians to detect the lesions related with this disease. Due to the increasing number of diabetic people, the amount of images to be manually analyzed is becoming unaffordable. Moreover, training new personnel for this type of image-based diagnosis is long, because it requires to acquire expertise by daily practice.

Deep Learning is a set of Machine Learning techniques for automatically constructing a model using multiple levels of representation from the underlying distribution of a large set of examples, with the final objective of mapping a high-multidimensional input into a smaller multidimensional output (f: $\mathbb{R}^n \mapsto \mathbb{R}^m, n \gg m$). This mapping allows the classification of multidimensional objects into a small number of categories. The model is composed by many neurons that are organized in layers and blocks of layers, using a cascade of layers in a hierarchical way. Every neuron receives the input from a predefined set of neurons. Every connection has a parameter that corresponds to the weight of the connection. The function of every neuron is to make a transformation of the received inputs into a calculated output value. For every incoming connection, the weight is multiplied by the input value received by the neuron and the aggregated value that used by an activation function that calculates the output of the neuron. The parameters are usually optimized using a stochastic gradient descent algorithm that minimizes a predefined loss function. The parameters of the network are updated after backpropagating the loss function gradients through the network. These hierarchical models are able to learn multiple levels of representation that correspond to different

[*] PhD advisor: Aïda Valls and Domènec Puig

levels of abstraction, which enables the representation of complex concepts in a compressed way [6], [8], [2], [1].

Quadratic Weighted Kappa (QWK) index is used in many medical diagnosis systems because the diseases have different degrees of severity, which are naturally ordered from mild to the most critical cases. If the diagnose is based on image analysis, the classification is even more difficult because in the interpretation of the image data normally is present some level of subjectivity that sometimes makes the conclusions of different experts to differ [5]. Quadratic Weighted Kappa is able to measure the level of discrepancy of a set of diagnosis made by different raters over the same population [9]. The strength of agreement between the raters is evaluated as a function of the distance between the prediction of both raters. For the case of diabetic retinopathy detection, human expert raters report inter-rater values of QWK about 0.80. This index has been used to evaluate the performance of the predictive model, in comparison with the human experts level.

The work done up to know in the thesis has been centered mainly in two studies. First, the construction of a classifier of diabetic retinopathy severity using the information encoded in the images of patient's retina using deep neural networks [4]. Second, the improvement of the classification quality using a learning approach based on ordinal information of the QWK [3].

2 Diabetic retinopathy detection using deep neural networks

The traditional model of pattern recognition has been based on extracting hand-crafted fixed engineered features or fixed kernels from the image and using a trainable classifier on top of those features to get the final classification. Using this scheme the problem of the DR detection has been based on hand engineering the features for the detection of microaneurism, haemorrhages and exhudate in retinal images that maximize the performance of the classifier. This type of approach requires a good understanding of the mechanism of the disease, requires a lot of labor time and is very task-specific and thereby not reusable for other different classification problems.

In this first work of the thesis we explore a completely different approach consisting on automatic feature learning. We use a deep convolutional neural network model for predicting the probability of every one of the five standardised DR severity levels [4]. The model is trained using a logarithmic loss function and stochastic gradient descend optimization based algorithms and a set of data augmentation techniques. The training procedure details can be found in [4]. The study was done with the Kaggle dataset of EyePACS. This image set has about 88.000 retina images, which are labeled by expert physicians.

To improve the classification rate, we use a probabilistic combination of the information that can be obtained from both eyes of the same patient. DR usually affects both eyes, specially when the illness is in high severity stages. The dataset used is big enough to infer from the frequencies of co-occurrence of the classes, the conditional probabilities of having one class in one eye given another class in the other, $P(Left|Right)$ and $P(Right|Left)$. Being $P(Left)$ and $P(Right)$, the probability distributions obtained by our predictive model with the left image and the right image, respectively, we can estimate P_L and P_R using $P_L = P(Left|Right)P(Right)$ and $P_R = P(Right|Left)P(Left)$. To merge the value obtained from the model with the estimation coming from the other eye, we calculate the arithmetic mean. The class with maximum value is the one selected for each eye.

3 Improving the classification rate with QWK loss function

The optimization of the neural networks for multi-class classification is traditionally done using the logarithmic loss. The logarithmic loss has a very robust probabilistic foundation: minimizing it, is the same as minimizing the logarithmic likelihood, that is equivalent to do a Maximum Likelihood Estimation (MLE) or equivalently, to find the Maximum a Posteriori Probability (MAP), given a uniform prior [7]. This loss function is designed to find perpendicular vectors in the output space. This model is suitable when the output classes are independent, but it may not be good in cases where classes are ordered. This is the case of some disease prediction, where an incremental severity scale is present. Normally in those cases a ordinal regression approach is better.

As Quadratic Weighted Kappa index is designed to evaluate a good ordinal rating, we explored the possibility of substituting the log-loss function by the QWK-loss function in deep neural networks training [3]. We defined the optimization procedure in terms of QWK and we showed that, for DR severity prediction, classification improves in more than a 5%. This method is directly generalizable to other multi-class classification problems where there is a prior known information about the predefined ordering of the classes.

4 Conclusions and future work

With the work done up to know we have been able to model the diabetic retinopathy detection using supervised deep learning techniques. Using the new QWK-loss function, we obtained up to a 5% increase in the classification rates over the standard approach. Moreover, thanks to the probabilistic combination of the results of both eyes we have been able to increase even further the results of the model, being able to reach human expert level performance.

The results of our study show that with the direct optimization of the QWK index allow the consecution of better generalization results in different datasets with ordered output classes. Log-loss has to learn the predefined ordering of the classes from data and this seems to be a disadvantage. Results showed that, depending on the use case, between 6-10% of improvement can be obtained from the direct optimization of QWK. This is a significant improvement that may be worth specially in medical diagnosis, since an accurate detection of the level of severity of a disease usually has great influence on the treatment prescription and the possibility of minimizing bad consequences of the illness.

Future work will be centered on the finding an human-understandable interpretation of the results given by the model and in the usage of unsupervised learning techniques to make the classification.

Acknowledgement. This work is supported by the URV grant 2015PFR-URV-B2-60 and the Spanish research projects PI15/01150 and PI12/01535 (Instituto Salud Carlos III).

References

[1] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, Aug. 2013.

[3] J. de la Torre, D. Puig, and A. Valls. Weighted kappa loss function for multiclass classification of ordinal data in deep learning. *Pattern Recognition Letters*, pages –, 2017.

[4] J. de la Torre, A. Valls, and D. Puig. Diabetic retinopathy detection through image analysis using deep convolutional neural networks. In *Artificial Intelligence Research and Development, Frontiers in Artificial Intelligence and Applications*, volume 288, pages 58–63. IOS Press, 2016.

[5] G. Hripcsak and D. F. Heitjan. Measuring agreement in medical informatics reliability studies. *Journal of biomedical informatics*, 35(2):99–110, 2002.

[6] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 5 2015.

[7] K. P. Murphy. *Machine Learning: A Probabilistic Perspective.* The MIT Press, 2012.

[8] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[9] A. J. Viera, J. M. Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363, 2005.

Sentiment Analysis of Tweets: from Traditional Machine Learning to Deep Learning

Mohammed Jabreel [⋆]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
mohammed.jabreel@urv.cat

1 Introduction

Sentiment Analysis (SA), also known as Opinion Mining, is the task of automatically identifying the opinions of customers about a product (or service, social event, etc.) from customer textual comments crawled from various social media resources. It normally involves the classification of text into categories such as "positive", "negative" and "neutral". SA can be done at different levels. Coarse-grained analysis attempt to extract the overall polarity on a document or sentence level, whereas, in a fine-grained level of analysis, the problem is to identify the sentiment polarity towards a certain target in a given text (*Target-dependent sentiment analysis*) [5, 9].

Most of the existing systems of SA and Targeted SA are inspired in the work presented in [7]. Machine learning techniques have been used to build a classifier from a set of tweets with manually annotated sentiment polarity. The success of the machine learning models is based on two main facts: a large amount of labeled data and the intelligent design of a set of features that can distinguish between the samples. In this approach most studies have focused on designing a set of efficient features to obtain a good classification performance. For instance, the authors in [6] used diverse sentiment lexicons and a variety of hand-crafted features. To leverage massive sets of tweets containing positive and negative emoticons for automatically learning the relevant features, several deep learning models have been proposed. For example, the work presented in [8] trained a convolutional neural network (CNN) to learn the best features and used it to classify the sentiment of the tweets.

Following those approaches we have developed systems to perform the sentiment analysis of tweets on two levels: coarse-grained and fine-grained. Based on the employed approach we categorise the systems into two groups: *Machine learning models* and *deep learning models*. The next subsections explain briefly those systems.

[⋆] PhD advisor: Antonio Moreno

2 Machine Learning Models for SA

Within this approach, we have developed three systems: SentiRich [1], En-SITAKA and Ar-SITAKA [2]. All these models are based on Support Vector Machines (SVMs). The basic steps of SentiRich are the following:

Preprocessing: in the pre-processing stage URLs and user mentions are removed, and all the text of the tweet is converted to lower case. After that, it is tokenized and POS-tagged using the tool Ark Tweet NLP. The suffix "*_NEG*" is added to all the words that appear in a negated context, which is a segment of a tweet which starts with a negation (e.g. no, don't) and ends with a punctuation mark.

Feature Extraction: the polarity of a tweet (positive, negative or neutral) is determined by SentiRich, which is fed with the following features:

- *Basic text features*: n-grams (contiguous sequences of n tokens, with n from to 1 to 4) and negated n-grams (the same information, but only with the tokens that appear in negated contexts).
- *Syntactic features*: number of occurrences of each POS and bi-tagged features (combination of bi-grams with their POS tags).
- *Lexicon features*: it includes the estimation of the polarity of the tweet according to seven popular opinion lexicons. The information about the positive/negative polarity of each word is combined, as described in [1], to obtain a global polarity of the tweet for each lexicon. Other lexicon-dependent features in this category include the average polarity of the positive/negative terms, the score of the last positive/negative term, and the maximum/minimum positive/negative score.
- *Semantic features*: each word of the tweet is mapped to a predefined cluster that groups together words that have similar meanings. Two sets of semantic clusters were used: the 1000 ones defined in the Ark Tweet NLP tool and the 4960 n-gram clusters obtained with the Word2vec tool.

Classification: SentiRich determines the polarity (positive, negative or neutral) at the tweet level. It is a classifier based on a SVM. This classifier was trained using the Twitter2013 train and development sets from SemEval2013, a well-known worldwide competition of natural language processing systems based on semantic analysis. The accuracy of the classifier was evaluated by comparing its performance with that of the top systems in the last SemEval competitions, using different data sets from these events. These results showed that the system obtained, in most of the cases, levels of accuracy that outperformed those of the state-of-the-art sentiment analysis systems (around 68% and 72%, depending on the input set).

En-SITAKA and Ar-SITAKA are extended versions of SentiRich where extra types of features have been added.

- *Embedding features*: *Word embeddings* are an approach for distributional semantics which represents words as vectors of real numbers. We used *sum*, *standard-deviation*, *min* and *max* pooling functions to obtain the tweet representation in the embedding space.

En-SITAKA was trained using the training sets of English-language tweets provided in SemEval13-16 whereas Ar-SITAKA was trained using the Arabic-language tweets provided by the organizers of SemEval17. The systems have been tested on 12,284 English-language tweets and 6100 Arabic-language tweets provided also by the organizers of SemEval2017. The golden answers of all test tweets were omitted by the organizers. En-SITAKA ranks 8th among 38 systems and Ar-SITAKA ranks 2nd among 8 systems in the SemEval2017 competition.

As a case study, SentiRihc has been used to analyse 3,000 tweets by local residents and 3,000 tweets by tourists at 10 major destinations in Europe with the aim of finding out the positivity, neutrality or negativity of their published tweets [4].

3 Deep Learning Models for SA

Deep learning techniques for SA have become very popular. They provide automatic feature extraction and both richer representation capabilities and better performance than traditional feature based techniques (i.e., hand-crafted features). We have leveraged Recurrent Neural Networks (RNNs) to solve the targeted SA problem.

We have developed a system called target-dependent bidirectional gated recurrent unit (TD-biGRU) [3]. It has the ability to represent the interaction between the target (an entity, like a person, organisation, product, object, etc., referred to in a text, about which an opinion is expressed) and its context (the text surrounding it). Its main steps are the following. First, the words of the input sentence are mapped to vectors of real numbers. This step is called vector representation of words or word embedding. Afterwards, the input sentence is represented by a real-valued vector using a bidirectional gated recurrent unit. This vector summarizes the input sentence and contains semantic, syntactic and/or sentimental information based on the word vectors. Finally, this vector is passed through a softmax classifier to classify the sentence into positive, negative or neutral.

We have evaluated the effectiveness of the proposed model on a benchmark dataset from Twitter. The experiments show that TD-biGRU outperforms the state-of-the-art methods for target-dependent sentiment analysis.

Acknowledgement. The authors acknowledge the support of Univ. Rovira i Virgili through a Martí i Franqués PhD grant and the Research Support Funds 2016PFR-URV-B2-60.

References

[1] Jabreel, M., & Moreno, A. Sentirich: Sentiment analysis of Tweets based on a Rich Set of Features. In *Artificial Intelligence Research and Development - Proceedings of the 19th International Conference of the Catalan Association for Artificial Intelligence, Barcelona, Catalonia, Spain, October 19-21, 2016*, pages 137–146, 2016.

[2] Jabreel, M., & Moreno, A. SiTAKA at SemEval-2017 Task 4: Sentiment Analysis in Twitter based on a Rich Set of Features. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[3] Jabreel, M., & Moreno, A. Target-dependent Sentiment Analysis of Tweets using a Bi-directional Gated Recurrent Unit. In *Proceedings of the 13th International Conference on Web Information Systems and Technologies - Volume 1: WEBIST,*, pages 80–87. INSTICC, ScitePress, 2017.

[4] Jabreel, M., Moreno, A., & Huertas, A. Do Local Residents and Visitors Express the Same Sentiments on Destinations Through Social Media? In *Information and Communication Technologies in Tourism 2017*, pages 655–668. Springer, 2017.

[5] Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics, 2011.

[6] Mohammad, S. M., Kiritchenko, S., & Zhu, X. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June 2013.

[7] Pang, B., Lee, L., & Vaithyanathan, S. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[8] Severyn, A., & Moschitti, A. Severyn, Aliaksei, and Alessandro Moschitti. "UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado*, pages 464–469, 2015.

[9] Vo, D. T., & Zhang, Y. Target-dependent Twitter Sentiment Classification with Rich Automatic Features. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 1347–1353, 2015.

Prioritization of Prefetching Traffic into Multicore Networks

Carles Aliagas [*] [**]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
carles.aliagas@urv.cat

1 Prefetching Overview

Cache holds very few data, so it is compulsory to decide which ones are best suited to be stored in. Their goal, is to increment the probability to find them in cache and avoid the main memory access. We have to decide when–how–where to copy a data into cache, when–how-where to look for data in cache and when-how-where to dismiss data from cache. This means hardware algorithms to *place*, *look–for* and *replace*.

Observing typical program pattern access, we can see a property of *data locality*. That means that after accessing a data it is very probable to access the same data, or a data near that one, in the near future. There are two types of locality: temporal locality (the same data) or spatial locality (data next to the accessed one). Caches designs wants to take profit from this property.

In order to take profit of spatial locality instead of bringing only one word (4, 8 bytes) caches brings a full block (16 to 128 bytes). This is a basic *prefetch* mechanism. This is very useful for scientific programs that have sequential accesses to vector and matrix, but if the block size is increased it can increase pollution.

Caches are initially empty. So at the beginning some accesses will produce a miss (also known as *compulsory miss*). This behaviour leads to a research challenge that tries to reduce the compulsory misses by bringing blocks to cache before the processor ask for them. This is the target of all prefetching mechanism.

Prefetching can be applied at any level of the memory hierarchy and from a software or hardware point of view. At the present moment, most of the processors are multicore/multithread and in this scenario caches can be shared between cores and threads. Several problems arise with prefetching in all those alternatives.

[*] Thanks to: M. Torrents (BSC), R. Martínez (Oracle)

[**] PhD Supervisors: C. Molina (URV), A. González (UPC)

2 Prefetching Challenges

With the idea of providing a good performance in terms of execution time, many prefetching[4] mechanism have been proposed for unicore processors with the main challenge to correctly choose the next useful block and the correct timing. Multicore prefetching[5] has also two main focus of prefetching, one is to bring data as near as possible to the core that would use it (privates caches) and the other is to bring data to the main shared cache of the processor from main memory. The first one will have to deal with the coherence cache protocol and the second one will have to deal with the shared resources of the cores (network, external connections,...). Many[7][8] uses the extra computation of multicores/multithreads to execute the prefetch mechanism but at the expense of not using them for raw computation. Different approaches have to deal with this dilemma and use complex prefetch alternatives that uses a lot of processors resources or simple prefetch alternatives that does not perform as good.

One resource that influence memory latency is the network that connects the cores with the shared cache and the external main memory. A good prefetching algorithm should not produce an increase of traffic in the memory interconnection. An increase in memory traffic entails higher power consumption and a higher degree of contention in the interconnection network. Note that the number of memory requests are in most of the cases going to be higher than in a system without prefetching. This is especially true if we are in a multicore system because it increases on-chip communication since coherence between the L1 caches of the tiled CMP must be ensured. An increase in the congestion of the network will most probably increase the latency of not only prefetching requests but also regular memory operations

Our proposal focus on network congestion in order to dynamically reorder data access to prioritize regular request in detrimental of prefetch request to avoid increments in latency of regular data access.

3 Prioritization into the network: Our proposal

Traditionally, memory systems do not differentiate between prefetch and regular requests. Recently a number of approaches[9][10][11] have appeared that give several priorities to both types of requests depending on the predicted behaviour, since it has been shown that delaying regular requests may degrade performance if prefetch requests are not accurate.

The prefetcher mechanism send requests to the network to bring data from external memory to internal caches or to move data between the internal caches. That requests have to coexist with regular data accesses and travel together within the network subsystem. One challenge is that prefetching does not have to penalize regular accesses but the network subsystem does not have information about the different types of requests.

In our proposal the prefetcher mechanism will add information to its data access in order to mark them as a prefetch requests. Also it will be add information related to the timing of the prefetch. Network routers will use that information to reorder requests in its queues and apply/modify the priority of them. Moreover, it will use dynamic information of congestion in order to apply different policies. regular accesses have maximum priority and prefetch accesses have variable priority depending on its time request.

Another possibility of the mechanism is to discard prefetch accesses when they are arriving after its time request, and also discard some of them when the network utilization is near full capacity.

Several experiments will determine optimal values of priorities and thresholds needed by the mechanism in order to modify priorities and discard some or all the prefetch access.

4 Experimental Framework

The SimpleScalar Tool Set[1]provide simulators ranging from a fast functional simulator to a detailed out-of-order issue processor that supports non-blocking caches, speculative execution, and state-of-the art branch prediction. In this work, we use the sim-outorder to obtain program statistics.

The Standard Performance Evaluation Corporation (SPEC)[2] is an organization founded in 1988 that provides several families of benchmarks to measure the performance of different computer systems. In this work we consider the CPU family, designed to provide performance measurements that can be used to compare compute-intensive workloads on different computer systems. In particular, we consider the following suite of the CPU family: SPEC CPU2006.

The Opnet Modeler Suite[3] provides a suite of protocols and technologies to design, model, and analyze communication networks.

5 Conclusions and Future Work

Prefetching in multicore processors shows new challenge designs that have to deal with sharing the available resources of the processor for demand requests and prefetch requests. We focus on network congestion as a measure to reorder, depriorize and also discard prefetch requests. In this way, tuning the correct values of our mechanism will reduce global average access time of memory accesses.

Acknowledgement. This work has been partially supported by the Spanish MCI and FEDER funds (EU) under contracts TIN2016-75344-R and TIN2016-77836-C2-1-R

References

[1] D.Burger, T.M.Austin and S.Bennet, Evaluating Future Microprocessors: The SimpleScalar Tool Set, CS-TR-96-1308. University of Wisconsin, July 1996.

[2] http://www.spec.org/, “The Standard Performance Evaluation Corporation”.

[3] https://www.riverbed.com/, ”Opnet Modeler Suite”

[4] N. Oren “A Survey of Prefetching Techniques”. TR CS-2000-10, University of the Witwatersrand, 2000.

[5] S.Byna, S.Chen and X-H.Sun Taxonomy of data prefetching for multiprocessors. Journal of Computer Science and Technology, 24(3):405–417, 2009.

[6] Hennessy J, Patterson D. Computer Architecture: A Quantitative Approach. The 4th Edition, Morgan Kaufmann, 2006.

[7] Kim D, Liao S S, Wang P H, et al. Physical experimenta- tion with prefetching helper threads on Intel’s hyper-threaded processors. In Proc. the International Symposium on Code Generation and Optimization: Feedback-Directed and Runtime Optimization, Palo Alto, USA, March 21–24, 2004, p.27.

[8] I. Ganusov and M. Burtscher, “Future Execution: A Hardware Prefetching Technique for Chip Multiprocessors”, Proceedings of the 14th Parallel Architectures and Compilation Techniques, 2005.

[9] A. Flores, J. L. Aragón, and M. E. Acacio “Energy-efficient hardware prefetching for CMPs using heterogeneous interconnects,” in Parallel, Distributed and Network-Based Processing (PDP), 2010 18th Euromicro International Conference on, pp. 147–154, IEEE, 2010.

[10] N. Chidambaram Nachiappan, A. K. Mishra, M. Kademir, A. Sivasubramaniam, O. Mutlu, and C. R. Das “Application-aware prefetch prioritization in on-chip net-works,” in Proceedings of the 21st international conference on Parallel architectures and compilation techniques, pp. 441–442, ACM, 2012.

[11] J. Lee, H. Kim, M. Shin, J.-H. Kim, and J. Huh “Mutually aware prefetcher and on-chip network designs for multi-cores,” Computers, IEEE Transactions on, vol. 63, no. 9, pp. 2316–2329, 2014.

Secure Interpolation in the Cloud

Jordi Ribes González [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
jordi.ribes@urv.cat

1 Introduction

Cost-effective third-party (cloud) service providers offer very convenient data storage and computation services at a low cost, thus providing an attractive alternative to other forms of storage. However, outsourcing potentially sensitive datasets to an external cloud service provider poses many security and privacy concerns.

A natural approach to address these security concerns consists in applying cryptographic techniques. However, traditional symmetric-key encryption techniques fail to provide an efficient solution. A trivial option consists on encrypting all data using a symmetric-key encryption scheme and using the server only for Storage-as-a-Service. To perform a computation, all relevant data is retrieved, decrypted and computed on locally. Unfortunately, this solution may not be efficient, particularly if client devices have limited computational power or storage capacity, and requires a high bandwidth during queries. Alternative cryptographic schemes must be developed in order to overcome this obstacle.

In recent years there have been important advances in cryptographic techniques that allow to take advantage of the economical and functional benefits of cloud computing while securing the data. Two of these techniques are Homomorphic Encryption (HE) and Secure Multi-party Computation (SMC), both of which allow for remote computations over encrypted data.

Secure Multi-party Computation protocols are interactive protocols that allow a set of parties to jointly compute a function over their inputs. In SMC, parties keep their inputs private and engage in an interactive protocol with each other, so that at the end of the protocol each party learns the function evaluation and nothing else about inputs from other parties.

Homomorphic Encryption schemes allow computations to be performed directly on encrypted data, and they are classified according to the operations they support. Additive HE (such as [6]) and multiplicative HE schemes

[*] PhD advisor: Oriol Farràs Ventura

efficiently support a single operation on ciphertexts, that is, addition and multiplication respectively. Somewhat Homomorphic Encryption (SHE) schemes support any number of additions, but a limited number of multiplications. Fully Homomorphic Encryption (FHE) schemes support an arbitrary number of additions and multiplications on ciphertexts. Unfortunately, all known FHE schemes are computationally very expensive, which hinders their applicability in practice.

2 Secure Interpolation in the Cloud

Interpolation and regression techniques such as generalized least squares, polynomial regression and Spline interpolation are often used in practical applications, for example in order to predict values of some phenomena given a set of samples. They have a vast amount of applications, ranging from computer graphics to data analysis or experiment designs.

Outsourcing such computations to the cloud can offer numerous cost-saving and practical benefits, since applications often involve massive datasets and expensive computations. Data ubiquity is also very convenient, as such computations can involve data owned by multiple organizations, or they can be requested by multiple parties.

However, outsourcing computations to the cloud can pose security and privacy concerns, since applications usually involve potentially sensitive datasets. Therefore, we aim to provide practical solutions to enable clients to efficiently delegate an encrypted dataset to a semi-trusted server, in such a way that interpolation computations can be performed directly over encrypted data.

Following the previous discussion, we may look for a solution involving HE schemes. The main obstacle to applying this approach is that the considered computations often involve complex operations, requiring many additions and products. Some interpolation techniques involve computations that are currently challenging even when using FHE, including the computation of square roots, natural exponentiations or solving systems of linear equations. In order to overcome this obstacle, we look for tailored adaptations of the interpolation computations, so that we can apply HE schemes and enable the delegation of interpolation computations to the cloud.

3 Private Outsourced Kriging Interpolation

Kriging [1, 3, 5, 8] is a well-recognized form of linear interpolation widely used with datasets involving spatially correlated data. It aims at predicting the value of some phenomena at an unobserved location in a two-dimensional region. This interpolation method was designed with geo-statistical applications in mind (*e.g.* to predict the best location to mine within a region, based on the mineral deposits found at previous boreholes), but has also found applications in a variety of settings including remote sensing, real-estate appraisal and computer simulations. Kriging has been identified as a good candidate process to be outsourced to the cloud, based on the practical and legislative requirements of industrial users [2, 4].

Based on a recent work carried out in conjunction with James Alderman, Benjamin Curtis, Oriol Farràs and Keith M. Martin, we present a method for the efficient private outsourcing of Kriging interpolation. The proposed solution uses a tailored modification of the Kriging algorithm in combination with additively homomorphic encryption, allowing crucial information relating to measurement values to be hidden from the cloud service provider. Moreover, with the exception of the high one-time cost of encrypting the dataset, the remaining client-side processes are very efficient. We evaluate the performance of our solution through an implementation in Python 3.4.3, using the PHE library [7].

Since the approach followed for Kriging interpolation is applicable to other interpolation techniques and statistical tools, a next step in this line of work is to develop solutions for other similar techniques.

The proposed results have been presented at the 5th Workshop on Encrypted Computing and Applied Homomorphic Cryptography (WAHC'17).

Acknowledgement. This work is partly supported by the European Comission through H2020-ICT-2014-1-644024 "CLARUS" and H2020-DS-2015-1-700540 "CANVAS", by the Government of Spain through TIN2014-57364-C2-1-R "SmartGlacis" and TIN2016-80250-R "Sec-MCloud", by the Government of Catalonia through Grant 2014 SGR 537, and by COST Action IC1306.

References

[1] J.-P. Chilès and P. Delfiner. Multivariate methods. *Geostatistics: Modeling Spatial Uncertainty, Second Edition*, pages 299–385, 1999.

[2] CLARUS: User centered privacy and security in the cloud. `http://clarussecure.eu`.

[3] N. Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992.

[4] InGeoCloudS: inspired geo-data cloud services. `https://www.ingeoclouds.eu/`. Accessed: 11/12/2016.

[5] D. Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951.

[6] P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pages 223–238. Springer, 1999.

[7] python-paillier: a library for partially homomorphic encryption in python, Data61|CSIRO. `https://github.com/NICTA/python-paillier`, 2016. Accessed: 11/12/2016.

[8] H. Wackernagel. *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media, 2013.

Classification of food related places using visual lifelogging

Md. Mostafa Kamal Sarker *

Department of Computer Engineering and Mathematics, Rovira i Virgili University
Tarragona, Spain
mdmostafakamal.sarker@urv.cat

1 Introduction

Classification of food related places is one of the most recent promising application in the area of scene recognition. It helps to analyze the nutrition intake based on the food related activity. Lifelogging is used to capture and analyze the data sources to record the events and patterns of a person's life by using wearable sensors, such as wearable cameras. In visual lifelogging images are captured by wearing a camera over a long period of time which shows the daily experience of the camera user. We aim to work in the area of food places recognition by analyzing the images, which were captured by the visual lifelogging. In this work, a computer vision based food related places recognition method will be introduced. Finally, we will develop and implement a fully automated food profiling system by analyzing the image of food related places. Currently, we used a deep learning based approach for classifying the food related places, which has shown promising results.

2 Proposed Approach

2.1 Visual Lifelogging Dataset

Lifelogging is a procedure in which personal data produced by human behavioral activities is tracked and recorded. It tracks personal activity data, such as exercising, sleeping and eating. For collecting visual lifelogging images, we used a wearable camera named "Narrative clip" [1] for developing our visual lifelogging dataset. Figure 1 shows the narrative clip camera.

This camera is able to collect big amount of images with respect to its continuous image collection capability (2-3 per minute and 1500 per day, 70000 per year). After gathering all of the images, we have labeled them with respect to their places class name (e.g., restaurant, supermarket, kitchen, etc.), and named it "Egoplaces". Egoplaces contains thirty three thousand images

* PhD advisor: Dr. Domenec Puig Valls and Dr. Petia Radeva

Fig. 1: Narrative clip camera

comprising of 28 food related places in total. The details about the dataset is shown in Table 1.

Table 1: Description of the “Egoplaces” dataset.

Class name	No. of images	Class name	No. of images	Class name	No. of images	Clsss Name	No. of images
bakery_shop	1038	cafeteria	1476	food_court	161	pizzeria	1005
balcony_interior	527	candy_store	154	greenhouse_indoor	55	pub_indoor	182
banquet_hall	206	cocktail	526	ice_cream_parlor	86	restaurant	4001
bar	1362	coffee_shop	1607	kitchen	4058	restaurant_patio	63
bazaar_indoor	217	delicatessen	745	market_indoor	1027	supermarket	3141
beer_hall	339	dining_room	3681	market_outdoor	1813	sushi_bar	127
butchers_shop	81	fastfood_restaurant	981	picnic_area	720	workplace_office	3654

In this dataset, we have 28 classes of food related places, which are collected by egocentric camera. It contains rich classes that cover various visual surroundings of our daily life experience.

2.2 Deep Learning

Deep learning is a part of machine learning that has revolutionized the area of artificial intelligence. It is widely used in computer vision and natural language processing, yielding best outcome and outperforming most of the state-of-the-art approaches. Neural Networks (NNs) is one of the well-known learning system in the area of machine learning. In deep learning, convolutional neural network (CNNs) is mainly used for object detection and image classification or recognition. Deep learning is mainly based of artificial neural networks (ANNs), which is able to learn (also called training) from data and apply the learned knowledge to new data (called testing). The core concept of ANNs comes from the human brain functionality. Deep Learning has become very powerful because of the recent technological advancement in graphics processing units (GPUs) and central processing units(CPUs), and the big amount of data which is publicly available nowadays. In this work, we used state-of-the-art CNNs models to classify food related places.

Currently, deep learning based models are also used in different mobile based applications such as, “Snap-n-Eat” [6] and “Lose it” [2] to detect and recognize different types of food items. However, no work has been done in the area of scene recognition related to the food places. A novel scene classification method using CNNs deep features was proposed by Bolei Zhou et al.

[7] to compare the density and diversity of images using the "Places" dataset. However, this method is proposed to classify the scenes based on different places such as, airfield, art studio, bathroom, classroom, etc.

3 Experimental Results

In the initial stage of the experimental setup, the dataset was splitting for the training and test phase. The dataset was split into 80% for training and the rest of the images for testing. Here we utilized NVIDIA GTX 1070 with 8GB memory size which facilitated to run a complex network architecture. The scheme for deep learning is the latest version of Keras 2.0.3 with Tensorflow backend. In this work, we proposed a deep learning based food related places recognition system to classify different types of food places by using the state-of-the-art CNN models. We used transfer learning method, which is fine tuning of ImageNet pre-trained models because currently it has the best accuracy for image classification task. We have used three different models: VGG16 [4], ResNet50 [3] and InceptionV3 [5]. The classification results show that the InceptionV3 model yields the highest accuracy among the used state-of-the-art models. Figure 2 shows the InceptionV3 model architecture for our 28 food related places classification task. We modified it by removing the intermediate auxiliary logits output to adopted with our problem.

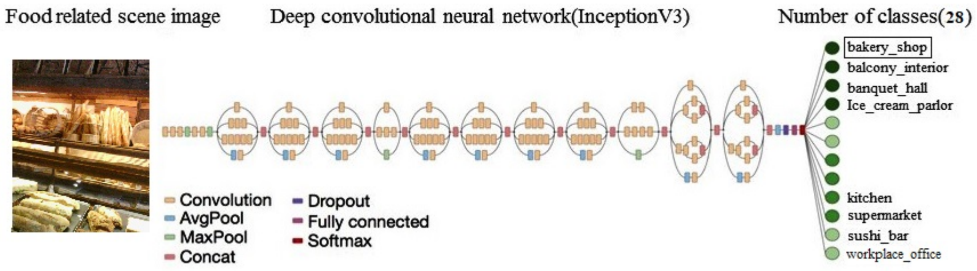


Fig. 2: InceptionV3 model architecture for the food related scene classification.

The number of classes are 1000 in ImageNet models classifier, but we have only 28 classes of food related places. So we have to tune the dimension of the last fully-connected (FC) layer of the networks with the number of 28 classes. We have tested the default model on Egoplaces and also retrained the final layer to get a model depending on the pre-trained model. We found that the default and the retrained last layer of CNNs models cannot produce performance higher than 40% and 60% respectively. To overcome this problem, we retrained the full models until the softmax layers of our pre-trained models and got much higher accuracy. It is necessary to fit and fine-tune the optimization parameters for example, weight decay, which avoids the over fitting problem and helps to provide the balance in between variance and bias. In the training process, we used s stochastic gradient descent with 0.9 momentum.

We used a learning rate of 0.001 as well as batch size 64 is also given to the network. The accuracy of the classification results for our models are listed in Table 2.

Table 2: Comparison of accuracy among the models

Deep models	Accuracy
VGG16	52.22
ResNet50	73.17
Inception V3	88.35

4 Conclusion and Future Work

In this work, we presented a method to classify food related places by using transfer learning of different CNNs models. In order to achieve a higher accuracy, we fine tuned all the network layers that increased the classification performance. The obtained results conclude that the InceptionV3 architecture is able to learn the features of food related places with a classification rate of 88.35%. In future work, we will increase the umber of images in the proposed dataset (especially, the categories that have few number of images) and create our own deep models for developing fully automated food profiling system.

Acknowledgement. This research is funded by the program Marti Franques under the agreement between Universitat Rovira Virgili and Fundació Catalunya La Pedrera.

References

[1] N. Clip. Narrative clip–a wearable. *Automatic Lifelogging.*

[2] M. Dohan and J. Tan. Lose it! *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, 6(2):60–65, 2011.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[6] W. Zhang, Q. Yu, B. Siddiquie, A. Divakaran, and H. Sawhney. "snap-n-eat" food recognition and nutrition estimation on a smartphone. *Journal of diabetes science and technology*, 9(3):525–533, 2015.

[7] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

Computational Methods for Classifying Breast Cancer Molecular Subtypes in Mammograms

Vivek kumar Singh*

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
vivekkumar.singh@urv.cat

1 Purpose

There is an unstoppable tendency towards personalized medicine in order to achieve both, diagnosis and treatment, and monitoring more effective for each patient. In this line, we propose the P-BreasTreat work, aimed at the personalized treatment of breast cancer by developing new computational techniques for image and data analysis. The ultimate purpose is to improve the effectiveness of current methods for determining the level of malignancy associated with that cancer tumors and also to propose models to prevent relapse and improve the quality of life of the patients.

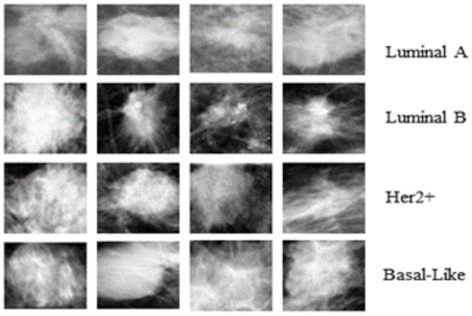


Fig. 1: 16 examples of input image samples, 4 for each of the 4 molecular subtypes of breast cancer

In proposed work, we will develop computer technologies for distinction and initial screening of the 4 molecular subtypes of breast cancer shown in figure 1 (Luminal A, Luminal B, Her2+ and Triple Negative) as advanced support to the traditional pathologic analysis. The impact will be focused

* PhD advisor: Dr. Santiago Romani Also and Dr. Domenec Puig Valls

on reducing both the number of biopsies and adverse psychological effects on patients. To do this, we will design specific methods of medical image analysis by using Computer Vision and Artificial Intelligence techniques, aimed at designing new adaptive biomarkers.

Once detected the molecular subtype, we will design customized models for the diagnosis and monitoring of patients treated with neoadjuvant therapy, conservative surgery and radiotherapy, in order to provide new tools for predicting relapse (either local or remote) of breast cancer, anticipating corrective measures to improve the rate of recovery. These models will also highlight critical points of the treatment or disagreements with the clinical standards (analysis of adherence). In order to do this, we will apply automatic process mining techniques to the evolutionary data of the patients.

2 Related Works

Numerous approaches have been proposed to classify the BC tumor subtypes based on histological information. The method designed by Perou et al. [2] performed a BC classification into certain "intrinsic" subtypes based on gene expression patterns. Herbeck et al. [3] presented the guidelines for the BC molecular subtype categorization based on several immunohistochemistry (IHC) biomarkers such as estrogen receptors (ER), progesterone receptors (PR), human epidermal growth factor receptor 2 (HER-2) and antigen KI-67 (Ki67).

Torrents-Barrena et al. [4] presented the first work to determine the feasibility of using a CAD system to differentiate among all BC molecular subtypes in mammograms. Authors designed two classification experiments: "Luminal A vs. Luminal B", and "Luminal A vs. Luminal B vs. Her-2+ vs. Basal Like". Support Vector Machines (SVM) and Local Binary Patterns (LBP) yielded the best accuracy: 75% and 52.17%, respectively. Moreover, they designed in [5] a new methodology based on fractal texture analysis and unsupervised / supervised classifiers. SVM also achieved the best performance (76.48% and 55.67%, respectively). The main drawback of both works was the limited number of Her-2+ and Basal Like samples.

3 Proposed Method

In this abstract, we propose a semi-automatic CAD system to classify the four molecular subtypes of BC from full-field digital mammograms (FFDM). A modified VGG16 [6] convolutional neural network architecture is presented to learn the underlying micro-texture patterns of the mammogram image pixels for each subtype.

Our hypothesis is that a CNN conveniently designed can learn the prototypical underlying micro-textures of each cancer subtype and that those

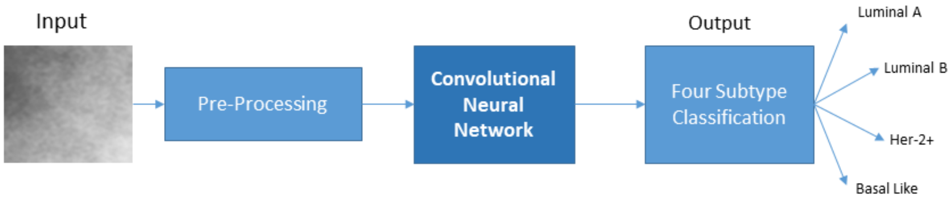


Fig. 2: Work flow process of four molecular subtypes classification of breast cancer

prototypes are characteristic of each subtype shown in figure 3, i.e., they are similar to all samples of the same subtype but different from the micro-texture prototypes of the other cancer subtypes. Hence, the trained CNN should be able to predict the subtype of any new breast tumor, given a ROI sample extracted from its corresponding segmented mammography.

We will base our design on the VGG_{16} architecture, since it uses small area filters (3x3) that we expect they are well suited for micro-texture prototype learning, in contrast to other CNN architectures (e.g. AlexNet) that use larger filters (11x11) to look for edges, macro-textures or other salient features of the objects.Since our CNN must learn just pixel-wide micro-texture prototypes (not the full tumor shapes) of only four classes of cancer subtype, we have checked several simplifications of the VGG_{16} original architecture.Concretely, we have defined smaller sets of filters and reduced the number of neuron layers.

4 Experimental Results

The experimental data are composed of 179 DICOM mammograms (CC and MLO views) distributed in 64, 63, 25 and 27 samples for classes Luminal A, Luminal B, Her-2+ and Basal-like, respectively. These medical image samples are provided by a Oncology Group in Spain.

Firstly, we have checked the performance of our model by training and validating the network with regards to the first two classes, Luminal A and Luminal B, which correspond to the less aggressive cancer subtypes. Our network has performed really well on Luminal A samples, achieving a 95% of accuracy. On the other hand, just 61% of Luminal B samples had been correctly classified, while the remaining 39% had been misclassified as belonging to Luminal A.Nevertheless, our network renders an overall accuracy around 78%, which is quite a good result taking into account the evident lack of visual patterns in the image samples. The second experiment corresponds to the full 4-class classification, i.e., including all breast cancer subtypes. From the individual accuracies, we can obtain an overall accuracy as the weighted average with respect to the number of test samples of each class, obtaining a fair 67% of good predictions.

5 Conclusion and Future work

In this abstract, we have presented a supervised BC molecular subtype classification method based on a CNN that analyse manually selected areas of breast tumors found in DICOM images of mammograms. To the best of our knowledge, this is the first effort to predict the molecular subtypes of malignant tumors just from image excerpts of digital mammograms using CNNs. Before, we tried other approaches to the same problem using classical texture descriptors (Uniform Local Binary Patterns, Histogram of Gradients, Gabor filters, Fractal dimension), but with less degree of accuracy ([6]: 75% — 52%; [8]: 76% — 56%; current approach: 78% — 67%). Other authors have only focused on automatic detection of tumors and determining if the tumor is benign or malignant. Future work will aim at validating our approach on larger datasets of MRI images, with the ultimate objective of gradually bringing computerized assistance to BC molecular subtypes classification into clinical practice.

Acknowledgement. This research has been partly supported by the University Rovira i Virgili through Marti-Franques Research Grant and Spanish Government through project DPI2016-77415-R.

References

[1] G. Shieh,,C. Bai, and C. Lee Identify breast cancer subtypes by gene expression profiles. *Journal of Data Science*, 12:165–75, 2004.

[2] C. M. Perou, T. Srlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees and Fluge. Molecular portraits of human breast tumours .*Nature*, 6797: 747–752, 2000.

[3] N.Harbeck, C. Thomssen, and Gnant Brief preliminary summary of the consensus discussion. *Breast care*, 8(2): 102–109, 2013.

[4] J. Torrents-Barrena , D. Puig, L. Diez-Presa, M. Arenas, and P. Radeva Assessment of a multidiscriminant supervised classifier driven by textural features to distinguish molecular subtypes of breast cancer. *In International Conference of Computer Assisted Radiology and Surgery (CARS)* 10(1): S31-S32, 2015.

[5] J. Torrents-Barrena, A. Valls, P. Radeva, M. Arenas and D. Puig Automatic Recognition of Molecular Subtypes of Breast Cancer in X-Ray images using Segmentation-based Fractal Texture Analysis. *In 18th International Conference of the Catalan Association of Artificial Intelligence (CCIA)* 277: 247-256, 2015.

[6] K. Simonyan, and Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv,* 277: 1409.1556, 2014.

Computational methods for breast density analysis

Nasibeh Saffari *

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
nasibeh.saffari@urv.cat

1 Introduction

Breast cancer is the most common cause of death among women which is prevalent in both developed and developing countries. It is one of the cancers which can be diagnosed at early stages and can be prevented if treated appropriately. Early detection of breast cancer improves the treatment process of the patients and increases their chance of survival. Several risk factors for breast cancer have been recognized, such as age, family profile, genetics, and breast density. The amount of fibro glandular tissue content in the breast as estimated mammographically, commonly referred to as breast percent density (PD %), is one of the most significant risk factors for developing breast cancer [1]. Breast density estimation is used to predict the presence of tumours at the early stage, which can help both doctors and radiologists to plan an appropriate treatment either chemotherapy or radiotherapy. Furthermore, the breast imaging reporting and data system standard (BIRADS) [2], presented by the American College of Cancer, provides the following breast density classification:

- BI-RADSI: Almost entirely fatty breast (0–25%).
- BI-RADSII: Some fibro glandular tissue (26–50%).
- BI-RADSIII: Heterogeneously dense breast (51–75%).
- BI-RADSIV: Extremely dense breast (76–100%).

Some of the breast density standards classify the breast tissues into fatty, glandular or dense.

Fig. 1 shows examples of breast density tissues classification in mammograms.

X-Ray mammography is considered as the most popular methods utilized by radiologists for screening and early detection. The common screening mammographic views are craniocaudal (CC) and mediolateral oblique (MLO). The

* PhD advisors: Dr. Domenec Puig, Dr. Blas Herrera

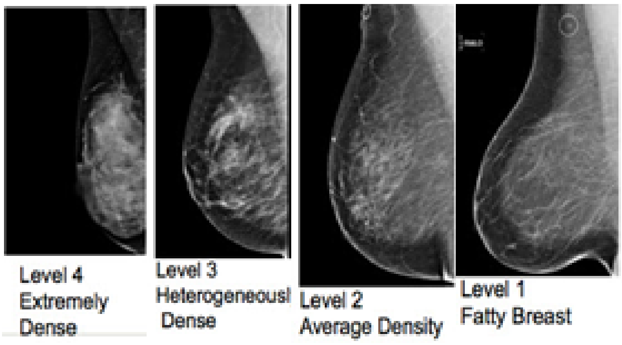


Fig. 1: Breast density tissues classification in mammograms.

CC mammographic view is captured from the superior view of a horizontally compressed breast. In turn, the MLO view is captured from the side of a diagonally compressed breast as illustrated in Fig. 2.

Analysis of mammograms is not easily feasible for every case. However, Computer techniques to diagnose and classify the breast density has attracted the researchers attention in recent decade. Hence, we tried to increase the accuracy of breast density estimation using multiple feature extraction and deep learning methods. Human brain has a unique capability for classification, thus scientists are endeavouring to accurately simulate the classification ability of the human brain using deep learning methods.

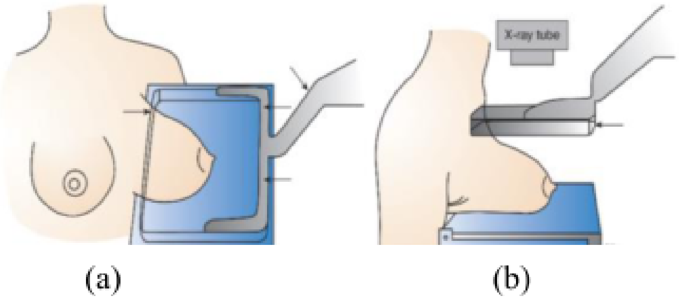


Fig. 2: Different view types in mammograms. (a) MLO view, (b) CC view.

2 Methodology

The overview of proposed method is illustrated in Fig. 3.

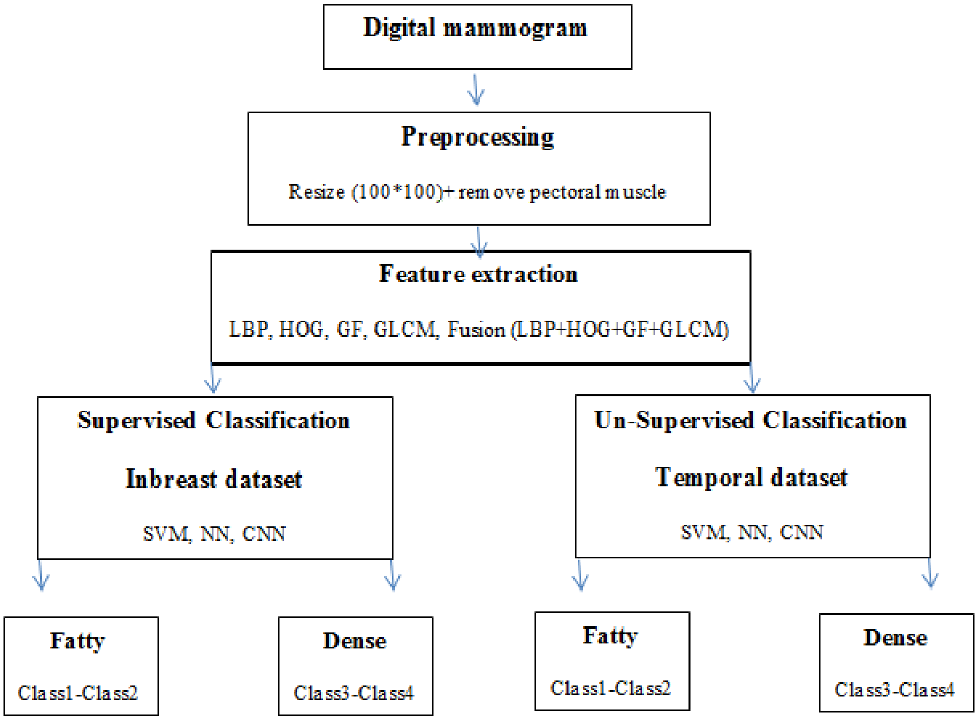


Fig. 3: Preprocessing stage.

2.1 Preprocessing

Resize (100×100)

Using full-size image processing, anomaly sample classification leads to increase the excessive memories and computational cost. To solve this problem, we resized the images.

Remove pectoral muscle

Fig. 4 shows an example of a mammogram image before and after the preprocessing step. We can notice the absence of the pectoral muscle and the labels in the processed image. In addition, the estimated breast boundary is shown in blue colour.

2.2 Feature extraction

Several feature extraction methods have been used, such as histogram oriented gradients (HOG), local binary patterns (LBP), gabor filter (GF) and finally we applied the fusion of HOG, LBP and GF to improve the accuracy.

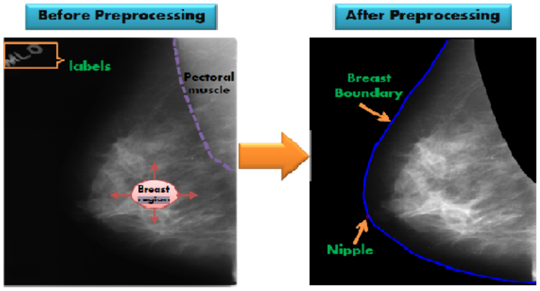


Fig. 4: Preprocessing stage.

2.3 Classification

First, we applied variety of supervised learning classification methods, for instance support vector machines (SVM), k-nearest neighbours (KNN), fuzzy c-means (FCM), radial basis function (RBF) and convolutional neural networks (CNN) on INbreast dataset for creating the model. In the next step, we utilized this model on the temporal images,which do not have labels to analyze the evolution of breast tissue density and classify it to dense or fatty. Finally,we present a novel methodology to create a standard, such as BIRADS for classification the breast density.

Acknowledgement. This work is supported by the Spanish project: DPI2016-77415-R and Marti Franques under the agreement between Universitat Rovira i Virgili and Fundaci Catalunya La Pedrera.

References

[1] B. M. Keller, D. L. Nathan, Y. Wang, Y. Zheng, J. C. Gee, E. F. Conant and D. Kontos. Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation. *Medical Physics*, 39(8):4903—4917, 2012.

[2] S. G. Orel, N. Kay, C. Reynolds and D. C. Sullivan. BI-RADS categorization as a predictor of malignancy. *Radiology*, 211(3): 845–850, 1999.

Process Mining: Extracting Valuable Knowledge from Event Log Data

Edgar Batista [⋆]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
edgar.batista@urv.cat

1 Motivation

The IT revolution has dramatically changed the infrastructures of the organisations and, especially, their information systems. The importance of such systems comes determined by the role that they play in the business processes executed in the organisations. A *business process* consists in a set of activities aiming at accomplishing a certain organisational goal (*e.g.* product or service) for their customers. Due to the rapidly change of services delivery and the fiercely competitive market, business processes need to be controlled, implying the appearance of a management field, called *Business Process Management*, focused on optimising the performance of the business processes from the organisations taking advantage of models to facilitate the identification of bottlenecks.

One of the most complex sectors within a welfare society is the healthcare sector. This sector and, consequently, its underlying business processes operate under a very dynamic and changeable context. Healthcare-related processes have some particular characteristics that rarely happen in processes from other sectors. Therefore, healthcare-related information systems play an interesting role in the gathering of information and the modelling, analysis and optimisation of processes in this particular context. This research fits perfectly under the umbrella of the Smart Health paradigm, defined as “the provision of health services by using the context-aware network and sensing infrastructure of smart cities” [10]. Several contributions have been made towards smart health: detecting wandering behaviour in people with dementia [4] [5], context-aware recommender systems [8] and channel wireless characterisation in medical scenarios [6] [7]. Combining the management of business processes together with the context-awareness that smart health promotes, further improvements can be made in an interesting novel research field: process mining.

[⋆] PhD advisor: Agusti Solanas

2 Process Mining

The increasing amount of the data stored in the information systems of the organisations is partly due to the records of events, as part of the activities performed by the business processes (*e.g.* the schedule of the visit of a patient with a particular doctor, the request of X-ray image or the submission of a certain documentation to another department). All this process-related information is usually stored in structured (or semi-structured) files, called *event log files*. Since these files contain all the actions and details of the business processes executed, the analysis of such files may determine the correctness of the business processes. Nevertheless, event log files might be extremely huge, especially in large organisations with thousands of processes.

Thus, the real challenge is to exploit event data in a meaningful way to obtain knowledge about the organisation, provide insights, identify bottlenecks, anticipate problems and apply appropriate countermeasures. This fact promoted the born of *process mining*. Process mining is a young research discipline that embraces machine learning and data mining techniques on the one hand, and process modelling and analysis on the other hand. The main idea of process mining is to "discover, monitor and improve real processes (*i.e.* not assumed processes) by extracting knowledge from event logs readily available in today's systems" [1].

The state-of-the-art distinguishes between three different types of process mining. The first (and most prominent) process mining type is *discovery*, a technique that produces a process model from an event log without using any a-priori information. The α-algorithm [3], one of the first process mining algorithms, aimed at extracting process models (represented as Petri nets) from event log files by detecting the relationship between events. The second type of process mining is *conformance*, involving the comparison of an existing process model with an event log of the same process. This case is particularly interesting to verify whether the reality (recorded in a log file) and the process model are aligned, and vice versa. The third type of process mining is *enhancement*, aiming at extending or improving an existing process model using information about the process recorded in the event log. This kind of process mining might be useful in highly dynamic environments, such as the healthcare sector, in order to provide up-to-date processes [2].

3 Time-efficient distributed computation for Process Mining

Event log files from organisations might reach large dimensions in short periods of time. Hence, the time required for obtaining valuable information from the event logs is crucial. Based on this assumption, and taking into account the importance of acquiring added-value knowledge rapidly to stand out from competitors, distributed and parallel computation may play a key role to reduce runtimes. Furthermore, the use of distributed programming models, such as MapReduce [9], can facilitate the processing and analysis of large event log files. At this point, the key point is to design a parallel process mining algorithm able to extract meaningful knowledge. Accordingly, this could be possible only if events from log files can be analysed in parallel.

The first step consists in transforming event log data from a (semi-) structured file into a matrix $\mathcal{M}^{n \times p}$, where n represents the number of events (rows) and p the number of characteristics (columns). After some transformations, matrix $\mathcal{M}$ may be similar to Equation 1, where three kinds of characteristics can be found: (1) the temporal characteristic $\vec{t}$, in which the events are ordered ascendantly (from the oldest to the newest), (2) the event characteristic $\vec{e}$ for determining the relationships between events (*e.g.* the identifier of a certain activity), and (3) the grouping characteristic $\vec{a}$ for generating submatrices $\mathcal{M}_i^a$. Grouping the events that are related between them into the same submatrix $\mathcal{M}_i^a$ opens the possibility of a parallel analysis of the relationships between events, in which each node from a distributed computing system evaluates a part of the events ($\mathcal{M}_i^a$).

$$\mathcal{M} = \left(\begin{array}{c|ccc|c|c} a_1 & m_{1,1} & \cdots & m_{1,p-2} & e_1 & t_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ a_n & m_{n,1} & \cdots & m_{n,p-2} & e_n & t_n \end{array} \right) \tag{1}$$

In this preliminary algorithm, the relationship between events is determined by a frequency approximation, providing knowledge on how events happen. This frequency-based approach determines the relative frequencies of succession between events represented by empirical probabilities, such that given an event e_i, the next event is e_j (in other words, compute $P(e_j|e_i)$). Once the algorithm concludes, a visual representation of such frequencies could be depicted as a graph $\mathcal{G} = (V, E)$, where V represents the events from $\vec{e}$ as vertices, and E represents the binary relations of the elements in V with a certain frequency or probability.

4 Conclusions and Future Work

Although process mining is a young research field, several contributions have been performed during the recent years. The need to add value to the products and services that organisations offer leads to the analysis of the event log data recorded in their information systems. The large dimensions that such files may achieve, conduct to the use of new techniques and programming paradigms that minimise the computational cost and time. Applying distributed and parallel models to process mining algorithms might facilitate the early extraction of knowledge.

Future research will extend the algorithm presented in this paper, as well as validating its robustness and efficiency on process discovery using distributed computing systems. Moreover, this system aims to be implemented in the Business Intelligence tool of an organisation from the healthcare sector.

Acknowledgement. The author is supported by an Industrial PhD grant (2017-DI-002) from the Generalitat de Catalunya.

References

[1] W.M.P. van der Aalst. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer Publishing Company, Berlin, 2011.

[2] W.M.P. van der Aalst, A. Adriansyah, A.K.A. de Medeiros, F. Arcieri, T. Baier, T. Blickle, and others. Process Mining Manifesto. *Business Process Management Workshops*, 169–194, 2012.

[3] W.M.P. van der Aalst, T. Weijters, Laura Maruster. Workflow Mining: Discovering Process Models from Event Logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1128–1142, 2004.

[4] E. Batista, F. Borras, F. Casino, A. Solanas. A study on the detection of wandering patterns in human trajectories. *Proc. 6th. Int. Conference on Information, Intelligence, Systems and Applications*, 1–6, Corfu, Greece, 2015.

[5] E. Batista, F. Casino, A. Solanas. Wandering detection methods in smart cities: Current and new approaches. *Proc. 1st. IEEE Int. Smart Cities Conference*, 1–2, Guadalajara, Mexico, 2015.

[6] F. Casino, L. Azpilicueta, P. Lopez-Iturri, E. Aguirre, F. Falcone, A. Solanas. Hybrid-based optimization of wireless channel characterization for health services in medical complex environments. *Proc. 6th. Int. Conference on Information, Intelligence, Systems and Applications*, 1–6, Corfu, Greece, 2015.

[7] F. Casino, L. Azpilicueta, P. Lopez-Iturri, E. Aguirre, F. Falcone, A. Solanas. Optimized Wireless Channel Characterization in Large Complex Environments by Hybrid Ray Launching-Collaborative Filtering Approach. *IEEE Antennas and Wireless Propagation Letters*, 16:780–783, 2017.

[8] F. Casino, E. Batista, C. Patsakis, A. Solanas. Context-aware recommender for smart health. *Proc. 1st. IEEE Int. Smart Cities Conference*, 1–2, Guadalajara, Mexico, 2015.

[9] J. Dean, S. Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[10] A. Solanas, C. Patsakis, M. Conti, I.S. Vlachos, V. Ramos, F. Falcone, and others. Smart Health: A Context-Aware Health Paradigm within Smart Cities. *IEEE Communications Magazine*, 52(8):74–81, 2014.

Learning the Insertion and Deletion Edit Costs for Graph Matching

Shaima Algabli [*]

Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
Tarragona, Spain
shaima.ahmed@estudiants.urv.cat

1 Introduction

We present a learning method to automatically deduce a penalty cost to each edit operation based on a ground truth graph matching. It uses a linear classifier applied to a bi-dimensional space in which each element represents a node-to-node mapping. The computational cost is cubic with respect to the number of pairs of graphs. Moreover, the algorithm does not need to compute any graph matching, which is the main obstacle of other methods due to its intrinsic exponential computational complexity. Experimental validation shows that the matching accuracy and the recognition ration of this method outperforms the current methods. Furthermore, there is a significant reduction in the runtime in the learning process.

2 Learning Graph Edit Distance

In [2], the human (or another system) interacts with the automatically obtained matching between nodes and imposes a new node-to-node mapping. Then, the method considers this new mapping and updates the whole matching. The method presented in [5] is the only one in the literature that learns the insertion and deletion edit cost on nodes and edges as constants (a Real number). These costs can be directly applied as the input parameters of the matching algorithms such as [1], [4] and [3]. It is an optimisation method that aims to minimise the distance between the edit cost obtained by the ground truth matching and the edit cost obtained by the current matching. Our method is compared to this one in the experimental section.

[*] PhD advisor: Francesc Serratosa

3 Defining two coordinate systems

Suppose that we have two bi-dimensional coordinate systems defined by axes $(\dot{x}, \dot{y})$ and $(\ddot{x}, \ddot{y})$. Moreover, suppose that a substitution (first option) of node v_i^p by node v_j^q or a deletion of node v_i^p (second option), (which is the same as substituting v_i^p to a null-node v_i^p) are represented by a point $(\dot{x}_{(i,j)}, \dot{y}_{(i,j)})$ in the coordinate system $(\dot{x}, \dot{y})$. Similarly, suppose that a substitution (first option) of node v_i^p by node v_j^q or a insertion of node v_i^q (third option), (which is the same as substituting a null-node v_i^p to v_i^p) are represented by a point $(\ddot{x}_{(i,j)}, \ddot{y}_{(i,j)})$ in the coordinate system $(\ddot{x}, \ddot{y})$.

Figure 1 schematically shows the position of mappings in the first option (S: substitute), the second option (D: deletion) and the third option (I: insertion). We also show the borders $C_(i,j) - C_(i,\varepsilon) = 0$ in the $(\dot{x}, \dot{y})$system and $C_(i,j) - C_(\varepsilon, j) = 0$ in the $(\ddot{x}, \ddot{y})$ system. Our learning method is based on finding each border in the coordinate systems $(\dot{x}, \dot{y})$ and $(\ddot{x}, \ddot{y})$ assuming these borders are a line. Then, parameters of K_v and K_e are deduced through the offset and slope of these lines. In the next section, we show how we define the coordinate systems $(\dot{x}, \dot{y})$ and $(\ddot{x}, \ddot{y})$ and then how the values of K_v and K_e are modelled.

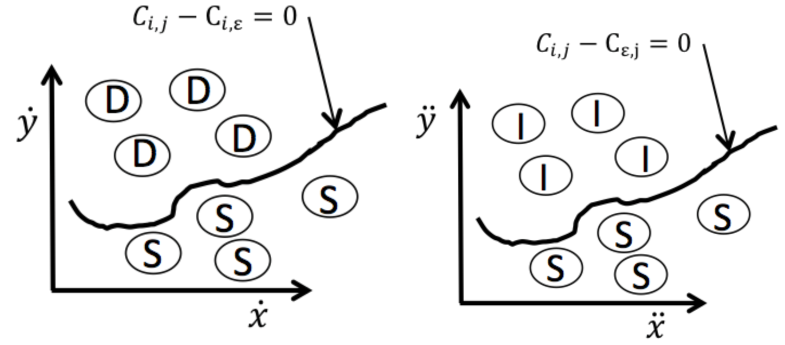

$C_{i,j} - C_{i,\varepsilon} = 0$
$\dot{y}$
D
D
D
D
S
S
S
S
$\dot{x}$
$C_{i,j} - C_{\varepsilon,j} = 0$
$\ddot{y}$
I
I
I
I
S
S
S
S
$\ddot{x}$

Fig. 1: Positions of mappings in the three options (S): Substitution, (D): deletion and (I): insertion and also two plausible borders between both sets given the coordinate systems $(\dot{x}, \dot{y})$ and $(\ddot{x}, \ddot{y})$.

3.1 Learning K_v and K_e

Since the border between elements of the first option$(\dot{x}, \dot{y})$ is $C_{(i,j)} - C_{(i,\varepsilon)} = 0$, it is already demonstrated that the cost of edit two stars is as below:

- Substituting two stars: $C_{(i,j)} = C_{vs}(v_i^p, v_j^q) + T_{ij} \cdot (K_v + K_e) + \breve{C}_{vs}(v_i^p, v_j^q)$
- Deleting a star: $C_{(i,\varepsilon)} = K_v + n_{N_i^p}(K_v + K_e)$
- Inserting a star:$C_{(\varepsilon,j)} = K_v + n_{N_j^q}(K_v + K_e)$

So

$$C_{vs}(v_i^p, v_j^q) + T_{ij} \cdot (K_v + K_e) + \breve{C}_{vs}(v_i^p, v_j^q) - (K_v + n_{N_i^p} \cdot (K_v + K_e)) = 0$$

Rearranging the terms we arrive at the expression,

$$\frac{C_{vs}(v_i^p, v_j^q) + \breve{C}_{vs}(v_i^p, v_j^q)}{n_{N_i^p} + 1 - T_{ij}} = K_e \cdot \frac{(T_{ij} - n_{N_i^p})}{T_{ij} - n_{N_i^p} - 1)} + K_v \tag{1}$$

Thus, given the mapping of node v_i^p to node v_j^q, which represents a substitution (both nodes are non-nulls) or it represents a deletion (the first one is non-null and the second one is a null node) then this mapping can be represented as a point in the coordinate system $(\dot{x}, \dot{y})$ as follows,

$$y_{(i,j)} \equiv \frac{C_{vs}(v_i^p, v_j^q) + \breve{C}_{vs}(v_i^p, v_j^q)}{n_{N_i^p} + 1 - T_{ij}} \tag{2}$$

$$x_{(i,j)} \equiv \frac{n_{N_i^p} - T_{ij})}{n_{N_i^p} + 1 - T_{ij}} \tag{3}$$

In this way, considering equations 1, 2 and 3, the border between mappings in the first and the second option can be approximated as the line in the coordinate system $(\dot{x}, \dot{y})$

$$y = \dot{K}_e x + \dot{K}_v$$

Similarly, the coordinate system $(\ddot{x}, \ddot{y})$ is defined considering mappings in the first and third option. Thus the border between these options is

$$y = \ddot{K}_e x + \ddot{K}_v$$

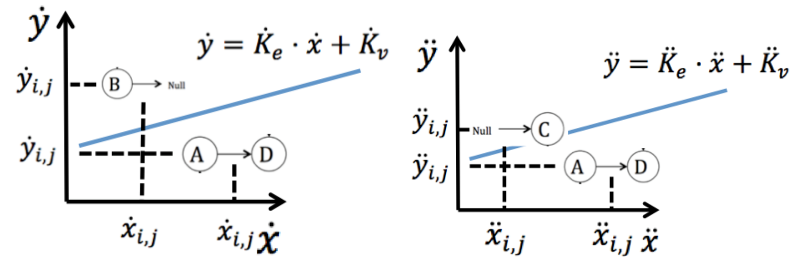


Fig. 2: Representation of the star substitution, deletion and insertion of the example in systems $(\dot{x}, \dot{y})$ and $(\ddot{x}, \ddot{y})$.

4 Conclusion

From the practical point of view, our method has three main advantages with respect to the one in the literature. First, it does not have parameters to be tuned, so as a weighting parameter to gauge the regularisation term. Second, it is not necessary to impose initial values of K_v and K_e. And third, the learning runtime is really much faster. We could consider as a drawback the need of having a ground truth correspondence. Nevertheless, in our databases, we have seen that the learned costs were almost the same with few samples than the whole learning set. This is because both classes in the domains $(\dot{x}, \dot{y})$ and $(\ddot{x}, \ddot{y})$ were nicely clustered. Only in the Rotation-Zoom database, classes were mixed up when the substitution weights w_v^t were not learned. Nevertheless, this problem was partially solved when the learned weights were applied.

Acknowledgement. This research is supported by the Spanish projects "TIN2016-77836-C2-1-R" and ColRobTransp MINECO "DPI2016-78957-R" AEI/FEDER EU.

References

[1] A. Fischer, C. Y. Suen, V. Frinken, K. Riesen and H. Bunke, Approximation of graph edit distance based on Hausdorff matching. *Pattern Recognition*, 48(2):331–343, 2015.

[2] F. Serratosa and X. Cortés, Interactive Graph-Matching using Active Query Strategies. *Pattern Recognition*, 48(4):1360–1369, 2015.

[3] F. Serratosa, Fast Computation of Bipartite Graph Matching. *Pattern Recognition Letters*, 45: 244–250, 2014.

[4] K. Riesen, H. Bunke, Approximate graph edit distance computation by means of bipartite graph matching. *Image and Vision Computation*, 27(7): 950–959, 2009.

[5] X. Cortés and F. Serratosa, Learning Graph-Matching Edit-Costs based on the Optimality of the Oracle's Node Correspondences. *Pattern Recognition Letters*, 56:22–29, 2015.

This proceeding book contains the contributions presented at the 4th URV Doctoral workshop in Computer Science and Mathematics. The main aim of this workshop is to promote the dissemination of the ideas, methods and results that are developed by the students of our PhD program.

[DΣIM] Departament d'Enginyeria
Informàtica i Matemàtiques

ETSE Escola Tècnica Superior
d'Enginyeria
Universitat Rovira i Virgili