PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Alfonso Cimasa

XXXII Cycle

Training and Research Activities Report – First Year

Tutor: prof. Anna Corazza

Information

PhD Candidate: Alfonso Cimasa

MSc Title: Master's degree in Computer Science (cum laude), University of Naples Federico II

Doctoral Cycle: XXXII – ITEE- University of Naples Federico II

Tutor: Prof. Anna Corazza

Study and Training activities

Courses

1. Ad hoc Module, "Le Imprese e la ricerca", lecturer: Dott. Marco Frizzarin, February, (4 CFU)
2. Ad hoc Module, "Introduction to Artificial and Computational Intelligence", lecturer: Prof. Fernando Buarque, February, (3 CFU)
3. Ad hoc Module, "Machine Learning", lecturer: Prof. Donatello Conte, Prof. Anna Corazza, Prof. Francesco Isgrò, Prof. Roberto Prevete, Prof. Carlo Sansone, May, (4 CFU)
4. Ad hoc Module, "Satellite Remote Sensing: Open challenges and opportunities", lecturer: Prof. Giuseppe Ruello, October, (3 CFU)
5. MS course, "Network Security", January, (6 CFU)

Seminars

1. *Cognitive computing and the da Vinci robot: research proposal and discussion*, organized by Prof. Bruno Siciliano, Università di Napoli Federico II, February 17th 2017, (0.20 CFU)
2. *IBM Cognitive computing: challenges and opportunities in building an artificial intelligence platform for business,* organized by Prof. Bruno Siciliano, Università di Napoli Federico II, February 17th 2017, (0.40 CFU)
3. *From mathematical formalization to artificial Visual-Attention: toward a human-like robot vision,* organized by Prof. Antonio M. Rinaldi, Università di Napoli Federico II, April 4th 2017, (0.40 CFU)
4. *Living bots and alter ego,* organized by Prof. Guglielmo Tamburrini, Università di Napoli Federico II, April 4th 2017, (0.40 CFU)
5. *Exploiting speech production knowledge for deep-learning based automatic speech recognition*, organized by Prof. Francesco Cutugno, Università di Napoli Federico II, May 9th 2017, (0.40 CFU)
6. *Using process mining and cloud technologies for dependability,* organized by Prof. Marcello Cinque, Università di Napoli Federico II, September 07th 2017, (0.40 CFU)
7. *Wireless opportunistic networking,* organized by Prof. Giorgio Ventre, Università di Napoli Federico II, September 28th 2017, (0.30 CFU)

8. *Effective machine learning in the time of Big Data,* organized by Prof. Anna Corazza, Università di Napoli Federico II, November 28th 2017, (0.20 CFU)
9. *Medical imaging: why not let the data speak for themselves?,* organized by Prof. Anna Corazza, Università di Napoli Federico II, November 28th 2017, (0.40 CFU)

Summer school

1. *Machine Learning Crash Course,* organized by Prof. Lorenzo Rosasco, Università degli studi di Genova, from 26th to June 30th 2017 (18 hours), (2 CFU)

Research Activity

Title

Development and testing of privacy preserving data mining (PPDM) algorithm

Research description

The entry into force of General Data Protection Regulation (GDPR - EU Regulation 2016/679) with which the European Commission intends to reinforce and make more homogeneous the protection of personal data of EU citizens and residents of the European Union (EU), creates new challenges for data scientist community and corporates. My research aims to evaluate and address the impact of new privacy requirements on automatic data processing procedures by focusing on new data mining approach named "*privacy preserving data mining*" (PPDM): these methods allow the knowledge extraction from data while preventing the disclosure of sensitive information.

The research activity (of the first year) is divided in two main stages: first, a long phase of review about literature on (pseudo)anonymization techniques of data has been carried out; this led me to categorize the most commonly used privacy models:

- Generalization: consisting in generalizing, or diluting, the attributes of data subjects (person in a dataset) by modifying the respective scale or order of magnitude (i.e. a range of ages instead the age);
- Randomization (data perturbation): consisting in altering the data, by adding noise, in order to remove the strong link between the data and the individual;
- Suppression: consisting in removing of some attribute values to prevent information disclosure.

Based on these operations, we identified and analyzed a set of privacy: ***k*-anonymity, *l*-diversity** and **ε-*differential privacy.*** Our analysis about literature led us to reject k-anonymity due to the fact that sensitive attributes are not taken into consideration when forming the k-anonymized dataset. This may lead to equivalent classes where the values of some sensitive attributes are equal for all the k records and consequently, disclosure of private information of any individual belonging to such groups. L-diversity model expands the k-anonymity model by requiring that every equivalence class is a set of entries such that at least l ''well-represented'' values exist for the sensitive attributes. The significance of "well-represented" is not so clear, not a concrete definition, that is why exist many instantiation of this model. One of the simplest instantiations considers that the sensitive attributes are ''well-represented'' if there are at least "l" distinct values in an equivalence class. Although the l-diversity model increases the diversity of sensitive values within

equivalence classes, it does not take into consideration the distribution of such values. This may present privacy breaches when the sensitive values are distributed in a not uniform way, which is generally true; so this leads to probabilistic inference attacks. The last model we considered is a randomization model, the ε-differential privacy. Presented by Dwork, the notion of differential privacy measure the difference, on individual privacy disclosure, between the presence and the absence of the individual's record. In this way, a person's privacy will not be affected by participating in the data collection since it will not make much difference in the outcome. Differential privacy measure privacy risk by a parameter ε that bounds the log-likelihood ratio of the output of a (private) algorithm under two databases differing in a single individual's data. When ε is small, the inferences that an adversary can make observing the output of the algorithm will be similar regardless of whether that individual is in the data set or not.

Because of the strong guarantees about the probability of information disclosure given by differential privacy framework, we decided to focus our attention on it.

Once we selected the privacy model, in the second phase of my research activity we focused on how to adopt DP framework for data mining tasks. In particular, I started a process of evaluation of (DP) compatible algorithms that can be summarized in three possible approaches (fig.1 for a brief scheme of possible noise adding):

- Input perturbation;
- Output Perturbation;
- Objective (or internal) Perturbation.

After some investigations on *output perturbation* and *objective (or internal) perturbation*, the research is now oriented on *input perturbation* techniques and in particular achieving DP requirements by dimensionality reductions. We made some preliminary experiments using Private PCA, a mechanism by Dwork et al., testing the impact of noise addition to the utility of data (where the utility, in our case, stands for the accuracy supervised learning task, Support Vector Machine (SVM)). Our objective is to minimize the data utility loss and, at same time, minimize the disclosure risk in the anonymized data.

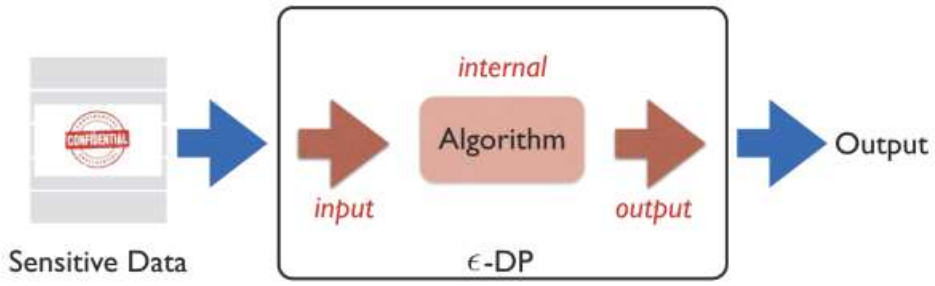

Figure 1 Private preserving data mining

Title

Assessing comment coherence using Word Embeddings

Research description

During software development, code commenting does not represent a crucial phase and are always neglected. Code comments are useful for several purposes. A code comment can explain what a particular function does, explain something that might not be obvious to the reader and serve as a reminder to change something in the future. Many studies proved a direct correlation between bad comments, bugs and other issues into code. In order to provide a tool to help software engineer and tester, we decided to investigate on automatic methods to assess comment-code coherence and in particular, we adopted some innovative techniques in word representation, called *word embeddings*, and tested their impact on classification accuracy.

In our study, we adopted embedding techniques like Word2Vec and GloVe, in contrast to popular term-weighting schemes: Tf-IDf (Term frequency – Invers Document frequency) with Bag of Word document representation. Furthermore, leveraging spatial properties of embedding techniques, we designed four documents models to enhance the difference between coherent and non-coherent couples (see Figure 2).

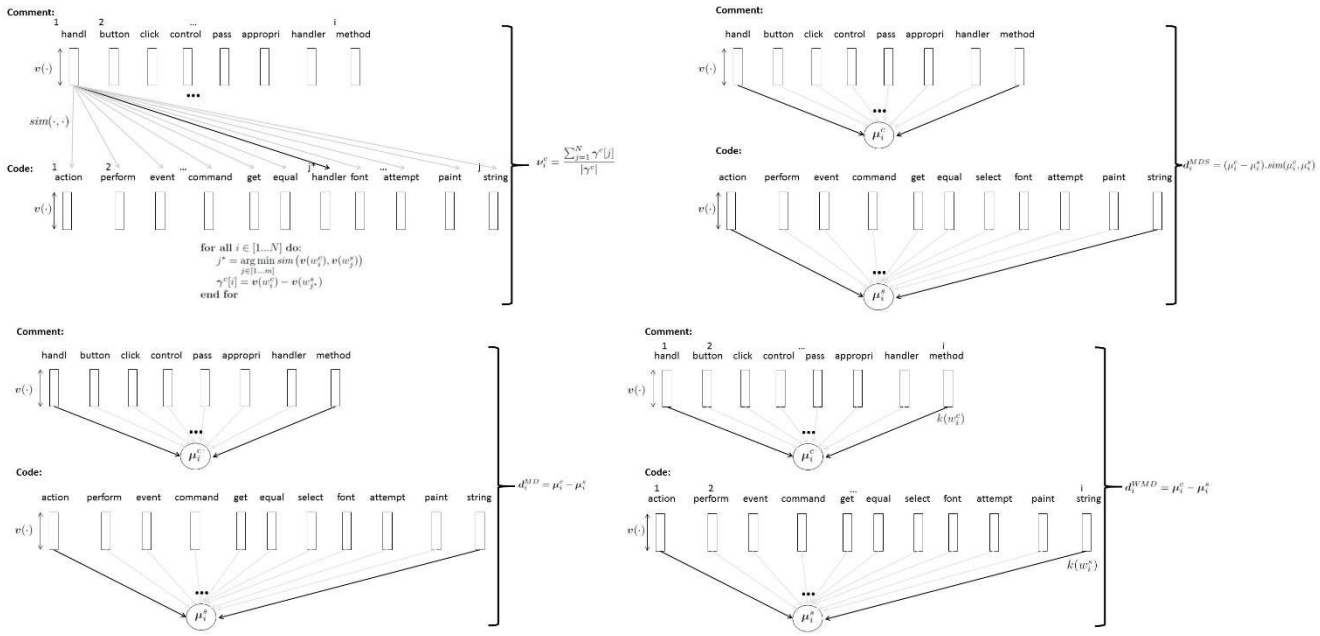

Figure 2 Documents representation models

We have chosen three different classifier, Support Vector Machine, Random Forest and K-Nearest Neighbour, and made several experiments to state a classification baseline and test our approaches. Our experiments

showed that, despite we were not able to exceed the baseline (Bag of Words with Tf-IDf), we achieved similar results by using an half of feature rather than the baseline: this was a huge improvement in terms of space needed to encode documents without losing so much in terms of classification accuracies.

We made several experiments to assess the best model and embedding couple. In particular, non-parametric statistical significance test (Friedmann test) to prove statistical significance.

This work is going to the end and now we are finalizing the experiments presentation phase and writing the conclusion of our study.

Product

1. Cimasa A., Corazza A., Scanniello G., *"Word Embeddings to decide Comment Coherence", - in preparation*

Tutorship

- "Laboratorio di programmazione"
  - Type: Course
  - Subject: C programming
  - Hours: 31 (ongoing)

Credits Summary

	Credits year 1							
		1	2	3	4	5	6	
	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary
Modules	20	7	4			3	6	**20**
Seminars	5	0,6	1,2	2	0,7	0,6		**5,1**
Research	35	4	6	6	5	7	6	**34**
	60	12	11	8	5,7	11	12	**59**

Year	Lecture/Activity	Type	Credits	Certification	Notes
1	Introduction to Artificial and Computational Intelligence	Ad hoc module	3	x	
1	Le Imprese e la Ricerca	Ad hoc module	4	x	
1	Machine Learning	Ad hoc module	4	x	
1	Satellite Remote Sensing: Open challenges and opportunities	Ad hoc module	3	x	
1	Network Security	MS Module	6	x	

1	Cognitive computign and da Vinci Robot: Research proposals and discussions	Seminar	0,2	x	
1	IBM Cognitive Computing: Challenges and Opportunites in building an anrtificial intelligence platform for business	Seminar	0,4	x	
1	Form mathematical formalization to artificial visual-attention: toward a human like robot vision	Seminar	0,4	x	
1	Living Bots and Alter ego	Seminar	0,4	x	
1	Exploiting Speech Production Knowledge for Deep Learningbased Automatic Speech Recognition	Seminar	0,4	x	
1	Using Process Mining and cloud technologies for dependability	Seminar	0,4	x	
1	Wireless Opportunistic Networking	Seminar	0,3	x	
1	Machine Learning Crash Course	External Seminar-Summer school	2	x	Link
1	Medial Imaging: why don't let data speack for themselves?	Seminar	0,4	x	
1	(Effective) Machine Learning in the time of big data	Seminar	0,2	x	