ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Automatic Crime Report Classification through a Weightless Neural Network

Rafael Adnet Pinho, Walkir A. T. Brito, Claudia L. R. Motta
and Priscila Vieira Lima

Federal University of Rio de Janeiro (UFRJ)
Pos-Graduation Program in Informatics (PPGI), Rio de Janeiro, RJ - Brazil
(rafaadnet, walkir.brito)@gmail.com, (claudiam, priscila.lima)@nce.ufrj.br

Abstract. Anonymous crime reporting is a tool that helps to reduce and prevent crime occurrences. The classification of the crime reports received by the call center is necessary for the data organization and also to stipulate the importance of a particular report and its relation to others. The objective of this work is to develop a system that assists the call center's operator by recommending classification to new reports. The system uses a weightless neural network that automatically attribute a class to a report. At the end of this work it was possible to observe that automatic classifications of crime reports with high accuracy are possible using a weightless neural network.

1 Introduction

The Brazilian state of *Rio de Janeiro* receives more than 100,000 anonymous crime reports each year through the main crime report call center named *Disque Denuncia*, this is a program pretty similar to the American *Crime Stoppers*. The information of those crime reports is extremely important to reduce crime levels by helping the police and other authorities to realize more efficient operations against criminals.

Such important data is constantly analyzed and thousand of studies and results are derived from it. However, all the reports are still been classified only by the call center's operators without any help of computer intelligence, leaving space for classifications mistakes. The text classification is a process of labeling text documents into one or more pre-specified categories. Due to the vast majority of web pages, emails, social networking information, and even corporate information that is available in digital form, automatic text classification has gained attention in many research domains. For example, it has been used to categorize newspaper articles into topics [1], and classify network intrusion attacks as Positive and Negative [2]. Current classification methods such as decision trees, k-nearest neighbors, neural networks, support vector machines, and Naïve Bayes have been successfully used in automated text classification. However, most similarity algorithms are not specially developed to compare and contrast crime reports, but general text reports. In addition, none of them, has used *Weightless Neural Networks* to perform this type of classification.

This work intends to improve the task of crime report classification by developing a Automatic Crime Report Classifier(ACRC) that uses computational intelligence in the report's text to assist the human classification process making

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

it more assertive. The automatic classification process was made using *Weightless Neural Networks*, trained with texts of real crime reports.

The results of this paper show that machine learning techniques such as *Weightless Neural Networks* can be very efficient on finding crime reports text similarities and consequently to the report's classification.

The remainder of this text is organized as follows: Section 2 explains the fundamental knowledge to understand the paper, Section 3 describes the dataset used and the built system, Section 4 presents the system results and Section 5 concludes this article and points at further research topics.

2 Fundamentals

It's essential to understand how the crime reports structure is divided. As illustrated by the figure 1 the crime report has three fundamental parts, "Report Description", "Main Classification" and "Secondary Classifications".

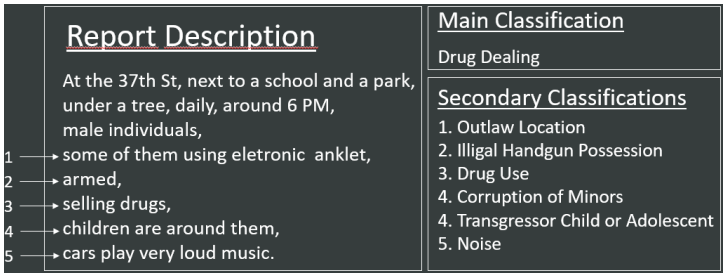

Fig. 1: Report Structure.

The *Report Description* is the denounce received by the call-center typed in natural language text by the operator, often describing the scene that was observed by the denunciator in which it is believed that a delict was being committed. The *Main Classification* defines what is the principal delict by the operator's perspective. The last part is the *Secondary Classification*. This characteristic of the report structure makes possible to assign other classifications to the report that were observed on the *Report Description*. The *Secondary Classification* is really important because it can determinate what authorities needs to be involved, it can change the importance of the report and help to determinate what kind of action needs to be take. To make the report's text usable to computers some *natural language processing* (NLP) techniques were used.

NLP is a field that intersects with artificial intelligence and linguistics. NLP techniques are frequently used to explore how computers can process and understand natural language text or speech. Generally, any system that uses NLP includes tasks such as tokenization, sentence splitting, phrase segmentation and information extraction [3]. The first three tasks are low-level tasks used to identify words, phrases, and sentences (described in the subsection 3.1), while the

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

last task is a high-level task used to extract relevant information in a domain.

As a large training dataset was available for this work, the information extraction was based on a neural network approach, using a weightless neural network (WNN) called WiSARD. Among all WNN, the WiSARD was chosen because it's the most representative model and it was the first to be produced as a machine for a commercial propose [4]. WiSARD is an acronym of *Wilkie, Stonham and Alksander's Recognition Device*, that is a weightless and multi-discriminator network that was originally used for image recognition [4]. In brief, the WiSARD is *online* (the training is not strictly separated from the recognition and they can alternate), parallelizable (it's possible to train different patterns at the same time, as long as they are different) and fast training neural model (it is possible to train it on a single round).

The WiSARD has been widely used in many applications for patterns recognition such as HIV-1 sub-types categorization [5], music real-time tracking [6] and credit analysis [7]. It works creating random relations on the original data and them checking its existence during the training and recognition phases. Every new pattern presented to the WiSARD is represented as a discriminator that has a set of RAMs. The function of the RAMs is to store what are the random relations that were seen during the training phase and to respond if a specific relation exists during the recognition phase. A discriminator is a structure that is basically responsible for the recognition of a single class among all others available classes. Each class has a maximum of 2^n neurons (or *RAMs*) where n is the number of bits that each of the neurons will have. The weightless neuron, or RAM, is a tuple <Address, Boolean> that flags if a specific part of a pattern was seen by the discriminator.

3 Methodology

3.1 Dataset and Text Processing

The dataset used in the experiments of this paper is a real dataset made of crime reports collected by *Disque Denuncia* during the year of 2015. It has 107,226 reports, each one with a text description, a main classification and zero or more secondary classifications. All addresses and sensitive information were hidden for security reasons.

Many techniques were used to treat the texts of the reports before using the WiSARD classifier. Basically five tasks were applied to the text of the reports: tokenizer, normalization, stopwords removal, sentence splitter and stemmer. The tokenization phase segments the text into individual tokens such as words and punctuation. The stemmer identifies the main part of tokens. For example, the stem of 'selling' and 'sales' is 'sale'. The stopwords removal is a process to remove words that are not relevant to the similarity measure between different texts, such as 'a', 'an' and 'the', it also removes duplicated entities and keeps relevant entities. Some specific words were manually added to the stopwords list because they appeared in the majority of the reports texts making them irrelevant for the text classification, for example, the word 'street' that appears

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

in 87% of the reports.

3.2 ACRC system

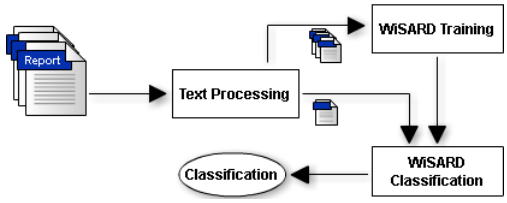

Fig. 2: ACRC system.

The Automatic Crime Report Classifier system is divided into three basic layers as shown by figure 2. The Text Processing layer is where all the texts are treated. The technical details of this process are described in section 3.1. On the WiSARD Training layer, a part of the crime reports are used to train the WiSARD with different types of crime classifications. The last layer is the WiSARD Classification where all other reports that were not used in the training step are submitted to the system classification through the trained WiSARD network. The ACRC code is available online [8] and all the details about the experiment are explained in the following section.

4 Experimental results

The first part of the experiment consisted of the selection of classes to be used since the 20 biggest classes contain 91% of all reports, they were the selected classes to be used. The counting of those classes can be found at table 1. On the experiment the k-fold cross-validation [9] was used with a k value of 10.

After the text-processing of all training reports, it was made a matrix that each line was a report, and each column represented a specific word from the whole group of unique words present in all training reports, in other words, each cell of the matrix represents the number of times that a word appears in a report.

The PyWANN [10] was the WiSARD implementation used, it's an open source stable version of the WiSARD developed in Python for general use. The WiSARD has been configure with 8 RAMs, each one using 3 bits of memory addressing. The Bleaching technique was used to overcome the saturation problem of the WiSARD [11]. With Bleaching, instead to mark if a certain part of the class pattern was observed or not, it counts how many times it was observed, a threshold value is set to determinate if the specific part of the pattern that is been analyzed will be used or not. The Bleaching is generally a dynamic threshold mechanism but in this experiment was used a simpler and easy-to-use version of the Bleaching calibrated with a fixed threshold value of 3.

As mentioned, 10% of each class reports were used to check the accuracy of the trained WiSARD. All text of the test reports were also treated by the same

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

Classification	Number of Reports	Accuracy (%)
Drug Dealing	37261	64%
Noise	8908	80%
Ilegal use of public service	5555	95%
Animal abuse	4319	91%
Extortion	3982	63%
Ilegal handgun possession	3756	72%
Abandoned Vehicles	3345	70%
Fraud	3216	63%
Pedestrian theft	3091	59%
Outlaw location	2651	79%
Drug use	2624	38%
Ilegal gambling	2569	74%
Violence against the elderly	2549	76%
Mistreatment	2460	45%
Store without license	2411	56%
Murder	1858	58%
Closure of public roads	1846	74%
Theft of automotive vehicles	1808	87%
Violence against women	1795	75%
Badern	1517	31%

Table 1: Classification Results

text-processing techniques described in section 3.1. Then the test reports were submitted to the trained WiSARD classification and one class was attributed by the WiSARD for each of the reports. Then the attributed class was compared with the classifications that were manually attributed by humans of the *Disque Denuncia*, considered the correct ones. The results detailed in table 1 shows that only 3 classes (Drug use, Mistreatment, Badern) had a correct classification rate (accuracy) below 50%. That is a comprehensive result since those classes content are related to others, for instance, on a logical thought is hard to imagine a "Violence against woman" scenario without the woman been mistreated. The same thought applies to other relations such as "Drug use" related to "Drug Dealing" and "Badern" related to "Noise".

The training set had a mean (μ) of 0.674, a standard deviation (σ) of 0.167 and a coefficient of variation (CV) of 0.25. Meaning that 75% of the results are concentrated around the mean (μ).

5 Conclusions

The experiments showed that weightless neural networks like the WiSARD can contribute to the classification of crime reports. Nevertheless, is recommended the use of this approach as an auxiliary process to improve the human classification, helping the call-center operator to classify the crime reports by reminding

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

them of a possible secondary classification that was not observed.

A limitation to this work was identified since the system might occasionally classify two reports describing different crimes as the same crime due to a high similarity of the text. This is because there is always some overlap in reports, e.g., a report of a murder that 90% of the text is used to describe the murder threat and only 10% to describe the actual murder, will be probably wrongly classified as "Murder Threat". More advanced text mining and classification techniques may be helpful in making the distinction between the different crimes. Also, the ACRC system is hard to configure and complex to use by people that are not familiar with Python code language. To make the system more practical and useful to law enforcement agencies, it must have an easy and user-friendly interface for interaction, this is part of a larger ongoing project.

In the future, we plan to try more different combinations on the fine-tune of the WiSARD and compare the WiSARD accuracy and performance to other machine learning methods from the state-of-art (e.g. decision trees). Future works also include the use of the addresses of the reports inside the report's description, in order to improve the classification accuracy by making the reports localization an impact factor for the classification and police statistical demographic analysis.

References

[1] G Mamakis, A G Malamos, and J A Ware. An alternative approach for statistical single-label document classification of newspaper articles. *Journal of Information Science*, page 0165551511403543, 2011.

[2] D C Wyld. Rubee: applying low-frequency technology for retail and medical uses. *Management Research News*, 31(7):549–554, 2008.

[3] X Wang, L Zhang, T Xie, J Anvik, and J Sun. An approach to detecting duplicate bug reports using natural language and execution information. In *Proceedings of the 30th international conference on Software engineering*, pages 461–470. ACM, 2008.

[4] I. Aleksander, W.V. Thomas, and P.A. Bowden. Wisard: a radical step forward in image recognition. *Sensor Review*, 4(3):120–124, 1984.

[5] C Souza, F Nobre, P M V Lima, R Silva, R Brindeiro, and F M G França. Recognition of hiv-1 subtypes and antiretroviral drug resistance using weightless neural networks. 2012.

[6] D F P De Souza, F M G Franca, and P M V Lima. Real-time music tracking based on a weightless neural network. pages 64–69, 2015.

[7] D de O Cardoso, D S Carvalho, D S F Alves, D F P de Souza, H C C Carneiro, C E Pedreira, P M V Lima, and F M G França. Credit analysis with a clustering ram-based neural classifier. In *ESANN*, 2014.

[8] R Adnet Pinho. Acrc system, www.github.com/radnet/acrc, 2016.

[9] P Refaeilzadeh, L Tang, and H Liu. Cross-validation. In *Encyclopedia of database systems*, pages 532–538. Springer, 2009.

[10] F Firmino. Pywann, www.github.com/firmino/pywann, 2016.

[11] D S Carvalho, H C C Carneiro, F M G França, and P M V Lima. B-bleaching: Agile overtraining avoidance in the wisard weightless neural classifier. 2013.