Automated Free Text Classification of Economic Activities using VG-RAM Weightless Neural Networks

Alberto F. De Souza[1], Felipe Pedroni[1], Elias Oliveira[2],
Patrick M. Ciarelli[3], Wallace F. Henrique[1], Lucas Veronese[1]
[1] *Departamento de Informática*
[2] *Departamento de Ciência da Informação*
[3] *Departamento de Engenharia Elétrica*
Av. Fernando Ferrari s/n
29060-970 – Vitória- ES, Brazil
{alberto, elias}@lcad.inf.ufes.br

Abstract

We tackle the problem of automating the categorization of companies according to their economic activities using business descriptions in free text format as input. This categorization is vital to fundamental aspects of national governmental administration such as short, medium and long term planning and taxation. As the number of categories considered is very large (more than 1000 in the Brazilian scenario), the automatic text categorization problem targeted here is challenging. We have applied and compared the use of two different techniques to deal with it: the Vector Space Model, a well known text categorization technique; and Virtual Generalizing Random Access Memory Weightless Neural Network, or VG-RAM WNN. To our knowledge, this is the first report on using VG-RAM WNN for text categorization.

1. Introduction

Automatic text classification and clustering are still very challenging computational problems to the information retrieval (IR) communities both in academic and industrial contexts. Currently, the majority of the work on IR one can find in the literature is focused on classification and clustering of webpages. However, there are many other important applications to which little attention has hitherto been paid, which are as well very difficult to deal with. One example of these applications is the classification of companies based on their statement of purpose, also called mission statement, which represent the business context of the companies' activities. The categorization of companies according to their economic activities is an important step of the process of obtaining information for performing statistical analyses of the economic activities within a city or country.

To easy and improve the quality of the categorization of companies according to their economic activities, the Brazilian government is creating a centralized digital library with the statement of purpose of all companies in the country. This library will help the three government levels – Federal, the 27 States, and the more than 5000 Brazilian Counties – in the task of categorizing the Brazilian companies in according to the Brazilian law. In order to categorize the statement of purpose of each company within this digital library into the economic activities recognized by law – more than 1000 possible activities – we estimate that the data related to more than 5 millions companies will have to be processed. Also, we estimate that at least 300000 statements of purpose of new companies, or of companies which are changing their statement of purpose, will have to be processed every year. It is important to note that the large number of possible categories makes this problem particularly complex when compared with others presented in the literature [11].

This work presents some preliminary experimental results on automatic categorization of a set of 3264 statements of purpose of Brazilian companies into a subset of 764 economic activities recognized by Brazilian law. We used two techniques in our experiments: Vector Space Model (VS) [8] and Weightless Neural Networks (WNN) [7]. The best performing technique, weightless neural network, has shown 68.29% accuracy in identifying a correct category for each of the 3264 statements of purpose. To our knowledge, this is the first report on using WNN for text categorization into as a large number of classes as that used in this work.

This paper is organized as follows. After this introduction, Section 2 discusses the problem of categorizing companies according to business activities by using their free text statement of purpose. Section 3 describes our experimental evaluation of VS and WNN as classifiers, and Section 4 or conclusions and directions for future work.

2. Categorization of companies according to business activities using their free text statement of purpose

In many countries, companies must have a contract (Articles of Incorporation or Corporate Charter, in USA) with the society where they can legally operate. In Brazil, this contract is called a social contract and must contain the statement of purpose of the company. This statement of purpose needs to be categorized into a legal business activity by Brazilian government officials; for that, all legal business activities are cataloged in a table called CNAE – *Classificação Nacional de Atividade Econômicas* (National Classification of Economic Activities) [5].

To perform the categorization, the government officials (at the Federal, State and County levels) must find the semantic correspondence between the company statement of purpose and one or more entries of the CNAE table. There is a numerical code for each entry of the CNAE table and, in the categorization task, the government official must attribute one or more of such codes (CNAE codes) to the company at hand. This can happen on the foundation of the company or in a change of its social contract, if that modifies its statement of purpose.

The computational problem addressed by us is that of automatically finding the semantic correspondence between a statement of purpose of a company and one or more items of the CNAE table. To do that, in this work we have employed two techniques: VS and WNN.

2.1. Text categorization with VS

In VS, documents are represented by multidimensional vectors were each element is a relevant word present in the documents. In our case, documents are statements of purpose of companies or items of the CNAE table, and they can be compared by computing the angle between the vectors representing them. To categorize companies according to their statement of purpose, we compare the vector representing the statement of purpose of one company to all vectors representing the entries of the CNAE table and give to the company the CNAE code of closest CNAE entry (as measured by the angle between its vector and the document's vector).

2.2. Text categorization with WNN

RAM-based neural networks, also known as n-tuple classifiers or weightless neural networks, do not store knowledge in their connections but in Random Access Memories (RAM) inside the network's nodes, or neurons. These neurons operate with binary input values and use RAMs as lookup tables: the synapses of each neuron collect a vector of bits from the network's inputs that is used as the RAM address, and the value stored at this address is the neuron's output. Training can be made in one shot and basically consists of storing the desired output in the address associated with the input vector of the neuron (Figure 1) [1].

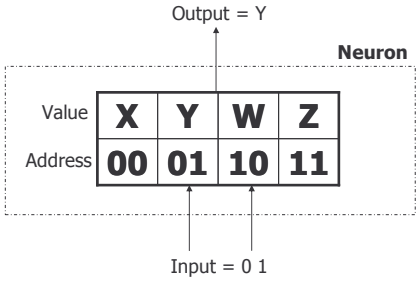

Figure 1: Weightless Neural Network

In spite of their remarkable simplicity, RAM-based neural networks are very effective as pattern recognition tools, offering fast training and easy implementation [3]. However, if the network input is too large, the memory size becomes prohibitive, since it must be equal to 2 to the power of the input size. Virtual Generalizing RAM (VG-RAM) networks are RAM-based neural networks that only require memory capacity to store the data related to the training set [7]. In the neurons of these networks, the memory stores the input-output pairs shown during training, instead of only the output. In the recall phase, the memory of VG-RAM neurons is searched associatively by comparing the input presented to the network with all inputs in the input-output pairs learned. The output of each VG-RAM neuron is taken from the pair whose input is nearest to the input presented – the distance function employed by VG-RAM neurons is the Hamming distance. If there is more than one pair at the same minimum distance from the input presented, the neuron's output is chosen randomly among these pairs.

Figure 2 shows the lookup table of a VG-RAM neuron with three inputs (X_1, X_2 and X_3). This lookup table contains three entries (input-output pairs), which were stored during the training phase (entry #1, entry

#2 and entry #3). During the recall phase, when an input vector (input) is presented to the neuron, its recall algorithm calculates the distance between this input vector and each input of the input-output pairs stored in the lookup table. In the example of Figure 2, the Hamming distance from the input to entry #1 is two, because both X_2 and X_3 bits do not match the input vector. The distance to entry #2 is one, because X_1 is the only non-matching bit. The distance to entry #3 is three, as the reader may easily verify. Hence, for this input vector, the algorithm evaluates the neuron's output, Y, as zero, since it is the output value stored in entry #2.

lookup table	**X_1**	**X_2**	**X_3**	**Y**
entry #1	1	1	0	1
entry #2	0	0	1	0
entry #3	0	1	0	0
input	1	0	1	**0**

Figure 2: Example of operation of a VG-RAM neuron

As we do in VS, to categorize companies according to their statement of purpose using VG-RAM we represent documents as multidimensional vectors were each element is a relevant word present in the documents. Again, documents are statements of purpose of companies or items of the CNAE table. To categorize companies according to their statement of purpose we use a single layer VG-RAM WNN whose neurons' inputs are feed with the vectors representing the documents. During training, for each CNAE table entry the inputs are connected to the vector representing it, and the outputs to its code – all neurons are trained to return the code of the CNAE entry. During recall, the inputs are connected to the vector representing a statement of purpose and the code returned by the majority of the neurons is taken as the desired categorization.

3. Experiments

In order to evaluate the performance of VS and WNN on automatic identification of CNAE categories in statement of purpose of companies, we conducted experiments with a database consisting of statements of purpose of 3264 Brazilian companies (an average of about 70 words each) and their associated CNAE codes. The CNAE codes of each company in the database were assigned by Brazilian government officials trained in this task; the number of codes assigned to each company varies from 1 to 12, and 764 different codes appear in the database.

The categorization performed by government officials is a multi-label classification (at-least-one) [11], but, in this preliminary work, both the VS and WNN classifiers assign only one label per test document. We have chosen to do this way because we are initially only interested in evaluating the viability of using VG-RAM WNNs to categorize text in the context of a large number of classes. In future works we will examine multi-label classification with VG-RAM WNN in the same context.

In addition to the statement of purpose of 3264 companies, the database also contains the official brief description of each one of the 764 CNAE codes (an average of 8 words and, in many cases, as small as 2 words [5]) associated with them.

The database was preprocessed in order to produce two term vs. document matrixes: one representing the 3264 statements of purpose, and the other the 764 CNAE entries descriptions. A total of 1001 terms were found in the database after removing stop words and trivial cases of gender and plural – only words appearing in the CNAE entries were considered. Therefore, the first matrix has dimensions (1001 x 3264) while the second, (1001 x 764), and the elements of both store the number of occurrences of each one of the 1001 terms.

3.1. Vector space model

To categorize the 3264 statements of purpose into the 764 CNAE codes using VS, the cosine similarity measure was used. For each vector (i), representing a statement of purpose in the first term vs. document matrix, and vector (j), representing each CNAE table entry in the second matrix, we computed the cosine of the angle between (i) and (j). The (j) for which the cosine (i, j) was the largest was selected as the category of (i).

3.2. Weightless Neural Network

To categorize using VG-RAM WNN, we employed networks consisting of 1, 2x2, 3x3, …, or 15x15 neurons with 64, 128, 256, 512, or 1024 synapses each (we have tested all 75 combinations). The synapses of the neurons, randomly connected to an input vector of 1001 elements, return 1 or 0 depending on the values in their inputs. A synapse returns 1 if the vector element to which it is connected contains a value larger than the value in the vector element to which the next synapse is connected, and 0 otherwise (the last synapse compares its value with that of the first). This synapse functionality, known as minchinton cell functionality, allows WNN to work with non binary inputs [9]. The output of the neurons can assume a value between 1 and 764.

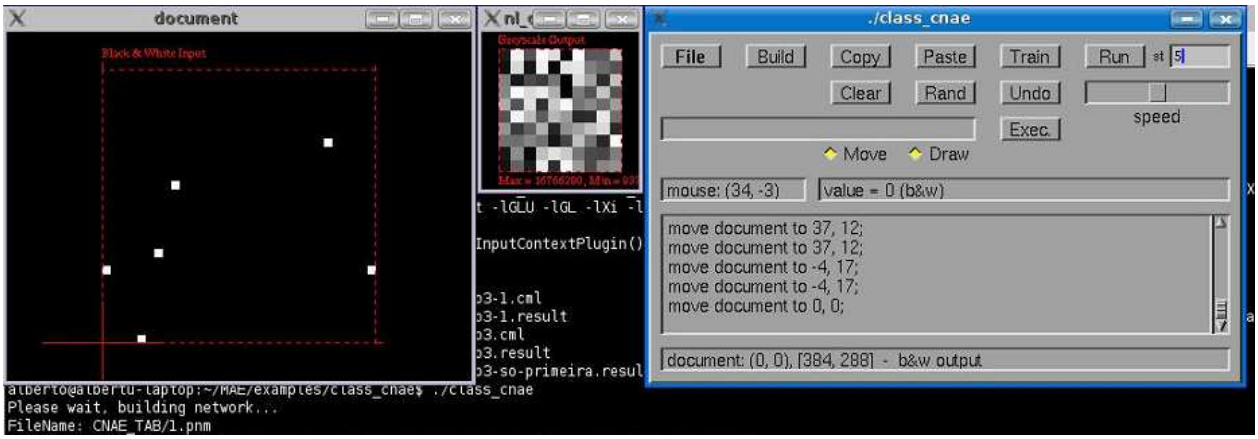

Figure 3: MAE Application

To implement these WNNs, we employed the Event Associative Machine (MAE), an open source framework for modeling VG-RAM neural networks developed at *Universidade Federal do Espírito Santo* [6]. MAE is similar to the Neural Representation Modeller (NRM), developed by the Neural Systems Engineering Group at Imperial College London and commercialized by Novel Technical Solutions [2, 10]. However, MAE differs from NRM on three main aspects: it is open source, runs on UNIX (currently, Linux), and uses a textual language to describe WNNs.

MAE allows designing, training and analyzing the behavior of modular WNNs whose basic modules are bidimensional neural layers and bidimensional filters. Neural layers' neurons may have several attributes (type, input sensitivity, memory size, etc) and the user can freely design filters using the C programming language. The MAE user specifies modular WNNs using the MAE Neural Architecture Description Language (NADL). NADL source files are compiled into MAE applications, which have a built-in graphical user interface and an interpreter of the MAE Control Script Language (CDL). The user can train, recall, and alter neural layers' contents using the graphical interface or scripts in CDL.

Figure 3 shows the MAE application we have built to run the experiments with the VG-RAM WNN configured 10x10 neurons with 256 synapses each. In the MAE application, the window named *document* shows the vectors representing the documents been trained or recalled. The 1001 elements of these vectors are transformed in a 32x32 input neuron layer (23 of the 32x32=1024 elements of this neuron layer are always filled with zero), which can be shown as a 32x32 pixel image. The outputs of the 100 (10x10) neurons of the network form the 10x10 window shown in the middle of Figure 3, while the left window, named *class_cnae*, is the graphical user interface of this MAE application.

During training, the VG-RAM WNN input vector was feed with the columns of the second matrix described above, and the output with a value equal to the order of each column of this matrix (an index to the CNAE table entry). During recall, the network was feed with each column of the second matrix, and all 100 outputs of the VG-RAM WNN were evaluated for each column. The value of the majority (the order of the column of first matrix, learned during training) was taken as the network's output. Therefore, in cases were a company has more than one CNAE code, only the code selected by the majority of the neurons is outputted. However, for networks with large number of neurons, the percentage of the number of neurons signaling each CNAE code can be taken as the network's estimate of the probability that this code is one of the codes associated with the company. Nevertheless, we only use the code with the highest estimate and left the problem of discovering all codes associated with each company for future work.

In Figure 4 we present the classification performance of each VG-RAM WNN configuration examined. In the graph of Figure 4, the horizontal axis is the number of neurons of the network: 1 (1x1), 4 (2x2), 9 (3x3), …, and 225 (15x15); while the vertical axis is the network performance as a percentage of correct CNAE code assignments to the 3264 statements of purpose. We have considered an assignment as correct when the technique under examination selected any one of the classes assigned by the human specialist to each statement of purpose.

There are five curves in the graph, one for each number of synapses per neuron we have employed for

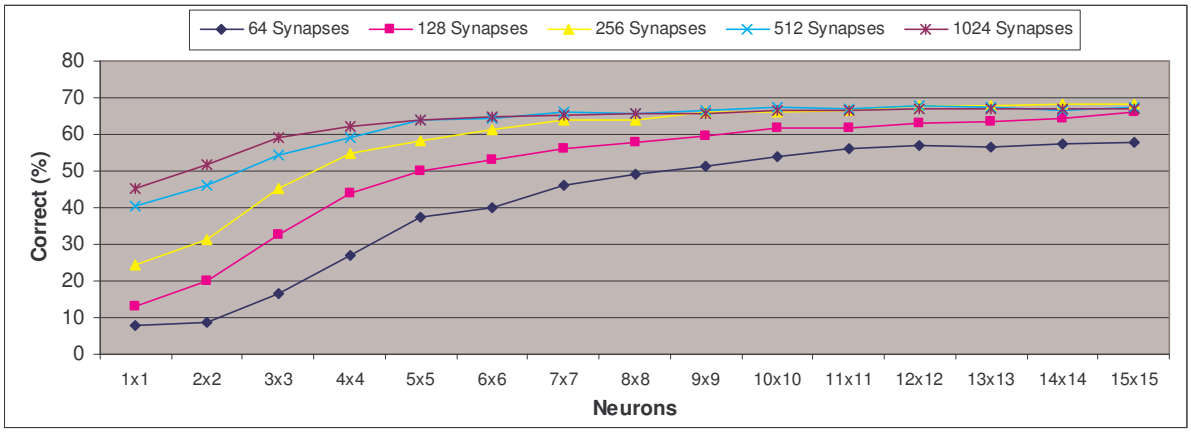

Figure 4: VG-RAM WNN classification performance

each network configuration (see legend in Figure 4).

As the graph of Figure 4 shows, the classification performance of the VG-RAM WNN increases with the number of neurons of the network, but levels off when the network has about 144 (12x14) neurons. We believe it is due to the fact that, with a small number of neurons, the network cannot discriminate well between the many CNAE classes. As the number of neurons increases above 144, additional neurons do not augment the discriminative power of the network. The performance also increases with the number of synapses per neuron, but again levels off at about 256 synapses.

Table 1 presents the categorization performance of VS and that of the best performing VG-RAM WNN (14x14 neurons with 256 synapses each). As Table 1 shows, VG-RAM WNN outperforms VSP by 4.93%.

Table 1: Percentage of correct CNAE code assignments of each technique

VS	Best VG-RAM WNN
63.36%	68.29%

4. Conclusions

This paper presented a preliminary experimental evaluation of the performance of VG-RAM WNN on automatic free text categorization into economic activities. We have trained VS and WNN systems with 764 brief official Brazilian descriptions of economic activities and use them to categorize 3264 companies into these economic activities according to the statements of purpose of each one of these companies. Our experiments showed that VG-RAM WNN can outperform VS for a significant margin: 68.29% x 63.36% accuracy, respectively. To our knowledge, this is the first time WNN is used for free text categorization. It is important to note the large number of categories used in the experiments, 764.

As future work, one of the improvements we are working on for the VG-RAM WNN is the use of knowledge correlation between the input-output pairs learned [4], while, for VS, the artificial centroid vector strategy for improving selectivity [8].

5. Acknowledgements

We would like to thank Receita Federal do Brasil for their support to this research work. We would like also to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq-Brasil (grants 308207/2004-1, 471898/2004-0, 620165/2006-5) and Financiadora de Estudos e Projetos – FINEP-Brasil (grants CT-INFRA-PRO-UFES/2005, CT-INFRA-PRO-UFES /2006).

6. References

[1] I. Aleksander. Self-adaptive Universal Logic Circuits (Design Principles and Block Diagrams of Self-adaptive Universal Logic Circuit with Trainable Elements). IEE Electronic Letters, (2):231–232, 1966.

[2] I. Aleksander, C. Browne, B. Dunmall, T. Wright. Towards Visual Awareness in a Neural System. In: Amari, S., Kasabov, N. (Editors): Brain-Like Computing and Intelligent Information Systems. Springer-Verlag, Berlin Heidelberg New York (1997) 513-533

[3] I. Aleksander. From WISARD to MAGNUS: a Family of Weightless Virtual Neural Machines. In: J. Austin (Editor). RAM-Based Neural Networks. pages 18–30. World Scientific, 1998.

[4] R. Carneiro, S. S. Dias, D. Fardin Jr., S. Oliveira, A. S. d. Garcez, and A. F. De Souza. Improving VG-RAM Neural

Networks Performance Using Knowledge Correlation. Lecture Notes on Computer Science, 4232:427–436, 2006.
[5] CNAE. Classificação Nacional de Atividades Econômicas - Fiscal. IBGE – Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, RJ, 1.1 edition, 2003.
[6] K. S. Komati, A. F. De Souza. Using Weightless Neural Networks for Vergence Control in an Artificial Vision System. Applied Bionics and Biomechanics 1: 21-32, 2003.
[7] T. B. Ludermir, A. C. P. L. F. Carvalho, A. P. Braga, and M. D. Souto. Weightless Neural Models: a Review of Current and Past Works. Neural Computing Surveys, 2:41–61, 1999.
[8] G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. Communications of the ACM, 18(11):613–620, 1975.
[9] R. J. Mitchell, J. M. Bishop, S. K. Box, J. F. Hawker. Comparison of Some Methods for Processing "Grey Level" Data in Weightless Networks. In: J. Austin (Editor). RAM-Based Neural Networks. pages 61–70. World Scientific, 1998.
[10] Novel Technical Solutions. Neural Representation Modeller. http://www.ntsu.com, accessed in June 13th, 2007.
[11] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1–47, 2002.