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

Alberto F De Souza , Felipe Pedroni , Elias Oliveira ,
Patrick M Ciarelli , Wallace F Henrique , Lucas Veronese
D partam nto In ormática
D partam nto Ci ncia a In ormaç o
D partam nto En n aria El trica
A F rnan o F rrari s n
9 6 9 – itória ES, Brazil
{alb rto, lias}@lca in u s br

Abstract

tackl t probl m o automatin t cat orization o compani s accor in to t ir conomic acti iti s usin busin ss scriptions in r t xt ormat as input is cat orization is ital to un am ntal asp cts o national o rnm ntal a ministration suc as s ort, m ium an lon t rm plannin an taxation As t numb r o cat ori s consi r is ry lar (mor t an in t Brazilian sc nario , t automatic t xt cat orization probl m tar t r is c all n in a appli an compar t us o two i r nt t c niqu s to al wit it: t ctor Spac Mo l, a w ll known t xt cat orization t c niqu ; an irtual G n ralizin Ran om Acc ss M mory i tl ss N ural N twork, or G RAM NN o our knowl , t is is t irst r port on usin G RAM NN or t xt cat orization

1. Introduction

Automatic text classi ication and clustering are still very c allenging computational problems to t e in ormation retrieval IR communities bot in academic and industrial contexts Currently, t e majority o t e work on IR one can ind in t e literature is ocused on classi ication and clustering o webpages However, t ere are many ot er important applications to w ic little attention as it erto been paid, w ic are as well very di icult to deal wit One example o t ese applications is t e classi ication o companies based on t eir statement o purpose, also called mission statement, w ic represent t e business context o t e companies activities T e categorization o companies according to t eir economic activities is an important step o t e process o obtaining in ormation or per orming statistical analyses o t e economic activities wit in a city or country

To easy and improve t e quality o t e categorization o companies according to t eir economic activities, t e Brazilian government is creating a centralized digital library wit t e statement o purpose o all companies in t e country T is library will elp t e t ree government levels – Federal, t e States, and t e more t an 5 Brazilian Counties – in t e task o categorizing t e Brazilian companies in according to t e Brazilian law In order to categorize t e statement o purpose o eac company wit in t is digital library into t e economic activities recognized by law – more t an possible activities – we estimate t at t e data related to more t an 5 millions companies will ave to be processed Also, we estimate t at at least statements o purpose o new companies, or o companies w ic are c anging t eir statement o purpose, will ave to be processed every year It is important to note t at t e large number o possible categories makes t is problem particularly complex w en compared wit ot ers presented in t e literature [ ]

T is work presents some preliminary experimental results on automatic categorization o a set o 4 statements o purpose o Brazilian companies into a subset o 4 economic activities recognized by Brazilian law We used two tec niques in our experiments: Vector Space Model VS [8] and Weig tless Neural Networks WNN [ ] T e best per orming tec nique, weig tless neural network, as s own 8 9% accuracy in identi ying a correct category or eac o t e 4 statements o purpose To our knowledge, t is is t e irst report on using WNN or text categorization into as a large number o classes as t at used in t is work

0-7695-2976-3/07 $25.00 © 2007 IEEE
DOI 10.1109/ISDA.2007.146

T is paper is organized as ollows A ter t is introduction, Section discusses t e problem o categorizing companies according to business activities by using t eir ree text statement o purpose Section describes our experimental evaluation o VS and WNN as classi iers, and Section 4 or conclusions and directions or uture work

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

In many countries, companies must ave a contract Articles o Incorporation or Corporate C arter, in USA wit t e society w ere t ey can legally operate In Brazil, t is contract is called a social contract and must contain t e statement o purpose o t e company T is statement o purpose needs to be categorized into a legal business activity by Brazilian government o icials; or t at, all legal business activities are cataloged in a table called CNAE – *Classi icaç o Nacional Ati i a Econ micas* National Classi ication o Economic Activities [5]

To per orm t e categorization, t e government o icials at t e Federal, State and County levels must ind t e semantic correspondence between t e company statement o purpose and one or more entries o t e CNAE table T ere is a numerical code or eac entry o t e CNAE table and, in t e categorization task, t e government o icial must attribute one or more o suc codes CNAE codes to t e company at and T is can appen on t e oundation o t e company or in a c ange o its social contract, i t at modi ies its statement o purpose

T e computational problem addressed by us is t at o automatically inding t e semantic correspondence between a statement o purpose o a company and one or more items o t e CNAE table To do t at, in t is work we ave employed two tec niques: VS and WNN

2.1. Text categorization with VS

In VS, documents are represented by multidimensional vectors were eac element is a relevant word present in t e documents In our case, documents are statements o purpose o companies or items o t e CNAE table, and t ey can be compared by computing t e angle between t e vectors representing t em To categorize companies according to t eir statement o purpose, we compare t e vector representing t e statement o purpose o one company to all vectors representing t e entries o t e CNAE table and give to t e company t e CNAE code o closest CNAE entry as measured by t e angle between its vector and t e document s vector

2.2. Text categorization with WNN

RAM based neural networks, also known as n tuple classi iers or weig tless neural networks, do not store knowledge in t eir connections but in Random Access Memories RAM inside t e network s nodes, or neurons T ese neurons operate wit binary input values and use RAMs as lookup tables: t e synapses o eac neuron collect a vector o bits rom t e network s inputs t at is used as t e RAM address, and t e value stored at t is address is t e neuron s output Training can be made in one s ot and basically consists o storing t e desired output in t e address associated wit t e input vector o t e neuron Figure [ ]

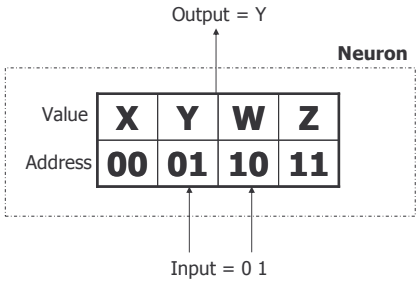

Figure 1: Weightless Neural Network

In spite o t eir remarkable simplicity, RAM based neural networks are very e ective as pattern recognition tools, o ering ast training and easy implementation [ ] However, i t e network input is too large, t e memory size becomes pro ibitive, since it must be equal to to t e power o t e input size Virtual Generalizing RAM VG RAM networks are RAM based neural networks t at only require memory capacity to store t e data related to t e training set [ ] In t e neurons o t ese networks, t e memory stores t e input output pairs s own during training, instead o only t e output In t e recall p ase, t e memory o VG RAM neurons is searc ed associatively by comparing t e input presented to t e network wit all inputs in t e input output pairs learned T e output o eac VG RAM neuron is taken rom t e pair w ose input is nearest to t e input presented – t e distance unction employed by VG RAM neurons is t e Hamming distance I t ere is more t an one pair at t e same minimum distance rom t e input presented, t e neuron s output is c osen randomly among t ese pairs

Figure s ows t e lookup table o a VG RAM neuron wit t ree inputs X , X and X T is lookup table contains t ree entries input output pairs , w ic were stored during t e training p ase entry # , entry

and entry # During t e recall p ase, w en an input vector input is presented to t e neuron, its recall algorit m calculates t e distance between t is input vector and eac input o t e input output pairs stored in t e lookup table In t e example o Figure , t e Hamming distance rom t e input to entry # is two, because bot X and X bits do not matc t e input vector T e distance to entry # is one, because X is t e only non matc ing bit T e distance to entry # is t ree, as t e reader may easily veri y Hence, or t is input vector, t e algorit m evaluates t e neuron s output, Y, as zero, since it is t e output value stored in entry

lookup table	X_1	X_2	X_3	Y
entry #				
entry #				
entry #				
input				**0**

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

As we do in VS, to categorize companies according to t eir statement o purpose using VG RAM we represent documents as multidimensional vectors were eac element is a relevant word present in t e documents Again, documents are statements o purpose o companies or items o t e CNAE table To categorize companies according to t eir statement o purpose we use a single layer VG RAM WNN w ose neurons inputs are eed wit t e vectors representing t e documents During training, or eac CNAE table entry t e inputs are connected to t e vector representing it, and t e outputs to its code – all neurons are trained to return t e code o t e CNAE entry During recall, t e inputs are connected to t e vector representing a statement o purpose and t e code returned by t e majority o t e neurons is taken as t e desired categorization

3. Experiments

In order to evaluate t e per ormance o VS and WNN on automatic identi ication o CNAE categories in statement o purpose o companies, we conducted experiments wit a database consisting o statements o purpose o 4 Brazilian companies an average o about words eac and t eir associated CNAE codes T e CNAE codes o eac company in t e database were assigned by Brazilian government o icials trained in t is task; t e number o codes assigned to eac company varies rom to , and 4 di erent codes appear in t e database

T e categorization per ormed by government o icials is a multi label classi ication at least one [ ], but, in t is preliminary work, bot t e VS and WNN classi iers assign only one label per test document We ave c osen to do t is way because we are initially only interested in evaluating t e viability o using VG RAM WNNs to categorize text in t e context o a large number o classes In uture works we will examine multi label classi ication wit VG RAM WNN in t e same context

In addition to t e statement o purpose o 4 companies, t e database also contains t e o icial brie description o eac one o t e 4 CNAE codes an average o 8 words and, in many cases, as small as words [5] associated wit t em

T e database was preprocessed in order to produce two term vs document matrixes: one representing t e 4 statements o purpose, and t e ot er t e 4 CNAE entries descriptions A total o terms were ound in t e database a ter removing stop words and trivial cases o gender and plural – only words appearing in t e CNAE entries were considered T ere ore, t e irst matrix as dimensions x 4 w ile t e second, x 4 , and t e elements o bot store t e number o occurrences o eac one o t e terms

3.1. Vector space model

To categorize t e 4 statements o purpose into t e 4 CNAE codes using VS, t e cosine similarity measure was used For eac vector i , representing a statement o purpose in t e irst term vs document matrix, and vector j , representing eac CNAE table entry in t e second matrix, we computed t e cosine o t e angle between i and j T e j or w ic t e cosine i, j was t e largest was selected as t e category o i

3.2. Weightless Neural Network

To categorize using VG RAM WNN, we employed networks consisting o , x , x , , or 5x 5 neurons wit 4, 8, 5 , 5 , or 4 synapses eac we ave tested all 5 combinations T e synapses o t e neurons, randomly connected to an input vector o elements, return or depending on t e values in t eir inputs A synapse returns i t e vector element to w ic it is connected contains a value larger t an t e value in t e vector element to w ic t e next synapse is connected, and ot erwise t e last synapse compares its value wit t at o t e irst T is synapse unctionality, known as minc inton cell unctionality, allows WNN to work wit non binary inputs [9] T e output o t e neurons can assume a value between and 4

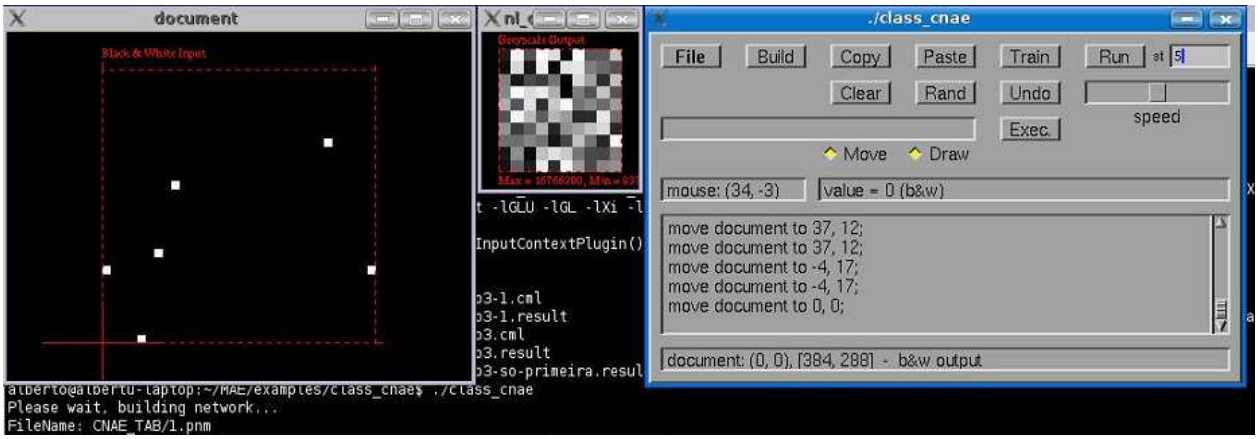

Figure 3: MAE Application

To implement t ese WNNs, we employed t e Event Associative Mac ine MAE , an open source ramework or modeling VG RAM neural networks developed at *Uni rsi a F ral o Esp rito Santo* [ ] MAE is similar to t e Neural Representation Modeller NRM , developed by t e Neural Systems Engineering Group at Imperial College London and commercialized by Novel Tec nical Solutions [,] However, MAE di ers rom NRM on t ree 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 t e be avior o modular WNNs w ose basic modules are bidimensional neural layers and bidimensional ilters Neural layers neurons may ave several attributes type, input sensitivity, memory size, etc and t e user can reely design ilters using t e C programming language T e MAE user speci ies modular WNNs using t e MAE Neural Arc itecture Description Language NADL NADL source iles are compiled into MAE applications, w ic ave a built in grap ical user inter ace and an interpreter o t e MAE Control Script Language CDL T e user can train, recall, and alter neural layers contents using t e grap ical inter ace or scripts in CDL

Figure s ows t e MAE application we ave built to run t e experiments wit t e VG RAM WNN con igured x neurons wit 5 synapses eac In t e MAE application, t e window named *ocum nt* s ows t e vectors representing t e documents been trained or recalled T e elements o t ese vectors are trans ormed in a x input neuron layer o t e x = 4 elements o t is neuron layer are always illed wit zero , w ic can be s own as a x pixel image T e outputs o t e x neurons o t e network orm t e x window s own in t e middle o Figure , w ile t e le t window, named *class_cna* , is t e grap ical user inter ace o t is MAE application

During training, t e VG RAM WNN input vector was eed wit t e columns o t e second matrix described above, and t e output wit a value equal to t e order o eac column o t is matrix an index to t e CNAE table entry During recall, t e network was eed wit eac column o t e second matrix, and all outputs o t e VG RAM WNN were evaluated or eac column T e value o t e majority t e order o t e column o irst matrix, learned during training was taken as t e network s output T ere ore, in cases were a company as more t an one CNAE code, only t e code selected by t e majority o t e neurons is outputted However, or networks wit large number o neurons, t e percentage o t e number o neurons signaling eac CNAE code can be taken as t e network s estimate o t e probability t at t is code is one o t e codes associated wit t e company Nevert eless, we only use t e code wit t e ig est estimate and le t t e problem o discovering all codes associated wit eac company or uture work

In Figure 4 we present t e classi ication per ormance o eac VG RAM WNN con iguration examined In t e grap o Figure 4, t e orizontal axis is t e number o neurons o t e network: x , 4 x , 9 x , , and 5 5x 5 ; w ile t e vertical axis is t e network per ormance as a percentage o correct CNAE code assignments to t e 4 statements o purpose We ave considered an assignment as correct w en t e tec nique under examination selected any one o t e classes assigned by t e uman specialist to eac statement o purpose

T ere are ive curves in t e grap , one or eac number o synapses per neuron we ave employed or

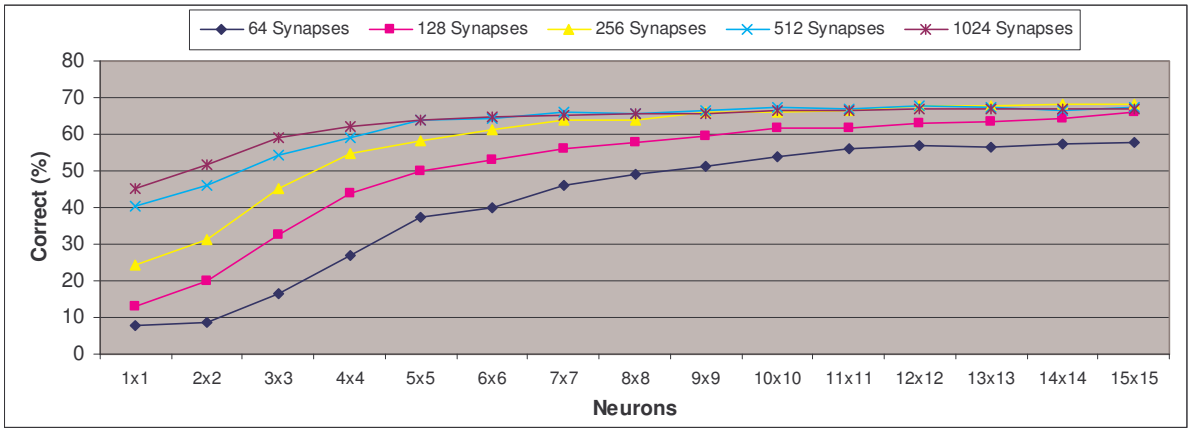

Figure 4: VG-RAM WNN classification performance

eac network con iguration see legend in Figure 4

As t e grap o Figure 4 s ows, t e classi ication per ormance o t e VG RAM WNN increases wit t e number o neurons o t e network, but levels o w en t e network as about 44 x 4 neurons We believe it is due to t e act t at, wit a small number o neurons, t e network cannot discriminate well between t e many CNAE classes As t e number o neurons increases above 44, additional neurons do not augment t e discriminative power o t e network T e per ormance also increases wit t e number o synapses per neuron, but again levels o at about 5 synapses

Table presents t e categorization per ormance o VS and t at o t e best per orming VG RAM WNN 4x 4 neurons wit 5 synapses eac As Table s ows, VG RAM WNN outper orms VSP by 4 9 %

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

VS	Best VG-RAM WNN
%	8 9%

4. Conclusions

T is paper presented a preliminary experimental evaluation o t e per ormance o VG RAM WNN on automatic ree text categorization into economic activities We ave trained VS and WNN systems wit 4 brie o icial Brazilian descriptions o economic activities and use t em to categorize 4 companies into t ese economic activities according to t e statements o purpose o eac one o t ese companies Our experiments s owed t at VG RAM WNN can outper orm VS or a signi icant margin: 8 9% x % accuracy, respectively To our knowledge, t is is t e irst time WNN is used or ree text categorization It is important to note t e large number o categories used in t e experiments, 4

As uture work, one o t e improvements we are working on or t e VG RAM WNN is t e use o knowledge correlation between t e input output pairs learned [4], w ile, or VS, t e arti icial centroid vector strategy or improving selectivity [8]

5. Acknowledgements

We would like to t ank Receita Federal do Brasil or t eir support to t is researc work We would like also to t ank Consel o Nacional de Desenvolvimento Cient ico e Tecnológico – CNPq Brasil grants 8 4 , 4 898 4 , 5 5 and Financiadora de Estudos e Projetos – FINEP Brasil grants CT INFRA PRO UFES 5, CT INFRA PRO UFES

6. References

[ ] I Aleksander Sel adaptive Universal Logic Circuits Design Principles and Block Diagrams o Sel adaptive Universal Logic Circuit wit Trainable Elements IEE Electronic Letters, : – , 9

[ ] I Aleksander, C Browne, B Dunmall, T Wrig t Towards Visual Awareness in a Neural System In: Amari, S , Kasabov, N Editors : Brain Like Computing and Intelligent In ormation Systems Springer Verlag, Berlin Heidelberg New York 99 5 5

[ ] I Aleksander From WISARD to MAGNUS: a Family o Weig tless Virtual Neural Mac ines In: J Austin Editor RAM Based Neural Networks pages 8– World Scienti ic, 998

[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 Per ormance Using Knowledge Correlation Lecture Notes on Computer Science, 4 :4 –4 ,
[5] CNAE Classi icação Nacional de Atividades Econ micas Fiscal IBGE – Instituto Brasileiro de Geogra ia e Estat stica, Rio de Janeiro, RJ, edition,
[ ] K S Komati, A F De Souza Using Weig tless Neural Networks or Vergence Control in an Arti icial Vision System Applied Bionics and Biomec anics : ,
[ ] T B Ludermir, A C P L F Carval o, A P Braga, and M D Souto Weig tless Neural Models: a Review o Current and Past Works Neural Computing Surveys, :4 – , 999
[8] G Salton, A Wong, and C S Yang A Vector Space Model or Automatic Indexing Communications o t e ACM, 8 : – , 9 5
[9] R J Mitc ell, J M Bis op, S K Box, J F Hawker Comparison o Some Met ods or Processing “Grey Level” Data in Weig tless Networks In: J Austin Editor RAM Based Neural Networks pages – World Scienti ic, 998
[ ] Novel Tec nical Solutions Neural Representation Modeller ttp: www ntsu com, accessed in June t ,
[ ] F Sebastiani Mac ine learning in automated text categorization ACM Computing Surveys, 4 : –4 ,