THE AUTOMATION OF THE CLASSIFICATION OF ECONOMIC ACTIVITIES FROM FREE TEXT DESCRIPTIONS USING AN ARRAY ARCHITECTURE OF PROBABILISTIC NEURAL NETWORK

Elias Oliveira*, Patrick Marques Ciarelli†, Fabio O. Lima‡

*Department of Information Science

†Department of Electrical Engineering

‡Department of Computer Science
Federal University of Espirito Santo
Av. Fernando Ferrari s/n
29060-970 – Vitoria, ES Brazil

Email: elias@inf.ufes.br

Abstract— The automation of the categorization of economic activities from business descriptions in free text format is a huge challenge for the Brazilian governmental administration in the present day. So far, this task has been carried out by humans, not all of them well trained for this job. When this problem is tackled by humans, the subjectivity on their classification brings another problem: different human classifiers can give different results when working on a set of same business descriptions. This can cause a serious distortion on the information for the planning and taxation of the governmental administrations on the three levels: County, Estate and Federal. Furthermore, the number of possible categories considered is very large, more than 1000 in the Brazilian scenario. The large number of categories makes the problem even more hard to be solved. We have applied an array of Probabilistic Neural Network nodes, each one of them specialized in dealing only with a small number of classes. We carried out a series of experiments with these probabilistic neural networks and the automatic classification were better than 70% in all classes tested.
Keywords— Text classification, business activities classification, clustering

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, a great effort of the work on IR one can find in the literature is focused on classification and clustering of generics content of text documents. 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 economic activities description, also called mission statements, which represent the business context of the companies' activities, in other words the business economic activities from free text description by the company's founders.

The categorization of companies according to their economic activities constitute a very important step towards building tools for obtaining information for performing statistical analyses of the economic activities within a city or country. With this goal, the Brazilian government is creating a centralized digital library with the business economic activities description of all companies in the country. This library will serve the three government levels: Federal; the 27 States; and more than 5.000 Brazilian counties. We estimate that the data related to more than 5 million companies will have to be processed every year (Ministério do Desenvolvimento, Indústria e Comércio Exterior – Secretaria do Desenvolvimento da Produção, Departamento Nacional de Registro do Comércio (DNRC), 2007) into more than 1.000 possible activities. 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 (Jain et al., 1999; Sebastiani, 2002).

This work presents some experimental results on automatic categorization of a set of 3696 business activities descriptions in free text format of Brazilian companies into a subset of 415 economic activities recognized by Brazilian law. We implemented an array of probabilistic neural network, each node of this array is in charge of classifying input documents into six classes. The first fifth classes are the economic activities, and the last one is a *passed document* flag, in other words, this flag says that that node-net is not able to determine a class for that document. This may happen whenever a text document is sent as an input to an inadequate node-net of the array.

Our approach has shown a performance of at least 70% of accuracy, in the worst case, in identifying correct categories for each one of the 3696 economic activities textual documents. In the average our approach yielded 76.88%. Besides improving the results presented in (Oliveira et al., 2007), to the knowledge of the authors, this is the first time a probabilistic neural network is applied to this sort of free text categorization problem. Hence, the results are very encouraging.

This work is organized as follows. In Section

2, we detailed more the characteristics of the problem and its importance for the government institutions in Brazil. We described the architecture of our probabilistic neural network array in Setion 3. In Section 4 the experimental results are discussed. We present our conclusions and indicate some future paths of this research in Section 5.

2 The Problem

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 tells the *business economic activities description* of that company and must be categorized into a legal business activity by Brazilian government officials. For that, all legal business activities are cataloged using a table called National Classification of Economic Activities – *Classificação Nacional de Atividade Econômicas*, (CNAE) (CNAE, 2003).

To perform the categorization, the government officials (at the Federal, State and County levels) must find *the semantic correspondence* between the company economic activities description 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 attributes one or more of such 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 economic activities.

The work of finding the semantic correspondence between the company economic activities description and a set of entries into the CNAE table is a very difficult task. This is because of the subjectivity of each local government officials who can focus on their own particular interests so that some codes may be assigned to a company, whereas in other regions, similar companies, may have a total different set of codes. Sometimes even inside of the same state, different level of government officials may count on a different number of codes for the same company for performing their work of assessing that company. Having inhomogeneous ways of classifying any company everywhere in all the three levels of the governmental administrations can cause a serious distortion on the key information for the long time planning and taxation. Additionally, the continental size of Brazil makes this problem of classification even worse.

Considering all that was previously said, the manual way of carrying out the economic activities classification of the companies is exhausted. Besides its human subjectivity cumbersome, another reason is the volume of companies that is registered every year in Brazil. Only in 2006 were registered 467.046 new companies, others 737.566 altered their current contract, possibly changing their economic activities description (Ministério do Desenvolvimento, Indústria e Comércio Exterior – Secretaria do Desenvolvimento da Produção, Departamento Nacional de Registro do Comércio (DNRC), 2007). In total, there were 1,204.612 companies which might have been classified by at least one of the three levels of the government officials. Due to this task is up to now decentralized, we might have the same job being performed many times by each of the three level of the government officials. Nevertheless, it is known that there are not enough employees to do this job. On top of that, we have the Brazilian economy steadily increasing, so will the number of companies each year onwards.

Hence, the computational problem addressed by us is mainly that of automatically *suggesting* the human classifier the semantic correspondence between an economic activities textual description of a company and one or more items of the CNAE table. Or, depends on the level of certainty the algorithms have on the automatic classification, we may consider bypassing thus the human classifier.

The number of codes assigned by the human specialist to a company can vary from 1 to 12 or, in some cases, even more. In the set of assigned codes, the first code is the main code of that company. The remaining codes have no order importance. These number of codes will be considered by the automatic system as a goal. Therefore, we will evaluate the quality of our system by the *Recall* criterion (Baeza-Yates and Ribiero-Neto, 1998). In the problem presented in this paper, the *Recall* will tell how many relevant codes the system will be able to *recovery*, for each company, over the total number of the relevant codes of that company. In our case, the total number of relevant codes of a given company is that assigned by a human specialist, including the first code. The aim in this evaluation is to have a good suggesting system, thus as much near as possible of 1 (*i.e*, 100%) is the *Recall*, the better.

We also consider evaluating our algorithm by the *Precision* criterion (Baeza-Yates and Ribiero-Neto, 1998). This evaluation measurement tells us how precise is the system on assigning relevant codes to a company. Hence, from those relevant codes that the system was able to *recovery* we divide by the total number of recovered codes for that company. Similarly, the aim in this evaluation is to have the *Precision* as much near as possible of 1, or 100%, for any company.

3 The Probabilistic Neural Network

The Probabilistic Neural Network was first proposed by Donald Specht in 1990 (Specht, 1990). This is an artificial neural network for nonlinear computing which approaches the Bayes optimal decision boundaries. This is done by estimating the *probability density function* of the training dataset using the Parzen (Duda et al., 2001) nonparametric estimator.

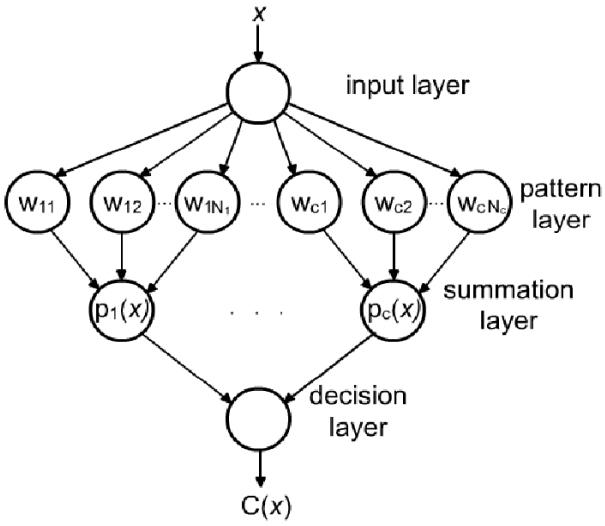

Figure 1: The Probabilistic Neural Network architecture.

Typically, this neural network is composed of 4 layers: the *input* layer, the *pattern* layer, the *summation* layer, and the *output* layer, or also called by the *decision* layer, as depicted in Figure 1. This neural network needs only one training step, thus its train is very fast comparing to the others feed-forward neural networks (Duda et al., 2001; Haykin, 1998). The train consists in assigning each training sample x_i of class C_i to a neuron of pattern layer of class C_i. Thus the weight vector of this neuron is the own characteristics vector of the sample.

For each pattern x passed by the input layer to a neuron in the pattern layer, it computes the output for the x. The computation is depicted as in Equation 1.

$$F_{k,i}(x) = exp(\frac{x^t w_{ki} - 1}{\sigma^2}), \qquad (1)$$

where the x is the pattern characteristics input vector, and the w_{ki} is the k^{th} weight for a neuron of class C_i, $k \in K$, whereas K is the number of neuron. Each neuron represents one input used during the train of the network. σ is the Gaussian standard deviation, which determines the receptive field of the Gaussian curve. N_i, in Equation 2, is the number of neuron in the pattern layer of C_i, corresponding the set of neurons of K, and $|C|$ is the total number of classes.

The next step is the summation layer. In this layer, all weight vectors are summed, Equation 2, in each cluster C_i producing $p_i(x)$ values. In Equation 3 we show how the decision layer chooses the maximum likelihood $p_i(x)$ to decide a class for the input x.

$$p_i(x) = \frac{1}{N_i} \sum_{k=1}^{N_i} F_{k,i}(x) \;\; k = 1, 2, ..., N_i = |C_i| \;\; (2)$$

$$C(x) = arg\ max\ p_i(x) \qquad (3)$$

Differently from other type of networks, such as those feed forward based (Haykin, 1998), probabilistic neural network needs only one parameter to be configured: the σ, (see in Equation 1) used to narrow the receptive field of the Gaussian curve in order to strictly select only the more likelihood inputs for a given class. Other advantages of the probabilistic neural networks is that it is easy to add new classes, or new training inputs, into the already running structure, which is good for the on-line applications (Duda et al., 2001). Moreover, it is reported in the literature that it is also easy to implement this type of neural network in parallel.

Besides its simplicity on the configuration, the literature has shown that this type of neural network can yield similar results in pattern recognition problems. One of its drawbacks is the great number of neurons in the pattern layer, which can be, nevertheless, mitigated by an optimization on the number of the neuron (Georgiou et al., n.d.; Mao et al., 2000). In the next section we present our approach which changed a little the way one can find the use of this type of network in the literature.

3.1 The Array Architecture

The standard structure presented in the literature (Georgiou et al., n.d.; Mao et al., 2000; Specht, 1990) (see Figure 1) usually suggest that every input pattern should equally be presented to all neurons in the pattern layer. In our problem that would enormously increase the number of characteristics the pattern layer should deal with, most of them unnecessarily. Because of this we proposed a preprocessor for the input layer in order to filter the unnecessary characteristics to be sent to the neurons in the pattern layer.

In Figure 2 we describe how the array architecture is implemented. The characteristics vector $\langle t_1, t_2, \ldots, t_n \rangle$ of words is presented to the input layer, but only the pertinent characteristics, in this case words, are passed through to each of the nodes in the pattern layer. By doing so, we also increase the performance of each node, as it has to compute less characteristics of its classes.

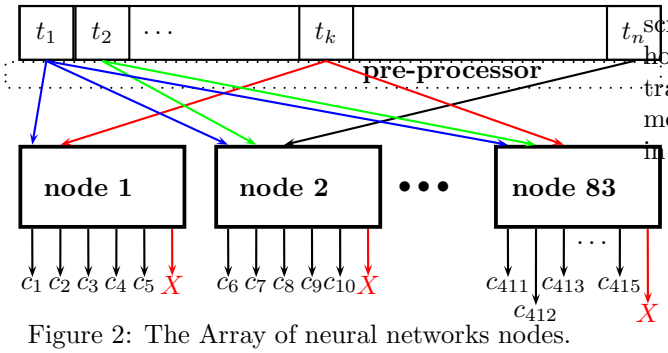

Figure 2: The Array of neural networks nodes.

The pre-processor task, in Figure 2, is to pass to *node 1* the characteristics $\langle t_1, t_k \rangle$ solely, whereas to *node 2* it passes $\langle t_1, t_2, t_n \rangle$ and to *node 83* $\langle t_1, t_2, t_k \rangle$, for example.

Giving to each node only the necessary characteristics to work with its classes we optimize each node so that it can improve its accuracy rates both on deciding an appropriate class for an input, and throwing out inputs that do not belong to that set of classes which that particular node is in charge.

Hence, to categorize the 3696 documents using the probabilistic neural network, we employed a network consisting of $K = 1723$ neurons into the pattern layer, 415 neurons into the summation layer all that clustered into 83 nodes of small networks as described in Figure 1 and 2. Each of these nodes is in charge of discriminating one input into an actual five classes. The sixth output of each of these nodes is to tell whenever an input is not one of the fifth classes performed by that node.

In order to choose a near-optimal σ's value for each of the 1723 neurons, we applied a genetic algorithm (Goldberg, 1989; Michalewicz, 1992; Mao et al., 2000) to evaluate good values based on our training dataset.

4 Experiments

The dataset we used for our experiments consists on a dataset of 3696 of documents containing each of them a business description of a real company placed in Vitoria County in Brazil. We used a set of 1723 documents for training and another 1973 documents for testing each node of the array of nets we specified, as described in Section 3. In the Vitoria's dataset there are 415 different CNAE codes, much more classes than one can find in problems tackled in the literature (Sebastiani, 2002). We implemented our algorithms in the MatLab software, version 6.1, and run the programs on a PC with Athlon 64 2200 processor.

The purpose of our experiments study is twofold: (1) to evaluate the proposed approach on determining a first code for a given business description within a document, and (2) to evaluate how good the approach is on pointing out the extra codes a human might have given for that document. The results of these experiments is depicted in Table 1.

Neural Network			
	$1^{\underline{st}}$ Code	**Recall**	**Precision**
1	76.88%	75.46%	3.36%
2		9.50%	0.07%

Table 1: Percentage of correct CNAE code assignments to the 3696 business descriptions and the recall and precision evaluation.

Table 1 presents the categorization performance of the approach proposed in this paper as a percentage of correct CNAE code assignments to the 3696 documents. The first line and first column of the table shows the performance of the algorithm on categorizing each document taking into account *only* the first code assigned to it. The algorithm was able to correctly determine 76.88%, on average, of the first codes category for each of the 3696 business information based on its textual proposal. In this case, we are considering a much harder problem than simply find any relevant code for that textual business description. The purpose of this part of the experiments was to evaluate how much the algorithms were capable of identifying the business' principal code within our dataset. In the *Recall* and the *Precision* columns, we were interested in evaluating how good our approach was on indicating the extra codes for that company. As discussed in Section 2, these extra codes are usually assigned by the human classifier with the aim to express the other activities a company may also performed.

Our approach performed quite well when considering the *Recall* evaluation measurement. The results show that, for this dataset, we can relay on this approach in 75.46% of times to help the human classifier to decide which additional codes to assigned to an economic activities textual description. On the other hand, our approach is still poor on given us a good *Precision* on the indicated codes. In the experiments the algorithm was precise on only 3.36% out of 100%. Hence, it still brings us many codes that are not, in principle, relevant to the economic activities description at hand. The second line of Table 1, in the *Reacall* and *Precision* columns, we present the variance of these measurements. These measure inform the window of certainty of our algorithm. It is informally known, among well trained human classifier, that a good specialist is able to achieve around 70%–80% of certainty on their classifications on the first code. Therefore, although we can continue working on improving these metrics,

the number of codes one can assign to a economic activities description is today an arbitrary choice that depends on the subjective of the classifier, the County or Estate in Brazil where this is done and the purpose of the administration on using this code. Hence, when thinking of this approach as part of a recommendation system, the higher is the *Recall* of any classifier we apply, the better, as the specialist will be able to accepted the suggestion in total, or make small changes if desired. Whereas the *Precision* gives us the idea of narrowing the window of theses suggestions.

5 Conclusions

The problem of classifying a huge number of economic activities description in free text format every day is a huge challenge for the Brazilian governmental administration. This problem is crucial for the long term planning in all three levels of the administration in Brazil. Therefore an automatic, or semi-automatic, manner of doing that is needed for make it possible and also for avoiding the problem of subjectivity introduced by the human classifier.

This paper presented an experimental evaluation of the performance of the probabilistic neural network on categorization of free text into economic activities classes. To our knowledge, this is the first report on using probabilistic neural network for text categorization into a large number of classes as that used in this work and the results are very encouraging. One of the advantages of probabilistic neural network is that it needs only one parameter to be configured. Thus, to specify near-optimal values for the 415 of each class we had, we applied a genetic algorithm to evaluate them. The appropriateness of these values yielded neurons with high level of efficiency on classifying inputs into their correct classes.

We have trained 83 small nodes of networks totalizing 1723 neurons. These nodes were organized as an array of specialized classifier for each of the 415 entries, a subset of the Brazilian CNAE table. A more accurate evaluation of this array optimization ought to be done in future work, but we already could empirically experienced the improvement on quality of the classifications of the sample by implementing the array structure. The number of neurons into the network is another subject for optimization. We also are planning to tackle this problem in the continuation of this research and compare the results with other well known techniques for text categorization.

6 Acknowledgments

We would like to thank Andréa Pimenta Mesquita, CNAE classifications coordinators at Vitoria City Hall, for providing us with the dataset we used in this work. In addition, we would also like to thank Alberto Ferreira de Souza, Hannu Ahonen, Felipe M. G. França and Priscila Machado Vieira Lima for their technical support and valuable comments on this work. This work is partially supported by the Internal Revenue Brazilian Service (*Receita Federal do Brasil*) and the CNPq, the Brazilian government research agency, under the project number 134830/2006-7.

References

Baeza-Yates, R. and Ribiero-Neto, B. (1998). *Modern Information Retrieval*, 1 edn, Addison-Wesley, New York.

CNAE (2003). *Classificação Nacional de Atividades Econômicas Fiscal*, 1.1 edn, IBGE – Intituto Brasileiro de Geografia e Estatística, Rio de Janeiro, RJ. http://www.ibge.gov.br/concla.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001). *Pattern Classification*, 2 edn, Wiley-Interscience, New York.

Georgiou, V., Pavlidis, N., Parsopoulos, K., Alevizos, P. and Vrahatis, M. (n.d.). Optimizing the Performance of Probabilistic Neural Networks in a Bionformatics Task.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing Company.

Haykin, S. (1998). *Neural Networks – A Comprehensive Foundation*, 2 edn, Prentice Hall, New Jersey.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). Data Clustering: a Review, *ACM Computing Surveys* **31**(3): 264–323.

Mao, K. Z., Tan, K. C. and Ser, W. (2000). Probabilistic Neural-Network Structure Determination for Pattern Classification, *IEEE Transactions on Neural Networks* **11**: 1009–1016.

Michalewicz, Z. (1992). *Genetic Algorithms + Data Structures = Evolution Programs*, 3 edn, Edited Springer-Verlag, Berlin Heidelberg, New York.

Ministério do Desenvolvimento, Indústria e Comércio Exterior – Secretaria do Desenvolvimento da Produção, Departamento Nacional de Registro do Comércio (DNRC) (2007). *Ranking* das Juntas Comerciais Segundo Movimento de Constituição, Alteração e Extinção e Cancelamento de Empresas. http://www.dnrc.gov.br/Estatisticas/ranking_2006.htm.

Oliveira, E., Ciarelli, P. M., Henrique, W. F., Veronese, L., Pedroni, F. and De Souza, A. F. (2007). Intelligent Classification of Economic Activities from Free Text Descriptions, *5º Workshop em Tecnologia da Informação e da Linguagem Humana*, Rio de Janeiro.

Sebastiani, F. (2002). Machine learning in automated text categorization, *ACM Computing Surveys* **34**(1): 1–47.

Specht, D. F. (1990). Probabilistic Neural Networks, *Neural Networks* **3**(1): 109–118.