Yet Another Optimization of the Combinatorial Neural Model

Rafael M. Noivo
Accenture Inc.
Brasília, Brazil, 70.741-640
rafael.moraes.noivo@accenture.com

Hércules A. do Prado
Brazilian Enterprise for Agricultural Research, Embrapa Cerrados,
Brasília, Brazil, 73.301-970
Catholic University of Brasília, Graduate Program in Knowledge and TI Management
Brasília, Brazil, 70.790-160
hercules@{cpac.embrapa.br, ucb.br}

and

Marcelo Ladeira
University of Brasilia, Computer Science Department,
Brasília, Brazil, 70.910-900
mladeira@cic.unb.br

Abstract

Combinatorial Neural Model (CNM) is a classification model that combines both symbolic and connectionist learning approaches. This model is able to recognize regularities from high-dimensional symbolic data, performing mappings from this input space to a set of classes. Due to its hybrid nature, it is possible to extract symbolic relations directly from CNM structure, making it a model of choice for applications that require the rule elicitation. However, this model presents an important drawback: the combinatorial explosion that occurs in its intermediate layer when building the network. To mitigate this problem, CNM has received many modifications that include parallel implementation and relaxation in the building algorithm. In this paper, we describe a new improvement over its architecture that leads to an expressive reduction in the intermediate layer. The model was implemented in the UnBMiner, a framework that provides a large amount of classes for model and data manipulation. An application was developed in the dactyloscopy recognition domain in order to evidence the advantages of our proposal. The space requirement used by our proposal was compared with the usual implementation, and the practical results make clear the gain achieved.
Keywords: Data Mining, Neural Networks, Combinatorial Neural Model, Dactyloscopy Recognition.

Resumo

O Modelo Neural Combinatório (CNM) é um modelo de classificação que combina as abordagens de aprendizado simbólica e conexionista. O modelo é capaz de reconhecer regularidades em dados multidimensionais, realizando mapeamentos entre o seu espaço de entrada e um conjunto de classes. Devido à sua natureza híbrida, é possível se extrair relações simbólicas da estrutura do CNM, o que o torna um modelo interessante para aplicações que requeiram a explicitação de regras. Entretanto, o modelo apresenta uma deficiência importante: a explosão combinatória que ocorre em sua camada intermediária na fase de construção da rede. Para reduzir este problema, o CNM tem recebido diversas modificações que incluem implementações paralelas e relaxamentos no algoritmo de construção. Descrevemos neste artigo uma modificação em sua arquitetura que leva a uma expressiva redução no tamanho da camada intermediária. O modelo foi implementado na plataforma UnBMiner que provê uma grande quantidade de classes para manipulação de dados e modelagem. Uma aplicação no domínio da identificação datiloscópica foi desenvolvida de modo a evidenciar as vantagens da nossa proposta. O requisito de espaço utilizado pela presente proposta foi comparado com a implementação usual e os mesmos resultados práticos obtidos tornaram claros os ganhos alcançados.
Palavras chave: Mineração de Dados, Redes Neurais, Modelo Neural Combinatório, Identificação Datiloscópica.

1. INTRODUCTION

To scale up machine learning algorithms to run over very large databases is an important concern for researchers in the Knowledge Discovery from Databases (KDD) field. In this sense, Combinatorial Neural Model (CNM) [1] has received many improvements ([2], [3], [4] and [5]), that have turned it suitable for KDD. This paper presents an alternative architecture to this model and reports the results achieved when applying it to the dactiloscopy recognition task. The results show that the proposed architecture can lead to a reduced growth rate in the model in comparison to the previous alternative. The implementation was developed in the UnBMiner open source platform. Next section presents a description of CNM. Section 3 explains the proposed architecture. Section 4 describes the dactiloscopy domain in which the application was developed. The last section presents the application and the results obtained.

2. THE COMBINATORIAL NEURAL MODEL

CNM is a hybrid-architecture for intelligent systems that integrates symbolic and connectionist learning paradigms. It has some significant issues, such as the ability to build a neural network from background knowledge; incremental learning by examples, solving the plasticity-stability dilemma [6]; a way to cope with the diversity of knowledge; knowledge extraction of an Artificial Neural Network; and the ability to deal with uncertainty. CNM is able to recognize regularities from high-dimensional symbolic data, performing mappings from this input space to a lower dimensional output space. CNM uses supervised learning and a feedforward topology with three layers: the input layer, the intermediate layer - also called combinatorial - and the output layer (see Figure 1). Each neuron of the input layer corresponds to a concept, which is a complete idea about an object of the domain, expressed by an object-attribute-value form. Their values represent the evidences of the domain application. On the combinatorial layer there are aggregator type neurons, each one connected to one or more neurons of the input layer by fuzzy AND arcs that represent logical concepts. The output layer contains one neuron for each possible class (also called hypothesis), linked to one or more neurons on the combinatorial layer by fuzzy OR arcs that also represent concepts. The synapses may be excitatory or inhibitory and they are characterized by a strength value (*weight*) from zero (not connected) to one (fully connected synapses) that can express the logical relations. For the sake of clarity, we consider only the learning of crisp relations, thus with strength value of synapses equal to one, when the concept is present, and zero, when the concept is not present. The proposed modification does not affect the functions of CNM, since it is applied only in the physical level of the model, having no consequences in the logical level.

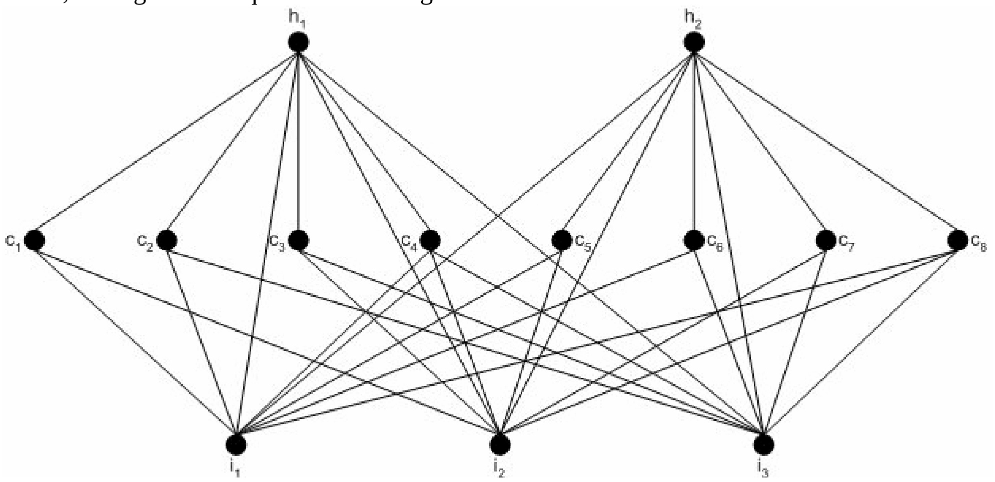

Figure 1 - The complete version of the combinatorial network for 3 input evidences and 2 hypotheses

The network is created completely uncommitted, according to the following steps: a) one neuron in the input layer for each evidence in the training set, b) a neuron in the output layer for each class in the training set, and c) for each neuron in the output layer, there is a complete set of neurons in the combinatorial layer which correspond to all possible combinations of connections with the input layer. This way, each combinatorial neuron represents a combination of connections among two to nine input neurons. This limit in the number of input neuron is a way to avoid combinatorial explosion. This limit is not a heavy restriction; it is only used because of the limitations for processing much information simultaneously [9]. There is no neuron in the combinatorial layer to represent combinations of single input neuron. In this case, each input neuron is connected directly to its hypothesis. The learning mechanism works in a single iteration, and it is described bellow:

```
PUNISHMENT_AND _REWARD_LEARNING_RULE
• Set an accumulator with initial value zero to each arc of the network;
• For each example case from the training database, do:
Propagate the evidence from input nodes until the hypotheses layer;
For each arc reaching a hypothesis node, do:
If the reached hypothesis node corresponds to the correct class of the case
Then backpropagate from this node until input nodes, increasing the accumulator of each traversed arc by its evidential flow (Reward)
Else backpropagate from the hypothesis node until input nodes, decreasing the accumulator of each traversed arc by its evidential flow (Punishment).
```

After training, the value of accumulators associated to each arc arriving to the output layer will be between [-T, T], where T is the number of all cases present in the training set. The last step is the network pruning. It is performed by the following actions:

- remove all arcs whose accumulator is lower than a threshold (specified by a specialist);
- remove all neurons from the input and combinatorial layers that became disconnected from all hypotheses in the output layer; and
- make weights of the arcs arriving at the output layer equal to the value obtained by dividing the arc accumulators by the largest arc accumulator value in the network. After this pruning, the network becomes operational for classification tasks.

Despite its simplicity, CNM has many worthy features, as seen in the previous section. However, it has some drawbacks that limit its use, like: (a) in the initial phase, the generation of the network completely empty, representing all possible combinations for each hypothesis, is clearly unfeasible; (b) as recognized by the authors of the model [1] the full generation of all combinations of attribute-values may create many unreal hypotheses with respect to the majority of the applications; and (c) as a consequence of its knowledge representation form, CNM has its expressivity limited to Propositional Logic. Machado [1] suggests a criterion based on the “magic number” of Miller [9], seven plus or minus two, to establish the upper bound to the order of combinations. Feldens [2] proposes to control the growth of the combinatorial layer by building incrementally the network. The mechanism starts with a low combination order and increases the order to an upper one until an arbitrary limit. Our approach is addressed to this problem and may be seen as an alternative that can rise the order of attribute-values combinations generated from the example cases while keep control of the network growth.

3. PROPOSAL OF OPTIMIZATION

In the usual algorithm to build CNM, for each combination related to a hypothesis, a node is created in the intermediate layer. Also, there are only one-to-one relations between neurons of the intermediate and the hypotheses layers. In practice more than one hypothesis related to the same combination may exist (e.g., in production rule form: *If is_holiday and has_not_to_do Then go_to_the_movie* or *If is_holiday and has_not_to_do Then go_to_swim*). However, this fact leads to a high level of redundancies in the intermediate layer, where the AND nodes are located. Our proposal consists of transform the CNM architecture to a full-connected network. This transformation is possible by changing the relation between the AND nodes and the hypotheses layer. In Figure 2 this change is illustrated. For each hypothesis in the output layer, it is introduced the list of all combinations related to it, and so, an AND node can be referred to any number of output nodes.

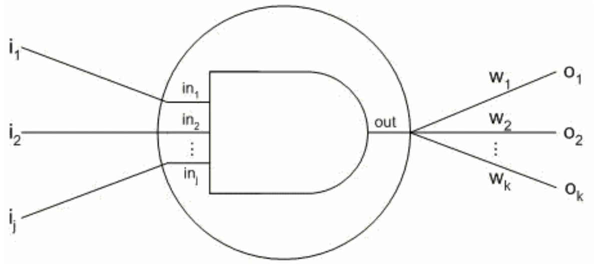

Figure 2 – The AND nodes shape

Considering Figure 1, let us suppose the following equivalences among attribute-value combinations C1=C5, C2=C6, C3=C7, and C4=C8. The resulting network with the proposed simplification is shown in Figure 3. It can be observed an expressive reduction in the network size, even in that simple example. It is expected a much bigger gain in real situations like the one presented in the fifth section that deals with dactyloscopy domain.

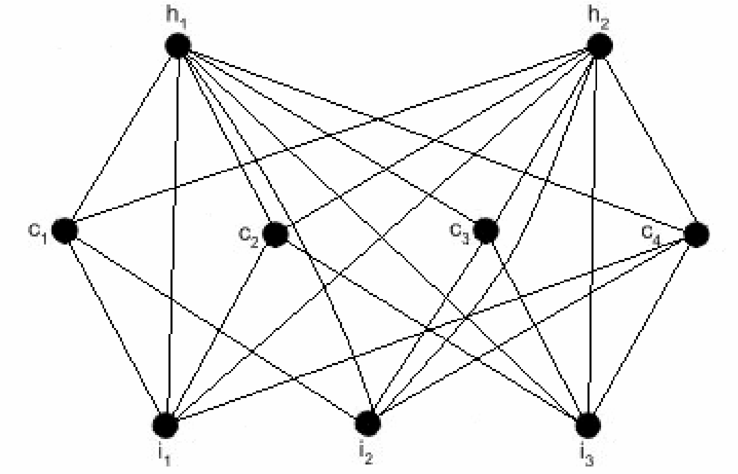

Figure 3 – Optimized architecture

4. THE DACTYLOSCOPY DOMAIN

The purpose of Biometrics is to identify an individual based on his or her physical characteristics. In conjunction with the resources offered by Information Technology, Biometrics offers interesting and effective solutions in the area of public safety, particularly in the identification of individuals involved in criminal activity. Dactyloscopy is a biometrics technique that has been widely used to identify criminals, given that it satisfies the requirements of the permanence, immutability, and singularity of fingerprints [8]. Dactyloscopy is the process by which individuals are identified through the examination of their fingerprints. The fingerprint is the mirror image of the digital pattern.

Dactyloscopic classification systems were developed to reduce the complexity of space and time required for the identification of fingerprints. Two major classification systems have been adopted around the world: Vucetich and Henry. The Brazilian police force employs the Vucetich system, the most widely used method in the world. The Vucetich system [8] defines four primary types of fingerprints with the following classifications: arches, internal loops, external loops, and whorls. Subsequently, the accidental, scar, and amputation types were added. These seven primary types are defined [7] according to the types below (see the primary types in Figure 4):

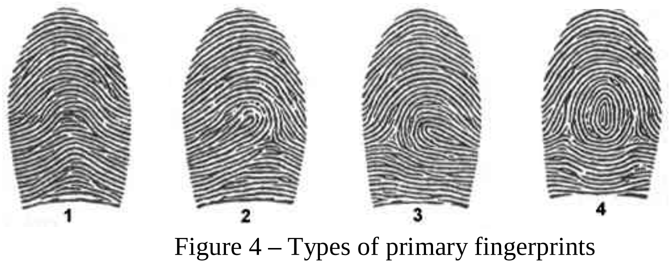

Figure 4 – Types of primary fingerprints

Arch - refers to the dactylogram made up of generally parallel and convex ridges that run or tend to run from one side of the print to the other and very often reveal angular or vertical ridges. The arch type is represented by number 1 or letter *A*.

Internal Loop: refers to the dactylogram that presents a delta to the observer’s right and a nucleus composed of one or more ridges, which run from the left of the print toward the center, recurving and returning, or tending to return, to the side from which they originated, thereby forming one or more loops. Loops involve the two-way movement of a papillary line, which must have perfect inflection. The internal loop type is represented by number 2 or letter *I*.

External Loop: refers to the dactylogram that reveals a delta to the observer's left and a nucleus composed of one or more ridges that run from the right of the print toward the center, recurving and returning, or tending to return, to the side from which they originated, thereby forming one or more loops. The external type is represented by number 3 or letter *E*.

Whorl: refers to the dactylogram characterized by the presence of a delta to the observer's left and right and a varied nucleus, which presents at least one curved ridge in front of each delta. The whorl type is represented by number 4 or letter W.

Accidental: refers to the dactylogram that does not fit in any of the four primary types cited above. It is represented by number 5.

Scar: refers to the dactylogram that presents a permanent mark caused by a cut, pustule, burn, or crushing, thereby making its classification within one of the 5 types cited above impossible and which is represented by number 6.

Amputation (or failure): refers to the type in which a total or partial loss of the phalange is evidenced, therefore compromising or even making impossible the classification of the primary type and which is represented by number 7.

If we create a fraction in which the numerator is the number formed by the algorisms that represent the pattern of the fingers of the right hand, extending from the thumb to the small finger, and the denominator constitutes the same number for the left hand, we arrive at what is known as *dactyloscopic formula*, or formula for short.

5. APPLICATION DESCRIPTION AND RESULTS

For the application developed it was used the data set described and pre-processed in [8]. In that work, the author had taken 855,000 records from "MECASinic" database, property of the Brazilian National Identification Institute. After discarding types 6 (scars) and 7 (amputations), 505,052 formulas remained that were segmented as follows:

- Subset A: formulas with a frequency near 1% or more (the formulas in this subset have more than of 5,000 observations each one). It corresponds to 67,364 observations and just 7 types of formulas;
- Subset B: 366,552 observations corresponding to formulas with frequencies between 0.0022% e 1% (the formulas in this subset have frequencies between 12 to 4,999 observations each one). There are 3.221 different formulas in this subset;
- Subset C: formulas with a frequency smaller than 0.0022% and less than 12 observations for each one. There are 68,136 observations and 34,906 different formulas in this subset;

Each record in the data sets contains one string of 10 positions representing the formula. One formula presents 10 values, one for each finger. For example, the string "3113431142" corresponds to a person with finger types (beginning with the right thumb) external loop, arch, arch, external loop, whorl, external loop, arch, arch, whorl, and internal loop. The classification task was defined as the prediction of the small finger. This task was chosen to illustrate one of the possible model applications, in this case to complement missing information of a dataset. Our experiments were carried out with the subsets A and B. No data mining techniques were applied to subset C because it has a sparse distribution of classes and does not follow any pattern, thus it is not statistically relevant. The identification of individuals involved in criminal activity, which have dactyloscopic formulae in the subset C were, handled ad-doc using a search algorithm because there were less than 12 individuals for each one formula.

Considering that our objective is to evaluate the effectiveness of our architecture in reducing the CNM intermediate layer growth, we adopted a simple scheme to build the model: 65% of each data set to the training phase and 35% to the test phase. Although it is a simple option, it is sufficient for our purpose.

The CNM intermediate sizes found is depicted in Table 1.

Table 1 – Comparing the proposed architecture with the usual model

Data set	Generated combinations (nodes in the intermediate layer)			Reduction rate (%)	Error rate (%)
	Usual model	Proposed architecture	Reduction		
A	99,156	63,762	35,394	35	18

B	439,641	224,303	21,5338	48	15

The results obtained show an important reduction in the growing rate of CNM that can strongly mitigate the combinatorial explosion of its intermediate layer, making the model viable to cope with bigger training sets.

Acknowledgements

We would like to thank Mr. Marcos Elias Cláudio de Araújo, a papilloscopist from the INI/DPF, who kindly provided the access to the “MECASinic” database.

References

[1] Machado, R. J, and Rocha, A. F., The combinatorial neural network: a connectionist model for knowledge based systems. In: Bouchon, B.; Yager, R. R.; Zadeh, L. A. *Uncertainty in Knowledge Bases*, Berlin, Germany: Springer Verlag, 1991, p.578-587.

[2] Feldens, M. A., *Engenharia da Descoberta de Conhecimento em Bases de Dados: Estudo e Aplicação na Área de Saúde*. Porto Alegre: CPGCC da UFRGS, Brazil, 1997. (M.Sc. Dissertation in Portuguese).

[3] Prado, H. A., Frigeri, S. R., Engel, P. M., A Parsimonious Generation of Combinatorial Neural Model. *Proc. of the IV Congreso Argentino de Ciencias de la Computación* (CACIC'98), Neuquén, Argentina, 1998.

[4] Beckenkamp, F. G., Feldens, M. A., Pree, W., Optimizations of the Combinatorial Neural Model. *Proc. of the 5th Brazilian Symposium on Neural Networks*. (SBRN'98), Belo Horizonte, Brazil, 1998.

[5] Prado, H. A. do; Machado, K.F.; Frigeri, S. R.; Engel, P. M. Accuracy Tuning in Combinatorial Neural Model. *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Bejing, China, 1999.

[6] Freeman, J. and Skapura, D. Adaptive Resonance Theory. In: *Neural Networks, Algorithms, Applications, and Program Techniques*. Reading: Addison-Wesley, 1992. 401p. p.292-339.

[7] INI - Instituto Nacional de Identificação, 1987. *Identificação Papiloscópica*, Departamento de Polícia Federal (DPF). Brasília. (Technical Report in Portuguese)

[8] Oliveira, M. G., *Otimização de Busca Decadactilar para Identificação de Impressões Digitais Utilizando Técnicas de Mineração de Dados*. CIC/UNB, Brasil, 2004. (M.Sc. Dissertation in Portuguese).

[9] Miller, G. A., The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, p.81-97, 1956.