Cluster Analysis in Science and Technology: An Application in Research Group Evaluation

Alexandre L. Gonçalves, Roberto Carlos dos Santos Pacheco, Aran Bey Tcholakian Morales, and Vinícius Medina Kern

The STELA Group, PPGEP/UFSC
Rua Lauro Linhares, 2123 torre B, salas 201 a 205
88036-002 Florianópolis-SC, Brasil
{alexl, pacheco, aran, kern}@eps.ufsc.br

Abstract

Parametric methods for science and technology evaluation are frequently rejected because evaluators prefer a subjective approach, usually achieved through peer review analysis. This paper takes a complementary approach to classify research activities by means of machine learning systems. We propose the use of a non-supervised neural network in the building of a ranking of Brazilian research groups. Two indexes are built, expressing productivity and qualification of research groups. The indexes and their relationship are used in the classification of research groups in five categories (*strata*). The results have been consistent with a parametric algorithm currently used by the Brazilian National Research Council (CNPq). In conclusion we suggest the plausibility of applying machine learning in knowledge extraction from science and technology databases.

Introduction

Evaluation of scientific and technological production is critical for science and technology management. It uses parameters for production measurement that can be applied in support of decision-making about the allocation of funding and other resources.

A usual measure of an author's scientific success is the amount of publications in archival journals. However, as pointed out by Price (1976), this measure is incomplete because it accounts only for quantity. Moreover, according to Patterson, Snyder, and Ullman (1999), this is harmful for Computer Science researchers because it works against the preference of the field, which is for conference publication.

Funding agencies use a number of production measurement *criteria*, taking into account the nature of the funding. They are usually more elaborate than simply adding up publications, including widely accepted evaluation policies and also particular methods.

Evaluation methodologies are usually classified into two groups (Kostoff 1997): qualitative (peer review) and quantitative (bibliometry, scientometry, econometric indexes). Schwartzman and Castro (1986) report on methodologies that combine both qualitative and quantitative approaches.

Peer review is widely used in the evaluation of science and technology. Smith (1990) gives a detailed account of the task of the reviewer, especially for article review, and Kostoff (1997) offers a comprehensive definition of this evaluation model: it is a process in which a person or a group evaluate the work of other person or group in the same category or area of knowledge.

Evaluators must be respected academics, expert in their specific knowledge area. The evaluation uses criteria such as quality and uniqueness of the work, scientific and technological impact, and the distinction between the work's revolutionary or evolutionary character.

The Brazilian National Research Council (CNPq), a government-funding agency, evaluates research groups every two years. Research groups usually gather a small to medium number of researchers – a graduate program usually comprises several research groups. The evaluation begins by a curricular analysis whenever a researcher applies to funding. Researchers' evaluations are combined, in order to produce a qualification index for the research group, with the rating assigned to the graduate program the researcher is associated to.

CNPq consultants, based on each applicant's productivity, produce the curricular analyses. The rating for the graduate program comes from an evaluation from CAPES, another government funding agency.

This paper uses data from the Research Group Directory of Brazil (CNPq) and its associated algorithm for the categorization or clustering of these groups.

Self-organizing neural networks

Self-organizing maps (SOM) belong to a class of artificial neural networks in which learning occurs in a non-supervised manner. These artificial networks, introduced

by Kohonen (1995), resemble biological structures found in hearing and visual cortex (Pandya *et al*. 1995). They have been used for pattern recognition and the identification of relevant characteristics.

The architecture of a SOM comprises two completely connected layers. The output or Kohonen layer can be of one or more dimensions. Training is based on competitive learning. Output neurons compete with each other to determine the winner, according to the input. The winner has its weights updated. The same may happen to its neighborhood, depending on the distance from the winner. Weight adjustment is done according to equation 1 or a variation, where *W* are the weights between neurons *i* and *j*, η is the learning rate, and *x* is (the index of) the input node.

$$W_{ij}^{new} = W_{ij}^{old} + \eta (x_I - W_{ij}^{old}) \quad (1)$$

A Kohonen network creates a map for weight adjustment from common input nodes to M output nodes organized in a matrix, usually bi-dimensional, as shown in fig. 1. A set of vector centers results at the end of this process, mapping the input space.

A usual method for the determination of the winner neuron is to minimize the euclidian distance from an input vector, according to equation 2, where *dj* is the distance, *xi* is the input node, and *wij* is the weight vector from the input node *i* to the output node *j*. An advantage of this method, according to Pandya *et al*. (1995), is that it doesn't require weights, or the input vector, to be normalized.

$$dj = ||xi-wij|| \quad (2)$$

Another feature of the architecture is the neighborhood topology, usually rectangular or hexagonal (fig. 1). Weight adjustment is done in terms of the distance from the winner. The topology may use, also, a Gauss function in which the influence of the winner cluster weakens as the distance increases.

Two other parameters may be considered: the neighborhood range *Ne* and the learning rate η. Both can be considered in terms of time, usually decreasing during the organization process.

Research group classification according to the Brazilian National Research Council

The Brazilian National Research Council (CNPq) evaluates the activity of research groups every two years. The evaluation is important for research groups' reputation, since CNPq is an important agency, and opens opportunities for funding, although funds allocation is not officially directed by the evaluation.

Researchers depend, to some extent, on their research groups' classification in order to qualify for a research grant from CNPq or CAPES, another federal funding agency.

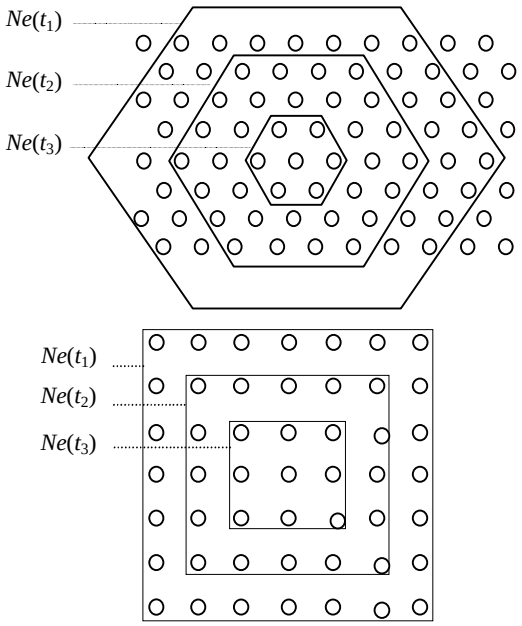

Fig. 1 – Hexagonal and rectangular neighborhood topology

Groups are classified in five *strata*: from A (top, excellence) to E. Researchers in groups of *stratum* A are usually grantees of CNPq's productivity grant system. Their graduate programs are usually best ranked by CAPES.

Parameters for evaluation

The classification of research groups at CNPq is based on two indexes: **qualification** (Q) and **productivity** (P). Qualification is a blend of quality levels of each graduate program (evaluated by CAPES) and their researchers who are grantees of CNPq, as illustrated by the weights in tables 1 and 2. The qualification index Q is normalized with average and standard deviation equal to 50 and 20, respectively.

CNPq assigns the researcher levels in table 1. Program levels, in table 2, are assigned by CAPES to graduate programs in the range [1-7], but only programs with level 3 or more have a pondering weight.

Researcher level	1A	1B	1C	2A	2B	2C
Weight for pondering	1	0.8	0.65	0.55	0.45	0.35

Table 1 – Weights for research group classification according to researchers' level (Guimarães *et al*. 1999)

Program level	7	6	5	4	3
Weight for pondering	1	0.7	0.5	0.35	0.25

Table 2 – Weights for research group classification according to graduate program level, 1997-98 (Guimarães *et al*. 1999)

Productivity is measured in terms of the quantities of several items produced by each researcher in a two-year period. Table 3 lists these items and the weights associated with them. Parameters in tables 1, 2, and 3 are used in the classification of research groups, described next.

Subset *i*	Y*ij*	Nature	Type *Tij*	Weight *v*
1	Y1j	Journal papers	11-National	0.3
			12-International	0.7
2	Y2j	Conference papers and others	21-Full	0.5
			22-Journal w/out editorial body	0.2
			23-Technical mag.	0.2
			24-Abstract	0.1
3	Y3j	Books and chapters	31-Book	0.7
			32-Chapter	0.3
4	Y4j	Technological items	41-Software	0.33
			42-Techn. product	0.33
			43-Process	0.33
5	Y5j	Defenses (advising)	51-Doctoral	0.7
			52-Masters	0.3

Table 3 – Weights for research group classification according to researchers' production (Guimarães *et al*. 1999)

Classification using current algorithm and SOM approaches

This section describes the evaluation of research groups using an algorithm developed by CNPq, and a neural network (SOM) approach for this evaluation. The SOM approach is implemented in parametric and non-parametric versions, i.e., using or not the parameters in tables 1 to 3.

We apply the SOM approach to data of Engineering and Computing research groups. In a first stage we build a classification according to both versions of the SOM. In a second stage, we compare the results given by the networks with the ones given by CNPq (Guimarães *et al*. 1999).

Two networks are built: one for the calculation of qualification, one for the calculation of productivity.

The P-SOM. The Productivity SOM (P-SOM) uses quantitative data from table 3. Equation 3, from CNPq's classification algorithm (Guimarães *et al*. 1999), was used for input data normalization. *Yi* is the productivity measure for the *i*-th subset, *Tij* is the type of work, *v* is the associated weight, and *n* is the number of doctors in the research group.

$$Y = \sqrt{\log\left(1 + \frac{\Sigma T_j * v}{n}\right)} \qquad (3)$$

The parametric version of P-SOM uses weights from table 3. The non-parametric version uses v=1. These are input data for the network, allowing for the building of a weight matrix for the calculation of P. The preliminary productivity index is given by equation 4, where *Ui* is a non-normalized productivity measure and the remaining symbols have the same meaning that in equations 1-3.

$$U_i = \frac{\sum Y_{ij} + w_{ij}}{x} \qquad (4)$$

A normalized index *z*, in the interval [-2.5; 2.5], is built from U using equation 5 (Guimarães *et al*. 1999). Finally, the normalized productivity P is given by equation 6 (Guimarães *et al*. 1999).

$$z_i = \frac{U_i - \mu(U)}{\sigma(U)} \qquad (5)$$

$$P_i = 50 + 20\, z_i \qquad (6)$$

The Q-SOM. The Qualification SOM (Q-SOM) uses qualitative data from tables 1 and 2. We used equations 7 and 8 for normalization (Guimarães *et al*. 1999), where bj is the number of doctors who are researchers and grantees of CNPq, dj is the number of doctors in graduate programs with level equal or greater than 3 (see table 2), n is the number of doctors in each graduate program, and v and w are weights according to tables 1 and 2 (for the parametric version), or v,w=1 (for the non-parametric version).

$$B_i = \frac{b_j * w}{n} \qquad (7) \qquad D_i = \frac{d_j * v}{n} \qquad (8)$$

The preliminary qualification index is given by equation 9, where *Xi* is a non-normalized qualification measure and the remaining symbols have the same meaning that in previous equations.

$$X_i = \frac{\sum Y_{ij} + w_{ij}}{x} \qquad (9)$$

A normalized index *z*, in the interval [-2.5; 2.5], is built from X using equation 10 (Guimarães *et al*. 1999). Finally, the normalized qualification Q is given by equation 11 (Guimarães *et al*. 1999).

$$z_i = \frac{X_i - \mu(X)}{\sigma(X)} \qquad (10)$$

$$Q_i = 50 + 20\, z_i \qquad (11)$$

Results

Fig. 2 shows the distribution of P and Q values for both versions (parametric and non-of parametric) of the SOM, and the distribution from the algorithm of CNPq (Guimarães *et al*. 1999). According to the algorithm, the distribution is organized in classes for Q intervals of 5 – the first class is related to research groups with Q < 20, the top class comprises research groups with Q > 85, adding up to 15 classes.

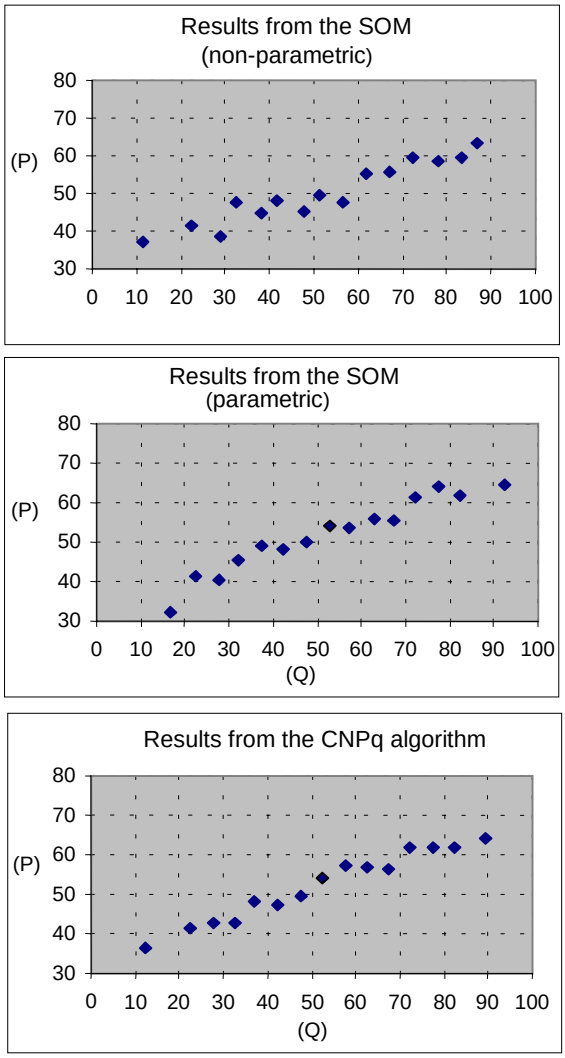

Fig. 2 – Distribution of P and Q indexes for Engineering and Computing research groups according to the SOM approach, non-parametric and parametric, and to the CNPq classification

As a general trend, it is possible to observe a positive correlation between P and Q. However, the SOM non-parametric analysis show some classes that don't follow that trend. This is due to the fact that the network does not consider different weights for the parameters.

Table 4 presents the classification of research groups in five categories according to the *criterion* of CNPq. The categories are formed according to decile intervals of Q. The classification is performed for the three measures in fig. 3.

Stratum	A	B	C	D	E
Interval	$\geq D_9$	D_6 to $<D_9$	D_3 to $<D_6$	D_1 to $<D_3$	$< D_1$

Table 4 – Research group categories (*strata*) according to classification intervals (deciles) (Guimarães *et al*. 1999)

Stratum	Groups (algorithm)	Groups (non-parametric)	Ratio n.-p. / alg.	Groups (parametric)	Ratio p. / alg.
A	*102*	113	1.11	102	1
B	*304*	277	0.91	304	1
C	*319*	333	1.04	315	0.99
D	*192*	187	0.97	195	1.02
E	*238*	245	1.03	239	1
total	*1155*	1155		1155	

Table 5 – Number of Engineering and Computing research groups in the SOM approach, and ratio in relation to the original parametric algorithm (non-parametric and parametric versions)

Table 5 shows a comparison of number of groups in each category, comparing the results of the SOM approach to those of CNPq. Tables 6 and 7 display a comparison of number of groups in each category compared to those with same classification by CNPq.

Stratum	Groups (non-parametric)	Groups (algorithm)	Percentage ratio
A	71	102	69,6
B	213	304	70,1
C	217	315	68,0
D	75	195	39,1
E	194	239	81,5
total	770	1155	66,7

Table 6 – Engineering and Computing research groups classified by the SOM, **non-parametric** version, compared to groups with same classification by the original algorithm

Stratum	Groups (parametric)	Groups (algorithm)	Percentage ratio
A	87	102	85,3
B	263	304	86,5
C	274	315	85,9
D	165	195	85,9
E	227	239	95,4
total	1016	1155	88,0

Table 7 – Engineering and Computing research groups classified by the SOM, **parametric** version, compared to groups with same classification by the original algorithm

The percentage of 66,7% of the non-parametric SOM indicates that it is useful for a preliminary study – for instance, for the establishment of weights. The parametric version reached a coincidence of 88%, very significant if we consider that in several cases the index was very close to the frontier of the correct (coincident) category.

Lastly, tables 8 and 9 exhibit the detailed comparison of each version of the SOM with the CNPq algorithm. The bold numbers show the coincident classifications. Focusing on the SOM classification as category C, for instance, from 333 groups classified as C by the non-parametric version, 217 were coincident with the original

NP Alg	A	B	C	D	E	total
A	**71**	42	0	0	0	113
B	29	**213**	35	0	0	277
C	0	33	**217**	81	0	333
D	0	16	52	**75**	11	187
E	0	0	15	36	**227**	245
total	102	304	319	192	238	1155

Table 8 – Distribution of research groups classified by the SOM, **non-parametric** version (NP), compared to the classification by the original algorithm (Alg)

P Alg	A	B	C	D	E	total
A	**87**	15	0	0	0	102
B	15	**263**	26	0	0	304
C	0	26	**274**	15	0	315
D	0	0	19	**165**	11	195
E	0	0	0	12	**227**	239
total	102	304	319	192	238	1155

Table 9 – Distribution of research groups classified by the SOM, **parametric** version (P), compared to the classification by the original algorithm (Alg)

algorithm, 81 were assigned to category D, 33 to category B, and 2 to category A. Similarly, the parametric version classified 315 groups as C, with 274 coincidences, assigning the remaining groups as 26 B and 15 D.

Conclusion

This paper presented an approach to research group evaluation based on automatic analysis using self-organizing maps. Self-organizing maps were briefly explained. The framework for research group evaluation of a Brazilian agency was outlined. This parametric evaluation measures research groups' productivity and qualification, and classify them in five categories.

A neural network approach to this evaluation was introduced, using two different networks for productivity and qualification measures. Each network was implemented in two versions: parametric and non-parametric.

The neural network approach resulted in a consistent evaluation when compared to the results produced using the agency's framework, with better results using the parametric version. This is due to invariance of weights of the non-parametric version. One strategy to improve the performance of the non-parametric evaluation could be the use, in a first stage, of a supervised network for the assignment of weights for input data normalization.

We consider these results satisfactory, but preliminary. In this stage we have the confirmation of results of well-established parametric algorithms. Further research is recommended, especially dealing with data normalization, similarity metrics, and data evaluation.

The results so far have opened new perspectives for machine learning application in science and technology databases. Considering that the results were consistent, it is plausible that we can apply machine learning in knowledge extraction from science and technology databases. Romão (2002) recently advanced in this direction, applying a fuzzy genetic algorithm for the discovery of interesting knowledge in science and technology databases.

Acknowledgments

This research was supported in part by CAPES, Brazil, in the form of a scholarship held by Mr. Gonçalves during his mastering.

References

Guimarães, R., Galvão, G., Cosac, S. M., Lourenço, R. S. *et al*. 1999. A pesquisa no Brasil: Perfil da pesquisa no Brasil e hierarquização dos grupos de pesquisa a partir dos dados do Diretório dos Grupos de Pesquisa no Brasil. Technical Report, CNPq (Brazilian National Research Council).

Johnson, R. A., Wichern, D. W. 1995. *Applied multivariate statistical analysis*, 4th ed. New Jersey: Prentice-Hall.

Kohonen, T. 1995. *Self-organizing maps*. Heidelberg, Germany: Springer Series in Information Sciences.

Kostoff, R. N. 2000. Research program peer review: Principles, practices, protocols. 1997. [Available on-line at http://www.dtic.mil/dtic/kostoff/index.html. Acess 2000.02.02]

Medler, D. A. 1998. A brief history of connectionism. Department of Psychology, University of Alberta, Alberta, Canada. *Neural Computing Survey* 1:61-101. [Available on-line at http://www.icsi.berkeley.edu/~jagota/NCS/vol1.html. Acess 2000.02.02]

Pandya, A. S., Macy, R. B. 1995. *Pattern recognition with neural networks in C++*. Boca Raton, Fla.: CRC Press, Florida Atlantic University.

Patterson, D., Snyder, L. and Ullman, J. 1999. Best practices memo: evaluating computer scientists and engineers for promotion and tenure. *Computing Research News*, p. A-B, Sept. 1999. Special insert.

Price, D. J. de S. 1976. *O desenvolvimento da ciência: análise histórica, filosófica, sociológica e econômica*. Rio de Janeiro, Brazil: Livros Técnicos e Científicos.

Romão, W. 2002. Descoberta de conhecimento relevante em bancos de dados sobre ciência e tecnologia. Doctoral thesis, Universidade Federal de Santa Catarina (Brazil), Programa de Pós-Graduação em Engenharia de Produção. [To be available at http://teses.eps.ufsc.br]

Schwartzman, S., and Castro, C. M. 1986. *Pesquisa universitária em questão*. Rio de Janeiro, Brazil: Ícone.

Smith, A.J. (1990) The task of the referee, *IEEE Computer* 23 (4):46-51, April 1990.