STUDY OF THE SELF-ORGANIZING MAP OF ONE LOCAL STELLAR SAMPLE

M.HERNANDEZ-PAJARES, R.CUBARSI
Departament de Matemàtica Aplicada i Telemàtica

and

E.MONTE
Departament de Teoría del Senyal i Comunicacions
ETSETB, Universitat Politècnica de Catalunya,
Apartat 30002 - Barcelona, Spain

ABSTRACT

The Self-Organizing Map (SOM) is a neural network algorithm that has the special property of creating spatially organized *representatives* of various features of input signals. The resulting maps resemble real neural structures found in the cortices of developed animal brains. Also the SOM has been successful in various pattern recognition tasks involving noisy signals as for instance speech recognition and for this reason we are studying its application to some astronomical problems. In this paper we present the 2-D mapping and subsequent study of one local sample of 12000 stars using SOM. The available attributes are 14: 3-D position and velocities, photometric indexes, spectral type and luminosity class. The possible location of halo, thick disk and thin disk stars is discussed.

1. Introduction

As [13] indicated, most of the methods currently used in observational Astronomy are rather old and need an urgent updating before they can be used with confidence for the treatment of high quality material provided by orbital observatories. So it is necessary to look carefully to the new trends and tools in the field of Statistics and Information Theory.

One relevant aspect of the studies that are carried out from stellar catalogues is the segregation of stars in populations in terms of spectral, photometric or kinematic criteria. For instance [16, 19, 20, 3, 4] have also worked on this subject recently. The viewpoints adopted by most of these have been statistical, numerical or dynamical approaches.

We present in this paper the application of one recent and powerful statistical tool to the problem of classifying one real stellar sample between several astronomical populations: *thin disk*, *thick disk* and *halo* (see [9]). This tool is the Self-Organizing Map (SOM): a classification scheme with an unsupervised competitive learning algorithm proposed by T.Kohonen in the 80's within the artificial neural network field

(for instance [14, 15]). The main advantage of the algorithm is that it arranges the resulting groups in an associated bidimensional map, where proximity means similarity between the *global* group properties.

The final suggested groups are studied and tested from a kinematic point of view in [5]. Indeed, assuming the superposition of gaussian distributions for the residual velocity it is possible to estimate properties of the mixed populations such the moments of the separate distributions [4].

2. The Self-Organizing Map

2.1. Fundamentals

SOM is an unsupervised neural classifier that has been applied to astronomical data in [11, 12]. The basic aim of this classifier is finding a smaller set $C = \{c_1, \ldots, c_p\}$ of p centroids that provides a good approximation of the original set S of n stars with m attributes, encoded as "vectors" $x \in S$. Intuitively, this should means that for each $x \in S$ the distance $\| x - c_{f(x)} \|$ between x and the closest centroid $c_{f(x)}$ shall be small. However, the main advantage of the algorithm is that it also arranges the centroids so that the associated mapping $f(.)$ from A to S maps the topology of the set S in a least distorting way. Usually A is a bidimensional set of indexes named *Kohonen map* where proximity between them means similarity between the global properties of the associated groups of stars.

From a detailed point of view, the neural network is composed by a set of p nodes or neurons. Every neuron will represent after training, a group of stars with similar features and the weight vector will be approximately the centroid of these associated stars. Following the concise description of [1], the training process consist of presenting sequentially all the training data in parallel to all nodes. For each training vector, each node computes the euclidean distance between its weight and that vector and only the node whose weight is closest to the vector, and its neighbours will update their weights by approaching them to the presented datum. So the nodes compete approaching as many as possible the training vectors. By also updating neighbours' weight instead of just that of winning mode, assures the ordering of the net [14] Finally we will have p good representatives of the input space after training with the associated p groups of input data. In addition, weights of nodes which are close within the grid will also be close within the input space.

The detailed algorithm scheme is:

1. We initialize the weights of the p nodes of the grid with small values: $C = \{c_1, \ldots, c_p\}$
2. For each of the n training vector of the overall database, x_i:
   (a) We find the node k whose weight c_k best approach x_i: $d(c_k, x_i) \leq d(c_l, x_i)$, $\forall l \in \{1, \ldots, p\}$.

(b) We update the weight of the winner node k and its neighbours, $N_k(i)$:

$$c_l(i) = \begin{cases} c_l(i-1) + \alpha(i)\,(x_i - c_l(i-1)) & l \in N_k(i) \\ c_l(i-1) & l \notin N_k(i) \end{cases} \quad l = 1 \ldots p$$

being:

- $\alpha(i)$ a suitable, monotonically decreasing sequence of scalar-valued gain coefficients, $0 < \alpha(i) < 1$. A good choice is a rapidly decreasing function during, let's say, the first 1000 iterations between 0.9 and 0.1 (ordering period); this function can be lineal. After the initial phase, $\alpha(i)$ should attain small values (≤ 0.01) over a long period. A valid dependence is $\alpha(i) \propto 1/i$.
- The radius of the activated neighbourhood $N_k(i)$, a monotonically decreasing function of the iteration i. It can begins with an initial fairly wide value, for $N_k(0)$ (e.g. more than half the diameter of the network), and letting it shrink with time during the ordering phase to, say, one unit; during the fine adjustment phase the radius can be zero (only the winner neuron is activated).

3. The process 2 is repeated for the overall database until a good final training is obtained. A *rule of thumb* is that for good statistical accuracy, the number of steps must be at least 500 times the number of nodes.

2.2. Calculations

The observational data considered is the [6] stellar catalogue (see [7]). It was made from the S.A.O. catalogue that contains all the kinematic and astrophysical information available about more than 250000 stars [17, 18]. The final catalogue contains 12824 stars with enough information to estimate the spatial velocity. The most relevant data for our purposes are the galactic longitude, latitude and the heliocentric distance; the spectral type and luminosity class; the Johnson photometric magnitude and indexes m_v, $B - V$, $U - B$; the spatial residual velocities taking out the simple rotation model in a galactic heliocentric reference frame; and finally the velocity components in the same reference system as the residual velocity components.

SOM has been applied working in a 14 dimensional characteristic space, i.e., the space formed by the 14 properties described above. We assume the symmetry referred to the galactic plane for the galactic latitude and for the perpendicular to galactic plane residual velocity component, that means considering its absolute values $|\,b\,|$ and $|\,W_1\,|$ respectively. In the calculations we have taken $8x8 = 64$ centroids to be determined after 4.10^6 training iterations of the neural network (≈ 330 presentations of the entire database). So, the resultant Kohonen map consists of a two-dimensional grid of $8x8 = 64$ neurons, with a 14 dimensional centroid vector and an associated group of stars for every node. To evaluate the results it is interesting to keep in

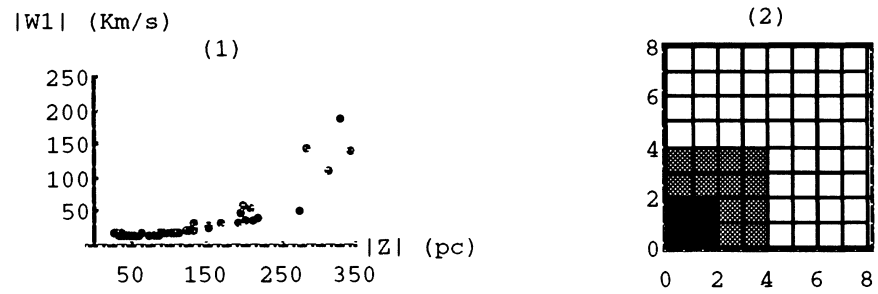

Figure 1: In figure (1) $|W_1|$ is plotted again the height above the galactic plane, $|Z|$, for the 64 centroids calculated. In figure (2) the distribution of the three families of centroids in the Kohonen map, regarding the $|W_1|$ characteristic is represented; (A) with $|W_1| \leq 24 Km/s$ (white squares), (B) with $24 < |W_1| \leq 60 Km/s$ (gray squares) and (C) with $|W_1| > 60 Km/s$ (black squares).

mind that if the j-th characteristic is significant in the segregation problem, then a systematic trend for that characteristic appears in the Kohonen map. The centroids obtained for the stellar catalogue present as the main significant characteristics, the distance and the absolute value of the residual velocity component perpendicular to the galactic plane $|W_1|$. The distance is directly correlated with other significant characteristics such as the spectral type and the luminosity class.

Using the Kohonen map we can segregate the catalogue from an astronomical point of view. Indeed, $|W_1|$ gives us the maximum perpendicular distance to which the star can climb away from the plane. This parameter is related directly with its metallicity and with the population to which the star can belong: disk and halo populations with low and high values, and the recent proposal of a third population, the *thick disk*, with intermediate values of $|W_1|$ [8].

In Figure 1.1 the $|W_1|$ in function of the distance perpendicular to the galactic plane $|Z|$ appears, calculated as $r \sin|b|$, for the 64 centroids obtained. We can distinguish between three groups of neighbouring centroids in the Kohonen map: (A) with $|W_1| \leq 24 Km/s$ (distances basically lower than 550 pc), (B) with $24 < |W_1| \leq 60 Km/s$ (distances between $\approx$ 380 and 1400 pc) and (C) with $|W_1| > 60 Km/s$ and distances in general greater than 1400 pc (Figure 1.2). These intervals agree with the kinematic bins considered by [2]; and related to the metallicity and to the Galaxy populations: region (A) with a predominant thin disk component, (B) with the thick disk and (C) with the halo population. By the other hand the residual velocity moments for the groups (A), (B) and (C) are listed in Table 1. The 2nd. order moments are compatible with the accepted values for the thin disk, thick disk and halo (for instance [4]). A detailed study of each group from these moments is

	Group A (a)		Group B (b)		Group C (c)	
	Moment	Error	Moment	Error	Moment	Error
U_o	10.87	0.34	4.5	1.3	27.1	4.7
V_o	18.09	0.26	13.0	1.1	3.6	3.7
W_o	7.65	0.20	5.0	1.2	21.6	5.7
μ_{11}	1257	31	2820	150	8000	700
μ_{22}	735	24	2110	130	5040	470
μ_{33}	435	16	2440	130	11900	890
μ_{12}	115	18	-	-	1200	400
μ_{112}	-237	23	-	-	-	-
μ_{222}	-363	39	-	-	-	-
μ_{233}	-87	14	-	-	-	-
μ_{1111}	1187	95	4820	580	24600	4300
μ_{1122}	328	30	1570	220	6180	880
μ_{2222}	674	82	3310	580	10600	2000
μ_{1133}	181	18	1310	190	8300	1100
μ_{2233}	158	18	1060	120	5230	640
μ_{3333}	305	60	3460	420	43400	5700

Table 1: The mean residual velocities, U_o, V_o, W_o and the non-vanishing central moments (3-sigma level) of order two, three and four,μ_{ij} , μ_{ijk} , μ_{ijkl} respectively, with the associated errors are listed for the groups A, B and C of stars (those with residual velocity greater than 300 Km/s have not been taken into account). The units for the 2nd, 3rd and 4th order moments are $(Km/s)^2$, $10^2(Km/s)^3$ and $10^4(Km/s)^4$ respectively.

done in [5].

3. Conclusions

In this paper we have applied the Self-Organizing Map method to the study of a stellar catalogue that contains 3D positions, jointly with spectral, photometric and kinematic data for a total of more than 12000 stars in the solar neighbourhood. We have found the existence of three regions of neighbouring centroids in the resulting Kohonen map from the $|W_1|$ attribute. Other important feature in the classification has been the distance. Hence the resulting groups present properties related with the *locality*. These three regions seem to correspond basically to the thin disk (A), thick disk (B) and halo populations (C) also taking into account the respective residual velocity moments. In order to characterize the efficiency of SOM in this kind of astronomical problems it is interesting to apply that algorithm to synthetic samples (see [10]).

4. Acknowledgments

This work has been supported by the D.G.C.I.C.I.T. of Spain under Grant No. PB90-0478.

References

1. A. Cabrera, J. Cid, and A. Hernández, Artificial Neural Networks, Lecture Notes in Computer Science, vol. 540, 401, Berlin, 1991, p. 401.
2. B.W. Carney, D.W. Latham, and J.B. Laird, Astron. J. **97** (1989), 423.
3. R. Cubarsí, Astron. J. **99** (1990), 1558.
4. _____, Astron. J. **103** (1992), 1608.
5. R. Cubarsí, M. Hernández-Pajares, and J. Conrado, In this conference, 1992.
6. F. Figueras, Ph.D. thesis, University of Barcelona, Barcelona, 1986.
7. F. Figueras and J. Núñez, Astrophys. and Space Sci. **177** (1991), 483.
8. G. Gilmore and N. Reid, Monthly Notices Roy. Astronom. Soc. **202** (1983), 1025.
9. G. Gilmore and R.F.G. Wyse, The Galaxy, 247, D.Reidel Publishing Company, Dordrecht, 1987, p. 247.
10. M. Hernández-Pajares, F. Comellas, E. Monte, and J. Floris, In this conference, 1992.
11. M. Hernández-Pajares and E. Monte, Artificial Neural Networks, Lecture Notes in Computer Science, vol. 540, 422, Berlin, 1991, p. 422.
12. _____, The Stellar Populations of Galaxies (B. Barbuy and A. Renzini, eds.), Kluwer Academic Press, Dordrecht, 1992, p. 430.
13. C. Jaschek, HIPPARCOS: Scientific Aspects of the Input Catalogue Preparation II (J. Torra and C. Turon, eds.), 1988, p. 97.
14. T. Kohonen, *Self organization and associative memory*, , Springer Series in Information Sciences, Springer-Verlag, Berlin-Heidelberg, 1989.
15. _____, Proceedings of the IEEE **78** (1990), no. 9, 1464.
16. B.P. Kondratev and L.M. Ozernoy, Astrophys. and Space Sci. **84** (1982), 431.
17. F. Ochsenbein, CDS Inf. Bull. **19** (1980), 74.
18. F. Ochsenbein, M. Bischoff, and D. Egret, Astronom. Astrophys. Suppl. Ser. **43** (1981), 259.
19. R.M. Ros, Rev. Mexicana Astronom. Astrofís. **11** (1985), 23.
20. R.F.G. Wyse and G. Gilmore, Astronom. and Astrophys. **60** (1986), 263.