On Analysis of HI Distribution Using Pattern Recognition Approach

S. Mashchenko

Département de Physique and Observatoire du Mégantic, Université Laval, Québec, PQ, Canada, G1K 7P4

1. Introduction

Pattern recognition (PR) can be viewed as a search for structure in data. A wide spectrum of the PR-based algorithms have been recently proposed for different astronomical applications — ranging from automated spectral classification (Vieira and Ponz 1998) to constructing of all-sky catalog of billions stars and galaxies (Greene et al. 1998). PR algorithms can be divided into two big groups: supervised, and unsupervised ones (Duda and Hart 1973, p.45). Algorithms belonging to the first group are designed to recognize objects of known nature. On the contrary, the unsupervised PR techniques are used when the nature of the objects is not known a priory. This paper describes two different PR algorithms aimed to make analysis and classification of the neutral hydrogen distribution in 21 cm spectral line data-cubes. The first algorithm (described in the section 2) is a supervised one; it deconvolves the HI distribution observed in external galaxies into a set of superimposed expanding superbubble structures (plus the slowly variating background) — on the basis of a lattice of the supershell models. Section 3 describes an unsupervised PR technique designed to classify HI features of unknown nature using the density estimation of the brightness temperature distribution.

2. Automated supershell recognition in external galaxies

For more than two decades the expanding HI supershell structures have been observed both in Milky Way (Heiles 1979), and in nearby spiral and irregular galaxies (see e.g. Brinks and Bajaja 1986, Puche et al. 1992). The size of the shells range from hundreds to more than a thousand parsecs, and the observed HI mass sometimes exceeds 10^6 $M_\odot$. It is generally believed that most of the expanding supershells are powered by a combined action of stellar winds and consequent supernova explosions from associations of O and B stars. Many analytical and numerical models have been developed to explain the phenomena of a supershell (see review by Bisnovatyi-Kogan and Silich, 1995).

Until recently the task of identifying expanding shells (often far from being spheric) in confusing HI background was solved solely with a use of recognition properties of human eye and brain. Thilker et al. (1998) were the first to propose for this purpose the classical PR approach — using a cross-correlation between data (observed HI distribution) and a template (projected supershell model) for all possible translations of the template. To make this procedure feasible for 3D images containing millions of pixels the pattern matching is performed in the

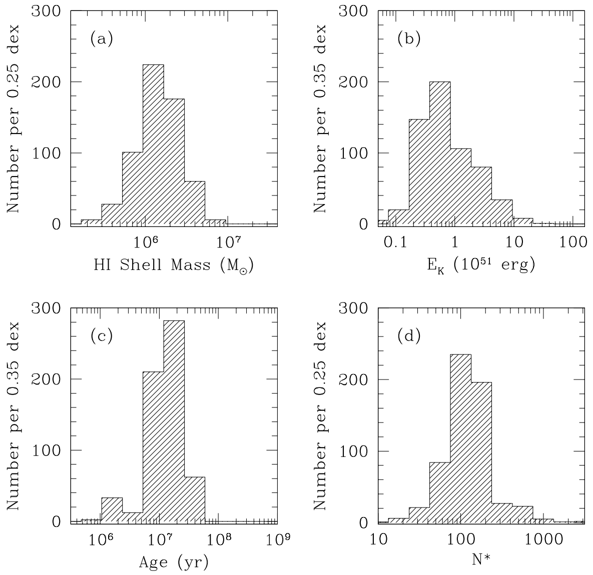

Figure 1. Histograms showing the characteristics of all HI supershells detected in NGC 2403. Panel (a) indicates an HI mass for each structure. Panel (b) demonstrates that our detections usually have modest kinetic energy. Panel (c) shows the distribution of dynamical age for each structure. Finally, Panel (d) presents the distribution of N_*, the number of stars having $M > 7\ M_\odot$ needed to power each expanding bubble.

frequency domain using the cross-correlation theorem (Ritter and Wilson 1996, p.229): $d \oplus t = F^{-1}[F(d)F^*(t)]$. Here the symbol $\oplus$ denotes cross-correlation, d is a data function, t is a template function, F and F^{-1} denote direct and inverse Fourier transforms, and the operation of complex conjugation is denoted by the asterisk. Local maxima of the cross-correlation function correspond to the regions where the observed HI distribution is well matched by the template model function. Applying this analysis for a lattice of supershell models the final catalog of the shell detections can be generated.

In Mashchenko et al. (1998b) we improved upon the original method by implementing the two-dimensional numerical hydrodynamical code for generating the model templates. Another major improvement was using a noise-corrected estimator of the normalized correlation coefficient $\tilde{r}$ as a robust measure of the quality of matching between model and observational data. It can be shown that $\tilde{r}$ is an invariant of linear transformations of the template function. This can have the following physical interpretation: $\tilde{r}$ is an invariant of both 1) presence of the locally homogeneous background emission (assuming that the gas is optically thin), and 2) density of the surrounding quiescent gas (it is valid if the mechanical luminosity of the source of energy is not known a priory, and if one can neglect all physical processes which have non-linear dependency on the gas density — such as a radiative cooling, and self-gravity of the gas).

We have applied the developed PR technique to detect supershells in the spiral galaxy NGC 2403 (Mashchenko et al. 1998b). As a result, we obtained a catalog of some 600 shells and supershells. Histograms for a few characteristics of the detected supershell population are shown in the Fig. 1.

3. HI features classification

The problem of dividing the features extracted from HI emission data-cubes into physically meaningful classes without any a priory assumptions about their nature belongs to the domain of unsupervised pattern recognition.

We define an HI feature either as a contiguous region in spectral line data-cube with all pixels having values above some specified threshold brightness temperature value (for an analysis of the objects which are believed to be complete), or as a small sub-cube (to deal with the small pieces of highly incomplete or diffuse objects).

Ghazzali et al. (1998) proposed to base the classification of HI features on the analysis of the density estimation of the brightness temperature distribution for all pixels belonging to the feature. In Mashchenko et al. (1998a) we proposed a PR algorithm, which divides features into different classes on the basis of their brightness temperature distributions ρ. As a measure of dissimilarity between any two features we use the following quantity:

$$d = 1 - \frac{\sum \rho_1(T_i)\rho_2(T_i)}{\sqrt{\sum \rho_1^2(T_i) \sum \rho_2^2(T_i)}},$$

where ρ_1 and ρ_2 are the density of the brightness temperature distribution for the first and second feature, and T is the brightness temperature. One can show, that $0 \leq d \leq 2$. For identical distributions $d = 0$.

We calculate the dissimilarity d taking into account two following PR invariants: (1) we allow for a presence of the locally homogeneous background (assuming, that the gas is optically thin), and (2) when our features are complete objects (not sub-cubes), the invariant of the distance to the object is used.

Having calculated the dissimilarity d for all pairs of features, the fuzzy clustering of the dissimilarity matrix can be carried out. Currently we use the algorithm `fanny` proposed by Kaufman and Rousseeuw (1990). The number of clusters to find K is an input parameter. For a given K the fuzzy clustering algorithm calculates for each feature K values of the cluster membership (or affiliation) coefficient C_m $(0 < C_m < 1)$.

To test whether the developed PR technique can perform a physically meaningful classification of HI features, we applied the algorithm to the region centered (both spatially and in velocity) at the location of Sh2–203 HII region in CGPS 21 cm spectral line data-cube (English et al. 1998). The size of the region is 2°.25x2°.5x28.84 km/s (450x500x35 pixels). The region was splitted into 9x10x7=630 smaller sub-cubes with the size 0°.25x0°.25x4.12 km/s (50x50x5 pixels). The Fig. 2 shows the smoothed distribution of the membership coefficient C_m as contours for one particular cluster (total number of clusters K is equal to 3). One can get an impression, that the pieces of the gas which brightness temperature distributions constitute this cluster follow the walls of the shocked material surrounding the HII region. The values of C_m for the front shell wall (right image) are comparable with C_m values for the rear wall (left image) notwithstanding the big difference in their HI emission brightness. The location of the maximum of the membership coefficient distribution for the front wall coincides both spatially and in velocity (Fich et al. 1990) with the location of the local maximum in ionized matter distribution. This (along with the lower

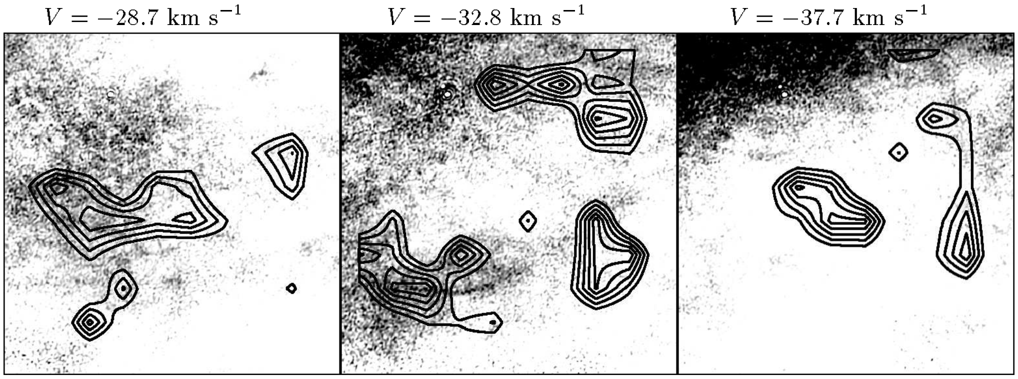

Figure 2. The smoothed distribution of the membership coefficient C_m (with levels at $C_m = 0.3, 0.4, \ldots 1.0$) calculated for $K = 3$, cluster No.3 (contours) superimposed with the HI distribution (grey-scale) for 3 different velocity channels

HI column density) can be explained by assuming that the front wall of the shell is partially ionized.

References

Bisnovatyi-Kogan, G.S. and Silich, S.A. 1995, Reviews of Modern Physics 67, 661

Brinks, E. and Bajaja, E. 1986, A&A 169, 14

Duda, R.O. and Hart, P.E. 1973, Pattern Classification and Scene Analysis (New-York: Wiley)

English, J. et al. 1998, PASA 15, 56

Fich, M., Dahl, G.P. and Treffers, R.R. 1990, AJ, 99, 622

Ghazzali, N., Joncas, G. and Jean, S. 1998, ApJ, in press

Greene, G., Mclean, B. and Volpicelli, A. 1998, BAAS 192, 5501

Heiles, C. 1979, ApJ 229, 533

Kaufman, L. and Rousseeuw, P.J. 1990, Finding Groups in Data: an Introduction to Cluster Analysis. (New York: Wiley)

Mashchenko, S., Joncas, G. and Ghazzali, N. 1998a, in preparation

Mashchenko, S.Y., Thilker, D.A. and Braun, R. 1998b, A&A, in press

Puche, D., Westpfahl, D. and Roy, J-R. 1992, AJ 103, 1841

Ritter, G.X. and Wilson, J.N. 1996, Handbook of Computer Vision Algorithms in Image Algebra (CRC Press: Boca Raton)

Thilker, D.A., Braun, R. and Walterbos, R.A.M. 1998, A&A 332, 429

Vieira, E.F. and Ponz, J.D. 1998, in: Astronomical Data Analysis Software and Systems VII, eds. R. Albrecht, R.N. Hook and H.A. Bushouse (ASP Conference Series, Vol. 145), p.508