Fisheries Research 200 (2018) 49–60

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Review

Robust clustering methodology for multi-frequency acoustic data: A review of standardization, initialization and cluster geometry

Marian Peña

Instituto Español de Oceanografía, Centre Oceanogràfic de Balears, Moll de Ponent s/n, 07015 Palma, Spain

ARTICLE INFO

Handled by Prof. George A. Rose

Keywords:
Cluster geometry
Correlation
Frequency response
Standardization
Variance
K-Means
EM clustering

ABSTRACT

Clustering is a useful unsupervised technique for the identification of acoustic groups in multi-frequency echograms based on frequency response. K-Means is the most well-known clustering technique but has significant requirements such as clusters of equal size and spherical shape. Initialization is a common problem in clustering as only local minima are usually guaranteed, and thus initialization must locate the centroids near the global minimum. Expectation-Maximization (EM) clustering also requires a good set of initial centroids but allows the identification of clusters with different statistical distributions. This work presents the comparison of these techniques applied to a case with several acoustic signatures presenting different cluster sizes and distributions. The main issues treated in this manuscript are: pre-processing of acoustic data for clustering, initialization of centroids with theoretical scattering models and the need to consider the geometry of the clusters in addition to means, including variance (spread around the mean), orientation (correlation between variables), spherical or ellipsoidal shape (difference in variance between variables) and cluster size (number of observations). EM clustering is the only technique that properly separates acoustic signatures (and noise) after using the supervised initialization presented in this study.

1. Introduction

Fisheries acoustics is a discipline that examines fishes and plankton species based on their scattering properties using the measured scattering intensity known as volume backscatter (S_v, dB re m^{-1}) (Simmonds and MacLennan, 2005). The identification of acoustic echotraces has traditionally been conducted through net sampling, known as 'ground-truthing'. However, linking acoustic and net data is complicated due to, among other things, net avoidance and acoustic shadowing of species with lower scatter. Net sampling of deep-distributed species such as mesopelagic fish often challenges the available logistics. In addition, sampling in acoustic surveys are often directed at schools/layers with higher scatter, as echotraces of lower numerical density or those that contain species with lower scatter are more difficult to spot. A priori knowledge of the location of different species or acoustic typologies in the echogram allows the proper sampling of all the desired targets (when biological information is also needed), and may be used to make commercial fishing more efficient, reducing bycatch. The identification of acoustic groups based on acoustic data without ground-truthing requires the employment of an unsupervised technique. Ideally, a quick and not very computationally demanding methodology is desired, such as clustering.

Clustering is an unsupervised machine learning technique that groups data according to similarity in the variables provided as input. As an unsupervised method, there is no training data with labels orientating the algorithm to a particular solution. Several papers have summarized the main clustering techniques (Banerjee and Davé, 2012; Xiao and Yu, 2012), which can be divided into hard-clustering, where one data point can only belong to one cluster, and fuzzy or soft clustering, where each data point may belong to several clusters through a membership function. The second group handles better overlapping clusters and is less sensitive to noise as noise influence is equally split among groups.

The most well-known clustering techniques have been designed for data without noise or outliers (Xiao and Yu, 2012). Robust variations have been posteriorly developed to adapt to real measurement data that contains noise. As shown in this paper, most clustering algorithms must also be robust for initialization (initial centroid estimation). Furthermore, the geometric characteristics of the data used is often overlooked, such as cluster size and shape. For instance, the most popular algorithm, K-Means, requires data with clusters of equal size and variance (spherical clusters). Different clustering algorithms or distance measures can lead to very different results (Jain et al., 2004). There is no single algorithm suitable for all applications and thus, data knowledge and requirement checking would reveal the most suitable. This work focuses on that analysis for fisheries acoustic data.

E-mail address: marian.pena@ba.ieo.es.

https://doi.org/10.1016/j.fishres.2017.12.013
Received 30 September 2017; Received in revised form 18 December 2017; Accepted 21 December 2017

The incorporation of several frequencies into fisheries and plankton acoustics gave birth to what it is known as multi-frequency methods, where the difference between frequencies is employed to identify acoustic groups, comparing their spectrum with theoretical scattering models. Species are categorized into three acoustic groups: gas-bearing (including a swim bladder or pneumatophore), fluid-like (with a weak acoustic signal, such as krill and copepods), and elastic shell (pteropod type) (Stanton et al., 1996). The first group presents a resonance peak at a particular frequency that depends on swim bladder size (near 18 kHz for lantern fish and around 4 kHz for small pelagic fish). The second and third groups present increasing scatter with frequency shifted in frequency with length. For vessel-borne echosounders, S_v is measured within a volume that increases with depth. Assuming only one acoustic typology is present in the volume, S_v is dependent on the scatter of one single organism (target strength, *TS*) and its numerical density ρ, following the equation $S_v = \mathrm{TS} + 10*\log_{10}(\rho)$. To remove numerical density dependence, each S_v is subtracted by the S_v of a reference frequency, usually 38 kHz for historical reasons (as it was the most common first frequency onboard research vessels). The results are known as Frequency Response $\mathrm{FR} = S_{vi} - S_{v38} = \mathrm{TS}_i - \mathrm{TS}_{38}$, which reduces the number of variables to the number of frequencies minus one (as *FR*(38) will be equal to 1 for all data points, and thus, will have little influence on the clustering; see discussion for further information). Typical working frequencies are 18, 38, 70, 120, 200 and 333 kHz but, as the usable range (depth if vertically orientated) decreases at higher frequencies, the number of frequencies that can be employed depend on the depth of the targeted species. S_v data are thus a type of curve data, like time series, where the trend (with frequency instead of time) is used to identify groups, but unlike time series, frequency is a dependent variable, while time is not (Pereira, 2013). The dependence of S_v values with frequency (serial correlation) has been modeled for the different acoustic groups. See, for example, Peña and Calise (2016) for the krill model adapted to short-length species and Peña et al. (2014) for mesopelagic fish models. As in time series, frequency shifts are bound to appear, due to length differences of organisms (reflected in the *TS* value), as well as vertical offset due to numerical density differences ($10 * log_{10}(\rho)$ term). Calculating the *FR* removes that offset and achieves some translation invariance, in a similar way that it is done for detrending in time series. The frequency shift is minimal for similar sizes, but could be the key to differentiate different species with similar *FR* tendency, but very different size, such as krill (~2–4 cm) and Mysidacea (~0.5–2.5 cm). The frequency spectrum (*FR* variation with frequency) has to be maintained in pre-processing and considered in the clustering.

In fisheries acoustics, data noise is often classified as background noise and impulse noise (Ryan et al., 2015). Background noise refers to ambient and vessel noise that affects all pings and varies in intensity and pattern with vessel speed, propeller pitch, bottom depth, number of vessels in the area, etc. (Peña, 2016). Impulse noise is usually caused by interferences with another acoustic device and affect a few pings. Several algorithms have been published to remove background and impulse noise (Ryan et al., 2015; Peña, 2016). Data with very low threshold also include white noise, a random signal having equal intensity at different frequencies. They are a sequence of serially uncorrelated random data with zero mean and finite variance. This noise needs to be accounted for when modeling acoustic data. The sample unit considered in this paper is the pixel, i.e. each data point in the 2D echogram as sampled by the echosounder. For an EK60 with 1 ms pulse duration, a pixel has a vertical length of ~19 cm. The horizontal length changes with beam width and depth due to the conical shape of the acoustic beam. For a 7° beam, the horizontal length is ~12 m at 100 m depth and ~30 m at 500 m. Each pixel represents a particular sampled volume that changes with distance to the transducer and beam angle. Differences in sampled volume between frequencies need to be accounted for when comparing pixels, particularly in cases of small echotraces.

The aim of this paper is to study the behavior of clustering techniques with multi-frequency acoustic data, very noisy data with clusters that can have very different sizes (proportion of echogram pixels). A very robust initialization procedure based on theoretical models that properly locates centroids and provides an estimation of the number of clusters is presented. The use of standardization is also analyzed. The paper is organized as follows: a short summary of clustering methods and their requisites is given, focusing on two techniques: K-Means (KM) and Expectation-Maximization (EM) clustering (also known as Gaussian Mixture Model or GMM). KM and EM clustering have already been used with acoustic data (see Section 1.3) and are both included in the top ten algorithms in data mining (Wu et al., 2008). The geometry of clusters is defined and shown with examples. A review of clustering applied to multi-frequency acoustic data is then given. The material and methods section presents the novel technique to initialize centroids. Finally, the two techniques are compared using a challenging example and the suggested initialization method.

1.1. Clustering review

Clustering techniques can be classified based on the clustering approach as center-based techniques, where one cluster is represented by its center, such as K-Means (Lloyd., 1982); density-based clustering like DBSCAN (Arlia and Coppola, 2001), where clusters are defined as areas of higher density surrounded by lower density areas; and distribution-based techniques, with clusters defined as objects belonging to the same distribution. Gaussian Mixture models fitted with an Expectation-Maximization (EM) algorithm (Krishnan and McLachlan, 1997) are included in the last category, and allow clusters to have different variances, density and size. Density-based clustering also allows the separation of clusters of different size, but requires the calculation of distances between all pair of data points, which is too computationally expensive with acoustic data.

Two of the critical aspects of clustering techniques are the pre-allocation of number of clusters and initialization of the centroids. Preselecting the number of clusters K is still a very challenging problem in clustering. The available techniques to estimate K are based on comparing different runs of the algorithm, which make them cumbersome. Even though several cluster validity indices (CVIs) exist, they are inefficient when clusters widely differ in density or size (Zalik, 2010). They are usually based on maximizing compactness and minimizing overlap among clusters, but in the presence of noise, overlapping is prone to appear. Distances between centroids do not take into account the cluster shape and dispersion; points from two neighboring but not dispersed clusters can be more separated than two spread clusters that overlap, despite the distance between the centroids being large. Using only centroid information (such as with the Davies-Bouldin measure (DB) (Davies and Bouldin, 1979), the Hartigan index (Ha) (Hartigan, 1975) or the Krzanowski-Lai index (KL) (Krzanowski and Lai, 1988)) is not sufficient to interpret the geometrical structure of the data, and therefore not sufficient for the separation between clusters. The elbow method, one of the most common CVIs based on the variance curve, was found to be unsuitable for several datasets in Milligan and Cooper (1985) and, as seen in Santos and Embrechts (2014) with 30 benchmark datasets, no cluster validation index is perfect.

In general, clustering algorithms guarantee convergence to the closest local minima, so the initial location of the centroids must ensure that this minimum is also the global minimum. MacQueen (1967) suggested choosing K random observations as initialized centroids, but different initialization runs may generate rather different clusters and more dense clusters have a higher probability to attract one or two centroids.

Center-based techniques assume all clusters are spherical (equal variance-covariance). Often standardization/normalization (centering each variable to 0 and scaling by its standard deviation or range) is used to equal the variance of all variables. Multifrequency echograms often