A Fuzzy Clustering Technique for Calorimetric Data Reconstruction

Radha Pyari Sandhir,* Sanjib Muhuri, and Tapan K. Nayak
Variable Energy Cyclotron Centre, Kolkata - 700064, INDIA

Introduction

In high energy physics experiments, electromagnetic calorimeters are used to measure shower particles produced in p-p or heavy-ion collisions. In order to extract information and reconstruct the characteristics of the various incoming particles, clustering is required to be performed on each of the calorimeter planes. Hard clustering techniques such as Local Maxima Search, Connected-cell Search and K-means Clustering simply assign a data point to a cluster. A data point either lies in a cluster or it does not, and so, overlapping clusters are hardly distinguishable.

Fuzzy c-means clustering is a version of the k-means algorithm that incorporates fuzzy logic, so that each point has a weak or strong association to the cluster, determined by the inverse distance to the center of the cluster. The term *fuzzy* is used because an observation may in fact lie in more than one cluster simultaneously, though to different degrees called 'memberships', as is the case with many high energy physics applications. The centers obtained using the FCM algorithm are based on the geometric locations of the data points.

The Fuzzy c-Means Algorithm

The fuzzy c-means (FCM) clustering algorithm [1] is one of the most widely used fuzzy clustering techniques. It seeks to minimize a sum of squared errors objective function. Optimization of the objective function is based on iteration through certain necessary conditions by using Fuzzy c-Means Theorem: if $D_{ik} = \|x_k - v_i\| > 0$ for all i and k, then (U,V) may minimize J_m only if, when $m > 1$,

$$u_{ik} = \left[\sum_{j=1}^{C} \left(\frac{D_{ik}}{D_{jk}} \right)^{\frac{2}{m-1}} \right]^{-1} \quad (1)$$

where $1 \leq i \leq C$ and $1 \leq k \leq n$, and

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m}. \quad (2)$$

where $V = (v_1, v_2, \ldots, v_C)$ is a vector of unknown cluster centers or prototypes, U consists of the memberships u_{ik} of the k^{th} point in the i^{th} cluster, and $D = \|x\| = \sqrt{x^T x}$ is any inner product norm. The algorithm loops through one cycle of estimates for $V_{t-1} \rightarrow U_t \rightarrow V_t$ until some error criteria is reached. An error threshold ϵ can be specified so that the error criteria is $\|V_{t-1} - V_t\|_{err} \leq \epsilon$.

Choosing a suitable number of clusters involves the evaluation of how well the cluster centres fit the data. This can be done with the help of a validity index. All simulations discussed in this paper made use of the Xie-Beni index, which is the ratio of the total variation of the cluster centres and memberships of the observations in the groupings to the separation between the cluster prototypes [2]. The most suitable clusters are the ones that minimize this index. Since FCM is a static algorithm taking all the data points into account at one time, it is sometimes unable to identify clusters with non-uniform data patterns, as will be shown. Therefore, a dynamic version of the algorithm has been studied.

The dynamic Fuzzy c-Means (dFCM) algorithm

The dFCM algorithm [3] is a modification of the fuzzy c-means algorithm, allowing cluster prototypes to be adaptively updated as data

*Electronic address: radha.pyari@gmail.com

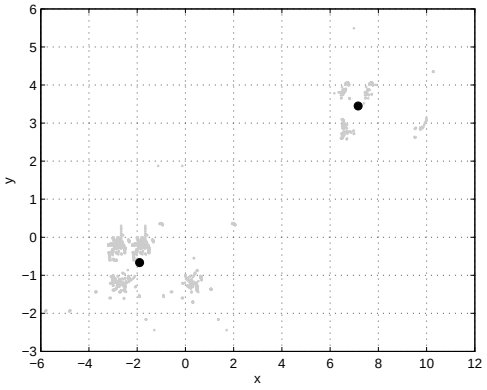

FIG. 1: Clusters found using the simple FCM algorithm. Only large clusters are found.

points keep streaming in. If a new cluster is formed, then a new cluster prototype is automatically generated. Similarly, if redundant clusters are present at a given instant, cluster centres are eliminated. Details of the algorithm can be found in [3].

Application to Calorimetric Data

In a sampling calorimeter with several sensitive planes, each particle deposits a large amount of energy in several pads or pixels. Clustering of data is needed for each plane in order to obtain the position and energy deposition of the cluster. Simulations based on GEANT4 have been performed on a calorimeter with 20 layers, out of which 3 were pixel layers with dimensions $0.5mm$ x $0.5mm$, and the rest were pad layers with cell dimensions $1cm$ x $1cm$. The clustering was performed in 3 dimensions, i.e., the x -position, the y - position and the energy deposited at each point, in order to take the energy deposition (keV) values into account. The final clusters were taken as the projections of the 3D clusters on the xy plane.

A sample of eight clusters on one of the pixel layers has been generated to study the clustering algorithms. Figure 1 shows the clusters obtained by the simple FCM algorithm. As the data appears to be two groups of four clusters each, the FCM algorithm could not resolve the individual clusters, and therefore the Xie-Beni index selected 2 clusters to be satisfactory. However, applying the dFCM algorithm, it was found that all 8 clusters could easily be identified, as shown in Figure 2.

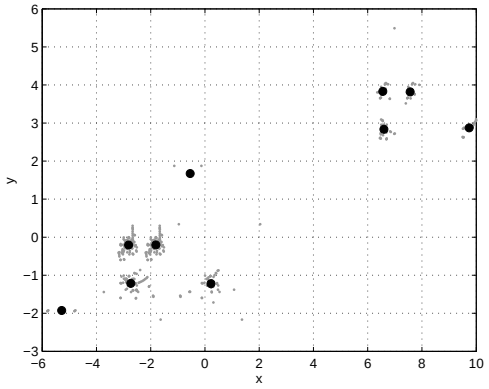

FIG. 2: Clusters using the dFCM algorithm. All the clusters are properly determined.

Clustering was also performed on a set of overlapping clusters, and it was found that those clusters were easily resolved as well.

Conclusions

For calorimetric data reconstruction, fuzzy clustering provides more accurate data than hard clustering. The Fuzzy c-Means algorithm is good for clusters that lie uniformly in space, while the dynamic Fuzzy c-Means algorithm can tackle data patterns that are non-uniform as well.

References

[1] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms.* New York: Plenum, 1981.

[2] N.R. Pal and J. C. Bezdek,"On cluster validity for the fuzzy c-means model" *IEEE Trans. Fuzzy Systems*, vol. 3, no. 3, 1995.

[3] R.P.Sandhir and S. Kumar,"Dynamic Fuzzy c-Means (dFCM) Clustering for Continuously Varying Data Environments" *Proc. IEEE World Congress on Computational Intelligence*, Barcelona, July 18-23, 2010.