Application of fuzzy-based pattern recognition techniques for cluster finding in a preshower detector in High Energy Heavy Ion Experiments

Susanta Kumar Pal,* Subhasis Chattopadhyay, and Y. P. Viyogi
Variable Energy Cyclotron Centre, 1/AF Bidhan Nagar, Kolkata 700064, INDIA

Introduction

In high energy heavy ion collision experiments, preshower detectors are employed for the measurement of photon multiplicities at forward rapidity. In a highly granular preshower detector of 3X0 thickness, photon clusters are affected mostly by 2 to 3 cells. In the forward rapidity region where the particle density is very high, the overlap between showers is too large to be separated, thereby requiring specilised cluster finding algorithm for identification of photons.

We have studied a method of cluster finding based on Fuzzy c-mean (FCM) algorithm. In the literature on fuzzy clustering, FCM clustering algorithm is one of the most well known and used methods [1] in many branches of science. In the FCM algorithm, the cluster-finding procedure involves two steps, which are iterated for proper validation. These steps are (a) initiation of the cluster finding procedure with a seed cluster number ≥ 2 and (b) use of proper validation procedure (called validity indexes) to finally obtain the optimum number of clusters. The performance of the FCM clustering depends on the use of proper validity index. A large number of validity indexes have been proposed and investigated in literature for different data sets with varying degree of success [2]. However in most of these tests, the data set used had fewer clusters, ranging within 2-10, and having little variation within the sets. In the present work, we have investigated the application of FCM clustering algorithm on the simulated data from a preshower detector built for the ALICE experiment [3], where the number of clusters in a set vary widely within a range of a few to a few hundred. This is expected to provide a more rigorous test for the validity indexes. We have also studied the performance of the clustering algorithm for fixed set of parameters with different particle density environments.

*Electronic address: sushant@vecc.gov.in

Results and Discussion

A. Clustering Efficiency, Ghost clusters and Track Loss

Cluster splitting and merging to some extent may occur in any clustering algorithm, where the extent depending on the nature of the data set. Due to high granularity of the detector in the present case the hit pattern may get distorted in such a way that the clustering algorithm may produce more than one cluster for one pattern originating from a single track. This leads to splitting of clusters. In the discussions to follow, whenever there are two or more clusters for the same track, we label all the clusters having distance more than 2 cm from the main cluster as 'deflected' and those within this distance as 'split'. Thus for any pattern having $N_{\rm inp}$ number of input tracks, we get the following number of clusters of different categories:

$N_{\rm raw}$ = optimum number of clusters as obtained by FCM using any given validity index

$N_{\rm spl}$ = number of split clusters

$N_{\rm def}$ = number of deflected clusters

The number $N_{\rm abs} = N_{\rm raw} - (N_{\rm spl} + N_{\rm def})$ gives the number of clusters in the set where there is only one cluster for a given track. Considering that the 'deflected' clusters are not the artifact of the clustering algorithm, we define the following set of clustering efficiency to compare the performance of various validity indexes:

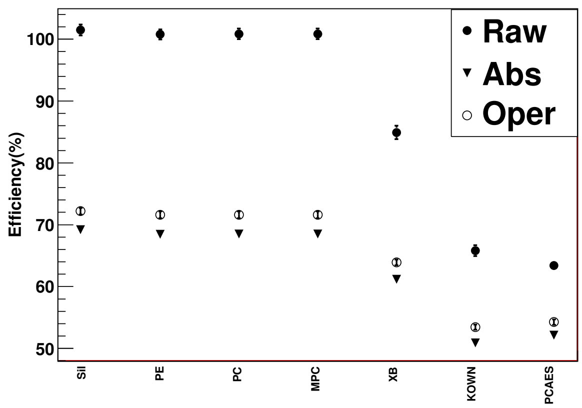

FIG. 1: The clustering efficiency for various validity indexes.

Raw efficiency ($\epsilon_{\rm raw}$) = $N_{\rm raw}/N_{\rm inp}$
Absolute efficiency ($\epsilon_{\rm abs}$) = $N_{\rm abs}/N_{\rm inp}$
Operational efficiency ($\epsilon_{\rm opr}$) = $(N_{\rm abs} + N_{\rm def})/(N_{\rm inp} + N_{\rm def})$

Fig. 1 shows the three types of clustering efficiencies defined above for the seven validity indexes. The raw efficiency goes above 100% in some cases. We find that the validity indexes PC, PE, MPC and Silhouette perform almost equally well, giving almost similar values of the three types of efficiency. Both the average absolute efficiency and the average operational efficiency are close to 70% in all the four cases, the difference between the two being less than 5% in all the cases. The validity index XB gives a performance somewhat worse than the above four but much better than PCAES and Kwon indexes. The later two indexes perform not so well, the average clustering efficiency being too low.

For further discussion given below, we consider only the three indexes : Silhouette, XB and Kwon giving large, intermediate and low values of efficiency for the present data set.

B. Cluster Properties

A good clustering algorithm should produce clusters having the following properties :

- The relative separation between the cluster centroid and the associated track position extrapolated on the detector plane should be close to zero. This is measured by calculating the difference $\delta\eta = \eta_{\rm track} - \eta_{\rm cluster}$ and $\delta\phi = \phi_{\rm track} - \phi_{\rm cluster}$, where (η, ϕ) are the pseudo-rapidity and azimuthal angle variables defined in the usual sense.
- The total cluster strength (sum of deposited energy in various member cells) should be close to the strength for the given track. This is measured by calculating the ratio $R_{\rm edep}$ of signal strength in the cluster to the signal strength in the associated track.
- The number of cells in the cluster should be close to the number of cells associated with the track. This is measured by calculating the ratio $R_{\rm ncell}$ of number of cells in the cluster to that in the track.

TABLE I: Various characterizing parameters (see text).

Validity	$\delta\eta$	$\delta\phi$	$R_{\rm edep}$		$R_{\rm ncell}$	
Index	RMS	RMS	Mean	RMS	Mean	RMS
Sil	0.12	0.14	1.42	2.9	1.29	1.12
XB	0.12	0.15	1.53	3.2	1.43	1.27
Kwon	0.13	0.15	1.68	3.4	1.65	1.52
Fixed	0.12	0.14	1.38	2.95	1.18	1.13

TABLE-1 shows the RMS values of the $\delta\eta$, $\delta\phi$ distributions and the mean and rms values of the $R_{\rm edep}$ and $R_{\rm ncell}$ distributions for various Validity Indexes. The mean of the $\delta\eta$ and $\delta\phi$ distribution is close to zero ($< 10^{-3}$) in all the three cases. The last entry corresponds to clustering using fixed parameters.

References

[1] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithm, Plenum, New York **(1981);**

[2] D.-W.Kim, K.H. Lee and D. Lee, Pattern Recogn. Lett.**24 (2003) 2561**, and references therein.

[3] ALICE Collaboration, K. Aamodt et al., JINST **3 (2008) S08002**.