Journal of Engineering and Applied Sciences 13 (21): 9189-9193, 2018
ISSN: 1816-949X

Using Random Forest Algorithm for Clustering

Laith Alzubaidi, Zinah Mohsin Arkah and Reem Ibrahim Hasan
University of Information Technology and Communications, Baghdad, Iraq

Abstract: Clustering is considered one of the most critical unsupervised learning problems. It endeavors to find an accurate structure in a collection of unlabeled data. In this study, we apply random forest clustering and density estimation for unsupervised decision. A dual assignment parameter will be used as a density estimator by combining random forest and Gaussian mixture model. Experiments were conducted using different datasets. Efficiency of using this algorithm is in capturing the underlying structure for a given set of data points. The random forest algorithm that is used in this research is robust and can discriminate between the complex features of data points among different clusters.

Key words: Random forest, clustering, Gaussian mixture, point, robust , complex

INTRODUCTION

Clustering is important process that has a big impact in unsupervised learning studies. It aims to find an accurate structure in a collection of unlabeled data. Then organizing objects into groups whose members have similar general properties. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

There are different algorithms have been used for clustering the data. Different datasets and applications have been addressed using different clustering algorithms. One of these algorithms is known as spectral clustering that is used to recognize the objects. Although, the spectral clustering is a widely-used technique but it is still facing some challenges such as the dimensionality of the data. For high dimensional data, it is hard to cluster similar objects by only depending on the distance among them. One of the famous algorithms to calculate the distance is called Euclidian distance method. This method depends on the precision of building the similarity matrix. (Stella and Shi, 2003; Chen *et al*., 2011; Shi and Malik, 2000) have mentioned that the eigenvectors can be driven from similarity matrix. A large extent of data dimensions is one of vectors that the accuracy of clustering matrix depends on. For high dimensional data, it is hard to use Gaussian Mixture Model (GMM) to estimate the density as well. To overcome these difficulties such as noise and redundant attributes that are related with high dimensional data. We used some techniques and algorithms that are utilized in this project and as explained:

- Building and assessing similar framework from the random forest (Albehadili and Islam, 2015)
- Combining Random Forest (RF) and GMM to get robust density estimation (Criminisi and Shotton, 2013)

MATERIALS AND METHODS

Random Forest (RF): It is a group of trees. Each tree has many nodes arranged hierarchically. The information transfers from the top to the bottom direction. Then doing checking in the opposite way from bottom to top. Plenty of tests at each node of the tree have been determined to recognize the dissimilar patterns at each split node. Our works are data clustering and density estimation which are related with unsupervised tasks. Discrimination between input patterns using unsupervised methods will be done. The assumption of input pattern is a feature vector = (1, 2, … ,) R . Each node has weak learner which partitions forthcoming patterns according to the following function (Aladjem, 2005):

$$h(v, \theta_j) = RdXT \rightarrow \{0, l\} \quad (1)$$

Where:

$\theta_j \epsilon$ = The parameters accompanied with each tree node
j = The jth node is the space for the split parameters
v = The incoming patterns

We endeavor ultimately to partition arriving data points at each node. It can be tackled using the following:

$$\theta_j = \arg\max \theta \in T_l\ (\mathcal{S}_j, \theta) \quad (2)$$

Corresponding Author: Laith Alzubaidi, University of Information Technology and Communications, Baghdad, Iraq

It is worth to maximize I in Eq. 2 to get high information gain by splitting samples points reaching the node as higher as possible. The split function can be trained using greedy search technique (Zhu *et al*., 2014). There are different method to maximize Eq. 2 either using Gini or entropy. The entropy can be describe as:

$$I = H(\mathcal{S}) - \frac{\Sigma i \in \{L,R\} |\mathcal{S}_i|}{|\mathcal{S}| H(\mathcal{S}_i)} \quad (3)$$

It is obvious that the higher the information gain is the better splitting node is. The optimization objective function induces partitioning arriving to either left or right channel of the node. Therefore, we consider that the weak learner is the cardinality of the decision forest. Constructing consolidate split functions for decision forest can formulate efficient clustering trees. Since unsupervised learning is demanded, following (Aladjem, 2005), below is the partitioning paradigm formulation used at each weak learner:

$$I(\mathcal{S}_j,\theta) = H(\mathcal{S}_j) - \frac{\Sigma i \in \{L,R\} |\mathcal{S}_{ji}|}{|\mathcal{S}_j| H(\mathcal{S}_{ji})} \quad (4)$$

Then, the entropy can be defined as:

$$H(\mathcal{S}) = \frac{1}{2\log\left((2\pi e)^{d|\Lambda(s)|}\right)} \quad (5)$$

Then the information gain can be obtained as following:

$$I(\mathcal{S}_j,\theta) = \log(|\Lambda(\mathcal{S}_j)|) - \frac{\Sigma i \in \{L,R\} |\mathcal{S}_{ji}|}{|\mathcal{S}_j| \log(|\Lambda(\mathcal{S}_j^i)|)} \quad (6)$$

Where:

Λ = Dxd covariance matrix

$|.|$ = A determinant for the matrix (Fig. 1)

We can sum up FR as following:

- The deeper the depth the better discrimination is
- Data are partitioned after each split
- Similar patterns transverse to the similar branches
- Dissimilar patterns travel into different branches

Decision Forest: In addition to the linear and nonlinear Model that are incorporated in our decision forest:

$$h(v,\theta) = [\tau 1 > \phi(v) ? > \tau 2] \quad (7)$$

$$h(v,\theta) = [\tau 1 > \phi T(v) ? \phi(v) > \tau 2] \quad (8)$$

We used GMM to be incorporated to the random forest because it is very robust method for unsupervised data clustering (Allili *et al*., 2010; Yu *et al*., 2012). In our implementation, we use each weak learner at a certain depth of random forest. The reason of using several weak learners is because diversity can lead to more generalization for capturing different data associations (Fig. 2).

Dual assignment to construct affinity matrix: In this part, imposing only the patterns that have the smallest similarity membership will be refrained and the patterns having highest similarity will be induced. We qualify the samples by adjusting a threshold. According to Eq. 8 an input pattern ($\varkappa^i$, $\varkappa^j$) can have a maxim similarity in defining affinity matrix if they are survived along a path together. However, they still also have some degree of similarities according to a used model even they are entirely from different clusters. Furthermore, imagine that we have a pair of patterns $\varkappa_i$ and $\varkappa_j$ and they belong to the same cluster then they will be assigned minimum degree of similarity in the constructed affinity matrix. Thus,

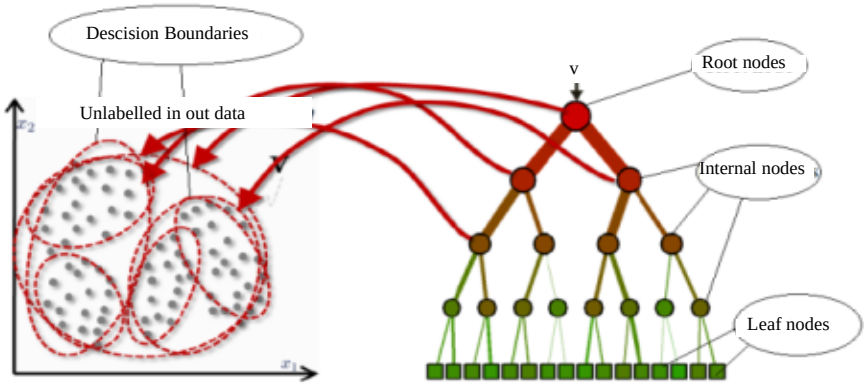

Fig. 1: Decision tree: deeper the depth the better discrimination

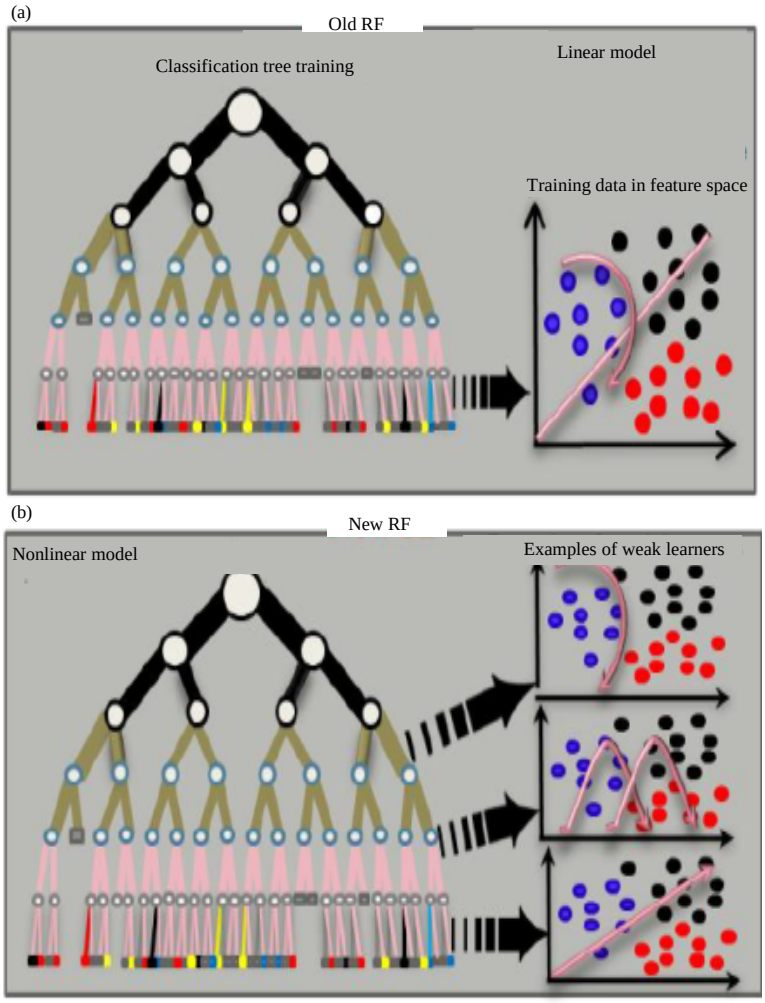

Fig. 2: The difference between old RF and new RF by using different functions

this assumption is not true, especially, for complex irregular patterns which cannot perfectly disassemble the similar examples related to the same clusters. This is the nature of the weak learners because they are only thresholds capturing inequalities trying to separate correlating samples affined to one cluster from other samples united to different clusters. Intuitively, since split function always not perfectly dissociate clusters, otherwise we would end with random forest with only just few nodes, then random forest will incorrectly cluster associated data points.

Updating equations for GMM and RF (GMM-RF): In this part we delve the relation between random forest and GMM and how we incorporate the two models. The intrinsic formula for GMM is give below:

$$P(x|\theta) = \Sigma mj=1 aj P(x|z_j, \theta_j) \quad (9)$$

Where:

x = An observation

z_j = The latent variables

θ_j = The associated parameters with GMM

Gaussian normal distribution is used in this implementation. The final distribution is given as: The latent/hidden variables can be obtained using the following Eq. 10:

$$Z_{ji} = \frac{(Z_{ji}|xi, \theta_{old}) = \Sigma j(xi|\theta_{jold})}{\Sigma mk = 1\varepsilon P(xi|\theta_{kold})} \quad (10)$$

To mitigate limitations of GMM because all hidden variables ji are not known, we alleviate the ambiguity inherited by absence all the latent variables, ensembles of

Fig. 3: CMU-PIE: samples of five persons with different poses and lightening

Fig. 4: MNIST datasets

these variables are afforded by decision forest. Revival GMM by decision forest deterministically tackles fogginess of conventional GMM. Assuming that the latent variables deriving from random forest embedded within RFji, therefore, the (Eq. 11) can be updated as below:

$$Z_{ji} \mid (Z_{ji} \neq RF_{ji}) = (Z_{ji} \mid Z_{ji} \neq RF_{ji} \mid x_i, \theta_{old}) = e_j(x_i \mid \theta_{jold}) S\, mk = 1ekP(x_i \mid \theta_{kold}) \quad (11)$$

Then, the updating equations for parameters of GMM can be update ed as following:

$$\varepsilon_{jnew} = \frac{1}{N} \Sigma N_i = 1(Z_{ji} \cup RF_{ji}) \quad (12)$$

$$\Sigma_{jnew} = \Sigma N_i = 1 \frac{(Z_{ji} \cup RF_{ji})(x_i - \mu_{jnew})(x_i - \mu_{jnew})}{\Sigma} N_i = 1(Z_{ji} \cup RF_{ji}) \quad (13)$$

$$\mu_{jnew} = \Sigma N_i = 1 \frac{(Z_{ji} \cup RF_{ji}) x_i}{(Z_{ji} \cup RF_{ji})} \quad (14)$$

RESULTS AND DISCUSSION

This study presents the data sets that are used then shows the results.

Datasets CMU-PIE: A gray (32×32 pixels) scale face images. In addition, the dataset has 68 persons with different illuminations and poses.

The MNIST: A hand written digits 0-9. The dataset consists of 10000 samples. All the samples have the same 28×28 pixels size (Fig. 3 and 4; Table 1).

Table 1: Statistics of the datasets used in this experiments

Dataset	No. samples	Dimensionality	No. of clusters
MNIST	10000	784	10
PIE	2856	1024	68

Table 2: Clustering accuracy on the two dataset

Dataset	CMU-PIE	MNIST
Random Forest [ours]	78.1	79.2

Table 3: Clustering performance on PIE dataset: accuracy metric

CMU-PIE dataset-accuracy (%)

K	RF
10	94.7
20	84.4
30	88.8
40	81.1
50	81.0
60	78.5
68	78.1

Table 4. Clustering performance on MNIST dataset: accuracy metric

MNIST dataset-accuracy (%)

k	RF
3	91.7
6	81.3
10	79.2

Best achieved results are 78.1 and 79.2 on CMU-PIE and MNIST respectively using RF. Comparing the accuracy results for CMU-PIE and MNIST datasets with different number of clusters (K) (Table 2-4).

CONCLUSION

Different functions are incorporated into split functions to induce more robust RF. Powerless learners are incorporated into both direct and nonlinear capacities circulated on specific levels on every tree of the irregular woodland. Furthermore, RF is consolidated by GMM inserted between linear and linear functions. Thus, strong RF is fittingly ready to segregate between uninformative elements since it finds semantic hidden structure information

REFERENCES

Aladjem, M., 2005. Projection pursuit mixture density estimation. IEEE. Trans. Signal Proc., 53: 4376-4383.

Albehadili, H. and N. Islam, 2015. Unsupervised decision forest for data clustering and density estimation. CoRR, 1: 1-7.

Allili, M.S., D. Ziou, N. Bouguila and S. Boutemedjet, 2010. Image and video segmentation by combining unsupervised generalized gaussian mixture modeling and feature selection. IEEE. Trans. Circuits Syst. Video Technol., 20: 1373-1377.

Chen, W.Y. Y. Song, H. Bai, C.J. Lin and E.Y. Chang, 2011. Parallel spectral clustering in distributed systems. IEEE. Trans. Pattern Anal. Mach. Intell., 33: 568-586.

Criminisi, A. and J. Shotton, 2013. Advances in Computer Vision and Pattern Recognition. Springer, London, England, UK., ISBN:978-1-4471-4928-6, Pages: 366.

Shi, J. and J. Malik, 2000. Normalized cuts and image segmentation. IEEE. Trans. Pattern Anal. Mach. Intell., 22: 888-905.

Stella, X.Y. and J. Shi, 2003. Multiclass spectral clustering. Proceedings of the 9th IEEE International Conference on Computer Vision, October 13-16, 2003, IEEE, Nice, France, pp: 313-319.

Yu, G., G. Sapiro and S. Mallat, 2012. Solving inverse problems with piecewise linear estimators: From Gaussian mixture models to structured sparsity. IEEE. Trans. Image Proc., 21: 2481-2499.

Zhu, X., C.C. Loy and S. Gong, 2014. Constructing robust affinity graphs for spectral clustering. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR14), June 23-28, 2014, IEEE, Columbus, Ohio, USA., ISBN:978-1-4799-5118-5, pp: 1450-1457.