COMPARISON OF PURITY AND ENTROPY OF K-MEANS CLUSTERING AND FUZZY C MEANS CLUSTERING

Satya Chaitanya Sripada

Department of Computer Science
JNTU, Hyderabad, 500085
chaitanya.ss@gmail.com

Dr. M.Sreenivasa Rao

Professor and Dean MSIT Department
JNTU, Hyderabad, 500085
srmeda@gmail.com

Abstract

Clustering is one the main area in data mining literature. There are various algorithms for clustering. The evaluation of the performance is done by validation measures. The external validation measures are used to measure the extent to which cluster labels affirm with the externally given class labels. The aim of this paper is to compare the for K-means and Fuzzy C means clustering using the Purity and Entropy. The data used for evaluating the external measures is medical data.

***Keywords*:** Purity; Entropy; K-means; Fuzzy C means; External validation measures; Contingency Matrix

1. Introduction

Clustering is the one of the vital areas in data mining. The evaluation of the performance of the clustering algorithm, we have to use the validation measures. There are two types of validation measures; they are internal validation measures and external validation measures. The internal validation measures use the information that is intrinsic to the data and measure the quality of the obtained clusters**.** One clustering output is given as input to external validation measures (E.g. the K-means clustering results) and examined in contrast with an existing known set of ground truth or class labels (‘true labels’) to determine the degree of accordance of ground truth and K-means clustering results. Ground truth or true label would be exclusive and imperforate that will exactly one class label is provided for every data item, and that the label is unequivocally defined.

2. K-Means and Fuzzy C means

The two popular partitional clustering algorithms are K-means and Fuzzy C means clustering. The popular and simplest probabilistic and unsupervised clustering algorithm is K-means algorithm. In K-means algorithm we initially decide the number of clusters let us say K number of clusters and hypothesize the centroid or clusters center point. The first initial centroid any randomly selected object is taken. The first K objects in progression could also be selected as the initial centroids. Given k the algorithm of K-means is carried out by distribution of the objects. The objects are distributed into k non-empty groups or subsets. The seed points are calculated as the centroids of the clusters of the current distribution of objects. The center point is the centroid of the distribution of objects and the mean point of that cluster. Now we allocate individual object to the cluster with the close by seed point .We return the seed points as centriods of the clusters and repeat until there is no more new assignment.

In the year 1973 Dunn developed the Fuzzy C Means algorithm and later in 1981 it was enhanced by Bezdek. Fuzzy C Means algorithm is extensively used in pattern recognition. Fuzzy C Means algorithm uses the iteratively process which rejuvenates cluster centers for individual data point. Fuzzy C Means algorithm repetitively iterates the cluster centers to the exact location with in data set elements. The performance of Fuzzy C Mean algorithm is based on the initial centroids selected. The mean of all data points in the Fuzzy C Means algorithm is calculated as the centroid of a cluster and is weighted by their degree corresponding to the cluster.

The degree of occurring in a specified cluster is related to the reciprocal of the distance to compute the weights of the cluster.

3. Contingency Matrix

The basis for the calculation and computation of these validation measures is the contingency matrix. This is the key area for these measures and is done by building the matrix and finding the values.

Table 1

	S_1	S_2	S_3	…	$S_{K'}$	Σ
D_1	n_{11}	n_{12}	n_{13}	…	$n_{1K'}$	$n_{1\cdot}$
D_2	n_{21}	n_{22}	n_{23}	…	$n_{2K'}$	$n_{2\cdot}$
D_3	n_{31}	n_{32}	n_{33}	…	$n_{3K'}$	$n_{3\cdot}$
.	.	.	.	.	.	.
D_K	n_{K1}	n_{K2}	n_{K3}		$n_{KK'}$	$n_{K\cdot}$
Σ	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$		$n_{\cdot K'}$	n

We have n objects in our data set Z and suppose that the partition D = {D_1, ・ ・ ・ , D_K} of Z, where $D_i \cap Dj = \phi$ and $U^K_{i=1}$ D_i = Z for 1 ≤i # j ≤K, and the count of clusters is K, it means there are K numbered clusters. Secondly for our data set Z we suppose that there is another partition S= {S_1, ・ ・ ・ ,$S_{K'}$} of Z, where $S_i \cap S_j = \phi$ and $U^{K'}_{i=1}$ S_i = D for 1 ≤ i # j ≤ K′ which constitutes the true class labels where count of clusters is K′, it means that K′ numbered of class clusters exist. The n_{ij} plays a very prominent role in our contingency matrix. This represents the count of objects present in D_i cluster from S_j class cluster, the data points that occur and the overlap between the D_i cluster and Sj class cluster are counted and this count is written as the n_{ij} in our contingency matrix. For suppose K =3 and K' = 3 which means that there are 3 clusters in K-means and we have 3 class cluster we represent a three by three contingency matrix. The values of p_{ij}, p_i and p_j are computed by pij = n_{ij} / n, pi = $n_{i\cdot}$ / n and pj = $n_{\cdot j}$ / n.

4. External Validation Measures

To measure the quality of any thing we require an index. Quality is very essential characteristic and distinguishing property of something. External validation is done class labels. Definition of the external validation indices: They are used to measure the extent to which cluster labels affirm with the externally given class labels. The class labels also known as ground truth are taken as base value. The external validation measures are extremely useful in deducing the ambit to which the clustering structure is ascertained by a clustering algorithm that matches some external structure. This is compared to the individual designated class labels. External validation measures criteria evaluate the final clustering output result with respect to a pre designated structure. For many applications these external validation measures are very much applicable. An illustration about the external validation measures exhibits that document clustering algorithm can pave to the results of clustering that match the distribution accomplishment by specialists.

There are many external validation measures like Purity, Rand Statistics, Entropy, Jaccard Coefficient, Mutual information, Fowlkes and Mallows Index, Minkowski score and Goodman – Kruskals Coefficient etc. We focus 3 external validation measures Purity, Entropy and F-measure. Here we discuss about the clustering methods, Fuzzy C-Means and K-Means. The distance between various data points of the clusters generated by the algorithms is determined and analyzed. The clusters formed are evaluated and interpreted according to the distance between data points and cluster centers of each cluster. The determination of cluster quality is done by the purity and entropy measures.

5. Purity and Entropy

The K-means algorithm is implemented and the respective clusters are obtained. These clusters are compared with the true label data set and the values of Purity and Entropy calculated clusters generated by K-means clusters and Fuzzy C means. The implementation is done in PERL. The Fuzzy C means is implemented and the respective clusters are obtained. The generated clusters are compared with the true label data set and the values of Purity and Entropy measures are calculated for the Fuzzy C means clusters. Purity is a one of very primary validation measure to determine the cluster quality. The entropy and purity are widely used measures. Entropy uses external information class labels in this case. The purity of the clusters is measured referencing to the class labels or ground truth is called as entropy. The lower entropy means better clustering. The Entropy amplifies when ground truth of objects in the cluster further diversifies. The greater entropy means that the clustering is not good. The quantity of disorder is found by using entropy. So we expect that every cluster should have low entropy to maintain the quality of our clustering. The purity is measured as $\Sigma_i\, p_i(\max_j\, (p_{ij} / p_i))$ where the values of pij , pi and pj are computed with respect to class labels. The entropy is measured as $\Sigma_i\, p_i(\Sigma_j\, (p_{ij} / p_i) \log (p_{ij} / p_i))$ where the values of pij , pi and pj are computed with respect to class labels.

Case 1 - Hepatitis

External validation measures	K-Means	Fuzzy C Means
Entropy	0.293299	0.289789
Purity	0.383776	0.391036

Case 2 - Diabetes

External validation measures	K-Means	Fuzzy C Means
Entropy	0.27078	0.251541
Purity	0.230571	0.261304

Case3 – Liver problem

External validation measures	K-Means	Fuzzy C Means
Entropy	0.302966	0.2993489
Purity	0.7665034	0.7756013

Case 4 – Heart disease

External validation measures	K-Means	Fuzzy C Means
Entropy	0.189828	0.1840025
Purity	0.5671967	0.5961377

Case 5 – Brain Tumor

External validation measures	K-Means	Fuzzy C Means
Entropy	0.298534	0.298373
Purity	0.6456010	0.646002

Case 6 –Eye lenses

External validation measures	K-Means	Fuzzy C Means
Entropy	0.2665893	0.2623207
Purity	0.3301948	0.3358693

6. Conclusions

The greater the value of purity indicates good clustering. The entropy is negative measure, the lower the entropy the better clustering it is. For the same data sets the Fuzzy C means and K-means algorithms were obtained and the clusters were generated. The above indicated values show that the Fuzzy C Means has high value of purity and low value of entropy. This indicates good clustering. The K-means has lower value of purity and high value of entropy compared to Fuzzy C Means. The Fuzzy C means clustering is more The Fuzzy C means clustering is more accommodating for medical data sets when compared to K means.

References

[1] Cao, X., Maloney, K.B. and Brusic, V. (2008). Data mining of cancer vaccine trials: a bird's-eye view. Immunome Research, 4:7. DOI:10.1186/1745-7580-4-7

[2] Cheng, T.H., Wei, C.P., Tseng, V.S. (2006) Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches. Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06).

[3] H. Xiong, J. Wu, and J. Chen. K-means clustering versus validation measures: A data distribution perspective. In KDD, 2006

[4] Health Grades, Inc. (2007). The Fourth Annual HealthGrades Patient Safety in American Hospitals Study.

[5] Kou, Y., Lu, C.-T., Sirwongwattana, S., and Huang, Y.-P. (2004). 7. Survey of fraud detection techniques. In Networking, Sensing and Control, 2004 IEEE International Conference on Networking, Sensing and Control

[6] Nightingale, F (1858). Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army.

[7] Shillabeer, A (29 July 2009). Lecture on Data Mining in the Health Care Industry. Carnegie Mellon University Australia.

[8] Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In Workshop on Text Mining, KDD, 2000.

[9] Strehl, J. Ghosh, and R.J. Mooney. Impact of similarity measures on web-page clustering. In Workshop on Artificial Intelligence for Web Search, AAAI, pages 58–64, 2000.

[10] S. van Dongen. Performance criteria for graph clustering and markov cluster experiments. TRINS= R0012, Centrum voor Wiskunde en Informatica. 2000.

[11] Wu, H. Xiong, J. Chen, and W. Zhou. A generalization of proximity functions for k-means. In ICDM, 2007.