Effective Purity Method for Measuring the Clustering Accuracy and its Illustration

Srinivasa Suresh Sikhakolli
Kirloskar Institute of Management, Pune, India

Asha Kiran Sikhakolli
Dr.D.Y. Patil B-School, Pune, India

ABSTRACT
Clustering is one of the commonly used model in business and scientific applications. Often, data science specialists and researchers apply clustering techniques for classification and optimization. Measuring the clustering accuracy is one of the key parameter. There are several extrinsic measures exists for measuring clustering quality. One of them is Purity. It indicates the level of homogeneity of the clusters. Purity computes the sum of frequencies of the dominant class in each cluster and then divides the sum by total number of records. In the existing purity method, total number of clusters is not taken into consideration. According to the researcher, number of clusters have significant effect on overall cluster quality. In this paper, the researcher proposed an algorithm with few changes to the existing purity method. The proposed algorithm is applied on machine learning data sets taken from UCI machine learning repository. Further, significant improvement in purity computation is observed when applied using FCM and K-means clustering. This paper explains proposed algorithm artificial illustration, results & analysis and comparative analysis between proposed purity method an existing purity method.

Keywords
Clustering, clustering accuracy, clustering extrinsic measure, clustering purity.

1. INTRODUCTION
Clustering is one of the well-known techniques for data partition. It is widely used statistical technique [1][2][3]. It has many application areas like fraud analysis if images, computational biology, medical sciences, market segmentation etc. In the literature, many clustering algorithms are proposed with varying degree of classifications [4]. Among all, K-means and Fuzzy-C means are the frequently used algorithms [4] [5]. On the other hand, Machine learning algorithms are also gaining its significance with clustering techniques. Machine learning applications are gaining significance in all areas of the development in the current days [6].

For quick understanding, we brief about clustering basics: Clustering is a process which partitions heterogeneous data into smaller homogenous groups based on attributes of data [7]. Clustering performs classification of the data. According to machine learning terminology, Clustering is one of the popular probabilistic and unsupervised learning techniques [7]. Unsupervised learning techniques are suitable for the data where class label (or output /predicted variable) is absent. These kinds of techniques (i.e unsupervised) are useful to classify the data [7].

Validating the cluster quality is one of the important criteria for cluster-based analysis. Cluster validation is defined as a technique to evaluate how best the clustering technique partitions the data into its natural partitions without class (or label) information. There are several measures to assess quality of clusters. Few methods measure how well the data fit into clusters, while some techniques help in checking if the clusters classify the data as per the ground truth values. Cluster validation techniques are broadly classified into three categories [8]:

- Intrinsic measures
- Extrinsic measures
- Relative measures

Intrinsic measures are based on the information internal to the data [7][8][9]. These measures do not require priori information about the data set. These measures have been further classified into two types collections. The first type measures the fitness between the data and the expected structure. The other group focuses on stability of the result [10]. General intrinsic measures are:

- Bic Index,
- Silhouette Index
- Davies-Boulding Index
- Dunn Index
- Calinski-Harabasz index
- Niva Index etc.

Extrinsic measures are based on previous knowledge about the data [10] and applied to measure if the existing cluster labels match with the externally specified class labels. These measures help us in solving problems by evaluating the results of a clustering algorithm with external data which is not contained in the dataset. Extrinsic measures are opted only if the ground truth (class labels) of the data exists. Ground truth is the ideal clustering that is often built by experts [14]. Hence, extrinsic measures are known as supervised measures and intrinsic measures are unsupervised measures [14]. Well-known extrinsic measures are: Purity, Entropy, Normalized mutual information, Rand index and F-measure [8]. Relative measures evaluate results by comparing the results with other clustering technique [8].

Each measure has clear scope and application. In this paper we consider purity, which is an extrinsic measure. Purity quantifies the extent to which the cluster C_i points only to one ground truth partition. Purity lies between 0 and 1 (1 indicates highest purity , whereas 0 is the lowest purity). In this paper we propose an effective purity method and compare the results with standard purity method. Here is the standard purity formula [11]:

Purity is computed as:

$$\boldsymbol{Purity = 1/N \sum_{i=1}^{k} max_j |c_i \cap t_j|} \text{--- (1)}$$

Where N= number of records (data points), k = number of clusters, c_i is one of the clusters in K clusters, t_j is one of the ground truth class. The formula (1) can be expressed using

contingency table and is filled by checking how many objects of each cluster Ci match with the ground truth classes Ti. See the example contingency table 1 given below:

Table 1. Contingency table

	T1	**T2**	**T3**
C1	5	50	5
C2	0	1	62
C3	0	17	0

To compute purity, fetch maximum value from each row and add them up and divide the result by the total number of observations.

From the table 1, Purity = (50 + 62 + 17) / 140 = 0.92142. Purity value is close to 1. This shows that clustering has classified the data 92% correctly. Here, we propose a different approach for computing purity of clustering.

The computational formula for the same is:

$$P = \sum_{q=1}^{k} \frac{1}{k} * \frac{\max_{1 \le i \le c} n_q^i}{n_q} \text{----(2)}$$

Detailed description of formula (2) is given in section III. In this paper, section II explains Literature Review of various clustering measures, section III describes Proposed Efficient Purity Algorithm, Section IV illustrates an artificial example, Section V describes results and analysis followed by conclusion.

2. LITERATURE REVIEW

This section will cover various formulas which are already existed in measuring quality of the clusters. They are Purity, Rand Index, Fowlkes and Mallows Index, Entropy, Jaccard Index, Mutual Information and F-measure [15].

2.1 Purity

Purity is a measure of the extent to which a cluster contains a single class. For each cluster, count the number of data points from the most common class in the said cluster and then sum over all clusters and divide by the total number of data points. In this paper, this purity method is referred as standard purity method. Considering M clusters and N data points, the formula of the purity is:

$$\frac{1}{N} \sum_{m \in M} \max_{d \in D} |m \cap d|$$

2.2 Rand Index

The Rand index computes how similar the clusters are to the benchmark classifications. Also, the Rand index as a measure of the percentage of correct decisions made by the algorithm. The computing formula is:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

Where TP represents True Positive, TN represents True Negative, FP represents False Positive and FN represents False Negative. The same notation has been used in subsequent formulas given below.

2.3 Fowlkes-Mallows Index

The Fowlkes–Mallows index computes the similarity between the clusters returned by the clustering algorithm and the benchmark classifications. The higher the value of the Fowlkes–Mallows index the more similar the clusters and the benchmark classifications are. The computing formula is:

$$FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}$$

2.4 Entropy

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 computing formula is:

$$H(\Omega) = -\sum_k P(\omega_k) \log P(\omega_k) = -\sum_k \frac{|\omega_k|}{N} \log \frac{|\omega_k|}{N}$$

2.5 Jaccard Index

The Jaccard index is used to quantify the similarity between two datasets. The Jaccard index takes on a value between 0 and 1. An index of 1 means that the two dataset are identical, and an index of 0 indicates that the datasets have no common elements. The computing formula is:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN}$$

2.6 Mutual Information

Mutual Information is an information theoretic measure of how much information is shared between a clustering and a ground-truth classification, that can detect a non-linear similarity between two clusters. Normalized mutual information is a family of corrected-for-chance variants of this that has a reduced bias for varying cluster numbers.

2.7 F-Measure

The F-measure can be used to balance the contribution of false negatives by weighting recall through a parameter. Let precision and recall (both external evaluation measures in themselves) be defined as follows:

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN}$$

These extrinsic measures are used according to the suitable context or problem in hand. This work focuses on Purity method. The standard purity method does not consider the total number of clusters while measuring the purity. There is a correlation between number of clusters & clustering accuracy and it has significant effect on cluster accuracy measure. Based on this fact, the researcher proposes an effective Purity method based on the standard Purity method.

3. PURITY ALGORITHM

The objective of this algorithm is to find the Purity of the clusters. It finds how best the actual class of the data affirms with predicted class of the data. Here are the algorithm steps:

Predominant_Class: It is the class label of the records which are predominant within the cluster.

Predominant_Count: It is the frequency of records belonging to predominant class within the Cluster.

Algorithm: Purity Algorithm
Input: Test_data;
Output: Purity;

Start

Step1: Find number of unique actual classes **c**, from test_data.

Step2: Apply standard clustering technique on *n* records to generate *k* clusters

Step3: Identify the actual class of each record in each cluster

Step4: ***Foreach*** of the $Cluster_q$, $(1 \leq q \leq k)$

a) Find total number of records n_q in $Cluster_q$
b) Group records of $Cluster_q$ by each actual class *i*, $(1 \leq i \leq c)$
c) Compute the Record_Count n_q^i for each actual class *i* in $Cluster_q$,
d) Find the Predominant_Count in $Cluster_q$, $\max_{1 \leq i \leq c} n_q^i$
e) Find the Predominant_Class in $Cluster_q$
f) *Map* < Predominant_ Class, Predominant_Count>

End for

Step5: Compute Purity

$$\sum_{q=1}^{k} \frac{1}{k} * \frac{\max_{1 \leq i \leq c} n_q^i}{n_q}$$

End

This proposed measure aggregates (sum) the purity of each cluster and then divides the sum value by number of clusters. This implies the measure of homogeneity. In brief, the proposed method computes purity by considering the weighted average of maximal precision values of each cluster. The following section iv explains the algorithm with an artificial example.

4. ILLUSTRATION WITH ARTIFICIAL EXAMPLE

This illustration is based on the proposed algorithm stated in the section 3. The following attributes are considered:

-Sample size of 20 records having class labels.

-No. of Records: 20,
-No. of Actual Classes: 4 (1,2,3,4),
-No. of Clusters: 5 (A, B, C, D, E)

Table 2: Records with actual class

Record No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Actual Class	3	2	1	4	3	2	1	3	3	2	1	4	3	4	1	4	1	4	1	3

Step1: Find number of unique actual classes **c**, in table 2 , C=4 (since there are 4 unique classes 1,2,3,4)
Step 2: Apply standard clustering technique on *n=20*records to generate *k=5* clusters

Table 3 Actual Vs Obtained class

Record No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Actual Class	3	2	1	4	3	2	1	3	3	2	1	4	3	4	1	4	1	4	1	3
Obtained Class	B	A	C	E	B	A	C	B	C	A	C	D	A	D	B	D	D	D	B	A

The table 3 shows obtained class after clustering**. T**he obtained class is a numeric number. For better understanding, it is considered A, B, C, D, E classes.

Step3: Identify the actual class of the records in each cluster.

The following figure shows the clusters formed after step 3.

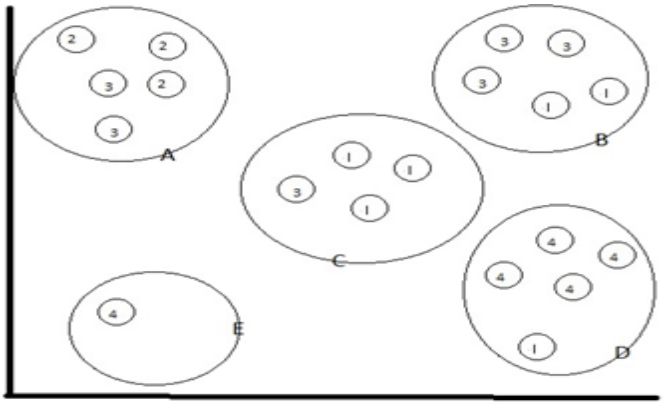

Fig. 1.Clusters with actual class labels

In the above figure 1, there are five clusters. Each cluster contains records of similar nature.

Step 4: For each cluster, the algorithm finds predominant class, predominant count by grouping records based on actual class. The following figure 2 shows application of step 4.

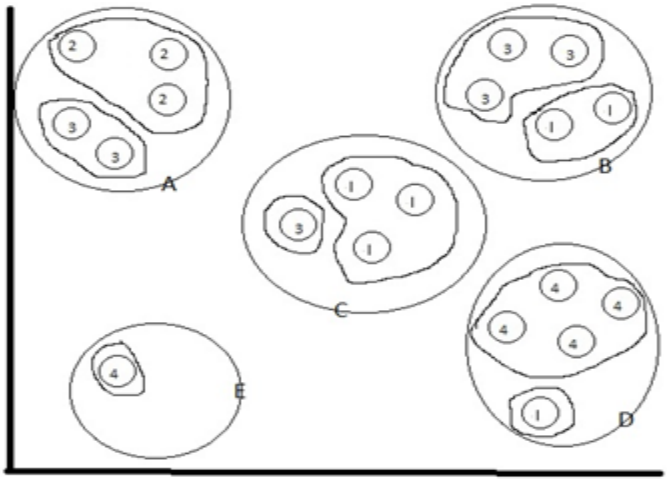

Fig 2 Sub-grouping of each cluster based on actual class labels

With reference to figure 2, purity values are computed using formula (2) and results of the same are shown in the Table 4

Table 4 Purity values

Cluster No.	**A**	**B**	**C**	**D**	**E**
No. of Records	5	5	4	5	1
Predominant Actual Class	2	3	1	4	4
Predominant Count	3	3	3	4	1
Purity	1/5 *3/5 = 0.12	1/5 * 3/5 = 0.12	1/5 *3/4=0.15	1/5 *4/5=0.16	1/5 * 1/1=0.2
Total Purity using formula (2)	0.12+0.12+0.15+0.16+0.2 = 0.75				
Total purity using formula (1)	(3+3+3+4+1)/20=0.7				

The table 4 shows the computed purity values using standard purity measure (formula-1) and proposed purity measure (formula-2). It is observed that proposed method obtained a value 0.75, which is greater than value computed through formula-1, which is 0.7. The proposed purity measure, computes cluster homogeneity better than the existing standard purity measure. The following section V shows the experimental results and analysis.

5. RESULTS AND ANALYSIS

This experiment conducted using K-means and Fuzzy C-means clustering techniques on the UCI machine learning data sets; Cleveland dataset and Switzerland dataset for validating the cluster accuracy [12]. After applying these techniques, purity is measured using standard method and proposed purity method. These datasets contain measurements of heart patients and contains ground truth values. Ground truth values are the ideal clustering values which are built by human experts [7]. The following tables 5 and 6 shows the purity comparison.

Table 5: Comparative Purity Measures on Cleveland Dataset

	K-Means		**Fuzzy C Means**	
Clusters	**Standard purity**	**Proposed Purity Method**	**Standard purity**	**Proposed Purity Method**
2	0.54098	0.53017	1	1
3	0.57377	0.61901	1	1
4	0.7377	0.73840	1	1
5	0.59106	0.61261	1	1
6	0.62295	0.59038	0.833330.9	0.9
7	0.57377	0.57393	1	1
8	0.54098	0.57924	0.75	0.83333
9	0.57377	0.5821	0.5555	0.73333
10	0.63934	0.61893	0.7	0.8

Table 6: Comparative Purity Measures on Switzerland Dataset

	K-Means		**Fuzzy C Means**	
Clusters	**Standard purity**	**Proposed Purity Method**	**Standard purity**	**Proposed Purity Method**
2	0.4	0.44565	1	1
3	0.4	0.58741	1	1
4	0.44	0.46548	1	1
5	0.56	0.55778	1	1
6	0.64	0.68333	0.66667	0.75
7	0.6	0.64524	0.71429	0.8
8	0.56	0.65774	0.875	0.93333
9	0.6	0.68704	0.88889	0.9
10	0.6	0.61667	0.9	0.96429

Tables 5 & 6 shows the results of standard purity and proposed purity. In this work, K-means and Fuzzy C Means (FCM) algorithms are applied to measure the overall clustering performance using Purity measure. The table 5 shows comparative results between Standard Purity and Proposed Purity Method. The clustering Purity is computed iteratively from clusters 2 to 10 for each of the datasets using K-means and FCM. Proposed purity method applied on Cleveland and Switzerland datasets. For both cases, proposed purity method showed better accuracy in majority of the iterations.

6. ACKNOWLEDGEMENT

Thanks to Moshe Lichman-UCI center (Machine Learning Repository), for timely response in clarifying doubts regarding datasets.

7. CONCLUSION

The objective of this experiment is to measure the clustering accuracy using improved Purity algorithm. Further it is tested using K-means and Fuzzy-C means clustering methods to check the accuracy of the clusters. The algorithm applied in iterative manner starting with number of clusters: 2 to 10. The proposed Purity method generated better results compared to standard purity method in majority of the iterations. The present paper included only K- means and Fuzzy C-means. However, it can be applied on wide variety of clustering methods like hierarchical clustering etc., The researcher primary research area is privacy data mining. Privacy data mining has wide research scope. For testing privacy accuracy, the proposed method will be applied with different parameter settings. The results stated in Table 5 and 6 are executed on MATLAB platform. It is observed that the execution time (for iterations) increasing enormously. So, for larger data sets, computational time increases. Actual time measurement is yet to done.

For testing privacy accuracy, the proposed method will be applied with different parameter settings. The results stated in Table 5 and 6 are executed on MATLAB platform. It is observed that the execution time (for iterations) increasing enormously. So, for larger data sets, computational time increases. Actual time measurement is yet to completed.

8. REFERENCES

[1] J. Vaidya and C.Clifton, "Privacy preserving k-means clustering over vertically partitioned data", the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003

[2] Luong The Dung and Ho TuBao, "Enhancing Privacy in Distributed Data Clustering", Journal of Computer Science and Cybernetics, Vol. 26, No. 2, pp. 1-15, 2010)

[3] Suzuki Kaoru, "Data Mining and the Case for Sampling", SAS Institute Best Practices Paper, SAS Institute, vol. 18, pp. 361-380, 1999

[4] P.Arabie, L.J Hubert, and G.Soete , Clustering and Classifications, World Scientific, 1996.

[5] S.Guha, Rastogi and K.Shim .Rock: A Robust Clustering Algorithm for Categorical attributes. In proc 1999 Int Conference: Data Engineering (ICD'99), PP 512-521, Sydney,Australia,Mar,1999.

[6] C.M.Bishop, Pattern recognition and Machine Machine Learning New York: Springer, 2006.

[7] J MichlineKamber, Jian Pei, "Data Mining Concepts and Techniques", ISBN: 978-93-80931-91-3 p.no.444,P.no.487.ELSEVIER, 2012.

[8] Erendira Rendon etal, "Internal Verses External Cluster validation Indexes", Issue 1, volume 5,2011, International Journal of Computers and communications.

[9] SatyaChaitanyaSripada, Comparision of purity and Entropy of K-means clustering and Fuzzy C means clustering. International journal of Computer Science and Engineering(IJCSE), Vol.2, No.3,June-July,2011, ISSN:0976-5166.

[10] Pacual D et al,Cluster validation using Information Stability Measures, Pattern Recognition, letter 31,2010, pp454-461.

[11] LeganyC,Cluster Validity Measurement Technique,Proceedings of the 5 th WSEAS International Conference on Artificial, Knowledge Engineering, and Data bases: Spain, Feb 15-17,2006,pp.388-393.

[12] Robert Detrano, M.D., Ph.D. Machine Learning Repository, Heart decease data sets available at http://archive.ics.uci.edu/ml/citation_policy.html, Cleveland Clinic Foundation.

[13] Asha kiran, ManimalaPuri, Srinivasa Suresh, PSO Enabled Privacy preservation, Indian Journal of Science and Technology, Vol 10(11), DOI: 10.17485/ijst/2017/v10i11/89318, March 2017, ISSN:0974-5645(online)

[14] Enrique Amigo et al, "A compariosion of Extrinsic clustering evaluation metrics based on formal constraints", UNED, Madrid, Spain, 2009. "APP purity method", APP method, Average Purity Method,

[15] Shaobin Huang, Yuan Cheng, * Dapeng Lang, Ronghua Chi, and Guofeng Liu Michal Zochowski, EditonA Formal Algorithm for Verifying the Validity of Clustering Results Based on Model Checking, 2014 Mar DOI: 10.1371/journal.pone.0090109.