# Improved Cluster Validity Using Gaussian Measure

R. Kavitha
Assistant Professor, Dept. of Computer Science, Sri Krishna Arts & Science College, Coimbatore, India

S. Sandhya
M.Phil Scholar, Dept. of Computer Science, Sri Krishna Arts & Science College, Coimbatore, India

**Abstract – Clustering is the unsupervised learning process which partitions objects into different groups. The quality of the clustering can be determined by the cluster validity index. One of the important initial parameter for the fuzzy clustering algorithm FCM is the number of clusters to be generated, which highly reflects the quality of the resulting partition. The validity indices proposed in the literature are dependent upon the membership and the data itself for validity calculation. After reviewing several validity indices a new validity index is proposed named Gauss index. The system proposed a Gauss index uses Gaussian measure which copes with the uncertainty issue associated with the current real data sets. Along with the membership degree, another measure that helps in proper evaluation result is the gaussian measure. It provides the necessary component in identifying the well-separated and compact cluster. The results show valid results when applied for the microarray data sets yeast, colon cancer, splice and leukemia.**

**Index Terms – k-NN with regression, missing value, data mining, microarray.**

## 1. INTRODUCTION

Fuzzy clustering aims at partitioning a data set into 'c' homogenous clusters. It suffers from the problem of assigning the number of clusters (c) in advance. In order to obtain good cluster, it is important to set the parameters of the algorithm right. It highly depends on the initial parameters and needs estimation of the number of clusters. The problem of finding an optimal c is called cluster validity [1]. It is essential to validate each of the fuzzy partition generated, since different numbers of initial cluster produce different clustering partitions. Several cluster validity indices are proposed in the literature with categories such as i)using only the membership values and ii) involves both the membership value and the data set. The commonly used validity indices in recent research are Bezdek's Partition Coefficient (PC) and Classification Entropy (CE)[2], Partition Index(SC) [3] , Separation Index(S), Xie-Beni's index(XB) [4] and Dunn's Index(DI) [5]. Compactness (closeness of cluster elements) and separation (distance between two different clusters) are the two major criteria proposed for evaluation and selection of the optimal clusters. The real-world clustering applications are stuck with the uncertainty in the localization of the feature vectors. Uncertainty, fuzziness and vagueness are the major elements in fuzzy clustering, that again adds indecision in defining the membership function of the object. The existing fuzzy validity index involves only the membership value and the data set in determining the optimality of the cluster number. In the proposed index the Gaussian measure along with the membership value is used to overcome the uncertainty in the real world application.

## 2. VALIDATION INDICES FOR FUZZY CLUSTERING

After finding a partition of data by a fuzzy clustering algorithm such as FCM, the objective is to determine whether the partition has presented the data structure correctly or not. The cluster validity problem is to determine the optimal number of clusters. Most of the fuzzy clustering methods assume an initial cluster number 'c' to describe the data structure completely. Cluster validity index method performs the validation of the generated fuzzy c-partition. $c_{min}$ and $c_{max}$ are the minimum and maximum number of partitions defined where each c range in $[c_{min}, c_{max}]$. The optimal cluster number is determined by minimum or maximum value of the validity index. Some of the validity indices are reviewed as follows.

a) Bezdeka (1974) proposed the validity index **partition coefficient(PC)** associated with FCM defined as

$$V_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2 \qquad (1.1)$$

where $\frac{1}{c} \leq V_{PC} \leq 1$. The PC index indicates the average contents of pairs of fuzzy subsets in fuzzy partition by combining into a single number. Most favorable cluster number c* can be obtained by solving $max_{2\leq c\leq n-1}V_{PC}$ to produce the best clustering performance for the data set X.

b) **Classification entropy(CE)** defined by Bezdek (1974;1981) as

$$V_{CE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}\, log_a\mu_{ij} \quad (1.2)$$

where a is the base of the logarithm. CE measures the fuzziness of the cluster partition similar to the Partition Coefficient. An optimal c* is obtained by minimizing $V_{CE}$to produce the best clustering performance for the data set X.

c) **Partition Index(SC)** indicates the relative amount of the sum of compactness and separation of the clusters. It takes the division of fuzzy cardinality of each partition to find the

sum of the individual cluster validity measures (Bensaid et al 1996).

$$SC(c) = \sum_{i=1}^{c} \frac{\sum_{j=1}^{N} (\mu_{ij})^m ||x_j - v_i||^2}{N_i \sum_{k=1}^{c} ||v_k - v_i||^2} \qquad (1.3)$$

SC is more suitable when equal number of clusters are produced by partitions. Better separation of SC can be obtained by taking minimum value.

d) **Separation Index(S)** uses a minimum-distance separation for partition validity (Bensaid et al 1996).

$$S(c) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ij})^2 ||x_j - v_i||^2}{N min_{i,k} ||v_k - v_i||^2} \qquad (1.4)$$

e) **Xie & Beni's Index** (1991) (XB) involves compactness and separation between clusters.

$$XB(c) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ij})^m ||x_j - v_i||^2}{N min_{i,j} ||x_j - v_i||^2} \qquad (1.5)$$

The numerator of the Equation (1.5) represents the compactness of the fuzzy partition and denominator denotes the strength between clusters. The optimal number of clusters should minimize the value of the index.

f) **Dunn's Index (DI)** is proposed to identify compact and well separated clusters (Dunn 1974). So the result of the clustering has to be recalculated as it is a hard partition algorithm.

$$DI(c) = min_{i \in c} \left\{ min_{j \in c, i \neq j} \left\{ \frac{min_{x \in C_i, y \in C_j} d(x,y)}{max_{k \in c} \{ max_{x,y \in C} d(x,y) \}} \right\} \right\} \qquad (1.6)$$

The main drawback of Dunn's index is computation since calculating becomes computationally very expensive as c and number of observations increase.

## 3. GAUSSIAN VALIDITY INDEX

Compactness and separation are the two measures that a good validation index should possess for a c-partition. Let A = $\{a_1, a_2, \ldots, a_n\}$ be a data set in $R^s$. Assume that $\mu = \{\mu_1, \ldots\ldots, \mu_c\}$ be the mean of the values in attribute associated with class c. Figure 4.1 shows the framework of the model. The preprocessed data partitions the data set into clusters. The results are validated using the proposed validity index Gaussian naïve bayes index.

In this work a reliable validation functional is proposed which provides a solution to the problem of validity associated with continuous values of each class according to the Gaussian distribution. This can help in obtaining an optimal cluster 'c' for the data set.

$$V_{Gauss} = \sum_{I=1}^{N} \frac{1}{\sqrt{2\pi a \sigma_c^2}} \frac{e^{\vartheta - \mu_c}}{n} \qquad (1.7)$$

$\mu$ Indicates the means of values in a, $a\sigma_c^2$ refers to the variance of values associated with 'c' cluster.

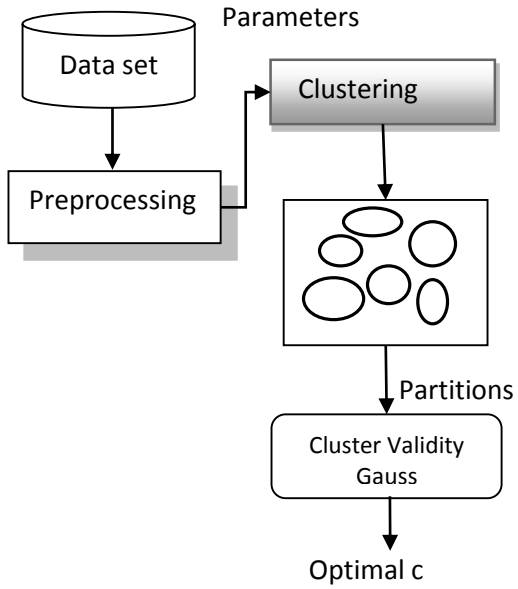


Figure 1.0 Framework of Gaussian index

The exponential function sets the compactness measure in the interval (0,1] and have the same degree(range) of measure. The total average of Gaussian detects the data structure with a compact partition and well-separated clusters. Thus, an optimal c* can be found by solving $min_{2 \leq c \leq n-1} V_{Gauss}$ to produce the best clustering performance for the dataset. The procedural steps for the validation of the Clustering using the proposed validity index $V_{Gauss}$, where $V_{Gauss}^{(\min)}$ denotes the minimum value of index is given as follows:

*Step 1:* Initialize the parameters related to the k-means and the validity index:

c=2, $c_{max}$ =10, $V_{GAUSS}^{(\min)}$=0, m=2, ε =0.001.

*Step 2:* With the initial assignment of weighting exponent 'm', the membership values are initialized such that $\sum_{i=1}^{c} \mu_{ij}$=1, where i=1, 2,…c, j=1,2…n.

*Step 3:* Update the fuzzy cluster centroid and fuzzy membership.

$$Cj = \frac{\sum_i [\mu_j(x_i)]^m x_i}{\sum_i [\mu_j(x_i)]^m} \qquad (1.8)$$

$$\mu_j(x_i) = \frac{[\frac{1}{d_{ji}}]^{1/m-1}}{\sum_{k=1}^{c}[\frac{1}{d_{ki}}]^{1/m-1}} \quad (1.9)$$

*Step 4:* If the improvement in objective function is less than a certain threshold ε, then go to step 5: otherwise go to step 3.

*Step 5:* Compute the non-membership value and Gaussian measure for the fuzzy partition obtained in step 4.

*Step 6:* Find $V_{Gauss}^{(\min)}$, and report the value of c that minimizes $V_{Gauss}$as the optimal number of clusters.

$$V_{Gauss} \leftarrow \min V_{Gauss} \quad (1.10)$$

The validation algorithm runs the FCM algorithm and computes the proposed validity index with respect to c=2,3,….,$c_{max}$.

## 4. EXPERIMENTAL STUDIES

Comparisons are made with various data sets to demonstrate the proposed validity index performance. The proposed index is compared with six fuzzy cluster validity indices such as Bezdek's Partition Coefficient(PC), Classification Entropy(CE), Partition index(SC), Separation Index(S), Xie-Beni's index(XB) and Dunn's Index(DI). In the experiments conducted the cluster validation is determined to obtain the optimal cluster number 'c'.

### 4.1. Validation performance

The cluster validity index is tested for four data sets. The microarray data set is collected from the public data set available at http://kzi.polsl.pl/~jbiesiada/Infosel/files/datasets.html, http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html. The validity indexes discussed for the study are implemented using MATLAB. The fuzzy cluster validity index performance varies with the fuzzy clustering algorithm. The FCM algorithm can easily able to discriminate the cluster validity with cluster number 'c' varying from 2 to $c_{max}$. The parameters of the FCM are set to a termination criterion ε=0.001, and weighting exponent m=2.0, and $||x_i-v_j||^2$ is the Euclidean norm. Random selection is made for the assignment of initial centroids. Four data sets are used to evaluate the validation performance of each index such as the yeast, colon cancer, splice and leukemia data sets. Table 1.1 - 1.4 gives the results of the evaluation of each index for the four data sets and the optimal value of c for each index is marked in bold face.

$V_{PC}$ and $V_{DI}$ take their maxima as optimal values, whereas the other indices take their minima as optimal values. Table 1.1 lists the results of validity indexes for yeast data set which contains 79 samples where, c=2,3,…,10. For each c≥2, index values are computed for each of the 8 validity indexes considered. The optimal c's of $V_{PC}$ and $V_{CE}$ are at c=2, whereas for $V_S$ and $V_{XB}$ are at c=10 and for the proposed index is at c=6.

Table 1.2 shows the validity indices values for colon cancer data set obtained from various validity indices with c=2,3,…,10. The optimal number of clusters c=2 is correctly identified by $V_{PC}$, $V_{CE}$ and $V_{Gauss}$ whereas $V_{XB}$ yielded the optimal partitions at c=4.The optimal values are identified at c=10 by $V_S$ and $V_{SC}$.

**Table 1.1 Values of Validity indices for yeast data set**

| c | PC | CE | SC | S | XB | DI | Gauss |
|---|---|---|---|---|---|---|---|
| **2** | **0.5000** | **0.6931** | 4.3617 | 2.6190 | 1.0146 | **0.1566** | 3.9483 |
| **3** | 0.3333 | 1.0986 | 4.2134 | 5.1674 | 0.6764 | 0.1542 | 5.9225 |
| **4** | 0.2500 | 1.3863 | 3.1654 | 5.7446 | 0.5073 | 0.1478 | 7.8966 |
| **5** | 0.2000 | 1.6094 | 3.7544 | 6.7196 | 0.4058 | 0.1542 | 9.8708 |
| **6** | 0.1667 | 1.7918 | 2.5185 | 4.6297 | 0.3382 | 0.1542 | **1.1845** |
| **7** | 0.1429 | 1.9459 | **1.8211** | 3.5533 | 0.2898 | 0.1494 | 1.3819 |
| **8** | 0.1250 | 2.0794 | 3.5534 | 3.2562 | 0.2536 | 0.1542 | 1.5793 |
| **9** | 0.1111 | 2.1972 | 4.2700 | 3.2861 | 0.2254 | 0.1542 | 1.7767 |
| **10** | 0.1000 | 2.3026 | 5.2058 | **2.1197** | **0.2029** | 0.1542 | 1.9742 |

Table 1.3 shows the performance of the validation methods for the splice data set of the various validity indices with c=2, 3,…,10. The optimal number of clusters c=6 is correctly identified by $V_{Gauss,}$ whereas $V_{PC}$, $V_{CE}$ and $V_S$ yielded the optimal partitions at c=2. The optimal values are identified at c=10 by $V_{XB}$ and $V_{SC}$. The results of the validity indices of leukemia data set are presented in table 4.4. It shows that $V_{Gauss,}$ $V_{PC}$ and $V_{CE}$ has yielded the optimal partitions at c=2, whereas $V_{XB}$ and $V_S$ gives optimal c at 10. Figures 1.1, 1.2, 1.3 and 1.4 shows the partitions on yeast, colon cancer, splice and leukemia data sets acquired by applying FCM with the number of clusters identified by the proposed cluster validity index, Gauss index respectively.

Table 1.2 Values of Validity indices for Colon cancer dataset

| c | PC | CE | SC | S | XB | DI | Gauss |
|---|---|---|---|---|---|---|---|
| **2** | **0.6233** | **0.5603** | 0.2184 | 0.0035 | 1.0279 | 0.3163 | **0.0187** |
| **3** | 0.4167 | 0.9610 | 0.1895 | 0.0048 | 0.6546 | 0.2792 | 0.0368 |
| **4** | 0.3144 | 1.2434 | 0.1805 | 0.0047 | **0.0523** | 0.3016 | 0.0523 |
| **5** | 0.2532 | 1.4647 | 0.1795 | 0.0044 | 0.3997 | 0.2784 | 0.0685 |
| **6** | 0.2117 | 1.6448 | 0.1767 | 0.0041 | 0.3374 | 0.3081 | 0.0703 |
| **7** | 0.1824 | 1.7972 | 0.1712 | 0.0040 | 0.2936 | 0.2785 | 0.0668 |
| **8** | 0.1612 | 1.9258 | 0.1654 | 0.0040 | 0.2597 | 0.3081 | 0.0715 |
| **9** | 0.1502 | 2.0250 | 0.1469 | 0.0036 | 0.2472 | 0.2939 | 0.0748 |
| **10** | 0.1377 | 2.1257 | **0.1446** | **0.0033** | 0.2314 | **0.3258** | 0.0817 |

Table 1.3 Values of Validity indices for splice dataset

| c | PC | CE | SC | S | XB | DI | Gauss |
|---|---|---|---|---|---|---|---|
| **2** | **0.5000** | **0.6931** | 4.5703 | **1.4327** | 0.7788 | **0.1107** | 3.6060 |
| **3** | 0.3333 | 2.0986 | 6.3082 | 2.9374 | 0.5192 | 0.0842 | 5.4090 |
| **4** | 0.2500 | 1.3863 | 8.2003 | 4.1635 | 0.3894 | 0.0842 | 7.2121 |
| **5** | 0.2000 | 1.6094 | **3.5018** | 1.5952 | 0.3115 | 0.0852 | 9.0151 |
| **6** | 0.1667 | 1.7918 | 5.9621 | 2.8283 | 0.2596 | 0.0595 | **1.0818** |
| **7** | 0.1429 | 1.9459 | 7.7484 | 3.2955 | 0.2225 | 0.0825 | 1.2621 |
| **8** | 0.1250 | 2.0794 | 8.6251 | 4.0114 | 0.1947 | 0.0484 | 1.4424 |
| **9** | 0.1111 | 2.1972 | 7.1447 | 3.8956 | 0.1731 | 0.0595 | 1.6227 |
| **10** | 0.1000 | 2.3026 | 4.4623 | 2.2648 | **0.1558** | 0.0593 | 1.8030 |

Table 1.4 Values of Validity indices for leukemia dataset

| c | PC | CE | SC | S | XB | DI | Gauss |
|---|---|---|---|---|---|---|---|
| **2** | **0.5000** | **0.6931** | 2.0552 | 5.4084 | 0.6378 | 0.5586 | **11.3831** |
| **3** | 0.3333 | 1.0986 | 6.3721 | 2.3904 | 0.4252 | **0.6128** | 17.0746 |
| **4** | 0.2500 | 1.3863 | 5.2264 | 2.0068 | 0.3189 | 0.5360 | 22.7662 |
| **5** | 0.2000 | 1.6094 | 5.8293 | 2.5136 | 0.2551 | 0.5882 | 28.4577 |
| **6** | 0.1667 | 1.7918 | 3.5899 | 1.4639 | 0.2126 | 0.5586 | 34.1493 |
| **7** | 0.1429 | 1.9459 | 2.0216 | 7.5091 | 0.1822 | 0.5998 | 39.8409 |
| **8** | 0.1250 | 2.0794 | 4.8351 | 1.9503 | 0.1595 | 0.5882 | 45.5323 |
| **9** | 0.1111 | 2.1972 | **1.3936** | 5.1390 | 0.1417 | 0.5126 | 51.2241 |
| **10** | 0.1000 | 2.3026 | 3.6922 | **1.4582** | **0.1276** | 0.5129 | 56.9155 |

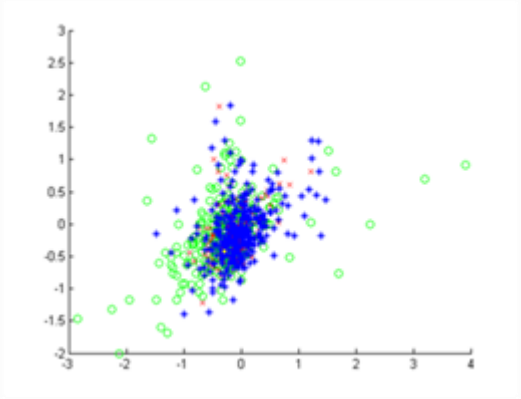

Figure 1.2 Clustered yeast data after application of FCM for c=6

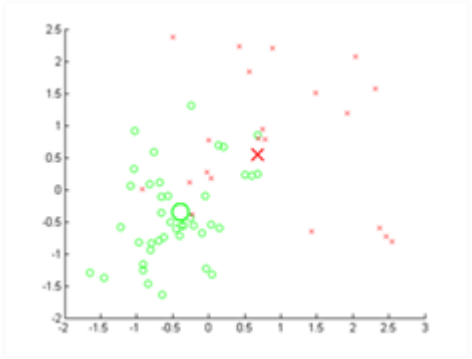

Figure 1.3 Clustered colon cancer data after application of FCM for c=2

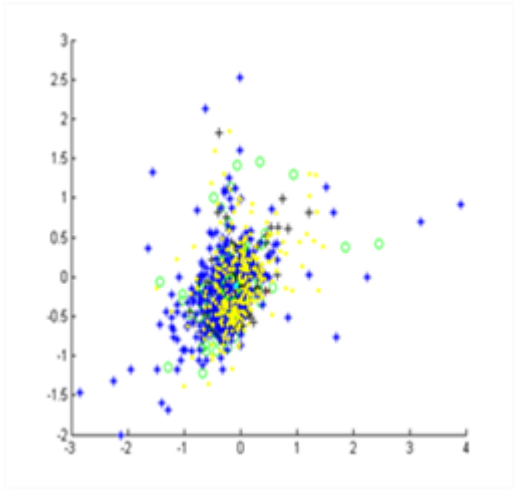

Figure 1.4 Clustered splice data after application of FCM for c=6

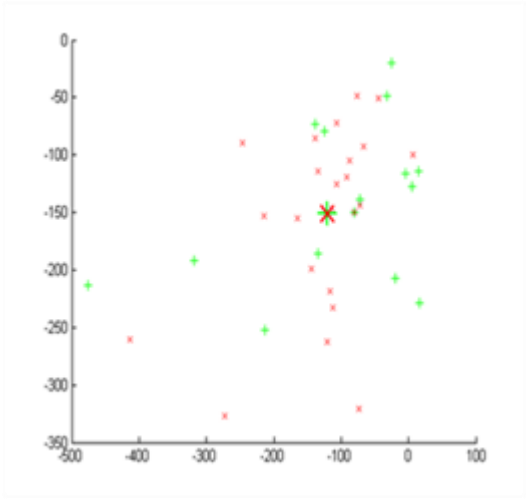

Figure 1.5 Clustered Leukemia data after application of FCM for c=2

### 4.2. Reliability

The FCM objective function $J_m$ (given in Equation (1.3)) depends on a weighting exponent 'm' which lies between 1 and ∞. The reliability of the validity index is determined by exploring the dependency of m. According to Pal & Bezdek (1995) best results are shown for FCM algorithm with the value of m varied between 1.5 and 2.5. The analysis is performed on the validity indices for four microarray data sets with changes in value for m. Tables 4.6 - 4.9 shows the validation results for different values of m between 1.5 and 2.5. The parameters of the FCM algorithm remain the same except for weighting exponent m. The experiments show that the proposed index Gauss recognizes the optimal c for all four gene microarray datasets. Moreover the optimality of the index is verified with varied values of m for the different validity indices.

Table 1.5 Validity index value for c=2,…10 and m= 1.5 to 2.5 for yeast data set

| Exponent(m) | PC | CE | SC | S | XB | DI | Gauss |
|---|---|---|---|---|---|---|---|
| 1.5 | 0.5000 | 0.6934 | 1.8214 | 2.1199 | 0.2030 | 0.1568 | 1.1846 |
| 1.6 | 0.5000 | 0.6934 | 1.8214 | 2.1199 | 0.2030 | 0.1568 | 1.1846 |
| 1.7 | 0.5000 | 0.6933 | 1.8213 | 2.1198 | 0.2030 | 0.1567 | 1.1845 |
| 1.8 | 0.5000 | 0.6932 | 1.8213 | 2.1198 | 0.2029 | 0.1567 | 1.1845 |
| 1.9 | 0.5000 | 0.6931 | 1.8212 | 2.1197 | 0.2029 | 0.1566 | 1.1845 |
| 2.0 | 0.5000 | 0.6931 | 1.8211 | 2.1197 | 0.2029 | 0.1566 | 1.1845 |
| 2.1 | 0.5001 | 0.6931 | 1.8211 | 2.1197 | 0.2029 | 0.1566 | 1.1845 |
| 2.2 | 0.5001 | 0.6930 | 1.8210 | 2.1196 | 0.2028 | 0.1566 | 1.1845 |
| 2.3 | 0.5002 | 0.6930 | 1.8210 | 2.1196 | 0.2028 | 0.1565 | 1.1845 |
| 2.4 | 0.5003 | 0.6930 | 1.8210 | 2.1196 | 0.2027 | 0.1565 | 1.1844 |
| 2.5 | 0.5003 | 0.6930 | 1.8210 | 2.1195 | 0.2027 | 0.1565 | 1.1844 |

Table 1.6 Validity index value for c=2,…10 and m= 1.5 to 2.5 for colon cancer dataset

| Exponent(m) | PC | CE | SC | S | XB | DI | Gauss |
|---|---|---|---|---|---|---|---|
| 1.5 | 0.6235 | 0.5605 | 0.1447 | 0.0034 | 0.0524 | 0.3259 | 0.0188 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1.6 | 0.6235 | 0.5604 | 0.1447 | 0.0034 | 0.0524 | 0.3259 | 0.0188 |
| 1.7 | 0.6234 | 0.5604 | 0.1446 | 0.0033 | 0.0523 | 0.3258 | 0.0188 |
| 1.8 | 0.6233 | 0.5603 | 0.1446 | 0.0033 | 0.0523 | 0.3258 | 0.0187 |
| 1.9 | 0.6233 | 0.5603 | 0.1446 | 0.0033 | 0.0523 | 0.3258 | 0.0187 |
| 2.0 | 0.6233 | 0.5603 | 0.1446 | 0.0033 | 0.0523 | 0.3258 | 0.0187 |
| 2.1 | 0.6233 | 0.5603 | 0.1446 | 0.0033 | 0.0523 | 0.3258 | 0.0187 |
| 2.2 | 0.6232 | 0.5603 | 0.1446 | 0.0033 | 0.0523 | 0.3258 | 0.0187 |
| 2.3 | 0.6232 | 0.5603 | 0.1445 | 0.0033 | 0.0523 | 0.3258 | 0.0187 |
| 2.4 | 0.6231 | 0.5603 | 0.1445 | 0.0033 | 0.0523 | 0.3258 | 0.0187 |
| 2.5 | 0.6231 | 0.560 | 0.1445 | 0.0032 | 0.0523 | 0.3258 | 0.0187 |

Table 1.7 Validity index value for c=2,…10 and m= 1.5 to 2.5 for leukemia data set

| Exponent(m) | PC | CE | SC | S | XB | DI | Gauss |
|---|---|---|---|---|---|---|---|
| 1.5 | 0.5000 | 0.6932 | 1.3937 | 1.4589 | 0.1277 | 0.6130 | 11.3834 |
| 1.6 | 0.5000 | 0.6932 | 1.3937 | 1.4587 | 0.1277 | 0.6130 | 11.3834 |
| 1.7 | 0.5000 | 0.6932 | 1.3936 | 1.4586 | 0.1276 | 0.6129 | 11.3833 |
| 1.8 | 0.5000 | 0.6931 | 1.3936 | 1.4586 | 0.1276 | 0.6129 | 11.3833 |
| 1.9 | 0.5000 | 0.6931 | 1.3936 | 1.4585 | 0.1276 | 0.6128 | 11.3831 |
| 2.0 | 0.5000 | 0.6931 | 1.3936 | 1.4582 | 0.1276 | 0.6128 | 11.3831 |
| 2.1 | 0.5000 | 0.6931 | 1.3936 | 1.4582 | 0.1276 | 0.6128 | 11.3831 |
| 2.2 | 0.5000 | 0.6931 | 1.3936 | 1.4582 | 0.1276 | 0.6128 | 11.3831 |
| 2.3 | 0.5000 | 0.6930 | 1.3935 | 1.4582 | 0.1276 | 0.6127 | 11.3830 |
| 2.4 | 0.5000 | 0.6930 | 1.3935 | 1.4581 | 0.1276 | 0.6127 | 11.3830 |
| 2.5 | 0.5000 | 0.6930 | 1.3935 | 1.4581 | 0.1276 | 0.6127 | 11.3830 |

Table 1.8 Validity index value for c=2,…,10 and m= 1.5 to 2.5 for splice data set

| Exponent(m) | PC | CE | SC | S | XB | DI | Gauss |
|---|---|---|---|---|---|---|---|
| 1.5 | 0.5000 | 0.6934 | 3.5020 | 1.4329 | 0.1560 | 0.1109 | 2.0431 |
| 1.6 | 0.5000 | 0.6934 | 3.5020 | 1.4329 | 0.1560 | 0.1109 | 2.0012 |
| 1.7 | 0.5000 | 0.6933 | 3.5019 | 1.4328 | 0.1559 | 0.1108 | 1.9921 |
| 1.8 | 0.5000 | 0.6933 | 3.5019 | 1.4328 | 0.1558 | 0.1108 | 1.5490 |
| 1.9 | 0.5000 | 0.6931 | 3.5019 | 1.4327 | 0.1558 | 0.1107 | 1.2318 |
| 2.0 | 0.5000 | 0.6931 | 3.5018 | 1.4327 | 0.1558 | 0.1107 | 1.0818 |
| 2.1 | 0.5000 | 0.6931 | 3.5018 | 1.4327 | 0.1558 | 0.1107 | 1.0437 |
| 2.2 | 0.5000 | 0.6931 | 3.5018 | 1.4327 | 0.1558 | 0.1107 | 0.0943 |
| 2.3 | 0.5000 | 0.6931 | 3.5018 | 1.4326 | 0.1557 | 0.1106 | 0.0547 |
| 2.4 | 0.5000 | 0.6930 | 3.5017 | 1.4326 | 0.1557 | 0.1106 | 0.0315 |
| 2.5 | 0.5000 | 0.6930 | 3.5017 | 1.4325 | 0.1557 | 0.1104 | 0.0018 |

Table 1.9 Value of c by each cluster validity index for 4 data sets (m [1.5,2.5], c=2...,10)

| Data set | PC | CE | SC | S | XB | DI | GA USS |
|---|---|---|---|---|---|---|---|
| **Yeast** | 2 | 2 | 7 | 10 | 10 | 2 | **6** |
| **Colon cancer** | **2** | **2** | 10 | 10 | 4 | 10 | **2** |
| **Leukemia** | **2** | **2** | 9 | 10 | 10 | 3 | **2** |
| **Splice** | 2 | 2 | 5 | 2 | 10 | 2 | **6** |

## 5. CONCLUSION

The quality of the partition can be determined by the cluster validity index. One of the important initial parameter for the fuzzy clustering algorithm FCM is the number of clusters to be generated, which highly reflects the quality of the resulting partition. The validity indices proposed in the literature are dependent upon the membership and the data itself for validity calculation. After reviewing several validity indices a new validity index is proposed named Gauss index. The proposed Gauss index uses Gaussian measure which copes with the uncertainty issue associated with the current real data sets. Along with the membership degree, another measure that helps in proper evaluation result is the gaussian measure. It provides the necessary component in identifying the well-separated and compact cluster. It shows valid results when applied for the microarray data sets yeast, colon cancer, splice and leukemia. The data sets are compared with the existing validity indices: PC, CE, SC, S, XB, DI, ADI and Gauss. The potential associated with the proposed index Gauss assess the validness of the partitions generated from the FCM clustering algorithm. The optimal fuzzy c-partition is obtained by minimizing $V_{Gauss}$ with respect to c. The results of the experimental tests in which various indices are used to determine the optimal number of clusters for microarray data sets showed that the proposed index delivers a reliable result. The reliability measure with varied values of m proves the optimality of the index.

## REFERENCES

[1] Bezdek, James C. "Numerical taxonomy with fuzzy sets." *Journal of Mathematical Biology* 1, no. 1 (1974): 57-71.

[2] Bezdek, James C. "Objective Function Clustering." In *Pattern recognition with fuzzy objective function algorithms*, pp. 43-93. Springer, Boston, MA, 1981.

[3] Bensaid, Amine M., Lawrence O. Hall, James C. Bezdek, Laurence P. Clarke, Martin L. Silbiger, John A. Arrington, and Reed F. Murtagh. "Validity-guided (re) clustering with applications to image segmentation." *IEEE Transactions on Fuzzy Systems* 4, no. 2 (1996): 112-123.

[4] Xie, Xuanli Lisa, and Gerardo Beni. "A validity measure for fuzzy clustering." *IEEE Transactions on pattern analysis and machine intelligence* 13, no. 8 (1991): 841-847.

[5] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern.,vol. 3,pp. 32-57,1973.