Proc. of the Intl. Conf. on Advances In Computing, Electronics and Electrical Technology - CEET 2014.
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-1-63248-005-7 doi: 10.15224/ 978-1-63248-005-7-96

# A Software Tool for Evaluating the Effect of Proximity Measures on Clustering Methods

Ahmet ELBIR, Vecdi Emre LEVENT, Fethullah KARABIBER

***Abstract* — In this study, commonly used proximity measures, in other words, distance and similarity functions are reviewed. An interactive software tool with graphical user interface is developed to examine the effect of proximity measures on the performance of clustering methods. The software named ClustProX allows the users to select a data set, to apply a clustering method and to observe the performance of the results. Xie-Beni and Kwon indices are used to evaluate the performance of the implemented clustering methods with different proximity measures and the number of clusters. In this way, the users will be able to choose an appropriate proximity function and the number of clusters to improve accuracy of clustering.**

***Keywords* — proximity, distance, similarity, clustering, cluster validation**

## I. Introduction

Data Mining, which is one of the most popular areas in Computer Science, retrieves some useful information from a data set [7]. Classification, clustering, association analysis, and anomaly detection are some subtopics of data mining. Clustering, which is an unsupervised learning technique, group the objects into the clusters employing similarity or dissimilarity functions. Measure of proximity is one of the most critical steps in the clustering algorithms. Because the distance based clustering methods group the objects based on proximity measure between attributes or features of the objects. Choosing an appropriate proximity function directly affect the performance of clustering. Therefore examination of different proximity functions and their effects on clustering methods are come into question.

In this study, commonly used proximity functions used in data mining, especially in clustering methods, are examined. Since there are many different proximity functions, we aim to develop an interactive software named *ClustProX* to help the users select the best appropriate proximity function for their datasets. The user can select a clustering method and then observe the results for different proximity functions. The user can also change the number of clusters to find the best clustering process.

Since there in prior knowledge about classes of the objects, it is a challenging problem to validate clustering. There are many techniques to evaluate the performance of clustering results. Some clustering validation indices such as Xie-Beni and Kwon index [3, 9] is implemented to evaluate performance of the clustering method. The software has been implemented by C# programming language is freely downloadable from www.cib.yildiz.edu.tr/software.

## II. Measures of Proximity

There are many different distance/similarity functions for different data types such as numerical, categorical and binary. Some of the most commonly used proximity measure for numerical attributes are Euclidian distance, Manhattan distance, Chebyshev distance, Bray Curtis distance, Canberra Distance, Cosine Distance and Pearson Correlation. In case of binary attributes, the proximity functions such as Hamming, Jaccard, Dice and Yule can be used. In this study, we focus on the proximity functions for numerical attributes.

In some cases similarity or distance values needs to be normalized. Generally the values are transformed to [0-1] range with some additional mathematical operations. In this study, some commonly used proximity functions for numerical attributes are given for two vectors of length N ( $X = [x_1 \ x_2 \dots x_N]$ and $Y=[y_1 \ y_2 \dots y_N]$ ) in the following.

### A- *Euclidean Distance*

Euclidean is the most popular distance function. Euclidean function, which is also called as L2 Norm distance, calculates the minimum distance among two instances [1]. The distance between two pairs of points can be calculated as follows.

$$d_{x,y} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + \dots + (x_N - y_N)^2} \qquad (1)$$

### B- *Manhattan Distance*

Manhattan distance is similar to Euclidean distance and it also known as L1 Norm or City Block distance. Manhattan distance is calculated by sum of absolute values of the difference between the pairs. Equation-2 finds Manhattan Distance value between two pairs of points [1].

$$d_{x,y} = |x_1 - y_1| + |x_2 - y_2| + .. + |x_N - y_N| \qquad (2)$$

Ahmet Elbir*, Vecdi Emre Levent*, Fethullah Karabiber*
*Yildiz Technical University, Computer Engineerin Department,
34220 Esenler - Istanbul / TURKEY

### C- *Chessboard (Chebyshev) Distance*

Chessboard distance is calculated by using the maximum value of absolute values among instances [6]. Chessboard distance function is defined as:

$$\mathrm{d}_{\mathrm{x,y}} = \max\left(|x_1 - y_1|, |x_2 - y_2|,.,|x_\mathrm{N} - y_\mathrm{N}|\right) \qquad (3)$$

### D- *Bray Curtis Distance*

Bray Curtis distance is modified from Manhattan distance. The Bray–Curtis distance is bounded between 0 and 1. The value 1 means that two vectors have same values and the value 0 means that two vectors have different values [2]. The Bray Curtis distance is defined by Equation-4.

$$\mathrm{d}_{\mathrm{x,y}} = \left(\frac{|x_1 - y_1| + |x_2 - y_2| + .. + |x_\mathrm{N} - y_\mathrm{N}|}{|x_1 + y_1| + |x_2 + y_2| + .. + |x_\mathrm{N} + y_\mathrm{N}|}\right) \qquad (4)$$

### E- *Canberra Distance*

The Canberra distance is used to compare ranked lists and to detect intrusion in computer security [5, 8]. Canberra distance is defined as in Equation-5.

$$\mathrm{d}_{\mathrm{x,y}} = \left(\frac{|x_1 - y_1|}{|\mathrm{x}_1| + |y_1|} + \frac{|x_2 - y_2|}{|x_2| + |y_2|} + \ldots + \frac{|x_\mathrm{N} - y_\mathrm{N}|}{|x_\mathrm{N}| + |y_\mathrm{N}|}\right) \qquad (5)$$

### F- *The Pearson Correlation Coefficient*

The Pearson correlation coefficient measures the strength and direction of the linear relationship of two variables *X* and *Y*. It gives a value between +1 and −1, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation. The sample Pearson's correlation coefficient represented by the character *r* between two random variables is defined as in Equation-6, which is the covariance of the two variables divided by the product of their standard deviations [1, 4].

$$r = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} \qquad (6)$$

### G- *Cosine Similarity*

Cosine similarity function finds the cosine of the angle between two vectors to measure the similarity [1,4]. If cosine of the angle between two vectors is 1, it means that the angle is 0° and the two vectors have the same orientation. The two vectors, which is orthogonal, have a similarity of 0 (Cos(90)=0), and two vectors diametrically opposed have a similarity of Cos(180)= -1. The cosine similarity bounded in [-1,1] for the two vectors is given by:

$$CosSim(x,y) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^{\,2}}\sqrt{\sum y_i^{\,2}}} \qquad (7)$$

## III. Clustering

Clustering algorithms partition the objects such that the objects within a cluster are similar to each other and dissimilar to the objects in other clusters. Since there is no prior knowledge about the classes, clustering is an unsupervised learning method. There are lots of clustering algorithms based on density, grid, partition, distance, etc.

In this study K-Means, which is the most popular clustering algorithm, is employed to examine the effects of proximity measures on clustering. The most critical step in K-Means algorithm is to calculate distance between the attributes. K-means algorithm is given below [7].

```
Algorithm 1: K-Means Clustering

1- Specify k, the number of clusters
2- Identify the initial centroid points of k
   clusters by choosing k objects randomly
3- Assign each object to its closest cluster
   using a distance function.
4- Calculate the centroid for each cluster
   to find new cluster center
5- Reassign all objects to the closest
   cluster centroid.
6- Iterate until the all centroids don't
   change.
```

Another popular distance based clustering algorithm is K-Medoids, which is derived from K-Means method. Since K-Means algorithm is sensitive to outliers, K-Medoids algorithm employs median value in sorted array to calculate the centroid value.

## IV. Cluster Validation Techniques

Since there is no prior knowledge about the dataset in unsupervised learning process, clustering is performed according to proximity values. The clustering process may have several problems such as anomaly samples and optimum number of clusters. There are two conditions for an ideal clustering process. Firstly, cluster centers must be away from each other. Secondly, each instance in the cluster must be close to each other. A good cluster should minimize within cluster variance and maximize between cluster variance.

There are many cluster validation techniques to evaluate performance of clustering methods. One of the most important approaches is Xie-Beni index. The Xie-Beni index is known as compactness or denseness and separation validity methods [3]. If Xie-Beni index value is close to 1, clustering is considered as more successful. The Xie-Beni index value is calculated as in (8).

$$V_{XB}(U,V;X) = \frac{\sum_{i=1}^{C}\sum_{j=1}^{R} \mu_{ij}^{\;2} \|x_j - v_i\|^2}{n \min_{i \neq j} \|v_i - v_j\|^2} \qquad (8)$$

where *X* is a set of *n* data points in *p*-dimensional space, *c* is the number of clusters, *V=[v₁,…,v_c]* is the cluster center

ISBN: 978-1-63248-005-7 doi: 10.15224/ 978-1-63248-005-7-96

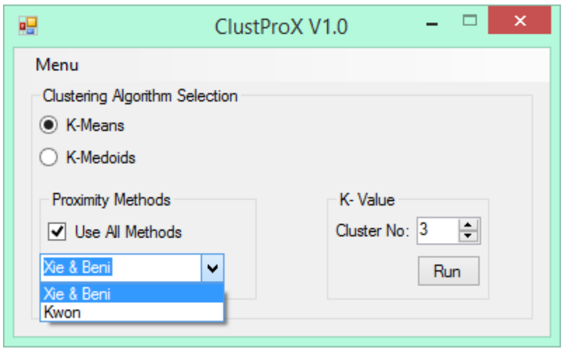
matrix, $\mu_{ij}$ is the membership value of $x_j$ belonging to $v_i$, $U=[u_{ij}]_{cxn}$ is the membership matrix.

Another popular cluster validity texhnique is Kwon index which is derived from Xie-Beni index. The Kwon index is calculated as in (9) [9, 10].

$$V_K(U,V;X) = \frac{\sum_{i=1}^{C}\sum_{j=1}^{R}\mu_{ij}^{2}\left\|x_j - v_i\right\|^2 + \frac{1}{c}\sum\left\|v_i - v'\right\|^2}{\min_{i \neq k}\left\|v_i - v_k\right\|^2} \qquad (9)$$

## Results and Discussion

An interactive software named ClustProX with graphical user interface was implemented by using C# language. Firstly, the users can import a dataset via menu interface and then select the number of clusters. Later, the users can select one of the implemented proximity functions to cluster by K-Means or K-Medoids. Main window of the application is given in Figure 1. The users are able to choose a proximity measure from the combo box and to define cluster numbers by the spin box.



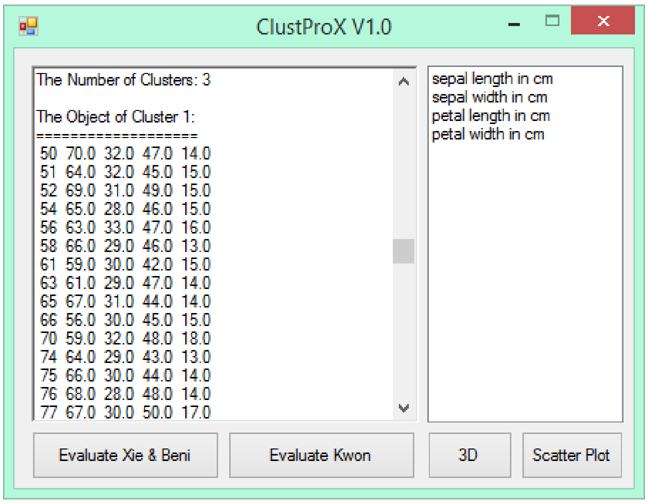
● K-Means
○ K-Medoids
☑ Use All Methods
Xie & Beni
Xie & Beni
Kwon
Cluster No: 3

Figure-1 Main Window. User choose a clustering method, select one or all proximity measures along with a validity index and specify the number of clusters.

After applying the clustering, there are three different options to visualize the results. These options are list view, three-dimensional view and scatter plot view.

The objects assigned to the clusters are listed sequentially as shown in Figure-2. All attributes of data are shown in right part. Users also can observe clustering validation results by pressing Xie-Beni and Evaluate Kwon buttons in Figure-2.

3D button is used to see three-dimensional view of the selected three attributes. Three features must be selected before clicking the 3D button. The software also shows the scatter plot by selecting two features of data. Figure-3 shows a scatterplot of the clustering result.



```
The Number of Clusters: 3

The Object of Cluster 1:
===================
50  70.0  32.0  47.0  14.0
51  64.0  32.0  45.0  15.0
52  69.0  31.0  49.0  15.0
54  65.0  28.0  46.0  15.0
56  63.0  33.0  47.0  16.0
58  66.0  29.0  46.0  13.0
61  59.0  30.0  42.0  15.0
63  61.0  29.0  47.0  14.0
65  67.0  31.0  44.0  14.0
66  56.0  30.0  45.0  15.0
70  59.0  32.0  48.0  18.0
74  64.0  29.0  43.0  13.0
75  66.0  30.0  44.0  14.0
76  68.0  28.0  48.0  14.0
77  67.0  30.0  50.0  17.0
```

```
sepal length in cm
sepal width in cm
petal length in cm
petal width in cm
```

**Figure-2** List View. Left panel shows the number of clusters and the objects assigned to each cluster. Right panel lists the attributes of the dataset. The buttons at bottom are used to observe results of the cluster validation techniques and to visualize data.

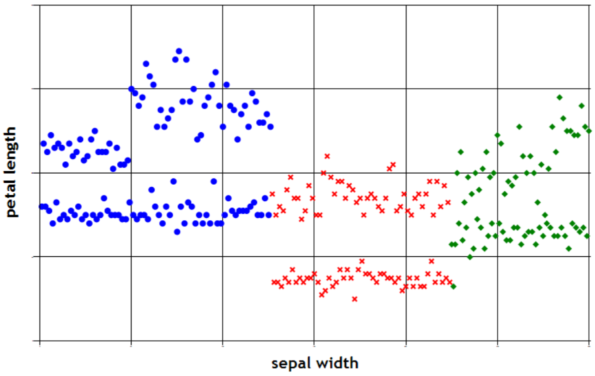
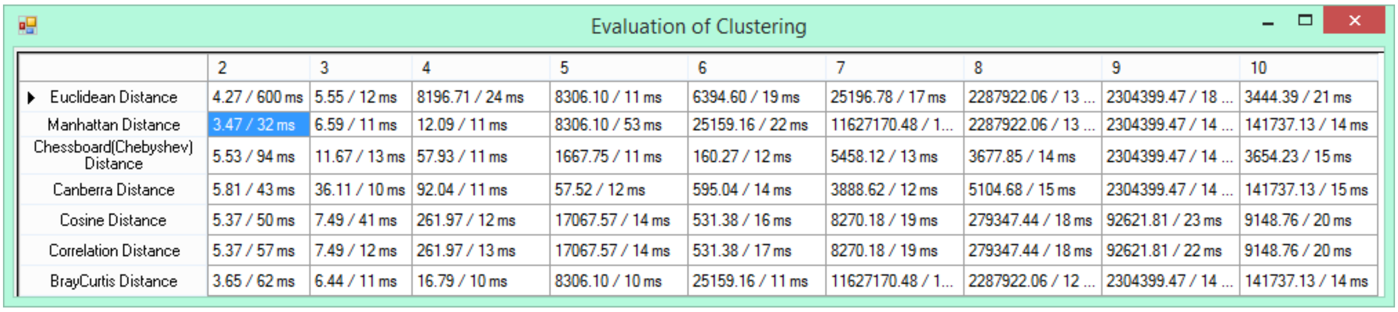


**Figure-3** Visualizing data in 2D using Scatter plot by selecting two attributes. Each cluster is shown in different colors.

In addition to visually comparison, the user can evaluate the clustering results for the different proximity functions by Xie-Beni and Kwon index values. The index values and corresponding execution times for different number of clusters and proximity function is displayed in Figure-4.

In order to analyze the effects of implemented proximity functions on clustering process, IRIS and Protein datasets are employed. Xie-Beni and Kwon indices for the datasets are obtained from K-Means algorithm using different number of cluster and proximity methods. Table-I and Table-II show the results of Xie-Beni and Kwon indices obtained from K-Means algorithm for Iris and Protein datasets, respectively. All tables include the number of clusters and seven different proximity measures. The best validation results are shown in bold in the tables. In this way, the users can select the most appropriate distance function and the number of clusters to increase accuracy of clustering process.

Proc. of the Intl. Conf. on Advances In Computing, Electronics and Electrical Technology - CEET 2014.
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-1-63248-005-7 doi: 10.15224/ 978-1-63248-005-7-96



Evaluation of Clustering

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| Euclidean Distance | 4.27 / 600 ms | 5.55 / 12 ms | 8196.71 / 24 ms | 8306.10 / 11 ms | 6394.60 / 19 ms | 25196.78 / 17 ms | 2287922.06 / 13 ... | 2304399.47 / 18 ... | 3444.39 / 21 ms |
| Manhattan Distance | 3.47 / 32 ms | 6.59 / 11 ms | 12.09 / 11 ms | 8306.10 / 53 ms | 25159.16 / 22 ms | 11627170.48 / 1... | 2287922.06 / 13 ... | 2304399.47 / 14 ... | 141737.13 / 14 ms |
| Chessboard(Chebyshev) Distance | 5.53 / 94 ms | 11.67 / 13 ms | 57.93 / 11 ms | 1667.75 / 11 ms | 160.27 / 12 ms | 5458.12 / 13 ms | 3677.85 / 14 ms | 2304399.47 / 14 ... | 3654.23 / 15 ms |
| Canberra Distance | 5.81 / 43 ms | 36.11 / 10 ms | 92.04 / 11 ms | 57.52 / 12 ms | 595.04 / 14 ms | 3888.62 / 12 ms | 5104.68 / 15 ms | 2304399.47 / 14 ... | 141737.13 / 15 ms |
| Cosine Distance | 5.37 / 50 ms | 7.49 / 41 ms | 261.97 / 12 ms | 17067.57 / 14 ms | 531.38 / 16 ms | 8270.18 / 19 ms | 279347.44 / 18 ms | 92621.81 / 23 ms | 9148.76 / 20 ms |
| Correlation Distance | 5.37 / 57 ms | 7.49 / 12 ms | 261.97 / 13 ms | 17067.57 / 14 ms | 531.38 / 17 ms | 8270.18 / 19 ms | 279347.44 / 18 ms | 92621.81 / 22 ms | 9148.76 / 20 ms |
| BrayCurtis Distance | 3.65 / 62 ms | 6.44 / 11 ms | 16.79 / 10 ms | 8306.10 / 10 ms | 25159.16 / 11 ms | 11627170.48 / 1... | 2287922.06 / 12 ... | 2304399.47 / 14 ... | 141737.13 / 14 ms |

**Figure-4** Evaluation of clustering. The results of selected cluster validity index for applied clustering algorithm are given separately for all implemented proximity functions and corresponding execution times. The best one is highlighted.

TABLE I. CLUSTER VALIDATION RESULTS OF K-MEANS ALGORITHM FOR IRIS DATASET.

| **Proximity Methods** | **Xie-Beni Index Value** | | | | **Kwon Index Value** | | | |
|---|---|---|---|---|---|---|---|---|
| | ***2*** | ***3*** | ***4*** | ***5*** | ***2*** | ***3*** | ***4*** | ***5*** |
| Euclidean | 4.27 | **5.55** | 8196.71 | 8306.10 | 4.33 | **5.71** | 8511.91 | 8306.75 |
| Manhattan | **3.47** | 6.59 | **12.09** | 8306.10 | **3.56** | 6.85 | **12.52** | 8306.75 |
| Chebyshev | 5.53 | 11.67 | 57.93 | 1667.75 | 5.57 | 11.81 | 59.05 | 1701.16 |
| Canberra | 5.81 | 36.11 | 92.04 | **57.52** | 5.86 | 36.40 | 92.68 | **61.02** |
| Cosine | 5.37 | 7.49 | 261.97 | 17067.57 | 5.42 | 7.76 | 271.66 | 17863.33 |
| Correlation | 5.37 | 7.49 | 261.97 | 17067.57 | 5.42 | 7.76 | 271.66 | 17863.33 |
| Bray-Curtis | 3.65 | 6.44 | 16.79 | 8306.10 | 3.74 | 6.69 | 17.44 | 8306.75 |

TABLE II. CLUSTER VALIDATION RESULTS OF K-MEANS ALGORITHM FOR PROTEIN DATASET.

| **Proximity Methods** | **Xie-Beni Index Value** | | | | **Kwon Index Value** | | | |
|---|---|---|---|---|---|---|---|---|
| | ***2*** | ***3*** | ***4*** | ***5*** | ***2*** | ***3*** | ***4*** | ***5*** |
| Euclidean | 1.57 | 2.08 | 28.78 | 586.02 | 1.60 | 2.15 | 31.20 | 630.10 |
| Manhattan | 1.57 | 2.08 | 12.94 | 78.18 | 1.60 | 2.15 | 13.51 | 84.60 |
| Chebyshev | **1.56** | **1.88** | 255.19 | 63.51 | **1.59** | **1.95** | 274.71 | 69.24 |
| Canberra | 2.04 | 5.48 | 11.95 | 600.74 | 2.10 | 5.77 | 12.57 | 687.69 |
| Cosine | 1.57 | 6.26 | **9.72** | 380.41 | 1.60 | 6.53 | **10.35** | 432.37 |
| Correlation | 1.57 | 6.26 | **9.72** | 380.41 | 1.60 | 6.53 | **10.35** | 432.37 |
| Bray-Curtis | 1.57 | 5.19 | 472.21 | **25.34** | 1.60 | 5.37 | 516.55 | **27.99** |

## V. Conclusion

In this study, the commonly used proximity functions in data mining are reviewed. One of the most critical steps in the clustering algorithms is measure of similarity or dissimilarity between two objects. Interactive software is created to examine the effect of proximity functions on clustering process. In order to select the best appropriate proximity function, the user can import a dataset, then apply a clustering method and finally observe the results visually. As a result, the user can improve the performance of clustering using the interactive software developed in this study. In future, additional proximity functions for different data types, some other popular clustering methods and clustering validation techniques will be implemented to improve capability of the software.

## *References*

[1] Monev V., Introduction to Similarity Searching in Chemistry, MATCH Commun. Math. Comput. Chem. 51 pp. 7-38 , 2004

[2] Bray, J. R. and J. T. Curtis. 1957. An ordination of upland forest communities of southern Wisconsin. Ecological Monographs 27:325-349.

[3] X. Xie and G. Beni, “A Validity Measure for Fuzzy Clustering,”IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI),
vol. 13, no. 8, pp.841 - 847, 1991.

[4] Rui Xu Survey of Clustering Algorithms IEEE Transactıons On Neural Networks, Vol. 16, No. 3, May 2005

[5] Jurman G, Riccadonna S, Visintainer R, Furlanello C: Canberra Distance on Ranked Lists. In Proceedings, Advances in Ranking – NIPS 09 Workshop Edited by Agrawal S, Burges C, Crammer K. 2009, 22–27.

[6] F. James, Fitting Tracks in Wire Chambers Using the Chebyshev Norm instead of Least Squares, Nucl. Instrum. Methods Phys. Res. 211 (1983) 145.

[7] Introduction to Data Mining Pang-Ning Tan, Michael Steinbach,Vipin Kumar Addison Wesley, ISBN:0-321-3236-7

[8] Lance, G. N.; Williams, W. T. (1966). "Computer programs for hierarchical polythetic classification ("similarity analysis").". Computer Journal 9 (1): 60–64. doi:10.1093/comjnl/9.1.60

[9] Y.G. Tang, F.C. Sun, Z.Q. Sun, Improved validation index for fuzzy clustering, in: American Control Conf., June 8–10, 2005, Portland, OR, USA.

[10] Weina Wanga, Yunjie Zhang On fuzzy cluster validity indices Fuzzy Sets and Systems 158 (2007) 2095 – 2117

[11] J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York