International Journal of Innovative Computing, Information and Control
Volume **8**, Number **2**, February **2012**

ICIC International ©2012 ISSN 1349-4198
pp. 1275–1297

SIMILARITY-DISSIMILARITY PLOT FOR HIGH DIMENSIONAL DATA OF DIFFERENT ATTRIBUTE TYPES IN BIOMEDICAL DATASETS

Muhammad Arif and Saleh Basalamah

College of Computer and Information Systems
Umm-Alqura University
Makkah, Kingdom of Saudi Arabia
{ mahamid; smbasalamah }@uqu.edu.sa

Received November 2010; revised March 2011

Abstract. *In real life biomedical classification applications, feature space may be of high dimension in which visualization of class distribution is impossible. Moreover, attributes of features may be numeric, ordinal, categorical or binary. Most of the time, features may be composed of mixed type of attributes. In this paper, the concept of similarity-dissimilarity is extended to various types of attributes. Similarity-dissimilarity plot projects the high dimensional feature space on two dimensional plot revealing the class separation in the feature space which may be continuous or discrete. Furthermore, effect of various distance measures proposed in the literature for different type of attributes is also studied. An index called percentage of data points above the similarity-dissimilarity line (PAS) is proposed which is the fraction of data points found near to its own class as compared to other classes. Several real life biomedical datasets are used to show the effectiveness of the proposed similarity-dissimilarity plot and the PAS index.*

Keywords: Visualization, High dimensional data, Pattern classification, Features quality, Nearest neighbors

1. **Introduction.**

In biomedical applications, data or feature set can be of numeric, nominal, ordinal or categorical type. Moreover, due to high dimensionality of the data, it is not possible to visualize the data and extract important information about the structure of the data in the context of pattern classification. In pattern classification problems [1, 2], raw data are measured from sensors, images or clinical tests and features are extracted so that different classes may be discriminated in the feature space in an appropriate way. Classifier of arbitrary type and settings are applied to these features to get best possible classification accuracy. Classifiers are optimized using classification accuracy as an optimization criterion. However, one question remains unanswered that whether certain classification accuracy is due to the selection of classifier or poor discrimination quality of the feature set. It is easy to visualize different clustering patterns belonging to different classes in low dimensional feature space (up to three dimensions). Hence, quality of features can be assessed easily. However, in the high dimensional feature space, specialized visualization tools are needed that can help in projecting the data on two or three dimensional space in a meaningful and more descriptive manner. Purpose of visualization is to study intra-class and inter-class relationships so that quality of the features in discrimination of different classes can be assessed.

In case of multivariate data sets where some or all variables are correlated to each other in some linear or nonlinear sense, projecting the high dimensional data to lower dimension by using Principle Component Analysis (PCA) [3], projection pursuit [4] or Kohonen's self organizing map (SOM) [5,6] is very popular and applied to many real life applications. In

Kohonen SOM, the number of output nodes is user defined and separation of the clusters belonging to different classes and their separation distances cannot be visualized. In the cases where different features are independent and contain different type of attributes, it is difficult to apply projection techniques like PCA and Projection pursuit.

Visualization of high dimensional data is very useful in many real life applications to extract interesting information about the data [7-9]. Scatter plots [10,11] show high dimensional data as pair-wise scatter plots on a scatter plot matrix. For feature space having n dimensions, scatter plot matrix generates $\frac{n(n-1)}{2}$ scatter plots. Certain variation of scatter plot like HyperSlice [12], HyperBox [13] and Prosection [14] is also proposed in the literature. These kinds of visualizations become inadequate for high dimensional feature space. Moreover, relationship between two dimensions can be understandable and clustering representation for different classes can be seen. However, relationship among all dimensions cannot be visualized and understood.

Parallel coordinates [15,16] translate a data point in the high dimensional space into a poly line intersecting each axis placed in parallel horizontally. Large dimensions can be represented by this technique for a limited number of data points in the feature space. However, arrangement of dimensions differently can change the patterns of poly lines. Zhou et al. [17] presented a framework to improve the visualization quality by optimally arranging the coordinates. Further variations of parallel coordinates are proposed in [18-20]. These techniques are good for visualization for limited number of data points of numeric type. As the number of data points increases, the number of poly lines will increase making the comprehension of clustering difficult.

Glyphs are another set of interesting visualization techniques in which complex symbols are used and the features of the symbols represent the values of the attributes. Examples of such techniques are Chernoff faces [21], trees [22], starts [23] and autoglyph [24], etc. Grand tour [25] is another set of projection technique in which two dimensional class preserving projections are displayed as a sequence. However, if the number of dimensions becomes large, this technique becomes computationally very expensive.

Several other techniques are proposed in the literature to visualize clustering patterns in the feature space. A hierarchical organization in which variables are placed at different levels of hierarchy is used to visualize the clustering structure of the data. Hierarchical clustering algorithms [26-28] represent hierarchical tree diagram (dendrogram) showing how close individual data points cluster together in the feature space. Hierarchical parallel coordinates [29] also used the concept of hierarchical organization and data are clustered according to user defined parameters. Visual Hierarchical Dimension Reduction (VHDR) [30] is another technique in which user can control the hierarchical organization to visualize certain aspects of the data.

For high dimensional feature space, Radviz [31,32] is a visualization technique that can be applied to visualize the data structure. It can be applied to feature space having continuous attributes normalized to interval $[0, 1]$. Discrete attributes have to be converted into continuous form before visualizing on Radviz. Radviz uses the concept of Hook's law from the physics. Dimension anchors are placed evenly on the circumference of a unit circle. Dimension anchors attract every data point towards itself with the strength proportional to the value of data point in the dimension corresponding to the dimension anchor. Data point is plotted in the unit circle where equilibrium of all the attraction forces is achieved. In Radviz, since dimensions are placed on the circumference of a unit circle, the number of dimensions placed on the unit circle becomes limited for meaningful visualization. Moreover, Radviz needs some optimization for arrangement of dimensions as dimension anchors (which dimension should be placed where?) on the unit circle. It is handled by Vizrank [32] and McCarthy et al. [33]. Freeviz [34] is another extension of Radviz which allows dimension anchors to be placed anywhere in the unit circle. Correlated features are placed near to each other and less important features are placed near the center of unit circle. These kinds of visualizations are not suitable for the feature space having dimensions greater than 100, as only 180 degrees are available on a unit circle. Vectorized Radviz [35] is proposed to better visualize the multi-clusters data by increasing the dimensions of the data through data flattening.

Hence, visualization tools explained above may perform well when the dimension of the feature space is small and as the number of dimension of the feature space increases, the visualization tools fail to provide useful information in the context of pattern classification. A good visualization tool in the context of pattern classification should exhibit following important information. Which are the good data points having high inter-class distances and low intra-class distances? Whether bad data points which are very near to some of the other classes and far away from their own classes can be identified? If a data point is near to a wrong class, then what is that wrong class? Are there any outliers in the data? Similarity-dissimilarity plot proposed in [36] answers all these three questions. Independent of the number of dimensions, the proposed plot can discriminate between good quality data points (producing good classification accuracy) and bad quality data points (creating confusion with other classes) on the feature space. Moreover, this plot will also show the class to whom bad quality data points are confusing. Outliers in the feature space can also be pointed out on this plot. In this paper, concept of similarity-dissimilarity plot is extended to different types of attributes which are very common in the biomedical applications. Biomedical dataset can consist of continuous attributes (outcome of a diagnostic test as numeric value, etc.), categorical attributes (presence or absence of certain characteristics, male/female, etc.) and ordinal attributes (discrete level of a certain disease or status of patient), etc. A visualization tool is required that can handle all types of attributes and provide visualization of the high dimensional data set in a meaningful way. By using an appropriate distance measure in the similarity-dissimilarity plot, all types of attributes can be handled simultaneously. Furthermore, an index called percentage of data points above similarity-dissimilarity line can predict the classification accuracy. Moreover, it can also provide a mechanism for selecting an appropriate distance measure to be used for a particular biomedical dataset.

2. Description of Similarity-dissimilarity Plot.

Let $X_i \in F$, where F is an n-dimensional feature space and $X_i = \{x_i^1, x_i^2, \ldots, x_i^n\}$ is a data point in the feature space. The data point may be a set of numeric, binary, categorical, ordinal or mixed attributes belonging to one of N_C classes. Total number of data points in the data set is $N = \sum\limits_{i=1}^{N_C} n_i$, where n_i, $i = 1, \ldots, N_C$ are the data points belonging to individual classes.

To remove the biasness of the data, all numeric attributes of the data set are normalized such that their means become zero and variances are set to one. Normalization can be done as follows,

$$X_i^j = \frac{X_i^j - \mu^j}{\sigma^j}, \; i = 1, 2, \ldots, N \text{ and } j = 1, 2, \ldots, n \tag{1}$$

where μ^j is the mean of a particular numeric attribute in the jth dimension and σ^j is the standard deviation of the numeric attribute in the jth dimension.

Algorithm of Similarity-Dissimilarity Plot is described in Algorithm 1. Similarity and dissimilarity distances of all the data points will be calculated and plotted on a two dimensional plot. Every class in the feature space will carry a particular marker shape and color. All data points of the ith class will be plotted on the Similarity-Dissimilarity plot by shape marker assigned to the ith class. Color of the marker of the data point will depend on the location of the data point in the feature space. If its similarity distance with the data points of its own class is less than the dissimilarity distance, the color of the marker of this data point will be the color of the class of the ith class. Otherwise the color of the marker of this data point will be assigned to the color of the majority class present in the nearest neighbors from other classes.

Algorithm 1: Algorithm of similarity-dissimilarity plot

```
Function Name: Similarity-dissimilarity-Plot
Input Parameters
Data           Input data (Instances × Attribrtes)
Labels         Class Labels (Instances × 1)
               $Labels = 1, 2, \ldots, N_c$ where $N_c$ is the number of classes
NN             Number of Nearest Neighbors
Dist-type      Type of Distance function used to find nearest neighbors
Markers        Style of markers for each class, shape and Color ($N_c \times 2$)
Output Parameters
PAS            Percentage of Data points above the Similarity-Dissimilarity Line

Function Name: [idx,D] = Find_KNN(A, query, NN, Dist_type)
Input Parameters
$A$            Data (Instances × Attribrtes)
query          Query Point for which nearest neighbors are required from $A$
NN             Number of Nearest Neighbors
Dist-type      Type of Distance function used to find nearest neighbors
Output Parameters
$Idx$          Index of NN nearest neighbors from $A$
$D$            Distances of NN nearest neighbors from the query point
Function PAS = Similarity-dissimilarity-plot (Data, Labels, NN,
               Dist-type, Markers)
$PAS = 0$
For $i = 1, \ldots, N_c$,
        $Class_i = Data(Labels = i, :)$
        $Class_{not\_i} = Data(Labels \neq i, :)$
For $m = 1, \ldots, n_i$
        $y_m = Class_i(m, :)$
        $[idx_1, D_1] = \text{Find\_KNN}(Class_i, y_m, NN, Dist\_type)$
        $[idx_2, D_2] = \text{Find\_KNN}(Class_{not\_i}, y_m, NN, Dist\_type)$
        $Sim\_dist(m) = mean(D_1)$
        $Dissim\_dist(m) = mean(D_2)$
If $Sim\_dist(m) < Dissim\_dist(m)$
        $Plot(Sim\_dist(m), Dissim\_dist(m), \text{Marker}(m, 1), \text{Marker}(m, 2))$
        $PAS = PAS + 1$
Else
        $New\_label = \text{Find\_}mode(Labels(idx_2, 1))$
        $Plot(Sim\_dist(m), Dissim\_dist(m), \text{Marker}(m, 1), \text{Marker}(New\_Label, 2))$
End
End $i$, $m$
               $PAS = \frac{PAS}{\text{Instances}} \times 100$
```

An example of three classes is plotted in Figure 1. Box is assigned to class 1, triangle to class 2 and circle to class 3 as marker shape. Red is the marker color of class 1, blue is for class 2 and green is for class 3. A similarity-Dissimilarity line is drawn as black showing the points where similarity distance is equals to dissimilarity distance. It can be observed from the figure that all data points which are above the similarity-dissimilarity line are drawn by the shape and color of the marker assigned to their respective classes. All the data points which are below the similarity-dissimilarity line are plotted by the shape of the marker assigned to their respective classes but their marker's color represent the majority class of the nearest neighbors from other classes. For example, triangle below the similarity-dissimilarity line is carrying the color of class 3 which is green. So, this data point is more nearer to class 3 as compared to its own class which is class 2. One data point of class 1 is far away from the rest of the data points and may be considered as an outlier.

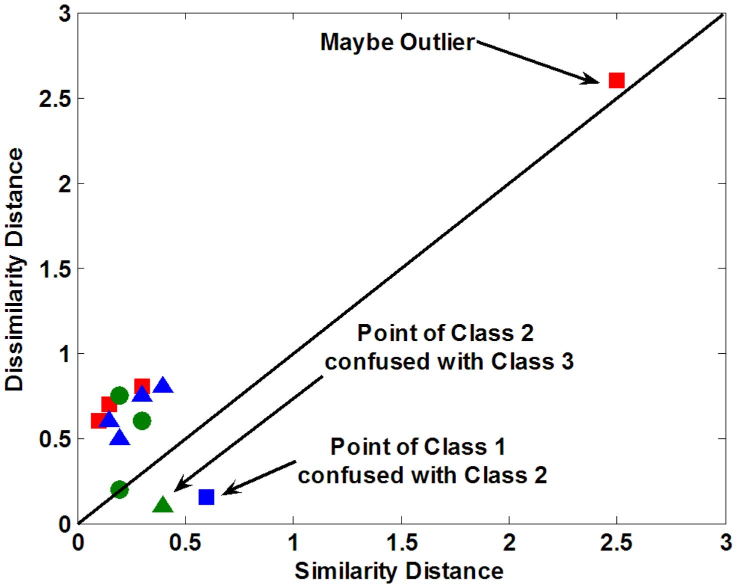


FIGURE 1. Example of similarity-dissimilarity plot (red box: class 1, blue triangle: class 2 and green circle: class 3)

Hence, all data points lying above this line will be considered good quality data points in the context of pattern classification. These data points are located near to each other for similar classes in the feature space and far away from rest of the classes. All those data points lying below this line will be considered as bad quality points. In pattern classification, these data points will be confused with other classes and in the feature space these data points represent overlapping regions among classes.

High dimensional data points which are impossible to visualize can be plotted on a two dimensional similarity-dissimilarity plot and many important characteristics can be extracted from this plot which may be very useful in pattern classification. Percentage of data points above the similarity-Dissimilarity line (PAS) shows the expected accuracy of the classifier using a particular feature set. PAS is defined as below,

$$PAS = \frac{100}{N} \sum_{i=1}^{N} \begin{cases} 1 & Sim\_dist < Dissim\_dist \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Visualization of high dimensional feature space on Similarity-Dissimilarity plot gives us very important information summarized below.

- Similarity-dissimilarity plot will not reveal the clustering structure of the classes in the feature space. But it will give the information about how far clusters of one class are located from rest of the classes. This information is very critical in pattern classification and helps a lot in predicting achievable classification accuracies.
- All data points plotted below similarity-dissimilarity line will be shown by the color of majority class of NN nearest neighbors. Hence, overlapping regions among different classes can be discovered. This will also helps us to know the class by which these data points will be confused by a classifier.
- Any outlier present in the feature space will be plotted on the similarity-dissimilarity plot as an outlier.
- Percentage of data points plotted above the similarity-dissimilarity line (PAS) gives an idea about the possible achievable accuracy of the classifier.
- If PAS is not deteriorating with the increase in the number of nearest neighbors to calculate the similarity and dissimilarity distances, clusters within class are compact or far away from the other classes.

3. Distance Measures for Similarity-dissimilarity Plot.

In pattern classification, many classifiers like K nearest neighbor, support vector machine, radial basis neural networks, and many more define the neighborhood or the optimizing criterion by calculating the distances among the data points in the feature space. A variety of distance metrics or measures are proposed in the literature for a pair of numeric, ordinal, nominal or categorical data points, a set of attributes or probability density measures. A distance measure or metric is defined such that it should be positive semi-definite, symmetric and satisfies the identity and triangular inequality. In this section, some of the distance measures used for different types of attributes are discussed.

3.1. Numeric type of attributes.

For numeric data, many distance measures like Euclidean distance, weighted Euclidean distance, City block (or Manhatten) distance, Minkowski distance, Chebyshev distance and Mahalanobis distance are widely used in the literature. Let $X = [x_1, x_2, \ldots, x_n]$ and $Y = [y_1, y_2, \ldots, y_n]$ are two n-dimensional data points in R^n. Minkowski distance measure of order p between X and Y is defined as follows,

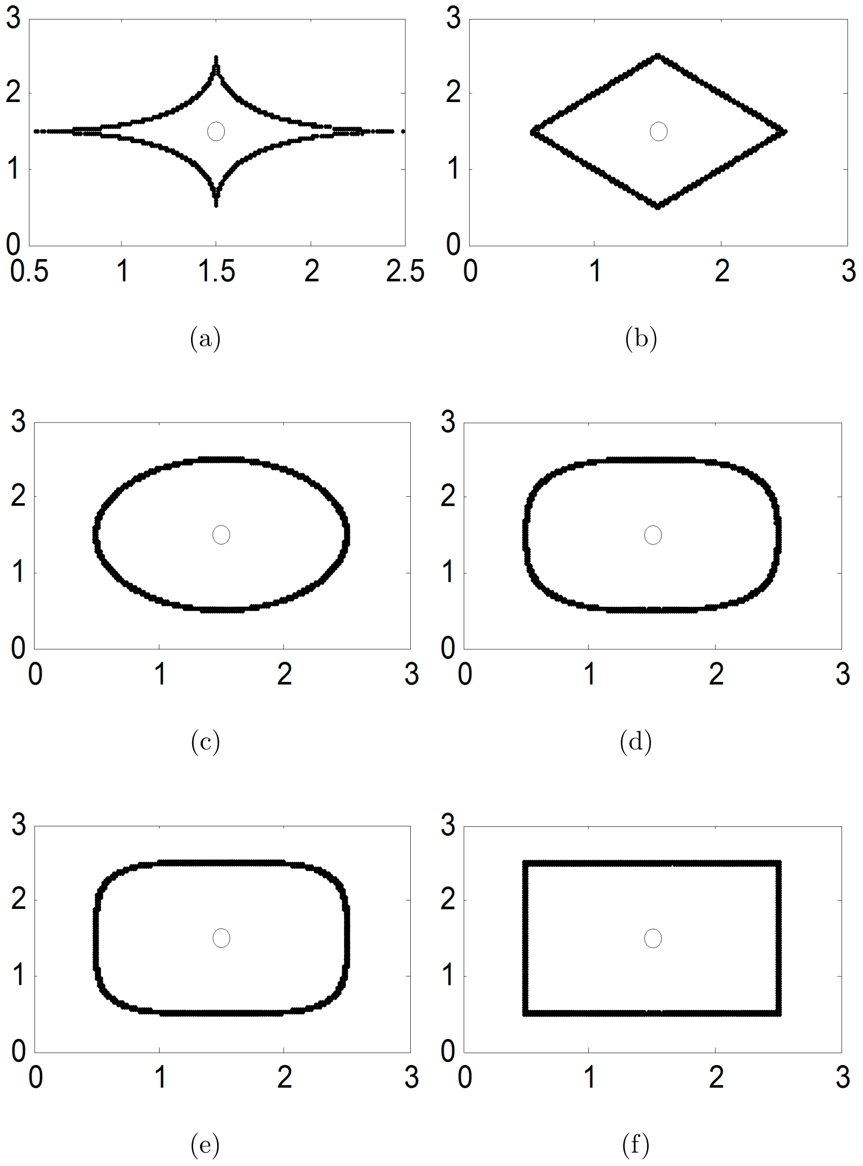
$$d_p(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \tag{3}$$

Euclidean distance is the Minkowski distance with $p = 2$ and Manhatten distance is another Minkowski distance with $p = 1$. Chebychev distance $d_{cheb}(X, Y)$ [37] is a special case of Minkowski distance with $p \to \infty$ as shown below,

$$d_{cheb}(X, Y) = \lim_{p \to \infty} \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} = \max_{i=1}^{n} |x_i - y_i| \tag{4}$$

Different values of p results in neighborhood of different sizes and shapes as shown in the Figure 2.

For a constant neighborhood of 1.0 around a data point of (1.5, 1.5) in the 2-dimensional space, the shape of neighborhood is diamond in case of $p = 1$, circular in case of $p = 2$ and square in case of $p \to \infty$. In the context of pattern classification, different distance metrics/measures generate different decision boundaries. Hence depending on the shapes of clusters within the data, classification accuracies will change with the selection of different distance measures. Moreover, clustering results in case of un-supervised learning will improve by selecting a distance metric appropriately. In the literature of clustering, lot of papers can be found on selection and adaptation of an appropriate distance measure [38-41].



(a) (b)

(c) (d)

(e) (f)

Figure 2. Plot of neighborhood of 1.0 from a point (1.5, 1.5) for different values of p in the Minkowski distance: (a) $p = 0.5$, (b) $p = 1$, (c) $p = 2$, (d) $p = 3$, (e) $p = 4$ and (f) $p \to \infty$

Mahalanobis distance measure [42] is also worth mentioning which consider the covariance of data and resolve the problem of correlation among different dimensions of the feature space. Mahalanobis distance is defined as,

$$d_{mahal}(X, Y) = \sqrt{(X - Y)^T \, \Sigma^{-1}(X - Y)} \qquad (5)$$

where Σ is the covariance matrix. A good survey can be found in [43].

3.2. **Binary valued attributes.** Biomedical datasets often contain binary valued attributes like gender (male/female) and presence or absence of an attribute (yes/no answers), etc. For n-dimensional binary data points, a contingency table [44] is defined as Table 1.

For binary type of data, hamming distance [45] is the most popular distance measure used to find out the dissimilarity between two binary strings. Hamming distance is defined as follows,

$$d_{ham}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{ll} 1 & x_i \neq y_i \\ 0 & \text{otherwise} \end{array}\right. = \frac{b+c}{n} \tag{6}$$

In real life, some matches are more important than others. But hamming distance measure gives equal weight to all matches. Kolbe et al. [46] have proposed Granularity-Enhanced Hamming (GEH) distance as defined below,

$$d_{ham}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{ll} 1 & x_i \neq y_i \\ 1-f(x_i) & \text{otherwise} \end{array}\right. \tag{7}$$

where $f(x_i)$ is the relative frequency of occurrence of x_i in the data. Hence, more occurring elements will contribute less in the distance calculation.

Some distance measures based on the contingency table shown in Table 2. A complete survey can be found in [47,48].

Table 1. Contingency table for two binary data points

	$Y=1$	$Y=0$
$X=1$	a	b
$X=0$	c	d

Table 2. Distance measures for binary attributes

	Distance Measure	Type
$d_{jaccard}$	$\frac{b+c}{a+b+c}$	Jaccard [49]
d_{dice}	$\frac{b+c}{2a+b+c}$	Dice [50]
$d_{s\&s}$	$\frac{b+c}{2a+2d+b+c}$	Sokal and Sneath [51]
$d_{hamman1}, d_{hamman2}$	$\frac{2(b+c)}{a+b+c+d}, \frac{2(b+c)}{a+b+c}$	Hamman I and II [52]
$d_{r\&t1}, d_{r\&t2}$	$\frac{2(b+c)}{a+d+2(b+c)}, \frac{2(b+c)}{a+2(b+c)}$	Rogers and Tanimoto I and II [53]

3.3. **Categorical type attributes.** Concept of similarity measures for binary type of attribute can easily be extended to two-valued categorical type of attributes also. It may include hamming distance, GEH distance and many more listed in the above section and in the literature [47]. For multi-valued categorical attributes type, the same idea of distance measures for binary attributes is used as follows [54].

Let X and Y are two categorical variables having q categories. Contingency table of X and Y can be formed as Table 3.

Table 3. Contingency table for two data points have categorical attributes of q levels

X/Y	0	1	2	...	q
0	m_{00}	m_{01}	m_{02}	...	m_{0q}
1	m_{10}	m_{11}	m_{12}	...	m_{1q}
2	m_{20}	m_{21}	m_{22}	...	m_{2q}
...	...	...	...	...	
q	m_{q0}	m_{q1}	m_{q2}	...	m_{qq}

The number of matches on diagonal entries are summed as $m^+ = \sum_{i=0}^{q} m_{ii}$ and number of matches on off-diagonal entries are summed as $m^- = \sum_{i=0}^{q} \sum_{\substack{j=0 \\ j \neq i}}^{q} m_{ij}$. Some of the distance measures for q-category variables are defined in Table 4 and as provided in [54].

Table 4. Distance measures for categorical attributes of q levels

	Distance Measure	Type
d^p_{hamm}	$\frac{m^-}{m^+ + m^-}$	Hamming
$d^p_{jaccard}$	$\frac{m^-}{m^+ - m_{00} + m^-}$	Jaccard [49]
d^p_{dice}	$\frac{m^-}{2(m^+ - m_{00}) + m^-}$	Dice [50]
$d^p_{s\&s}$	$\frac{m^-}{2m^+ + m^-}$	Sokal and Sneath [51]
$d^p_{r\&t1}$	$\frac{2m^-}{m^+ + 2m^-}$	Rogers and Tanimoto I [53]

Value Difference Metric (VDM) was proposed by [55] and also reported in [56] is also an appropriate distance measure for categorical type of attributes. VDM is defined as follows,

$$VDM(X,Y) = \sqrt{\sum_{i=1}^{n} vdm_i(x_i, y_i)} \tag{8}$$

where

$$vdm_i(x,y) = \sum_{c=1}^{C} \left| \frac{N_{i,x,c}}{N_{i,x}} - \frac{N_{i,y,c}}{N_{i,y}} \right|^s = \sum_{c=1}^{C} |P_{i,x,c} - P_{i,y,c}|^s \tag{9}$$

here C is the total number of classes present in the data set. $N_{i,x}$ is the number of data points having value of x for ith attribute and $N_{i,x,c}$ is the number of data points belonging to class c and having value of x for ith attribute. A constant s can take the value as 1 or 2.

3.4. **Ordinal type attributes.** In case of ordinal attributes, following distance measure can be used,

$$d_{ord}(x,y) = \frac{|x-y|}{\psi} \tag{10}$$

where ψ is the range of ordinal attribute.

3.5. **Mixed type attributes.** Most of the biomedical datasets consist of mixed type of attributes including numeric, binary or categorical types. For mixed type of attributes, distance measure should incorporate distances of all types of attributes. Hence, it is defined as follows,

$$d_{mixed}(X,Y) = d_{numeric}(\eta_x,\eta_y) + d_{binary}(\kappa_x,\kappa_y) + d_{cat}(\varsigma_x,\varsigma_y) \tag{11}$$

$X = (x_1, x_2, \ldots, x_n)$ is a n-dimensional data point and η_x, κ_x, ς_x are the subset of X comprising of numeric, binary and categorical type of attributes. Sum of the dimensions of η_x, κ_x, ς_x is n. The distance measures $d_{numeric}(\eta_x,\eta_y)$, $d_{binary}(\kappa_x,\kappa_y)$, $d_{cat}(\varsigma_x,\varsigma_y)$ should be scaled on a uniform scale, for example $[0,1]$ and it can be chosen from above mentioned distance measures for numeric, binary or categorical type of attributes.

4. **Results and Discussion.** In this section, similarity-dissimilarity plots are described and explained for four cases, namely, when data set is comprised of numeric type of attributes only, when data set is comprised of binary type of attributes only, when data set contains categorical type of attributes only and when dataset has all three types of attributes.

4.1. **Numeric attributes only.** In this sub-section, three different cancer databases are used to illustrate the effectiveness of similarity-dissimilarity plot.

Acute leukemia dataset [57] is based on the analysis of bone marrow samples of adult patients. Whole dataset consists of total of 72 leukemia patients out of which 25 suffer from acute myeloid leukemia (AML) and 47 from acute lymphoblastic leukemia (ALL). Number of attributes (number of gene expression levels) are 7129. Minkowski distance is used in the similarity-dissimilarity plot for various values of p. PAS is calculated for k nearest neighbor ($k = 1, 2, 3, \ldots, 8$) and values of p ranging from 1 to 8.

Table 5 shows the PAS values in percentage. It can be observed from the table that PAS is changing by changing the value of p in the calculation of Minkowski distance. Similarly, PAS is also changing as number of nearest neighbors (NN) is increased to calculate similarity and dissimilarity distances. Highest value of PAS is for $k = 1$ and $p = 1$ and 2. However, only one nearest neighbor is not good for calculation of similarity and dissimilarity distances to get a proper idea of clustering of the classes.

Table 5. PAS values for acute leukemia dataset

k/p	1	2	3	4	5	6	7	8
1	88.89	88.89	87.50	86.11	86.11	84.72	86.11	86.11
2	87.50	86.11	84.72	84.72	84.72	81.94	83.33	83.33
3	86.11	87.50	83.33	83.33	81.94	79.17	80.56	80.56
4	80.56	84.72	79.17	79.17	80.56	80.56	80.56	79.17
5	79.17	81.94	81.94	80.56	77.78	77.78	77.78	76.39
6	79.17	80.56	81.94	81.94	79.17	79.17	77.78	77.78
7	77.78	81.94	80.56	81.94	79.17	79.17	77.78	77.78
8	77.78	80.56	80.56	80.56	80.56	79.17	77.78	77.78

Figure 3 shows similarity-dissimilarity plot for $k = 3$ and $p = 2$ (PAS = 87.5%). For fixed value of $p = 2$, PAS does not change much as NN is increased from 1 to 8. It shows that clusters of the classes are dense enough to give PAS equals to 81% in case of NN = 8. Data points of class ALL are mostly above the similarity-dissimilarity line contributing in high value of PAS. Out of 25 data points of AML, 10 are below the line and may be confused with the majority class (ALL) and are shown with the color of ALL class. Three data points (2 of ALL and 1 of AML) are far away from the rest of the data points. By looking at the similarity-dissimilarity plot, it can be predicted that true classification of ALL will be higher than AML class. Li et al. [58] reported accuracy of different classifiers for Acute Leukemia dataset ranging from 80% to 95%. It validates the predicted accuracy by using similarity-dissimilarity plot.

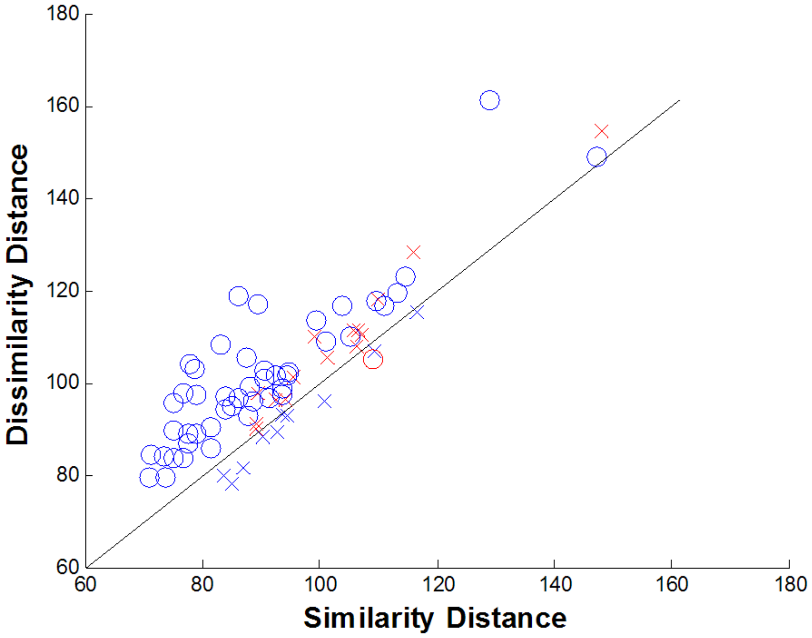


Figure 3. Similarity-dissimilarity plot of acute leukemia database for $p = 2$ and $k = 3$. (O is ALL and X is AML).

Colon adenocarcinoma cancer dataset [59] consists of 2000 gene expressions of 62 subjects including 40 tumor (cancerous) and 22 normal (non-cancerous) colon tissue samples. Similarity and dissimilarity distances are calculated for nearest neighbors ranging from 1 to 8 and using minkowski distance measure with the value of p varying from 1 to 8. PAS of Colon database is tabulated in Table 6.

Table 6. PAS values for colon adenocarcinoma cancer dataset

k/p	1	2	3	4	5	6	7	8
1	83.87	79.03	79.03	80.65	75.81	74.19	72.58	69.36
2	82.26	82.26	82.26	80.65	77.42	74.19	72.58	70.97
3	74.19	75.81	75.81	77.42	72.58	74.19	72.58	70.97
4	79.03	72.58	72.58	70.97	67.74	67.74	64.52	62.90
5	70.97	70.97	69.36	64.52	64.52	64.52	61.29	61.29
6	72.58	67.74	62.90	61.29	61.29	61.29	61.29	61.29
7	69.36	64.52	62.90	61.29	61.29	61.29	61.29	61.29
8	64.52	66.13	59.68	61.29	61.29	61.29	61.29	61.29

Maximum value of PAS (83%) is achieved with $p = 1$ and $k = 1$. For a fixed value of p, PAS decreases rapidly with the increase of nearest neighbors to calculate the similarity and dissimilarity distances. It shows the sparseness of the database and data points are well spread in the feature space.

Figure 4 shows similarity-dissimilarity plot for the colon cancer database for nearest neighbors of 2 and value of p equals to 1 (city block distance). PAS value for both city block distance ($p = 1$) measure and Euclidean distance measure ($p = 2$) is same. Huynh et al. [60] have reported the classification accuracy of 64% to 83.6% on the testing dataset of Colon adenocarcinoma datasets using various classifiers. It support our predicted accuracy calculated from PAS value for $p = 1$ and $k = 1$.

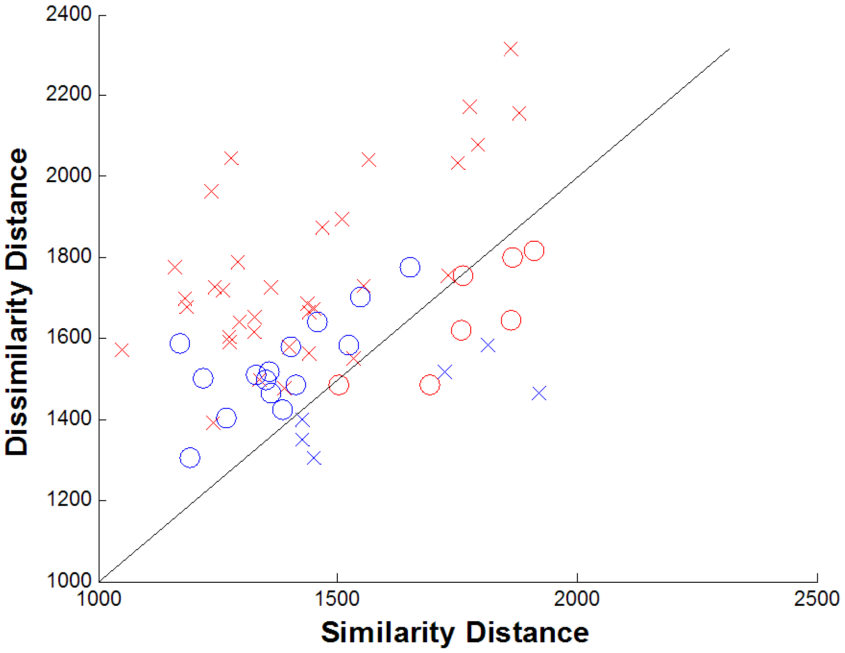


FIGURE 4. Similarity-dissimilarity plot of colon cancer database for $p = 1$ and $k = 2$. (O is Normal and X is Cancerous), PAS = 82.26%.

High grade glioma dataset [61] consists of 50 samples described by 12625 gene expressions. So, it is very high dimensional feature space and task is to classify the glioblastomas and the anaplastic oligodendrogliomas. PAS values for different nearest neighbors and values of p in the minkowski distance measure are tabulated in Table 7.

TABLE 7. PAS values for high grade glioma dataset

k/p	1	2	3	4	5	6	7	8
1	70.0	78.0	80.0	76.0	78.0	78.0	76.0	78.0
2	74.0	76.0	76.0	76.0	74.0	76.0	76.0	78.0
3	74.0	76.0	74.0	78.0	76.0	76.0	76.0	78.0
4	82.0	78.0	74.0	78.0	74.0	74.0	72.0	74.0
5	74.0	76.0	78.0	74.0	68.0	70.0	76.0	76.0
6	74.0	68.0	70.0	72.0	70.0	68.0	66.0	72.0
7	72.0	72.0	68.0	70.0	68.0	68.0	68.0	66.0
8	66.0	72.0	70.0	68.0	66.0	68.0	66.0	68.0

Highest value of PAS equals to 82% is observed with city block distance ($p = 1$) and nearest neighbors k equals to 4. Similarity-dissimilarity plot is shown in Figure 5 for k and p equal to 4 and 1, respectively.

Although 82% of the data points are above the similarity-dissimilarity line, all the data points are very near to the line showing closeness of both classes. One data point of oligodendrogliomas is far away from rest of the data points which can be checked for the outlier. Li et al. [58] have reported accuracies for glioma dataset in the range of 65%

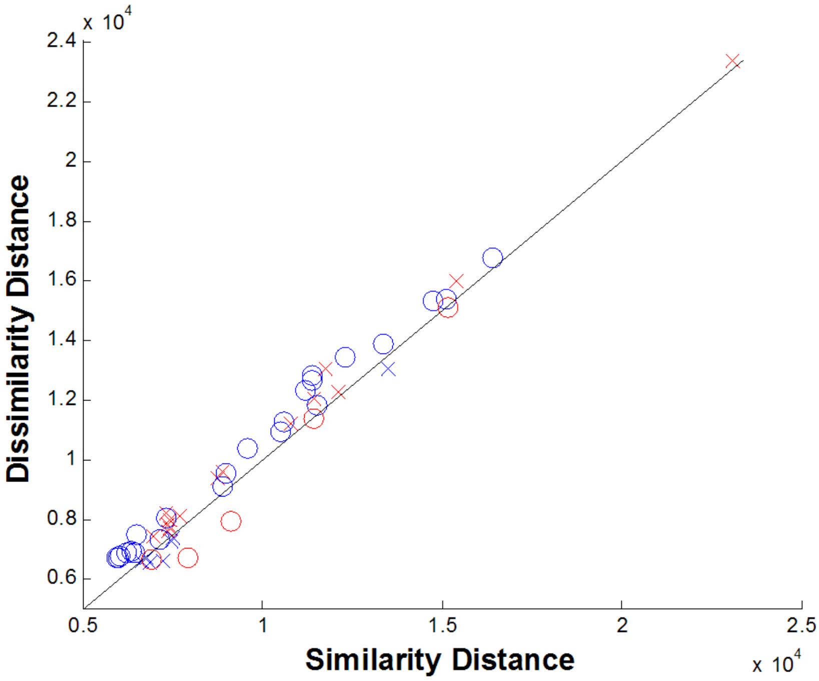


Figure 5. Similarity-dissimilarity plot of high grade glioma database for $p = 1$ and $k = 4$. (O is glioblastomas and X is anaplastic oligodendrogliomas), PAS = 82%.

to 76% for different classifiers. The reported results support the observations made by using similarity-dissimilarity plot (Although PAS values is about 82% but most of the data point are very close to similarity-dissimilarity line).

Special care must be taken for the datasets having data imbalance. If minority class is also sparsely located in the feature space and scattered near the majority class, calculation of similarity and dissimilarity distances may lead to wrong placement of data points on the similarity-dissimilarity plot. There are many techniques reported in the literature like SMOTE [62] and others to solve the data imbalance problem first.

4.2. **Binary attributes only.** In this sub-section, similarity-dissimilarity plot is used to study the effect of various types of distance measures for binary categorical attributes. A list of 8 distance measures is listed in the Table 2.

Two real life databases are used to study the effect of distance measures on similarity-dissimilarity. Spect Heart database [66] consists of cardiac Single Proton Emission Computed Tomography (SPECT) images and the feature space consists of 22 binary types of attributes. There are 267 instances out of which 55 are normal and 212 are having heart disease. Table 8 shows PAS values for different type of distance measure and different values of nearest neighbors (k).

Distance measure proposed by [69] performs better than other distance measures and PAS value is 90.5%. CLIP3 algorithm [68] generated rules that have given 84% classification accuracy. Another important observation can be made from Table 8 that if number of nearest neighbors is increased to 7, PAS value increases considerably. It shows that for a particular data point, 7 nearest neighbors of same class is nearer to the data point as compared to the 7 nearest neighbors from other classes.

Figure 6 shows Similarity-dissimilarity plot of Spect Heart Database for distance measure $d_{r\&t1}$ and number of nearest neighbors equal to 7. Almost 90% data points are above

Table 8. PAS values for spect heart dataset

Distance Type	Number of Nearest Neighbors (k)							
	1	2	3	4	5	6	7	8
d_{ham}	81.65	83.15	81.65	82.02	81.65	82.40	89.51	89.14
$d_{s\&s}$	81.65	83.15	81.65	82.02	81.65	82.40	89.51	88.76
$d_{r\&t1}$	81.65	84.27	83.52	83.90	83.52	83.90	90.64	90.26
$d_{hamman1}$	81.65	83.15	81.65	82.02	81.65	82.40	89.51	89.14
$d_{jaccard}$	81.65	84.27	82.40	83.52	82.77	82.40	88.02	86.89
d_{dice}	81.65	83.52	82.02	82.77	82.02	80.15	86.52	86.14
$d_{r\&t2}$	81.65	84.64	84.27	84.27	83.90	83.52	89.89	89.14
$d_{hamman2}$	81.65	84.27	82.40	83.52	82.77	82.40	88.02	86.89

the similarity-dissimilarity line (PAS = 90.5%). Almost all positive (disease is present) examples are above the similarity-dissimilarity line. Figure 7 shows Similarity-dissimilarity plot of Spect Heart Database for distance measure d_{dice} and number of nearest neighbors equal to 7 with PAS value of 86%. In this figure, it can be observed that by using distance measure d_{dice}, more negative examples (absence of disease) will be confused with positive example. Asadi et al. [63] reported classification accuracies of various classifiers for Spect Heart database and the classification accuracies are ranging from 70% to 92% supporting PAS predicted best possible accuracy of 90.5%.

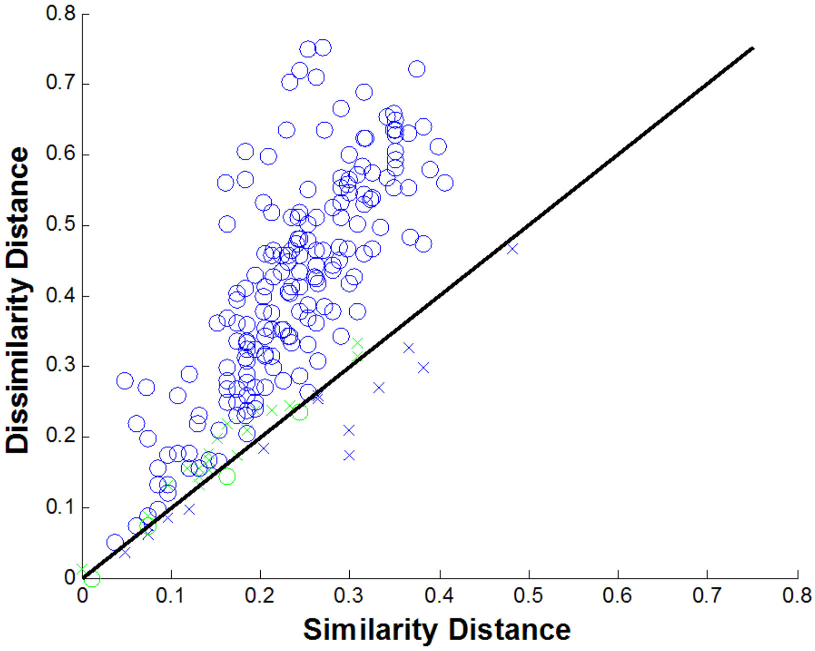


Figure 6. Similarity-dissimilarity plot of spect heart database for distance measure $d_{r\&t1}$ and $k = 7$. (O is presence of disease and X is absence of disease), PAS = 90.6%.

Acute inflammation dataset [66,69] comprises of 120 examples described by five binary attributes and one numeric attribute. For the analysis, numeric attribute is dropped and only binary attributes are considered. The dataset is related to the diagnosis of two diseases of urinary system; diagnosis of the acute inflammations of urinary bladder and acute nephritises. In the dataset, decision is presence or absence of the any of the above mentioned disease. The feature set for Acute Inflammation dataset is of very high quality

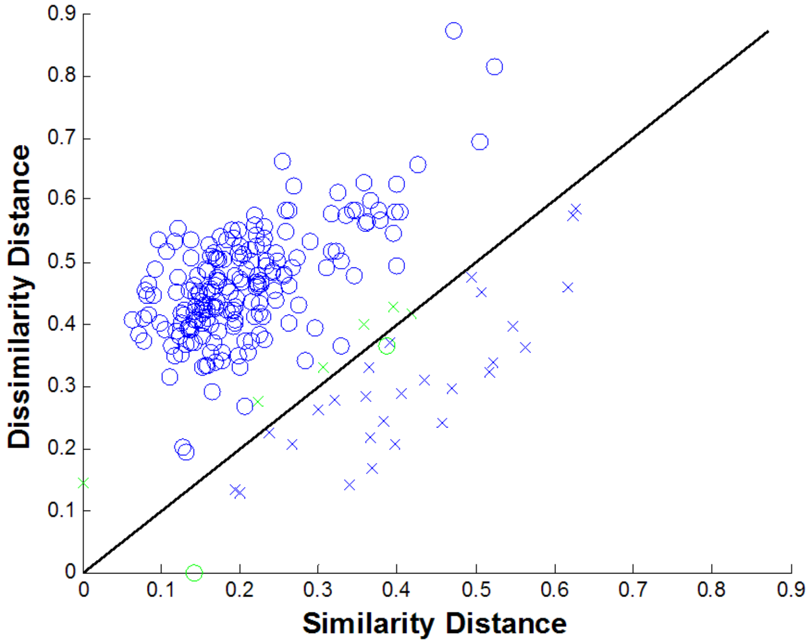


Figure 7. Similarity-dissimilarity plot of spect heart database for distance measure d_{dice} and NN = 7. (O is presence of disease and X is absence of disease), PAS = 86%.

and 100% classification accuracy is possible for all type of distance measures as evident from Table 9 and Table 10.

Table 9. PAS values for acute inflammation (acute inflammations of urinary bladder)

Distance Type	Number of Nearest Neighbors (k)							
	1	2	3	4	5	15	25	35
d_{ham}	100	100	100	100	100	100	100	100
$d_{s\&s}$	100	100	100	100	100	100	100	100
$d_{r\&t1}$	100	100	100	100	100	100	100	100
$d_{hamman1}$	100	100	100	100	100	100	100	100
$d_{jaccard}$	100	100	100	100	100	100	100	100
d_{dice}	100	100	100	100	100	100	91.7	91.7
$d_{r\&t2}$	100	100	100	100	100	100	100	100
$d_{hamman2}$	100	100	100	100	100	100	100	100

Figure 8 shows similarity-dissimilarity plot for the Absence or presence of acute inflammations of urinary bladder for number of nearest neighbors equals to 5 and distance measure of $d_{s\&s}$. There are 61 examples of absence of the disease and 59 examples of presence of disease. It can be observed from the figure that all data points are mapped on two points on similarity-dissimilarity plot. It shows that there are at least five nearest neighbors having exactly same similarity distance of zero. To investigate it further, number of nearest neighbors is increased from 5 to 30 and similarity-dissimilarity plot is shown in Figure 9. Now the data points are somewhat spread but still above the similarity-dissimilarity line.

Table 10. PAS values for acute inflammation (acute nephritises)

Distance Type	Number of Nearest Neighbors (k)							
	1	2	3	4	5	15	25	35
d_{ham}	100	100	100	100	100	100	100	100
$d_{s\&s}$	100	100	100	100	100	100	100	100
$d_{r\&t1}$	100	100	100	100	100	100	100	100
$d_{hamman1}$	100	100	100	100	100	100	100	100
$d_{jaccard}$	100	100	100	100	100	100	100	100
d_{dice}	100	100	100	100	100	100	100	100
$d_{r\&t2}$	100	100	100	100	100	100	100	100
$d_{hamman2}$	100	100	100	100	100	100	100	100

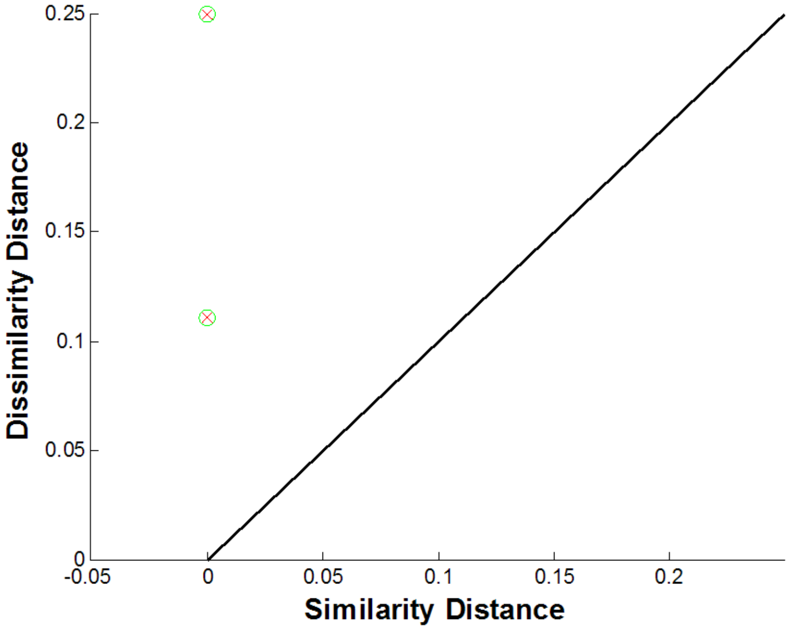


Figure 8. Similarity-dissimilarity plot of acute inflammation data set for distance measure $d_{s\&s}$ and $k = 5$. (O is presence of disease and X is absence of disease), PAS = 100%.

Table 10 shows PAS values for absence or presence of acute nephritises. Again, both classes are well separated in the feature space as evident from PAS value which is equals to 100% for all distance measures and even for large number of nearest neighbors.

Figure 10 shows that all the data points are well above the similarity-dissimilarity line and feature set can give excellent classification accuracy.

4.3. **Categorical attributes of q levels.** In dermatology, differential diagnosis of erythemato-squamous diseases is difficult with little difference among the clinical features of erythema and scaling. In the dermatology dataset [66], there are 12 attributes related to clinical features of erythema and scaling. Out of twelve attributes, ten attributes takes the value of 0, 1, 2 and 3 where as one attribute takes the value of 0 or 1. Age is not considered in our analysis. Similarly, there are 22 attributes which take the value of 0, 1, 2 and 3 from histo-pathological features of biopsy. Class distribution is given in Table 11.

Hamming distance is used for the dermatology data set to calculate PAS values. Table 12 gives the value of PAS for various number of nearest neighbors. Value of PAS is not changing much with the increase in the number of nearest neighbors.

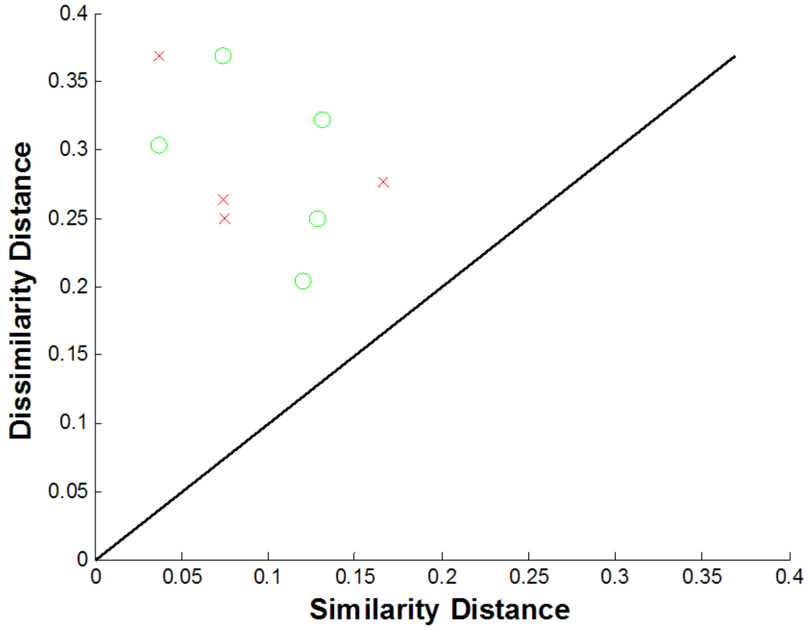


FIGURE 9. Similarity-dissimilarity plot of acute inflammation data set for distance measure $d_{s\&s}$ and $k = 30$. (O is presence of disease and X is absence of disease), PAS = 100%.

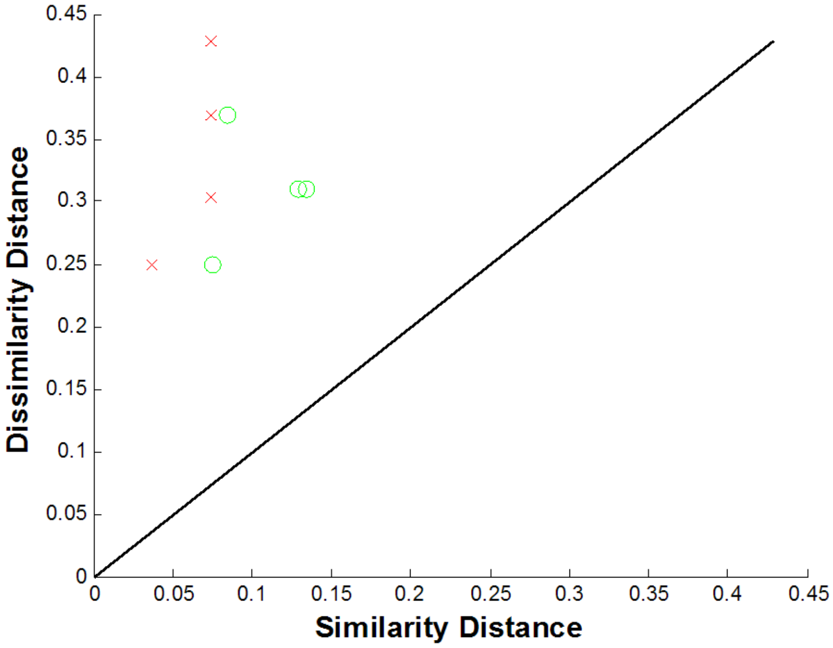


FIGURE 10. Similarity-dissimilarity plot of nephritises data set for distance measure $d_{s\&s}$ and $k = 30$. (O is presence of disease and X is absence of disease), PAS = 100%.

Figure 11 shows similarity-dissimilarity plot for dermatology data set using number of nearest neighbors equal to 5 and hamming distance as the distance measure. Most of the data points are above the similarity-dissimilarity line.

Some of the data points of Seboreic dermatitis are below the similarity-dissimilarity line and it may be confused with Pityriasis. All data points below similarity-dissimilarity

TABLE 11. Distribution of six disease classes in the dermatology dataset

Disease type	Number of examples
Psoriasis	112
Seboreic dermatitis	61
Lichen planus	72
Pityriasis rosea	49
Cronic dermatitis	52
Pityriasis rubra pilaris	20

TABLE 12. PAS values of dermatology dataset for different values of nearest neighbors

	Number of Nearest Neighbors (k)							
	1	2	3	4	5	6	7	8
Dermatology	92.35	93.99	93.99	94.54	95.08	95.08	95.08	95.08

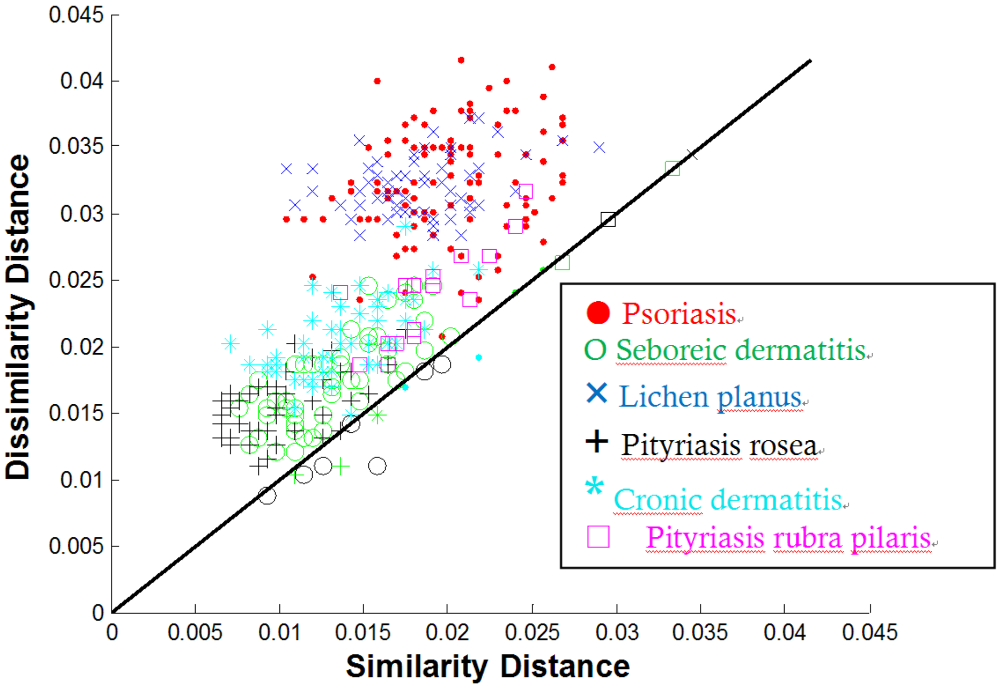


FIGURE 11. Similarity-dissimilarity plot of dermatology data set for distance measure d_{ham} and $k = 5$, PAS = 95%

line carries the shape of the marker associated with their classes and color of the marker is of the class to whom they may confuse.

Maximum predicted accuracy by PAS value is about 95%. Song et al. [64] reported best classification accuracy in the range of 90% ± 5% using ASSEMBLE classifier. Hasan et al. [65] used various architectures of neural network to obtain classification accuracy from 89% to 92%. Agreement of the reported results with predicted accuracy by PAS shows effectiveness of proposed method.

4.4. **Mixed attributes.** Statlog_heart dataset [61] is used to demonstrate the similarity-dissimilarity plot for mixed type of attributes. In this dataset, there are 270 instances having 13 attributes. Distribution of type of attributes is given in Table 13. In Table 14, PAS values for different number of nearest neighbors are given. For numeric attributes, minkowski distance is used for three values of p, i.e., 1, 2 and 3 as given in Table 14. It can be observed from Table 14 that city block distance is better than Euclidean distance.

Table 13. Type of attributes for statlog heart dataset

Attribute Type	Number of Attributes
Numeric	6
Ordinal	1
Binary	3
Nominal	3

Table 14. PAS values of statlog heart dataset for different values of nearest neighbors

	Number of Nearest Neighbors (k)							
	1	2	3	4	5	6	7	8
Numeric ($p = 1$)	77.04	80.37	80.74	81.48	83.70	84.44	84.07	84.07
Numeric ($p = 2$)	77.78	78.15	80.37	80.74	82.22	82.22	82.22	82.22
Numeric ($p = 3$)	77.778	78.15	79.63	80.74	81.11	81.48	81.11	81.11

Figure 12 shows similarity-dissimilarity plot for statlog heart dataset using number of nearest neighbors equals to 5, Euclidean distance as distance measure for numeric attributes and hamming distance for nominal and binary attributes. For ordinal attribute, distance measure given in Equation (10) is used. Overall distance measure given by Equation (11) is used to compare two data points of the dataset. It can be observed from the figure that some of the data points are near to the opposite class (below similarity-dissimilarity line).

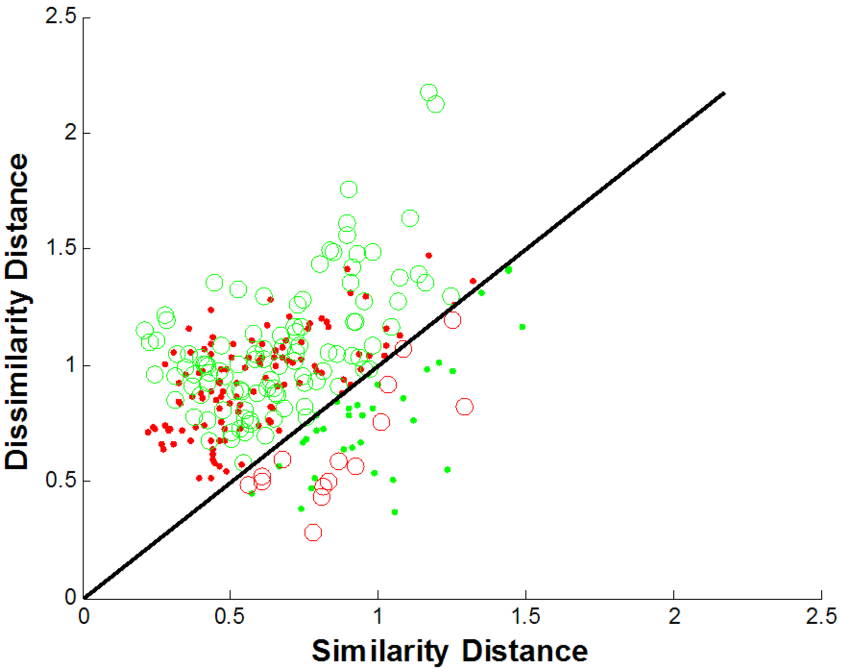


Figure 12. Similarity-dissimilarity plot of statlog heart database for NN = 5. (O is Absent of heart disease and X is presence of heart disease), PAS = 82%.

It can be seen from Table 14, maximum classification accuracy of 85% is predicted by using PAS. Kumar et al. [67] reported average classification accuracy of 85% for Statlog heart dataset using various settings of support vector machine which is in agreement of the predicted accuracy.

To summarize all of the above results mentioned in sections IV-A to IV-D, following observations can be made. For the classification problem in a feature space containing numerical attributes only, Euclidean distance with number of nearest neighbors equals to 3 can be used to get an overall idea about the discrimination ability of the feature space in different classes. If distance based classifiers are used to classify the feature space into different classes, discrimination ability of feature space can be explored for different type of distance metrics to get an idea about the suitability of a particular distance metric before applying the classifier. Compactness of clusters within class can also be studied by calculating PAS value for different number of nearest neighbors to calculate similarity and dissimilarity distances.

Similarly for the feature space containing attributes having binary or multi-level values, hamming distance metric can be used as distance metric. Discrimination characteristics of the feature space can further be studied by using different distance metrics to propose proper distance metric to be used in the distance based classifiers. For feature space having mixed type of attributes, general guideline for plotting the similarity-dissimilarity plot is to use Euclidean distance for numeric type of attributes and hamming distance can be used for categorical type of attributes.

Furthermore, above results also demonstrated the effectiveness in finding the data points in the feature space which may confuse with other classes and identification of the class to whom they may confuse. This information is very important for pattern classification as it can provide some guidelines to the researcher to study the effectiveness of a particular feature set in identifying separable classes in the feature space.

5. Conclusions.

In this paper, the concept of similarity-dissimilarity plot is generalized for different types of attributes including numeric, binary, categorical, ordinal and mixed. It is shown by various examples of real life biomedical data sets that similarity-dissimilarity plot is very effective in studying the features discrimination quality by projecting the high dimensional feature space to two dimensional similarity-dissimilarity plot in the context of pattern classification. Furthermore, it is shown that selection of an appropriate distance measure in distance based classifier can play a significant role. Effect of number of nearest neighbors on percentage of data points above similarity-dissimilarity line (PAS) is studied to get an idea about the compactness of class clusters. Sensitivity of PAS for different distance measures is also studied. It is concluded that similarity-dissimilarity plot is very effective in studying the quality of features not only for numeric but also for other attributes like binary, categorical and ordinal.

REFERENCES

[1] C.-Y. Yeh, C.-W. Huang and S.-J. Lee, Multi-kernel support vector clustering for multi-class classification, *International Journal of Innovative Computing, Information and Control*, vol.6, no.5, pp.2245-2262, 2010.

[2] B. Chen, L. Ma and J. Hu, An improved multi-label classification method based on SVM with delicate decision boundary, *International Journal of Innovative Computing, Information and Control*, vol.6, no.4, pp.1605-1614, 2010.

[3] I. T. Jolliffe, *Principal Component Analysis*, Springer Verlag, New York, 1986.

[4] D. Cook, A. Buja, J. Cabrera and H. Hurley, Grand tour and projection pursuit, *Journal of Computational and Graphical Statistics*, vol.4, no.3, pp.155-172, 1995.

[5] T. Kohonen, *Self-Organizing Maps*, 3rd Extended Edition, Springer Series in Information Sciences, Springer, New York, NY, USA, 2001.

[6] T. Samatsu, K. Tachikawa and Y. Shi, Visualization for fuzzy retrieval using self-organizing maps, *ICIC Express Letters*, vol.3, no.4(B), pp.1345-1350, 2009.

[7] Y. Li and K. Horio, Visualization and analysis of mental states based on photoplethysmogram, *ICIC Express Letters*, vol.4, no.3(B), pp.923-928, 2010.

[8] M. Murata, T. Shirado, K. Torisawa, M. Iwatate, K. Ichii, Q. Ma and T. Kanamaru, Extraction and visualization of numerical and named entity information from a very large number of documents using natural language processing, *International Journal of Innovative Computing, Information and Control*, vol.6, no.3(B), pp.1549-1568, 2010.

[9] S. Ishimitsu, K. Sakamoto, T. Yoshimi, Y. Fujimoto and K. Kawasaki, Study on the visualization of the impression of button sounds, *International Journal of Innovative Computing, Information and Control*, vol.5, no.11(B), pp.4189-4203, 2009.

[10] D. F. Andrews, Plot of high dimensional data, *Biometrics*, vol.28, no.1, pp.125-136, 1972.

[11] J. M. Chambers, W. S. Cleveland, B. Kleiner and P. A. Tukey, *Graphical Methods for Data Analysis*, Chapman and Hall, New York, 1983.

[12] J. J. V. Wijk and R. V. Liere, HyperSlice – Visualization of scalar functions of many variables, *Proc. of IEEE Visualization*, Los Alamitos, CA, pp.119-125, 1993.

[13] B. Alpern and L. Carter, Hyperbox, *Proc. of IEEE Visualization*, pp.133-139, 1991.

[14] R. Spence, L. Tweedie, H. Dawkes and H. Su, Visualisation for functional design, *Proc. of IEEE Visualization*, pp.4-10, 1995.

[15] A. Inselberg, The plane with parallel coordinates, *The Visual Computer*, pp.69-92, 1985.

[16] B. Inselberg and B. Dimsdale, Parallel coordinates: A tool for visualization high dimensional geometry, *Proc. of IEEE Visualization*, pp.361-378, 1990.

[17] H. Zhou, X. Yuan, H. Qu, W. Cui and B. Chen, Visual clustering in parallel coordinates, *IEEE-VGTC Symposium on Visualization*, vol.27, no.3, 2008.

[18] W. Peng, M. O. Ward and E. A. Rundensteiner, Cluster reduction in multi-dimensional data visualization using dimension reordering, *Proc. of IEEE Symposium on Information Visualization*, pp.89-96, 2004.

[19] J. Johansson, P. Ljung, M. Jern and M. Cooper, Revealing structures within clustered parallel coordinates display, *Proc. of IEEE Symposium on Information Visualization*, pp.125-132, 2005.

[20] H. Siirtola, Direct manipulation of parallel coordinates, *Proc. of the 4th IEEE International Conference on Information Visualization*, pp.373-378, 2000.

[21] H. Chernoff, The use of faces to represent points in k-dimensional space graphically, *Journal of the American Statistical Association*, vol.68, pp.361-368, 1973.

[22] B. Kleiner and J. A. Hartigan, Representing points in many dimensions by trees and castles, *Journal of the American Statistical Association*, vol.76, no.374, pp.260-269, 1981.

[23] J. H. Siegel, E. J. Farrel, R. M. Goldwyn and H. P. Friedman, The surgical implication of physiologic patterns in myocardial infarction shock, *Surgery*, vol.72, pp.126-141, 1972.

[24] J. Beddow, Shape coding of multidimensional data on a microcomputer display, *Proc. of the 1st IEEE Conference on Visualization*, San Francisco, pp.238-246, 1990.

[25] A. Buja, D. Cook, D. Asimov and C. Hurley, *Theory and Computational Methods for Dynamic Projections in High-dimensional Data Visualization*, 1999.

[26] F. Murtagh, A survey of recent advances in hierarchical clustering algorithms, *The Computer Journal*, vol.26, no.4, pp.354-359, 1983.

[27] E. Boudaillier and G. Hebrial, Interactive interpretation of hierarchical clustering, *Intelligent Data Analysis*, vol.2, no.3, pp.229-244, 1998.

[28] P. Willet, Recent trends in hierarchical document clustering: A critical review, *Information Processing and Management*, vol.24, no.5, pp.577-597, 1988.

[29] Y.-H. Fua, M. O. Ward and E. A. Rundensteiner, Hierarchical parallel coordinates for exploration of large datasets, *Proc. of the 10th IEEE Conference on Visualization*, pp.43-50, 1999.

[30] J. Yang, M. O. Ward, E. A. Rundensteiner and S. Huang, Visual hierarchical dimension reduction for exploration of high dimensional datasets, *Proc. of the Joint Eurographics/IEEE TVCG Symposium on Data Visualization*, pp.19-28, 2003.

[31] C. Brunsdon, A. S. Fotheringham and M. E. Charlton, An investigation of methods for visualising highly multivariate datasets, In *Case Studies of Visualization in Social Sciences*, D. Unwin and P. Fisher (eds.), 1998.

[32] G. Leban, I. Bratko, U. Petrovic, T. Curk and B. Zupan, Vizrank: Finding informative data projections in functional genomics by machine learning, *Bioinformatics*, vol.21, no.3, pp.413-414, 2005.

[33] J. F. McCarthy, K. A. Marx, P. E. Hoffman, A. G. Gee, P. O'Neil, M. L. Ujwal and J. Hotchkiss, Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis and management, *Annals of New York Academy of Sciences*, vol.1020, no.1, pp.239-262, 2004.

[34] B. Zupan, FreeViz – An intelligent multivariate visualization approach to explorative analysis of biomedical data, *Journal of Biomedical Informatics*, vol.40, no.6, pp.661-671, 2007.

[35] J. Sharko, G. Grinstein and K. A. Marx, Vectorized radviz and its application to multiple cluster datasets, *IEEE Transactions on Visualization and Computer Graphics*, vol.1, no.6, pp.1444-1451, 2008.

[36] M. Arif, Similarity-dissimilarity plot for visualization of high dimensional data in biomedical pattern classification, *Journal of Medical Systems*, 2010.

[37] C. D. Cantrell, *Modern Mathematical Methods for Physicists and Engineers*, Cambridge University Press, New York, NY, USA, 2000.

[38] E. P. Xing, A. Y. Ng, M. I. Jordan and S. Russell, Distance metric learning, with application to clustering with side-information, in *Advances in Neural Information Processing Systems 15*, S. Becker and S. Thrun (eds.), MIT Press, 2002.

[39] S. Xiang, F. Nie and C. Zhang, Learning a Mahalanobis distance metric for data clustering and classification, *Pattern Recognition*, vol.41, no.12, pp.3600-3612, 2008.

[40] J. Ye, Z. Zhao and H. Liu, Adaptive distance metric learning for clustering, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1-7, 2007.

[41] K. Q. Weinberger, F. Sha and L. K. Saul, Convex optimizations for distance metric learning and pattern classification, *IEEE Signal Processing Magazine*, vol.27, no.3, pp.146-158, 2010.

[42] P. C. Mahalanobis, On the generalised distance in statistics, *Proc. of the National Institute of Sciences of India*, vol.2, no.1, pp.49-55, 1936.

[43] M.-J. Lesot and M. Rifqi, Similarity measures for binary and numerical data: A survey, *Int. J. Knowledge Engineering and Soft Data Paradigms*, vol.1, no.1, pp.63-84, 2009.

[44] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2001.

[45] R. Hamming, Error-detecting and error-correcting codes, *Bell System Technical Journal*, vol.29, no.2, pp.147-160, 1950.

[46] D. Kolbe, Q. Zhu and S. Pramanik, On k-nearest neighbor searching in non-ordered discrete data spaces, *Proc. of the 23rd IEEE International Conference on Data Engineering*, pp.426-435, 2007.

[47] S.-H. Cha, C. C. Tappert and S. Yoon, Enhancing binary feature vector similarity measures, *Journal of Pattern Recognition Research*, vol.1, no.1, pp.63-77, 2006.

[48] S. S. Choi, S. H. Cha and C. Tappert, A survey of binary similarity and distance measures, *Journal on Systemics, Cybernetics and Informatics*, vol.8, no.1, pp.43-48, 2010.

[49] P. Jaccard, Nouvelles recherches sur la distribution florale, *Bulletin de la Societe Vaudoise de Science Naturelle*, vol.44, pp.223-270, 1908.

[50] L. R. Dice, Measures of the amount of ecologic association between species, *Ecology*, vol.26, no.3, pp.297-302, 1945.

[51] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, W. H. Freeman, San Francisco, CA, USA, 1973.

[52] V. Hamann, Merkmalbestand und verwandtschaft sbeziehungen der farinosae, *Ein Beitragzum System der Monokotyledonen*, Willdenowia, vol.2, pp.639-768, 1961.

[53] D. J. Rogers and T. T. Tanimoto, A computer program for classifying plants, *Science*, vol.132, pp.1115-1118, 1960.

[54] S. Selinski and K. Ickstadt, Similarity measures for clustering SNP data, *Technical Report, SFB 475*, Universität Dortmund, 2005.

[55] C. Stanfill and D. Waltz, Toward memory-based reasoning, *Communications of the ACM*, vol.29, no.12, pp.1213-1228, 1986.

[56] H. Wang, Nearest neighbors by neighborhood counting, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, no.6, pp.942-953, 2006.

[57] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, vol.286, 531-537, 1999.

[58] B. Li, C.-H. Zheng, D.-S. Huang, L. Zhang and K. Han, Gene expression data classification using locally linear discriminant embedding, *Computers in Biology and Medicine*, vol.40, pp.802-810, 2010.

[59] A. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack and A. J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA*, vol.96, pp.6745-6750, 1999.

[60] H. T. Huynh, J. Kim and Y. Won, Classification study on DNA microarray with feedforward neural network trained by singular value decomposition, *International journal of Bio-science and Biotechnology*, vol.1, no.1, pp.17-24, 2009.

[61] C. L. Nutt, D. R. Mani, R. A. Betensky, P. Tamayo, J. G. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M. E. McLaughlin, T. T. Batchelor, P. M. Black, A. von Deimling, S. L. Pomeroy, T. R. Golub and D. N. Louis, Gene expression-based classification of malignant gliomas correlates better with survival than histological classification, *Cancer Research*, vol.63, no.7, pp.1602-1607, 2003.

[62] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, vol.16, pp.321-357, 2002.

[63] R. Asadi, N. Mustapha, N. sulaiman and N. Shiri, New supervised multi layer feed forward neural network model to accelerate classification with high accuracy, *European Journal of Scientific Research*, vol.33, no.1, pp.163-178, 2009.

[64] E. Song, D. Huang, G. Ma and C. C. Hung, Semi-supervised multi-class adaboost by exploiting unlabeled data, *Expert Systems with Applications*, vol.38, pp.6720-6726, 2011.

[65] H. Hasan and H. Bal, Comparing performances of backpropagation and genetic algorithms in the data classification, *Expert Systems with Applications*, vol.38, pp.3703-3709, 2011.

[66] A. Frank and A. Asuncion, *UCI Machine Learning Repository*, School of Information and Computer Science, University of California, Irvine, http://archive.ics.uci.edu/ml, 2010.

[67] M. A. Kumar and M. Gopal, Least squares twin support vector machines for pattern classification, *Expert Systems with Applications*, vol.36, pp.7535-7543, 2009.

[68] K. J. Cios, D. K. Wedding and N. Liu, CLIP3: Cover learning using integer programming, *Kybernetes*, vol.26, no.4-5, pp.513-536, 1997.

[69] J. Czerniak and H. Zarzycki, Application of rough sets in the presumptive diagnosis of urinary system diseases, in *Artifical Inteligence and Security in Computing Systems*, J. Soldek and L. Drobiazgiewicz (eds.), Massachusetts, USA, Kluwer Academic Publishers, 2003.