A Novel Similarity Measure for Clustering Algorithm

Jamil Al Shaqsi, Wenjia Wang and VJ Rayward-Smith

School of Computing Sciences, University of East Anglia, UK

Abstract

This paper presents a new clustering similarity measure. The new similarity measure is defined particularly to reflect the degree of the relative change between data samples and to cope with both numerical and categorical variables. To evaluate the performance of the proposed similarity measure, we embedded it into the traditional k-means and tested it on 3 different benchmark datasets. We then compared the obtained results with 3 other similarity definitions with each of them is implemented in the similar k-means based clustering algorithm. The experimental results showed the strengths our proposed similarity measure over the compared similarity measures in terms of clustering accuracy.

1. Introduction

Similarity is a measure used in clustering analysis to calculate the similarity value between two data samples and then decide whether they should be assigned to the same cluster or not. Measuring the similarity between data samples plays an essential role in all clustering techniques and can determine their performance. This paper presents a novel similarity definition with an intention to measure the similarity of the relative changes between data samples. The details of this similarity measure will be described in Section 3 after reviewing the related work in Section 2. Section 4 gives a brief description of the modified k-means algorithm (mk-means). Section 5 presents the experiments and evaluation of the results, before drawing conclusions highlighting the fundamental issues and the future research in the final Section.

2. Related Work

There are many similarity measures that have been defined from various point of view to represent similarity degree between data points. This section reviews some common similarity definitions.

2.1 Distance-Based Similarity Measures

The fundamental concept of the distance-based similarity measure is that samples that are close to each other are considered as similar to that are far apart are considered as dissimilar.

Euclidean Distance. The Euclidean Distance is the distance between any two points in the coordinate plane that can be measured with a ruler. Equation 1 presents the definition of this measure as the square root of the sum of the squared differences between the values of samples drawn from the same feature. In clustering, this measure is commonly used to measure the distance between a sample i and a centroid k. It is usually effective, when a particular dataset has compact and well-separated; but less effective when clusters are nested [1].

\[ d(x_i, y_k) = \sqrt{\sum_{j=1}^{n} (x_{ij} - y_{kj})^2} \]  

where \( d \) is the abbreviation of distance, \( x \) and \( y \) represent two different samples, \( j \) is the samples index, \( k \) is the cluster index, and \( i \) is the features index.

Minkowski Metrics. In this measure, the distance between \( n \) samples is the \( p \)th root of the sum of the absolute differences to the \( p \)th power between the values for the samples. This measure is defined as:

\[ p(x_i, y_i) = \left( \sum_{j=1}^{n} |x_{ij} - y_{ij}|^p \right)^{1/p} \]  

where \( p \geq 0 \). Euclidean distance is a special case of Minkowski Metrics when \( p = 2 \) whereas in Manhattan Distance \( p = 1 \).

Mahalanobis Distance. It is one of the distance-based measures. It has an advantage over the Euclidean Distance as in calculating distances it takes into consideration the covariance among the features and applies a weight transformation for the features. Thus, with this measure, the problems of correlation and scale are no longer an issue. The definition of this measure is presented in Equation (3)

\[ d_n(x_i, y_i) = \left( x_i - y_i \right) \sum^{-1} \left( x_i - y_i \right)^T \]  

where \( d_n \) is the different weights that is assigned to features based on their variances and correlations, \( \sum \) is the sample covariance matrix.

Manhattan Distance (City-Block), also known as City-block distance, it is the distance that would be travelled between any two points in a city (down 2 blocks and up 4 blocks for a total of 6 blocks). In clustering, the distance between two samples by Manhattan is the sum of the absolute differences
between the values of their corresponding features. The Manhattan Distance is defined as:

$$\text{Block} (x, y) = \sum_{j=1}^{n} |x_j - y_j| \quad (4)$$

where $n$ is the number of features.

There are four most significant properties of distance-based similarity measure known as distance axioms. The four distance axioms are:

- Symmetry. $d(x, y) = d(y, x)$ for all points $x$ and $y$.
- Triangle Inequality. $d(x, y) + d(y, e) > d(x, e)$ for all points $x, y$, and $e$. In the Triangle Inequality, if $x$ and $y$ are similar, and $y$ and $e$ are similar, then $x$ and $e$ should also be similar.
- Minimality $d(x, y) > d(x, x)$ for all points $x$ and $y$. Thus, $s(x, y) = s(x, x)$ for all stimuli $x$ and $y$.
- Equal self-similarity. $d(x, y) = d(y, x)$ for all points $x$ and $y$. Thus, $s(x, x) = s(y, y)$ for all stimuli $x$ and $y$.

### 2.2 Feature-Based Similarity Measures

In this method, the similarity between any two samples is the result of feature-matching process that differentially weights common and distinct sample features [8]. To better understand this, consider the following example. Let $g(A \cap B)$ represent the features that are common to samples A and B, and let $g(A - B)$ represent the features that are unique to sample A. By employing Tversky’s feature contrast model [8], the similarity between samples A and B is defined as:

$$s(A, B) = \alpha g(A \cap B) - \beta g(A - B) - \gamma g(B - A) \quad (5)$$

where $\alpha$, $\beta$, $\gamma$ are constants which may vary across individual, context, and instructions. Based on this model, more features in common means high similarity, whereas features that are unique to one sample decrease the similarity.

A Simple Feature-Matching Measure. It is another example of Feature-Based Similarity Measures. It is a simple matching dissimilarity measure in which the similarity between a data sample and a particular mode is determined by the number of corresponding features that have identical values [2]. The similarity measure of $k$-modes Algorithm is presented in Equation 6 [2]. Note that in this similarity measure, the smaller value indicates the higher the similarity.

$$d_j (X, Y) = \sum_{j=1}^{n} \delta(x_j, y_j) \quad (6)$$

where

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \quad (7)$$

By taking into consideration the frequencies of categories in a dataset, the dissimilarity measure can be defined as [2]:

$$d_j (X, Y) = \frac{\sum_{j=1}^{n} (n_{x,j} + n_{y,j})}{n_{x,j} n_{y,j}} \delta(x_j, y_j) \quad (8)$$

where $n_{x,j}$ and $n_{y,j}$ are the number of samples, $j$ is the feature index, and $x$ and $y$ are the category for feature $j$.

### 2.3 Correlation-Based Similarity Measures

#### Correlation Measure

It is a similarity measure that is used to determine the degree of relationship between the two samples and it is defined as:

$$\text{Correlation} (x, y) = \frac{\sum (z_x z_y)^2}{N - 1} \quad (9)$$

where $Z_{ij}$ is the $z$ score (standardized) value of $x$ for the $i$th case or variable, and $N$ is the number of cases or variables.

### 2.4 Probabilistic Similarity Measures

Many probabilistic models have been proposed and used in a variety of applications [9]. Some of these models are based on the assumption that similarity is inversely related to psychological distance [10], while other models are based on the signal-detection notion of a decision bound [11].

**Mutual Information.** The mutual information (MI) is the process of measuring the shared information between two variables. It measures the probability of determining one variable by knowing the information about another. The $MI$ between two different variables is defined as:

$$MI (x, y) = \sum_{y} \sum_{x} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right) \quad (10)$$

In information theory, $MI$ is a symmetric measure to calculate the information shared between two distributions.

The $NMI$ is the normalized form of $MI$ [0 to 1] and is defined in (6) [3, 4]. Thus, zero mutual information means that the two variables are independent; thus, knowing one of them does not give any information about the other. If the two variables are identical then $MI$ will have the value of 1.

$$\phi^{(NMI)} (\lambda^{(a)}, \lambda^{(b) =}) = 2 \sum_{x} \sum_{y} p^{(a)} (x, y) \log \left( \frac{p^{(a)} (x, y)}{n^{(a)} n^{(b)}} \right) \quad (11)$$

where $\lambda^{(a)}$ and $\lambda^{(b)}$ are the cluster labels for variables $A$ and $B$, respectively. $n^{(a)}$ is the number of samples in cluster $C_a$ according to $\lambda^{(a)}$, $n^{(b)}$ is the number of samples in cluster $C_b$ according to $\lambda^{(b)}$, $n^{(a)}$ denotes the number of samples in cluster $C_a$ according to $\lambda^{(a)}$, and in cluster $C_b$ according to $\lambda^{(b)}$.

The average normalized mutual information (ANMI) is defined as:

$$\phi^{(ANMI)} (\lambda) = \frac{1}{|\lambda|} \sum_{x=1}^{\lambda} \phi^{(NMI)} (\lambda, \lambda^{(x)}) \quad (12)$$
where Λ is a set of r labeling.

2.5 A Classification of Similarity Models

In case the same response is made by given the same information, the decision process is of type deterministic. On other hand, if a response is selected randomly for each information state by sampling some probability distribution, the decision process is of type probabilistic. In is important to point that with a probabilistic decision process, for two separate occasions with identical percept, different decision might be made.

The common major weaknesses of Equations (1) to (12) are:

1. Unable to handle categorical features: Equations (1), (2), (3), (4) and (8).
2. Unsuitable for unweighted features: Equations (1) to (9). Therefore, one feature with large value might dominate the overall distance measure.
3. Unable to reflect the degree of change between data samples: Equations (1) to (12).

To address these weaknesses we proposed a novel similarity measure described in the next section.

3. A Novel Similarity Measure

The new similarity measure, represented by Equation (14), was defined particularly to reflect the degree of the relative change between samples and to cope with both numerical and categorical variables. For numerical variables, Term 1 is used. For the categorical variable, the similarity between two data samples is the number of the variables that have same categorical values in two considering data samples, and is calculated by Term 2. Where Sim is an abbreviation of similarity; N is the number of features; x represents the sample; i is sample index; j is feature index; R and Cat represent the numerical and categorical features respectively; B the BASE, k the index for clusters and BASES. A BASE is a sample that act like a centroid or medoids. Initially it is a real sample, but as the clustering algorithm goes for more iterations it become an artificial sample. Based on the dataset type, a BASE may consists of modes only, means only or combination of both. More detail about this will be presented in section 4.

Note that, in this similarity measure, features contribute equally to the computation of the similarity measure and no one feature can dominate the overall similarity measure. This is because the similarity value between input x_i and a BASE_{ij} is scaled to [0, 1]. This definition can be extended to measure similarity between any two samples not only limited to the BASE.

4. mk-means Algorithm

Modified k-means is a version of the traditional k-means algorithm. As the name indicates, in this algorithm we modified the traditional k-means algorithm by replacing its similarity measuring, Euclidian Distance, with our proposed similarity measure. As the proposed similarity measure is capable of handling both categorical and numerical variables, mk-means is suitable for clustering numerical, categorical and mixed data types.

The objective function of mk-means is to minimize the total cost of the partition by maximizing the similarity between the clusters’ BASES and the mean/mode of cluster m. To illustrate, let \{R_1, R_2, R_3, \ldots, R_k\} be the clustering results of dataset D, where \(R_m \neq \emptyset\) for \(1 \leq m \leq k\), and \{Q_1, Q_2, Q_3, \ldots, Q_k\} the mean/mode of \{R_1, R_2, R_3, \ldots, R_k\}, respectively. The total cost function of the partition is defined as:

\[ E = \sum_{i=1}^{n} \sum_{j=1}^{k} \text{sim} \left( B_i, Q_{a} \right) \]

where \(n\) is the number of samples in the dataset \(D\), \(B_i\) represents the BASE of cluster \(i\), and \text{sim} is the proposed similarity measure which is defined in Equation (14).

The major steps of mk-means include:

1. Choosing a number of the cluster randomly (2 <= k <= N, where k is the number of cluster and N is the number of samples)
2. Initializing seeds for the each of the chosen clusters.
3. Calculating the similarity between the chosen seeds and the remaining samples.
4. Assigning samples to the most similar cluster
   1. Find a mode or a centroid
      • Find a mode (medoids) for each categorical feature. Calculate the frequency of each category for all the categorical features and then take the most frequent category in each feature as its mode.
      • Calculate the average (centroid) for each numerical feature.
   2. Construct a sample with the modes and centroids.
5. Calculate the similarity between each cluster’s initial seeds and the constructed samples. If they are not the same, the constructed samples become the new seeds.
6. Steps 3 through 6 are repeated until the similarity between the BASES and the candidate BASES is met (BASES no longer change).

\[
\text{Sim}(x_i, B_{ij}) = 1 - \frac{1}{N} \left( \sum_{j=1}^{n} \max_{x_i \in \Lambda} \left\{ \frac{|x_i - B_{ij}|}{\max_{x \in \Lambda} \{x, B_{ij}\}} \right\} + \sum_{j=1}^{n} 1 \text{ if } x_i = B_{ij}, 0 \text{ otherwise} \right) \]
Like the traditional \( k \)-means algorithm, \( mk \)-means produces locally optimal clustering results that mainly depend on the initial seeds, which are selected randomly. Thus, if appropriate seeds are selected, good clustering results can be obtained; otherwise, poor clustering results might be generated.

5. Experiments and Evaluation

To evaluate the accuracy and the effectiveness of the new similarity measure, we compared with \( k \)-means and other two clustering algorithm. To make the comparison as fair as possible, we modified the traditional \( k \)-means by replacing its similarity measuring, Euclidian Distance, with our proposed similarity measure. The new modified algorithm is called \( mk \)-means. It worth mentioning that, the other two compared algorithms work the same as \( k \)-means but each one of them has its own similarity measure. This way, we ensured that the comparison will be mainly based on the similarity measures.

After conducting several experiments on the some benchmark datasets, the obtained clustering results of \( mk \)-means where then compared with the results generated by chosen clustering algorithms. The basic strategy of our comparison is to use \( k \)-means as a baseline. Before presenting the experimental results and carrying out the intended comparison, we present the compared algorithm is section 4.1, give the criteria and carrying out the intended comparison, we present the method of using data in section 4.2 and present the method of using data in section 4.3.

5.1 Compared Similarity Measures

\( k \)-means algorithm is the most commonly used algorithm because it is simple and computationally efficient. To calculate the similarity value between a dataset samples, \( k \)-means algorithm uses the Euclidean Distance similarity measure.

\( k \)-modes Algorithm. It is an extension of the traditional \( k \)-means and has been developed to handle categorical datasets [2, 5]. Unlike the traditional \( k \)-means this algorithm uses modes instead of means for clusters. It also uses a frequency-based method to update the modes. Equation (2) presents the similarity measure that is used by \( k \)-modes Algorithm.

\( k \)-ANMI Algorithm [6]. This algorithm is viewed to be suitable for clustering data of categorical variables. It works in the same way as \( k \)-means algorithm. However, instead of using the Euclidean Distance similarity measure, it uses the ANMI similarity measures.

5.2 Measuring the Clustering Accuracy

One of the most important issues in clustering is how to measure and evaluate the clustering performance, usually in terms of accuracy. In unsupervised clustering, there is no absolute criterion of measuring the accuracy of different sets of clusters. However, in some cases where the class labels are available, the quality of a partition can be assessed by measuring how close the clustering results are to the known groupings in the dataset. Thus, the correct clusters should be those clusters that have all the samples with the same labels within their own cluster. It should be noted that, with such a strategy, the class label is not included during the clustering process but just used at the end of the clustering procedure to assess the partition quality.

In practice, two measurements are commonly used for estimating clustering accuracy or error. The accuracy \( r \) can be measured by \( r = \frac{\sum_{i=1}^{k} a_i}{n} \) [5], where \( a_i \) is the number of majority samples with the same label in cluster \( i \), and \( n \) the total number of samples in the dataset. Hence, the clustering error can be obtained by \( e = 1 - r \).

5.3 Testing Datasets

We used a total of 3 other datasets in our testing experiments. Mushroom dataset, Congressional Votes dataset, Wisconsin Breast Caner dataset. These three dataset has been chosen to test the proposed similarity measure due to their publicity as they have been used quite often in many clustering algorithm [5, 12-15]. Table 1 show the demographic details of 3 benchmark datasets (obtained from UCI Machine learning Repository [7]) that have been used in our experiment.

- **Mushroom dataset:** it includes the descriptions of hypothetical mushroom that was drawn from The Audubon Society Field Guide to North American Mushrooms in 1981. It consist of 8124 samples each being described by 22 features. All features are of type categorical and each one of them represents physical characteristics of a single mushroom, which is identified as either edible or poisonous. The number of edible samples is 4208, while the number of poisonous sample is 3916.
- **Congressional Votes dataset:** it is the Voting records of the United States in 1984. It has 435 votes with 16 attributes. All attributes are Boolean with Yes and NO values. The Yes is donated as Y and the NO is donated as N. The samples are classified as follows: 267 samples with Democrats, and 168 samples with Republicans.
- **Wisconsin Breast Caner dataset:** it consists of 699 samples and 9 features. Each sample identified as benign or malignant. Note that samples are of type numeric. However, as they have been considered as categorical in [6], in our experiments we will use them as categorical as well.

Following a commonly used strategy in clustering experiments, the whole dataset is used in experiments...
5.4 Experimental Results

In this section, we compared the performance of the \textit{mk}-means with other similarity measures that have been employed in other commonly used clustering algorithms. The, \textit{k}-modes, and \textit{k}-ANMI algorithms have been used by He et al. [6] to cluster the Votes, Cancer and Mushroom datasets. As \textit{k}-means algorithm lacks the ability to identify the appropriate number of clusters, different numbers of clusters from 2 to 9 were chosen to cluster the datasets. For a given value of \( k \), we run the experiments 10 times with different random seeds and then take the average of the clustering accuracy as the a final result of that particular \( k \) value.

In clustering Vote dataset, according to Table 2 the \textit{mk}-means is the best algorithm. It is the only algorithm that scored the first position in 4 cases. Beside this, it achieved the best average clustering accuracy. \textit{k}-means algorithm scored the first position only in 3 cases and it overall average clustering accuracy is the third, after the \textit{k}-ANMI algorithm. \textit{k}-modes algorithm is the worst one as it scored the last position in 5 cases. Therefore, it has the lowest average clustering accuracy.

In clustering Mushroom dataset, as shown in Table 3 and Figure 2 that \textit{mk}-means and \textit{k}-ANMI equally shared the first position as each one of them won in 4 cases and lost in 4 case. However, in term of the average clustering accuracy, \textit{mk}-means beat all the other algorithm. \textit{mk}-means is the only algorithm that has never performed less than 87.4\%. In contract, at \( k = 2 \), none of the algorithm achieved an accuracy of 70\%. At \( k = 3 \), \textit{k}-ANMI, is the only algorithm, beside \textit{mk}-means, which obtained an accuracy more than 70\%. The reaming algorithms started acceding the boundary of 70\% at \( k = 4 \). \textit{k}-modes algorithm is the worst algorithm as it lost in all cases. More importantly, when it lost, it lost by a significant amount especially at \( k = 2 \) and 3. Overall, \textit{mk}-means, \textit{k}-means, and \textit{k}-ANMI are the only algorithms that achieved average clustering accuracy higher than 80\%.

Concerning Cancer dataset, none of the algorithm-achieved an accuracy less than 80\% (see Table 4 and Figure 3). Although \textit{mk}-means did not win in all cases, its accuracy is always comparable to the other algorithms and less by a tiny fraction than the \textit{k}-ANMI, which achieved the best in most cases. In general, like cases in Vote and Mushroom dataset, \textit{mk}-means beat the other algorithm in terms of the average clustering accuracy. It is important to note that in all cases the accuracy of \textit{mk}-means is higher than 95\%.

Based on Tables 2, 3, and 4 the clustering performance of \textit{mk}-means is superior to the traditional \textit{k}-means and the compared algorithm. Thus, we can conclude that by embedding our proposed similarity measure in the traditional \textit{k}-means the clustering accuracy is increased. This indicates our new similarity measure is better in representing the degree of similarity.

### Table 1: Details of the benchmark datasets

<table>
<thead>
<tr>
<th>No.</th>
<th>Datasets</th>
<th>No. of classes</th>
<th>No. of Samples</th>
<th>No. of Features</th>
<th>Missing Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mushroom</td>
<td>2</td>
<td>8124</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Votes</td>
<td>2</td>
<td>435</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Cancer</td>
<td>2</td>
<td>699</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>

### Table 2: Experimental Result of Vote dataset:

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Number of Clusters</th>
<th>Average</th>
<th>Stddev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>0.864</td>
<td>0.872</td>
<td>0.924</td>
</tr>
<tr>
<td>k-ANMI</td>
<td>0.869</td>
<td>0.872</td>
<td>0.883</td>
</tr>
<tr>
<td>k-means</td>
<td>0.880</td>
<td>0.883</td>
<td>0.892</td>
</tr>
<tr>
<td>mk-means</td>
<td>0.867</td>
<td>0.89</td>
<td>0.933</td>
</tr>
</tbody>
</table>
Figure 1: Experimental Result of Votes Dataset

Table 3: Experimental Result of Mushroom Dataset

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Number of Clusters</th>
<th>Average</th>
<th>Stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>k-modes</td>
<td>0.565</td>
<td>0.567</td>
<td>0.737</td>
</tr>
<tr>
<td>k-ANMI</td>
<td>0.587</td>
<td>0.636</td>
<td>0.756</td>
</tr>
<tr>
<td>k-means</td>
<td>0.677</td>
<td>0.894</td>
<td>0.899</td>
</tr>
<tr>
<td>mk-means</td>
<td>0.891</td>
<td>0.896</td>
<td>0.894</td>
</tr>
</tbody>
</table>

Figure 2: Experimental Result of Mushroom Dataset

Table 4: Cancer Dataset: comparison between 6 clustering algorithms

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Number of Clusters</th>
<th>Average</th>
<th>Stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>k-modes</td>
<td>0.915</td>
<td>0.914</td>
<td>0.933</td>
</tr>
<tr>
<td>k-ANMI</td>
<td>0.978</td>
<td>0.956</td>
<td>0.927</td>
</tr>
<tr>
<td>k-means</td>
<td>0.956</td>
<td>0.969</td>
<td>0.973</td>
</tr>
<tr>
<td>mk-means</td>
<td>0.953</td>
<td>0.97</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Figure 3: Cancer Dataset: comparison between 6 clustering algorithms
5.5 Evaluation and Discussion

In terms of average clustering accuracy, mk-means scored the first position in all datasets (refer to Tables 2, 3 and 4). As k-means is used as a target algorithm, it is worth noting that our algorithm performed better than k-means, won in all cases. k-means managed to score the first position only in three cases and all of them occurred in Votes dataset at k = 2, 8, 9.

k-modes has never ever scored the first position in any of the datasets. By referring to Table 2, we can see that it was always the worst algorithm apart from k = 4, 5, and 6. In clustering Mushroom dataset, out of 8 experiments, it achieved the worst performance in 7 cases (see Table 3). As for Cancer dataset, k-modes could not compete with other algorithm and scored the lowest position in 5 cases (see Table 4).

Regarding k-ANMI algorithm, it performed the first position in one occasion only, at k = 5 (see Table 2 and Figure 4). In clustering Mushroom dataset, it achieved the first position in four cases, at k = 5, 7, 8 and 9 (see Table 3 and Figure 2). In clustering Cancer dataset, it achieved the first position in five cases, at k = 1, 6 and 7 (see Table 4 and Figure 3).

Table 5 summarized the accuracies differences between mk-means and the compared algorithms in clustering the Vote dataset. As shown that when the proposed algorithm achieved the best clustering results the highest difference between its accuracy and the second highest accuracy was 1.8% at which k = 6 while the lowest accuracy difference was 0.9% at which k = 4. When the proposed algorithm achieved the third position at k = 8, although its accuracy was 1.9% less than the algorithm, which achieved the first position, it was 2.7% higher than the algorithm, which achieved the lowest accuracy.

In clustering the Mushroom dataset, there was a significant difference between the mk-means and the compared algorithm at k = 2. As presented in Table 6, mk-means was 21.4% higher than the second best algorithm and 32.6% than the worst algorithm. After k = 5, the performance of the mk-means starts decreasing and become, at least, 5% less than the best algorithm. However, as discussed earlier in section 4.4 that it maintained the best average clustering accuracy.

Regarding Cancer dataset, there is no big difference between the first and the second best algorithm in all cases (see Table 7). For example, when the proposed algorithm achieved the best clustering results the highest difference between its accuracy and the second highest accuracy was 0.2% at which k = 5 while the smallest accuracy difference was 0.1% at which k = 3. When the mk-means algorithm achieved the second position, its accuracy was only 0.3%, 0.4%, and 0.1% less than the algorithm the best algorithm at which k = 4, 6 and 7, respectively. mk-means scored the third position in two cases and its accuracy was about 2.5% and 0.3% at k = 2 and 8, respectively, less than the best algorithm. However, in both cases it was better than the worst algorithm by, at least, 3.8%.

**Table 5. Summary of Experimental Result: Vote dataset**

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking of mk-means</td>
<td>3rd</td>
<td>1st</td>
<td>1st</td>
<td>2nd</td>
<td>1st</td>
<td>1st</td>
<td>3rd</td>
<td>2nd</td>
</tr>
<tr>
<td>mk-means - Min.</td>
<td>+0.30%</td>
<td>+2.30%</td>
<td>+5.00%</td>
<td>+5.10%</td>
<td>+3.20%</td>
<td>+2.30%</td>
<td>+2.70%</td>
<td>+3.00%</td>
</tr>
<tr>
<td>mk-means - Max</td>
<td>-1.30%</td>
<td>+0.70%</td>
<td>+0.90%</td>
<td>-2.00%</td>
<td>+1.80%</td>
<td>+0.70%</td>
<td>-1.90%</td>
<td>-1.30%</td>
</tr>
</tbody>
</table>

**Table 6. Summary of Experimental Result: Mushroom dataset**

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking of mk-means</td>
<td>1st</td>
<td>1st</td>
<td>2nd</td>
<td>2nd</td>
<td>1st</td>
<td>4th</td>
<td>4th</td>
<td>4th</td>
</tr>
<tr>
<td>mk-means - Min.</td>
<td>+32.6%</td>
<td>+32.9%</td>
<td>+15.7%</td>
<td>+1.8%</td>
<td>+4.1%</td>
<td>-1.20%</td>
<td>-0.80%</td>
<td>-0.40%</td>
</tr>
<tr>
<td>mk-means - Max</td>
<td>+21.40%</td>
<td>+0.20%</td>
<td>-0.50%</td>
<td>-1.40%</td>
<td>-0.30%</td>
<td>-5.20%</td>
<td>-8.50%</td>
<td>-7.20%</td>
</tr>
</tbody>
</table>

**Table 7. Summary of Experimental Result: Cancer dataset**

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking of mk-means</td>
<td>3rd</td>
<td>1st</td>
<td>2nd</td>
<td>1st</td>
<td>2nd</td>
<td>2nd</td>
<td>3rd</td>
<td>2nd</td>
</tr>
<tr>
<td>mk-means - Min.</td>
<td>+3.8%</td>
<td>+5.6%</td>
<td>+4.3%</td>
<td>+4.1%</td>
<td>+3.6%</td>
<td>+2.4%</td>
<td>+9.5%</td>
<td>+4.4%</td>
</tr>
<tr>
<td>mk-means - Max</td>
<td>-2.50%</td>
<td>0.10%</td>
<td>-0.30%</td>
<td>0.20%</td>
<td>-0.40%</td>
<td>-0.10%</td>
<td>-0.30%</td>
<td>-1.20%</td>
</tr>
</tbody>
</table>

**Conclusion**

In this paper, we proposed a novel similarity definition for clustering. To evaluate the accuracy of the proposed similarity measure we embedded it into the traditional k-means and then compared its clustering performances with the traditional k-means and other two clustering algorithm. The criterion of selecting the compared algorithm was under the condition that they must all work the same, like the traditional k-means. This is to ensure
that the comparison will be not be affected by any other factor apart from the similarity measures. With respect to the clustering accuracy, the experimental results show that our proposed similarity measure has always been ranked highly among 3 similarity measures compared. This indicates that the proposed similarity measure is accurate and reliable in general. Thus, we concluded that by embedding our proposed similarity measure in the traditional $k$-means the clustering accuracy is increased. As the similarity value between features is scaled to $[0, 1]$ all features will have the same weight in calculating the overall similarity value.

**Future Work**

Future work will involve re-implementing the traditional $k$-means with other similarity measures, and then conducting more experiment to compare the clustering results with $mk$-means. It will also involve using some statistical tests such as T-test to indicate the significance of the overall improvement.

**References**


