Similarity Measures Project Report

Lamjed Lounissi

1. Introduction. The main purpose of the project is helping dairy agriculture in decision-making process. Clustering analysis played an import role and it is considered as a precious tool to achieve this goal. Traditionally, clustering approaches concentrate on purely numerical or categorical data only. An important area of cluster analysis deals with mixed data, composed by both numerical and categorical attributes. Clustering mixed data is not simple, because there is a strong gap between the similarity metrics for these two kinds of data. In this study we provide a mixed similarity measure (MSM) that could be used with mixed-data types.

2. Main results. The proposed mixed similarity measure (MSM) uses a specific similarity measure for each type of data attribute. Weights for numeric features are calculated, Euclidean distance for numeric features and co- occurrence-based distance measure for categorical. The measures also take into account the significance of an attribute towards the clustering process. [3] [1][2]

2.1. The distance measure for mixed datasets. This distance measure uses the distances between categorical attribute values. The proposed distance measure with m attributes, in which m_r attributes are numeric whereas m_c attributes are categorical is presented in Eq. (2.1). It has two components; the first component computes the distance for numeric attributes and the second component for categorical attributes.

Definition 2.1. *Let us assume $D1$ and $D2$ are two objects in a mixed data set defined by a total of m attributes. The two objects may be represented as $D1 = X1, X2, ..., Xm$ and $D2 = Y1, Y2, ..., Ym$ where first m_r attributes are numeric, next m_c are categorical attributes and $m_r + m_c = m$. Distance between $D1$ and $D2$, denoted by $Dist(D1, D2)$ is computed as follows:*

$$Dist(D1, D2) = \sum_{t=1}^{m_r} (w_(t)(X_t - Y_t))^2 + \sum_{t=1}^{m_c} \delta(X_t, Y_t)^2 \tag{2.1}$$

In the first component, w_t is the weight of the tth numeric attribute. The second component represents the distance between categorical attribute values. The distance between two attribute values of a categorical attribute is calculated by computing the co-occurrence of these attribute values with attribute values of other categorical attributes. This algorithm has following steps;

1. Discretize all the numeric attributes to make the dataset pure categorical. This step is required to compute the distances between each pair of attribute values of every categorical attribute and the weight of each numeric attribute. We would like to note

ANM_ID	Animal Status	Condition Code	TEMPERAMENT	LAC_NO
Cow01	Dry	Mastitis	Nervous	AM
Cow02	Left Herd	Mastitis	Nervous	AM
Cow03	Milking	Injury	Average	PM
Cow04	Milking	Healthy	Average	PM
Cow05	Entered milking	Healthy	Quiet	PM
Cow06	Entered milking	Injury	Quiet	PM

Table 1

that while calculating distances the original attribute values are used.

2. For every categorical attribute, distances between every pair of attribute values are calculated. The distance between attribute values, x and y, of categorical attribute A_i with respect to attribute categorical A_j, for a given subset W of attribute A_j values, is defined by δ_w^{ij} and computed by using following formula;

$$\delta_W^{ij} = P(W/x) + P_i(\sim W/x) - 1 \tag{2.2}$$

$P(W|x)$ is the probability estimates of data points with attribute value equal to x belong to a class contained in W. Probability estimates are computed by computing the frequencies of pairs of attribute values and class values.

$P(\sim W|y)$ is the probability estimates of data points with attribute value equal to y belong to a class not contained in W. Probability $P(W|y)$ is the probability estimates of data points with attribute estimates are computed by computing the frequencies of pairs of attribute values and class values.

The distance, $\delta^{ij}(x, y)$, between x and y with respect to attribute A_j is given by $\delta_w^{ij} = P(w/x) + P_i(\sim w/x) - 1$, where w is the of subset of values of attribute A_j that maximizes the quantity $P(W/x) + P_i(\sim W/x) - 1$.

This distance is calculated against every other distance between x and y in the dataset. For example we consider the cows dataset in the Table 1. Table 2 shows the conditional probability values that are used in computing the distance between two values. In the same way distance between every pair of values for a categorical attribute are computed and the results are shown in Table 3.

2.2. Computing the dissimilarity matrix. The distance function defined in Eq.1 can be used to compute an $n \times n$ dissimilarity matrix, which represents pair-wise distance between objects of a data set. We illustrate the accuracy of the distance function in capturing object similarities computed for 2-class data sets. Fig. 1. represents similarities in terms of distances among elements picked up from the cows data set. The dissimilarity matrix encodes the pair-wise distance between data elements. The more similar two elements are, the lower is the distance between the two. Fig 2. represents 5 groups of cows using MSM measure.

Probability table

P(Dry /Mastitis)= 1/2	P(Entered milking /Injury)= 1/2
P(Left Herd /Mastitis)= 1/2	P(Nervous / Mastitis)= 1
P(Milking/Mastitis)= 0	P(Average / Mastitis)= 0
P(Entered milking /Mastitis)= 0	P(Quiet / Mastitis)= 0
P(Dry /Injury)= 0	P(Nervous / Injury)= 0
P(Left Herd /Injury)= 0	P(Average / Injury)= 1/2
P(Milking/Injury)= 1/2	P(Comedy / Injury)= 1/2

Table 2

Distance between pairs of categorical values of cows data

δ (Dry, Left herd)=0	δ(Mastitis, Injury)= 1	δ(Nervous, Quiet)= 1
δ (Dry, Milking)=1	δ(Mastitis, Healthy)= 1	δ(Nervous, Average)= 1
δ (Dry, Entered milking)=1	δ(Injury, Healthy)= 0	δ(Nervous, Quiet)= 1/2
δ (Milking, Left herd)=1		
δ(Entered milking, Left herd)=1		
δ(Entered milking, Milking)= 1/2		

Table 3

3. Conclusion and future work. In this part of project, we have proposed a novel mixed similarity measures for mixed datasets. More sophisticated methods for discretizing numeric valued attributes can definitely better results. Other implementations of k-mean algorithm and K-Harmonic algorithm may also be examined to achieve more optimized performances.

REFERENCES

[1] A. Ahmad and L. Dey, *A k-mean clustering algorithm for mixed numeric and categorical data*, Data & Knowledge Engineering, 63 (2007), pp. 503–527.
[2] A. Ahmad and S. Hashmi, *K-harmonic means type clustering algorithm for mixed datasets*, Applied Soft Computing, 48 (2016), pp. 39–49.
[3] A. Ahmad and S. S. Khan, *Survey of state-of-the-art mixed data clustering algorithms*, IEEE Access, 7 (2019), pp. 31883–31902.

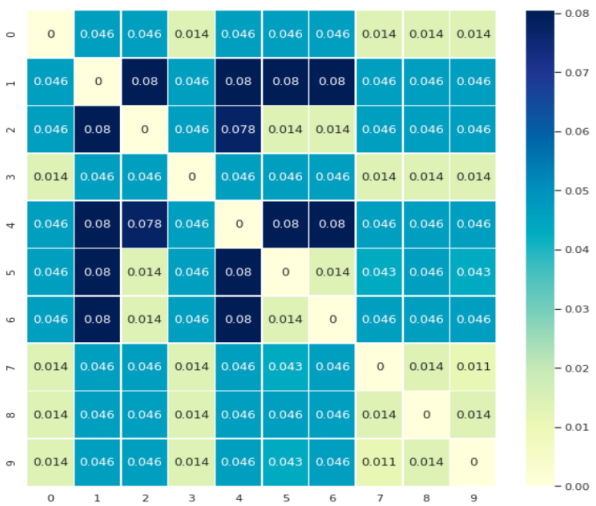

Figure 1. *Dissimilarity matrix for 10 cows*

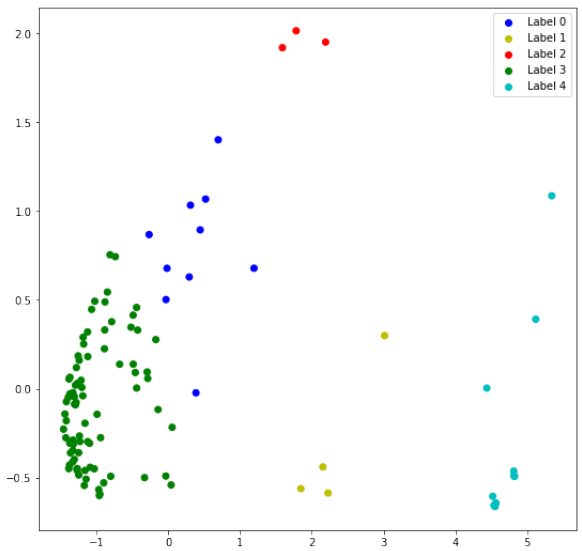

Figure 2. *5 clusters for herd 2: 114 cows*