Prediction of the Recognition Reliability using Clustering Results

Otar Verulava, Ramaz Khurodze, Tea Todua, Otar Tavdishvili, Taliko Zhvania
Department of Informatics and Control Systems
Georgian Technical University (GTU)
Tbilisi, Georgia
verulava@gtu.ge

***Abstract*. The recognition reliability problem when realization of the recognition experiments is connected to some difficulties (some operative conditions or material expenditures) is considered.**

For estimation of recognition reliability clustering process is used. Particularly, clustering results must unambiguously determine the number of clusters and their contents. The main requirement is validity of clustering results, which depends on the cardinal number of set of realizations in learning sample. The set of realizations must be representative as far as possible. In case of satisfaction of the above-mentioned conditions it is possible using of other procedures differing from the presented.

Keywords: cluster, prediction, rank links, recognition reliability, Hausdorff distance

I. INTRODUCTION

For Prediction of the recognition reliability clustering process with Rank Links is used [1,2,3]. It gives us possibility to establish the following cluster characteristics:

1. The number of clusters;
2. The number and list of realizations in each cluster;
3. Characteristic feature of cluster construction – a value of cluster construction rank;
4. Indicator of the clusters isolation – the number of missing ranks;

All four characteristics are scalars. Their values are the more valid the more are the number of realizations in learning sample of each pattern. Condition of representativeness of the realizations is a quite strong requirement and sometimes it is difficult to satisfy it for some practical tasks. Therefore empirical criteria is used. It implies that the more dimension of receptive field the more should be the number of realizations in the learning samples of each pattern.

Below are presented some concepts for estimation of clustering results:

Definition 1. A cluster is compact if it contains only one kind of realizations, otherwise cluster is incompact.

Definition 2. A pattern is compact if it is presented by only compact clusters otherwise pattern is incompact.

II. PREDICTION OF THE RECOGNITION RELIABILITY FOR COMPACT PATTERNS

Let's assume that set of realizations X are clustered and the values for all four clustering characteristics are obtained. Assume for simplicity that two elements A_i and A_j from a set of patterns A are taken. The task is to determine the probabilities of correct or false recognition for these elements.

Let's assume that patterns A_i and A_j are compact and their corresponding clusters are located as shown on Figure 1.

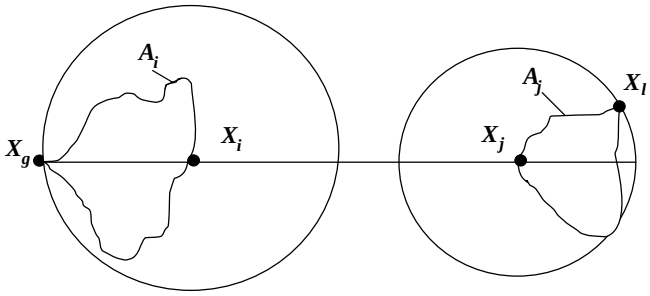

Figure 1.

Define a minimum of the distances between the realizations of the different clusters. This distance represents a Hausdorff metric between the sets. Let's assume that Hausdorff metric is realized for realizations $X_i \in A_i$ and $X_j \in A_j$ (Fig.1). Define a maximum of distances from the realizations X_i and X_j to the realizations of their cluster, which are represented by the points-realizations X_g and X_l (Fig.1). Let's circumscribe the hyperspheres (circles) of radiuses $X_i X_g$ and $X_j X_l$ from points-realizations X_i and X_j.

Definition 3. A part of feature space, which is circumscribed by the hypersphere of radius $X_i X_g$ is called a pattern A_i influence zone on pattern A_j.

ISSN : 0975-3397

Definition 4. A part of feature space, which is circumscribed by the hypersphere of radius X_jX_l is called a pattern A_j influence zone on pattern A_i.

Take into consideration that the above-mentioned definitions can be used for any pairs of the set A elements.

Let's consider some alternate versions of the influence zone locations:

1. Influence zones are disjointed.

It means for patterns A_i and A_j that hyperspheres with radiuses X_iX_g and X_jX_l are disjointed (Fig.1). This case can be described by following expression:

$$X_iX_g + X_jX_l < X_iX_g \qquad (1)$$

According to Rank Links we'll have the following inequality:

$$Rank(X_i;X_g) + Rank(X_j;X_l) < Rank(X_i;X_j) \quad (1')$$

2. The influence zones are intersected (Fig. 2)

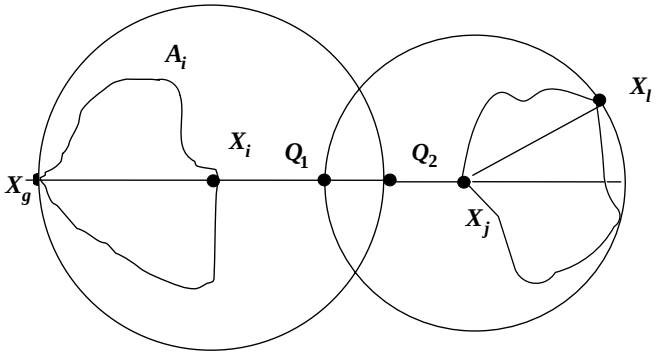

Figure 2.

It's obvious, that intersection of the influence zones doesn't imply intersection of the clusters, because in this case the condition of patterns compactness would not be satisfied.

An alternate version shown on Fig. 2 can be presented by the following inequality:

$$X_iX_g + X_iX_l > X_iX_g \qquad (2)$$

According to Rank Links we'll have:

$$Rank(X_i;X_g) + Rank(X_j;X_l) > Rank(X_i;X_g) \quad (2')$$

Prediction of the recognition reliability is based on the following axioms:

Axiom 1. The realizations of any pattern A_j with respect to some pattern A_i can be appeared in pattern A_j influence zone only with respect to pattern A_i.

Axiom 2. Predictable recognition reliability is defined as a result of division of the number of realizations belong to the influence zone of the given pattern on the general number of realizations of the cluster.

Let's denote the number of realizations of the pattern A_i by M_i, while the number of the realizations which has been located in the influence zone of the pattern A_j by M_{ij}. According to axiom 2 for estimation of the recognition error P_{ij}^- we'll have:

$$P_{ij}^- = \frac{M_{ij}}{M_i} \qquad (3)$$

where $i, j = \overline{1,I}$, $i \neq j$.

Based on the axiom 1, if the influence zones are not intersected, then the probability of appearance of other pattern realizations in these zones is equal to zero. Correspondingly, we'll have $M_{ij} = 0$. According to (3) we'll obtain:

$$P_{ij}^- = 0 \;\Rightarrow\; P_{ij}^+ = 1$$

Let's use for estimation of the recognition reliability a division of a maximal width Q_1Q_2 of the influence zones intersection on the Hausdorff distance between the clusters (Fig. 2). For calculation of Q_1Q_2 we'll apply the following expression:

$$Q_1Q_2 = X_iX_j - (X_iQ_1 + Q_2X_j) \qquad (4)$$

where

$X_iQ_1 = X_iX_j - X_jQ_1$, $Q_2X_j = X_iX_j - X_iQ_2$.

Substitute the obtained values of X_iQ_1 and Q_2X_j into (4) and take into consideration that $X_iQ_1 = X_iX_g$ and $X_jQ_2 = X_jX_l$, we'll obtain

$$Q_1Q_2 = -X_iX_j + X_jX_l + X_iX_g \qquad (5)$$

Q_1Q_2 is a distance and consequently is positive. The value of reliability is nonnegative scalar also. Hence we can write:

$$Q_1Q_2 = |-X_iX_j + X_jX_l + X_iX_g| \qquad (6)$$

Then the recognition reliability will be:

$$P_{ij}^- = \frac{|-X_iX_j + X_jX_l + X_iX_g|}{X_iX_j} \qquad (7)$$

The same result we can obtain by (2`), written in a different way:

$$Rank(X_iX_j) - (Rank(X_i;X_g) + Rank(X_j;X_l)) < 0 \qquad (8)$$

As we see from (8), the obtained difference is a negative value. Therefore, similar to (7), let's use absolute value of the

ISSN : 0975-3397

difference again. Respectively, for estimation of the recognition reliability by Rank Links we'll have:

$$P_{ij}^{-} = \frac{|Rank(X_i; X_j) - (Rank(X_i; X_j) + Rank(X_j; X_i))|}{Rank(X_i; X_j)} \qquad (9)$$

where $i, j = \overline{1; I}$, $i \neq j$.

Hence, we can use (9) in case when clustering is realized by Rank Links and (7), otherwise.

III. Prediction of the Recognition Reliability for Non-compact Patterns

Based on definition 2 we have non-compact patterns, if more than one kind of realizations are united in cluster. Let's assume that A_i and A_j patterns are non-compact. Denote their corresponding non-compact cluster by CL_{ij}, which is presented by bold contour line on Fig. 3. In the same place, the realizations of patterns A_i and A_j are shown by little circles and triangles respectively. In this case it is necessary to define an intersection area, i.e. a list of realizations (points) fell into the area. Solution of this task is possible if we apply neighborhoods' principle developed in Rank Links, which is based on notion of cluster construction rank r_{ij} [1, 3]. It allows us to define for all cluster points those realizations, which have less or equal to r_{ij} closed Rank Link with respect to the given point. In the same way we'll construct the subclusters for each realization and then select only those in which the realizations of the both A_i and A_j patterns are united. A union of such subclusters will give us intersection area including the list of realizations.

Let's for the pattern A_i realizations located in intersection area define maximal distance which on Fig. 3 is shown for the points X_h and X_k. Repeat the

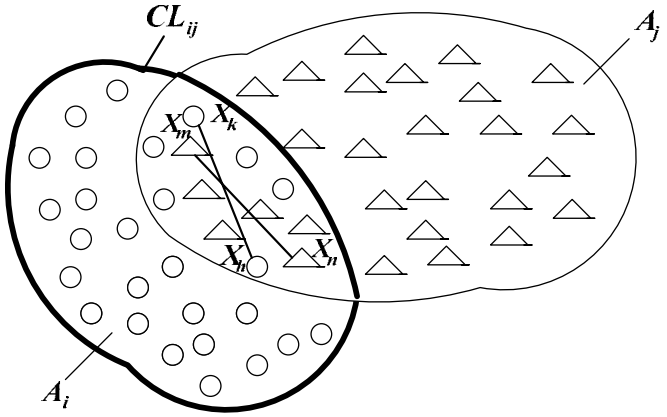

Figure 3

same operation for the realizations of pattern A_j included in the intersection area. As a result we'll obtain the points X_m and X_n. From these points circumscribe the hyperspheres of radiuses X_hX_k and X_mX_n respectively. Let's define the number of the realizations of pattern A_i included in radius X_hX_k (Fig.3). It is important to take into account that the realizations which belong to the both hyperspheres should be counted once. Denote the number of counted in the hyperspheres realizations by M_{ji}, while the numbers of the realizations of pattern X_j united in hyperspheres of radius X_kX_h by M_{ij}. Then the probability of recognition error will be:

$$P_{ij}^{-} = \frac{M_{ij}}{M_i};\ P_{ji}^{-} = \frac{M_{ji}}{M_j} \qquad (10)$$

where $i, j = \overline{i, I}$, $i \neq j$.

In that case the estimation of recognition reliability will be:

$$P_{ij}^{+} = 1 - P_{ij}^{-};\ P_{ji}^{+} = 1 - P_{ji}^{-}.$$

Conclusions

For estimation of recognition reliability so-called the pattern influence zones are used. They can be determined using clustering results. Clustering algorithms should satisfy condition of determinacy. In the proposed paper clustering by Rank Links method is considered. Possibility of using other algorithms for prediction of recognition reliability also is shown.

References

[1]. O. Verulava. Clustering Analysis by "Rank of Links" Method. Transactions of GTU, #3(414), Tbilisi, 1997, pp.277-287.

[2]. O. Verulava, R. Khurodze. Clustering Analysis and Decision-making by "Rank of Links". Mathematical Problems in Engineering, 2002, vol. 8(4-5), pp. 475-492;

[3]. O. Verulava, R. Khurodze. Theory of "Rank of Links" – Modeling and Recognition process, GTU Press, Tbilisi, 2002, pp. 346.

.

ISSN : 0975-3397