Mathematical Problems in Engineering, 2002, Vol. 8(4–5), pp. 475–492

Clustering Analysis and Decision-making by "Rank of Links"

O. VERULAVA* and R. KHURODZE

Machine Intelligence Center (N 347), Tbilisi 380075, Kostava str. 77, Georgia

(Received 13 March 2002; In final form 15 May 2002)

In this paper, formalism for cluster analysis, based on the "Rank of Links"-theory, is suggested. It tackles resemble measures, cross-distance matrices, "rank of links"-metric and some other cluster characteristics. Using these notions, an algorithm of clustering has been designed. Its application to estimation and prognosis of decision-making process shows nice workability and reliability.

Key words: Cluster analysis; Rank of links; Decision-making

AMS Classification Code: 68P, 68U, 68W30

1 INTRODUCTION

The cluster analysis, in general, is a technique for the classification of objects into groups based on their similarities. There is a wide choice of methods with different requirements in computer resources. Below we present a short survey of last publications within this area. So, in the paper of Bouguettaya and Le Viet [1] the authors present the result of a fairly exhaustive study to evaluate three commonly used clustering algorithms, namely, single linkage, complete linkage, and centroid. The cluster analysis study is conducted in the two dimensional (2-D) space. Three types of statistical distribution are used. Two different types of distances to compare lists of objects are also used. The results point to some startling, similarities in the behavior and stability of all clustering methods.

Kollios *et al.* [6] present a new approach for indexing animated objects and efficiently answering queries about their position in time and space. In particular, they consider an animated movie as a spatiotemporal evolution. A movie is viewed as an ordered sequence of frames, where each frame is a 2D space occupied by the objects that appear in that frame. The queries of interest are range queries of the form, "find the objects that appear in area S between frames $f(i)$ and $f(j)$" as well as nearest neighbor queries such as, "find the q nearest objects to a given position A between frames $f(i)$ and $f(j)$". The straightforward approach to index such objects considers the frame sequence as another dimension and uses a 3D access method (such as, an R-Tree or its variants), This, however, assigns long "lifetime" intervals to objects that appear through many consecutive frames. Long intervals are difficult to cluster efficiently in a 3D index. Instead, the authors propose to reduce the problem to a partial-persistence problem. Namely, a 2D access method is used that is made partially persistent. It is

* Corresponding author. E-mail: verulava@dtu.edu.ge

ISSN 1024-123X print; ISSN 1563-5147 online

shown that this approach leads to faster query performance while still using storage proportional to the total number of changes in the frame evolution. What differentiates this problem from traditional temporal indexing approaches is that objects are allowed to move and/or change their extent continuously between frames. Some novel methods are presented to approximate such object evolutions. The authors formulate an optimization problem for which an optimal solution is provided for the case where objects move linearly. Finally, an extensive experimental study of the proposed methods is presented. While it is concentrated on animated movies, the suggested approach is general and can be applied to other spatiotemporal applications as well. The paper of Weir *et al.* [11] describes the design and implementation of a software system for producing, managing, and analyzing catalogs from the digital scans of the Second Palomar Observatory Sky Survey. The system (SKICAT) integrates new and existing packages for performing the full sequence of tasks from raw pixel processing, to object classification, to the matching of multiple, overlapping Schmidt plates and CCD calibration frames. The authors describe the relevant details of constructing SKICAT plate, CCD, matched, and object catalogs. Plate and CCD catalogs are generated from images, while the latter are derived from existing catalogs. A pair of programs complete the majority of plate and CCD processing in an automated, pipeline fashion, with the user required to execute a minimal number of pre- and post-processing procedures. They apply a modified version of FOCAS for the detection and photometry, and new software for matching catalogs on an object-by-object basis. SKICAT employs modem machine-learning techniques, such as decision trees, to perform automatic star-galaxy-artifact classification with a $>90\%$ accuracy down to similar to 1 mag above the plate detection limit. The suggested system also provides a variety of tools for interactively querying and analyzing the resulting object catalogs. In making a prediction, Hartigan [5] divides his attention between objects presently perceived and previously experienced objects. The present objects are recognized as similar to objects previously experienced, and the qualities remembered from previous examination are predicted for the present objects. Prediction is fallible, in that the author may make errors in recognizing the present object, or in past observations of the experienced object, or in assigning qualities to the present object which hold for the experienced objects similar to the present object, but not for the present object. Probability to quantify these errors is used. The classification in organizing the experiences, and in recognizing present objects as being similar to some species of experienced objects, is represented. Prabhakar and Jain [7] have proposed a scheme for classifier combination at decision level which stresses the importance of classifier selection during combination. The proposed scheme is optimal (in the Neyman–Pearson sense) when sufficient data are available to obtain reasonable estimates of the join densities of classifier outputs. Four different fingerprint-matching algorithms are combined using the proposed scheme to improve the accuracy of a fingerprint verification system. Experiments conducted on a large fingerprint database (similar to 2700 fingerprints) confirm the effectiveness of the proposed integration scheme. An overall matching performance increase of similar to 3% is achieved. Further there is shown that a combination of multiple impressions or multiple fingers improves the verification performance by more than 4% and 5%, respectively. Analysis of the results provides some insight into the various decision-level classifier combination strategies. A software signal and image processing laboratory, which has proved effective both as an educational "workbench" and in practical operational use, is suggested by Campbell *et al.* [2]. It requires a pedagogical tool, a research environment, and a fully operational data analysis system, *i.e.*, it is used not only in undergraduate engineering courses, but also in graduate study and general research. The system must be easily extendable, *e.g.*, to allow undergraduates to perform practical programming of standard digital filters and image processing algorithms, or to provide a realistic platform upon which novel algorithms can be implemented. On a further dimension, the system must handle seamlessly and efficiently three

broad data types: digital signals (sequences), images (possibly multiband), and multivariate data sets.

In this paper, following to [8–10], the formalism for cluster analysis, based on the "Rank of Links"-theory, is developed. It operates with resemble measures, cross-distance matrices, "rank of links"-metric and some other cluster characteristics. Then an algorithm of clustering, using these notions, has been designed and applied to estimation and prognosis of decision-making reliability.

2 SUGGESTED FORMALISM FOR CLUSTER ANALYSIS

2.1 Resemblance Measure

Measuring of some characteristics, called below *features*, may designate a material plant. The measurements of n features may correspond to m plants. Let us represent the results of these measurements as a $(m \times n)$ matrix (see Fig. 1).

Every row of this matrix corresponds to one plant. Let E be a set of m plants with n features, that is $E = \{X_1, X_2, \ldots, X_m\}$. The space of such plants is called the n-dimensional *feature space* where every plant is associated with a point.

The *resemblance measure* between two plants X_i, X_j we will characterize by the corresponding distance function $d(X_i, X_j)$ (below, for the simplicity, $d(i,j)$), which, as any metrics, satisfies three axioms: reflectivity $(d(i,i) = 0)$, symmetry $(d(i,j) = d(j,i))$ and the triangle inequality $(d(i,j) \leq d(i,k) + d(k,j)$ valid for any i, k and $j)$. Sometimes, an *ultra* metric $d(i,j)$ [3], satisfying the axioms of reflectivity, symmetry and the generalized triangularity $(\partial(i,j) \leq \text{Sup}[\partial(i,k), \partial(k,j)])$, is used.

Cluster analysis basically tackles two problems: clustering process itself and clustering identification. The main purpose of the clustering process is to construct a partition of the set $E = \{X_1, X_2, \ldots, X_m\}$ into disjoint subsets, called *clusters*, using a resemblance measure. The main goal of the clustering identification is puting a new submitted plant X in to one of these clusters.

2.2 Rank of Links

Let us calculate the distances between every pair of m plants within the subset $E = \{X_i\}$ obtaining the *cross-distance matrices* $\|D_{i,j}\| i = 1\text{–}m, j = 1\text{–}m$. The symmetry property of the

		F	E	A	T	U	R	E	S
		1	2	3	.	.	.	.	n
O	X1	x_{11}	x_{12}	x_{13}	.	.	.	.	x_{1n}
B	X2	x_{21}	x_{22}	x_{23}	.	.	.	.	x_{2n}
J	X3	x_{31}	x_{32}	x_{33}	.	.	.	.	x_{3n}
E	.					.			
C	.					.			
T	.					.			
S	Xm	x_{m1}	x_{m2}	x_{m3}		.	.	.	x_{mn}

FIGURE 1 A feature table.

metric $d(i,j) = d(j,i)$ permits us to calculate only one half of the distance matrix. The matrix $\|D_{i,j}\|$ defines the relative positions of the points within the subset $\{X\}$ of the given features space.

Align every row of this matrix by growing. Note that the first element of the aligned row is zero, that is, $D_{i,1} = 0$. Let us use the following rule for numbering of the aligned row:

1. The number of the first element of the aligned row is 0, that is, $\text{Num}(D_{i,1}) = 0$.
2. The number of the element $D_{i,j}$, which is the next to the element $D_{i,k}$, is calculated as follows:

$$\text{Num}(D_{i,j}) = \text{Num}(D_{i,k}) + 1 \quad \text{If } D_{i,k} < D_{i,j}$$
$$\text{Num}(D_{i,j}) = \text{Num}(D_{i,k}) \quad \text{If } D_{i,k} = D_{i,j}$$

Based on this rule and using $\|D_{i,j}\|$, we may construct a new matrix $\|\text{Num}_{i,j}\|$ with non-negative and bounded ($\text{NUM}_{i,j} < m$) elements.

DEFINITION 1 *The operator "rank of link" between X_i and X_j elements is defined as*

$$\text{rank}(X_i, X_j) = \text{NUM}_{i,j} \tag{1}$$

The matrix of distances between 4 points is given below as an example. Figure 2 shows the process of the rank matrix construction from the distances matrix for X_2.

Denote the obtained rank-of-links matrix by $\|\text{rank}(X_i, X_j)\| i, j = 1\text{–}m$. This matrix is invariant to the choice of the origin point and has the following properties.

LEMMA 1 *The operator "rank of link" is invariant to a linear transformations of a compression and stretching.*

Proof A linear compression and stretching is realized by multiplying the matrix of distances $\|D_{ij}\|$ by $\lambda = \text{const}$. Hence, since the multiplication by the scalar λ the elements of the matrix $\|D_{ij}\|$ doesn't change regularity of distances in the rows, thus the rank of the links matrix is not changed. ■

Remark 1 The operator "rank of links" is reflexive, that is, $\text{rank}(X_i, X_i = 0) = 0$.

Remark 2 The operator "rank of links" is not symmetric. Indeed, consider 3 points in Figure 2 for which $\text{rank}(X_1, X_2) = \text{rank}(X_2, X_1) = 1$. It is symmetry, but the $\text{rank}(X_2, X_3) = 2$ and $\text{rank}(X_3, X_2) = 1$. It means that the points X_2, X_3 have not the property of symmetry.

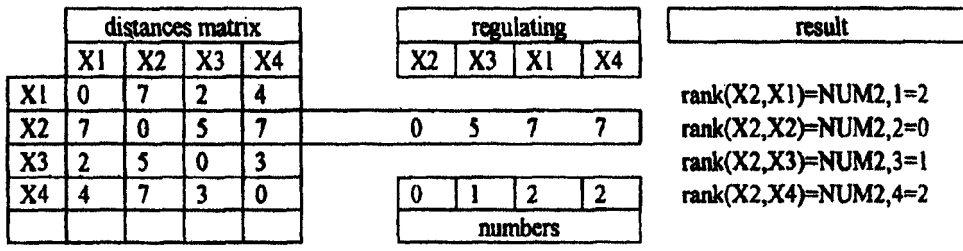


FIGURE 2 The rank matrix construction from the distance matrix for X_2.

DEFINITION 2 *Value of the "closed rank of link" between* $X_i, X_j \in E$ *points is equal to* $\mathrm{Sup}\{\mathrm{rank}(X_i, X_j), \mathrm{rank}(X_j, X_i)\}$ *and marked as* $\mathrm{Rank}\lfloor X_i, X_j \rfloor$.

Closed rank of link satisfies the conditions of reflectivity and symmetry:

$$\mathrm{Rank}[X_i, X_i] = 0, \quad \mathrm{Rank}\lfloor X_i, X_j \rfloor = \mathrm{Rank}\lfloor X_j, X_i \rfloor.$$

Example 1 For the simplest situation depicted at Figure 3 we have:

$$\mathrm{Rank}[X_1, X_2] = 1; \quad \mathrm{Rank}[X_3, X_2] = \mathrm{Rank}[X_2, X_3] = 2.$$

2.3 Clusters Representation

DEFINITION 3 *The subset* $\{X_1\} \in \{X\}$ *is said to be "closed by r-rank" if it fulfills 3 following conditions:*

1. *For every* $X_{1i} \in \{X_1\}$ *point there exists at least one* $X_{1j} \in \{X_1\}$ *point such that* $\mathrm{Rank}\lfloor X_{1j}, X_{1j} \rfloor \leq r$;
2. *For every* $X_{1i} \in \{X_i\}$ *and* $X_j \in \{X\}\backslash\{X_1\}$ *points* $\mathrm{Rank}\lfloor X_{1j}, X_{1j} \rfloor > r$;
3. *There exists at least one pair of* $X_{1j}, X_{1j} \in \{X_i\}$ *points such that*\ $\mathrm{Rank}\lfloor X_{1j}, X_{1j} = r$.

The number r is called *the closing rank* for $\{X_1\}$ subset. Consider some subset of $\{X\}$ set, closed by rank 1. Denote it by $\{X[1]\}$. Form a new subset by adding new points to $\{X[1]\}$ such that for every new X_{new} point $\mathrm{Rank}[X_{\mathrm{new}}, X_i] \leq 2$ ($X_i \in \{X[1]\}$). Denote this new set by $\{X[2]\}$. Iterating this procedure, we can get $\{X[3]\}$ from $\{X[2]\}$, $\{X[4]\}$ from $\{X[3]\}$, etc.

DEFINITION 4 *The rank of links are called "non-break" by r on the subset* $\{X[r]\}$ *if*

$$\{X[i]\}\backslash\{X[i-1]\} \neq \emptyset, \quad i = 2\text{–}r \tag{2}$$

The value l is called a "missing of ranks" for the subset $\{X[r]\}$ *if* $\{X[i]\}\backslash\{X[i-l]\} \neq \emptyset$ $(i = 2\text{–}r)$, $\{X[r]\}\backslash\{X[r+j]\} \neq \emptyset (j = 1\text{–}l)$, *and* $\{X[r]\}\backslash\{X[r+l+1]\} \neq \emptyset$.

If condition (2) does not fulfill we will say that the "breaking" of rank of links on the subset $\{X[r]\}$ takes place.

DEFINITION 5 *A subset* $\{X_c\} \in \{X\}$ *of points is called cluster if there is "hon-break" r rank of links on the* $\{X_c\}$ *subset and beginning from* $r + l$ *there exists at least one "missing of rank".*

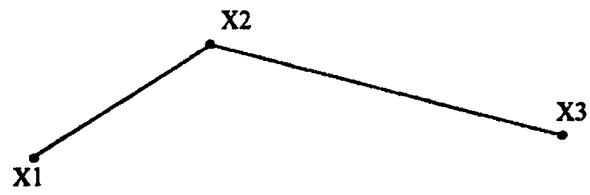


FIGURE 3 Rank of links illustration.

The fact that $\{X_c\}$ subset is cluster we will mark as $C = \{X_c\}$. Two properties hold for the existing cluster: $\{X_c[r + l]\} \backslash \{X_c[r]\} \neq \emptyset$ and

$$\{X_c[i]\} \backslash \{X_c[i - 1]\} \neq \emptyset \quad (i = 2\text{–}r) \tag{3}$$

The corresponding number r is called *the rank of cluster forming*. For every cluster C the following descriptions are suggested to be in use: set of the cluster's points – $\{X_c\}$, number of the cluster's points – m, rank of forming of cluster – r, number of missing of rank – l.

DEFINITION 6 *The cluster C_1 is said to be strictly isolated on set of $\{X\}$ if*

$$r_1 + l_1 \geq m_i - 1; \quad i_1 \geq r_1 \tag{4}$$

It means that the cluster C_1 is strictly isolated, if the closed rank of links between any pair of $X_i, X_i \in C_1$ is not greater than m_1.

2.4 Identity and Isolation of Clusters

If the cluster C consists only coincident points, then the rank of forming of cluster is equal to zero. If the cluster contains only points of the even grid (Fig. 4a), rank of forming is equal to 1.

If any X_i point form the even grid (Fig. 4a) is moved and distance between this point and the nearest neighbor is less than step of grid (the degree of identity is growing (Fig. 4b)), then rank of forming of this cluster is equal to two. Consider the Figure 5 containing an odd set of the points on the one-dimensional space such that $d_{1,j+1} > d_{1j}$; $i = 1\text{–}m$ (m is the number of points).

The neighbors of these points $m - 1$ rank of links are required. Note that $m - 1$ rank of link is the maximum possible one for every set of points. One can see that if the unevenness of the cluster is growing then the forming rank of the cluster is growing too. Thus, the rank of a cluster forming can be used to describe unevenness of cluster.

THEOREM 1 *More missing of rank for a given cluster corresponds to a more degree of isolated clusters quality from other set of points.*

Proof Let there be two clusters C_1 and C_2 on the set of $\{X\}$ with parameters $(r_1; l_1, m_1)$ and (r_2, l_2, m_2) (Fig. 6). The minimum distance between points of these clusters are $d\{X_i, X_j\}$, $X_i \in C_1$; $X_j \in X_2$, thus

$$d(C_1, C_2) = d(X_i; X_j) \tag{5}$$

Besides, we have

$$m_1 > r_1 + l_1; \quad m_2 > r_2 + l_2 \tag{6}$$

FIGURE 4a, b Moving within a point grid.

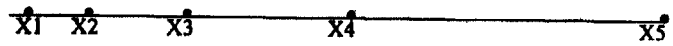


FIGURE 5 Aligned set of points.

Let us construct two spheres with 2 centers X_i, X_j and diameters $d(X_i, X_j)$ and mark by $X_n \in C_1$ the point, which is nearest to the first sphere, and then mark by $X_k \in C_2$ the point, which is nearest to the second sphere. The distances between these points and its sphere are $\Delta X_n = d(X_i, X_n) - d(X_i, X_j)$ and $\Delta X_k = d(X_k, X_j) - d(X_i, X_j)$. According to the previous definitions, $r_1 + l_1 + l = \text{rank}(X_i, X_j)$ and $r_2 + l_2 + l = \text{rank}(X_j, X_i)$. Let move cluster C_1 such that its structure will remain unchangeable. Then the distance between clusters will increase by $\max(\Delta X_k, \Delta X_n)$. Mark new parameters with *. So, after moving we obtain $d^*(i,j) = d(i,j) + \max(\Delta X_k, \Delta X_n)$; $r_1^* + i_1^* + l = \text{rank}^*(X_i, X_j)$; $r_2^* + i_2^* + 1 = \text{rank}^*(X_j, X_i)$; that implies, $\text{rank}^*(X_i, X_j) = \text{rank}(X_i, X_j) + 1$; $\text{rank}^*(X_j, X_i) = \text{rank}(X_j, X_i) + 1$; and $r_1^* + i_1^* + l = r_1 + l_1 + 1$; $r_2^* + i_2^* + l = r_2 + l_2 + 1$. According to the Lemma 1, $r_1^* = r_1$; $r_2^* = r_2$; and, hence, $l_1^* = l_1 + 1$; $r_2^* = r_2 + 1$, that completes the proof. ■

In the same way, we can show that if the distance between the nearest points of different clusters decreases, the "missing of rank" of each cluster will also decrease. An important result follows from Theorem 1: we can define the isolating degree of cluster by value of "missing of ranks".

2.5 Some Cluster Characteristics

The descriptive cluster parameters l, r and m provide the possibility to define correctly *the degree* of isolated cluster.

THEOREM 2 *If Cluster C_1 is strictly isolated, then the maximum distance between points of Cluster C_1 is smaller than distance between them and other points of $\{X\}\backslash C_1$ set.*

Proof Consider Figure 7, where $X_j \in \{X\}\backslash C_1$ point is the nearest to the X_i point of Cluster C_1, and $X_k \in C_1$ point is the nearest to the point X_i. We wish to show that the following inequality is valid: $d(X_i, X_j) > d(X_i, X_k)$, where $X_i, X_k \in C_1$. By Definition 6, in Cluster C_1 it is possible to fulfill $(r_1 + l_1)$ number rank of links for any point of cluster. But, if the number of points m_1 in C_1 cluster is less than value $(r_1 + l_1)$,

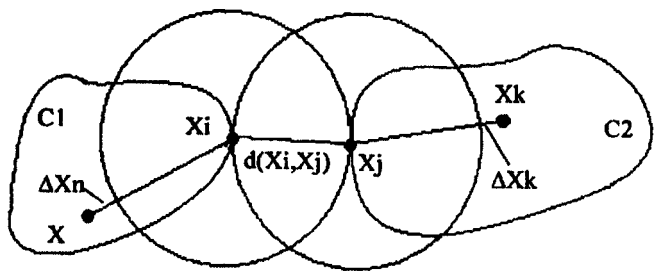


FIGURE 6 Clusters and a set of points.

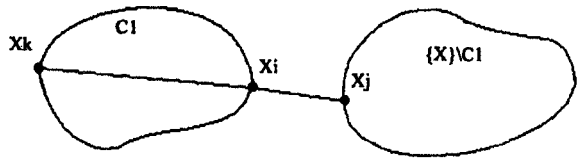


FIGURE 7 Illustration to Theorem 2.

then fulfilled ranks of links values are not more than $(m_1 - 1)$ value. Let us suppose that $\text{rank}(X_i, X_k) = m_1 - 1$. Considering that X_i point is the nearest for the point X_i from the set of points $\{X\}/C_1$, we obtain $\text{rank}(X_i, X_j) = r_1 + l_1 + l > \text{rank}(X_i, X_k)$. The definition of rank of links together with the image above implies $d(X_i, X_k) < (X_i, X_j)$. ■

THEOREM 3 *If C_α cluster is strictly isolated, then the inequality $l_\alpha \geq (m_\alpha - 1)/2$ holds.*

Proof For the strictly isolated clusters we have the following system of inequalities:

$$r_\alpha + l_\alpha = m_\alpha - 1, \quad l_\alpha \geq r_\alpha$$

Summing both inequalities implies:

$$r_\alpha + 2l_\alpha \geq m_\alpha - 1 + r_\alpha \Rightarrow \frac{m_\alpha - 1}{2}.$$ ■

We will say that C_α cluster is *well isolated* in $\{X\}$ set of points, if the following inequalities hold:

$$l_\alpha \geq r_\alpha; \quad r_\alpha + l_\alpha < m_\alpha - 1$$

Let us suppose that all other clusters are simply isolated. Then a cluster, which consists of one point, is strictly isolated (well isolated or isolated) from the set of other points, if the nearest cluster is strictly isolated (well isolated or isolated). A set $\{X\}$ of points is usually said to be *unstructured* (doesn't contain individual clusters), if after the process of clustering there is only one cluster.

To solve the tasks of pattern recognition, the following question is of great importance: *how are different kinds of objects isolated from each other*? This represents a problem of compactness.

DEFINITION 7 *Any cluster is compact if it consists of only one kind of objects. A cluster is called strictly compact, compact or weakly compact if it is strictly isolated, well isolated or isolated.*

By definition of compactness in the given feature space a pattern could consist of more than one cluster. These kinds of clusters could be situated near each other (Fig. 8) or could be separated by other kinds of clusters (Fig. 9). These pictures show compact cluster C_1 of A_1 pattern and the clusters C_2, C_3 of A_2 pattern.

In the n-dimensional space a reciprocal situation of clusters is not so obvious as in the two-dimensional space. Thus, it is necessary to detect situation with one kind of clusters. Is it possible to unite these clusters and group from then one cluster? It is clear that for the situation, which is given in Figure 8, it is possible, and for that given in Figure 9 it is impossible.

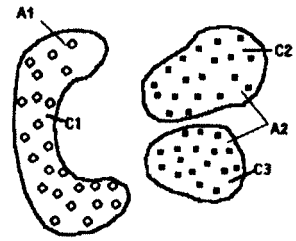


FIGURE 8 A pattern with several clusters.

Let us increase value of rank of links for the points of one of the cluster from $r_1 + l_1$ rank. In this case there will be new points in the given cluster. If these points are the points of the same kind of clusters, then we can unite them and this cluster is called "*consisted cluster*". If new points are the points of different kind of clusters, then this kind of uniting is impossible.

For the determination of missing of ranks for the "consisted cluster" it is necessary to increase the value of ranks until other kinds of point are united in the "consisted cluster". Let us mark rank of forming of "consisted" cluster as R^*. For the value of missing of ranks let us mark it as L. We have $L = R^* - R$, where R represents the rank of forming of the first cluster. So, the number of points in the consisted cluster is equal to the sum of the number of the points of the initial clusters.

2.6 Algorithm of Clustering

The process of clustering, using rank of links, has some specificity, that must be provided for in the process of building of algorithms, that is, for any set of points there exists at least one couple of points connected with closed rank of link equal to one. Note that this kind of couple of points is situated in one cluster; it gives us possibility to start the process of clustering from this couple of points.

In view of this note, we may suggest the following algorithm of clustering which consists of the next stages:

Step 1 Calculate matrices of distances and ranks of links.
Step 2 Define $C_1, C_2, \ldots, C_n$ subsets of points closed with zero and one rank.
Step 3 Define new subsets closed by next rank, adding new points to the $C_1, C_2, \ldots, C_n$ subsets.

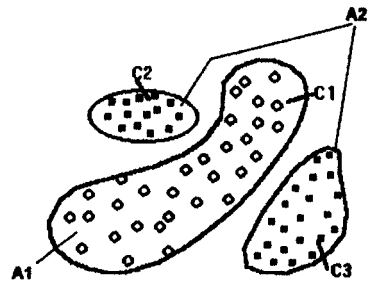


FIGURE 9 Ungrouped clusters.

Step 4 Define subsets number of points, which have not grown. Form new clusters for these subsets and exclude relative points from further review.

Step 5 If the sum of number of all formed cluster points is less than number of initial points go to step 3.

Step 6 Stop.

In the next sections we will discuss the application of the suggested technique to some concrete problems.

3 ESTIMATION AND PROGNOSIS OF DECISION MAKING RELIABILITY BY "RANK OF LINKS"

The correct decision-making in pattern recognition depends on variety of factors such as choosing or formation of feature space. It should be mentioned, that in this section we do not consider the feature choosing or estimation problem in relation to recognition reliability or in any other point of view. The feature space is assumed to be given and the recognition reliability estimation and prognosis problem is considered according to the rules and methods of the decision-making. Besides, we consider only the decision making rules formed by the rank of links method that are presented by Duda and Hart [4]. In the sequent we will call "*realizations*" the outcomes obtained by measuring the pattern features. Thus, a point corresponds to each realization in the feature space, and a cluster in this space represents a collection of the points (realizations). The realization interlocation in the feature space gives the complete information of the patterns isolation or intersection. It is natural, that the farther the realizations of a pattern from another pattern's realizations, the less the possibility of error in recognition of unknown realization of the pattern, and, conversely, if the realizations of different patterns are located together, *i.e.*, the patterns intercast, then in the intersection area the error making possibility increases. Therefore it is possible, in principle, to make prognosis of the recognition results if the isolating degree is known and what really matters, it is represented in quantitative terms. In such cases the reliability prognosis problem may be treated as the determination of regularities, which link the pattern isolating degree and the error making possibilities in the recognition process. Here, certain circumstances should be taken into account, which, as a rule, are of heuristic nature and are known beforehand. *Example* whether it is admissible to make multivalued decisions, rejection of the recognition, whether the admissible level of errors exists, etc. It is natural that, according to these circumstances, we will have various methods and indicators of estimation and prognosis.

3.1 The Problem Formalization

Let a pattern set $\{A\}$ be given for which the realization representative set $\{X\}$ exists. Any realization from the space $\{X\}$ is obtained by measuring the feature set $\{X\}$. Denoting by N the number of features, for the realization set $\{X\}$ we will have N-dimensional Euclidean feature space where each realization may be represented by one of points. The collection of points, isolated from other points in the sense given above, is a cluster. Denote by $\{C\}$ the set of clusters obtained by grouping the points of set $\{X\}$, by r_k the rank of cluster creation for any $C_k \in \{C\}$, by l_k the number of rank omissions, and by m_k the number of realizations in this cluster. The set of clusters, contained in cluster C_k, denote by $\{X_k\}$, and the set of realizations of type $A_j \in \{A\}$, denote by $\{X_j\}$. If the cluster C_k contains only realizations of type A_j, then we have $\{X_k\} \subset \{X_j\}$. If, moreover, the condition $\{X_k\} = \{X_j\}$ is valid, then the pattern A_j is compact. Note that compact patterns may have several clusters consisting of only the

realizations of this type. If the realizations of several patterns are integrated in any cluster then these patterns are noncompact.

Figure 10 shows the clusters C_1, C_2, C_{22} that correspond to the compact patterns A_1 and A_2, and the cluster C_{34} corresponding to the noncompact patterns A_3 and A_4. The shaded area of the cluster C_{34} gives the intersection of patterns A_3 and A_4. Any number of patterns from the set $\{A\}$ may intersect.

Denote by "CLS" the procedure by means of which the cluster set $\{C\}$ is obtained from the elements of the set $\{X\}$. Then the clustering process may be represented by the following expression:

$$\text{CLS: } \{X\} \rightarrow \{C\} \tag{7}$$

If we take into consideration that the set $\{X\}$ represents the realizations of the pattern set $\{A\}$, denoting this fact by $\{A[X]\}$ and that the cluster set $\{C\}$ may be represented by the set of its parameters $\{m, r, l\}$, then (7) takes the following form

$$\text{CLS: } \{A[X]\} \rightarrow \{C[m, r, l]\} \tag{8}$$

Denote by $\{k\}$ the elements of the set the prognosis estimations of the correct recognition of each pattern from the pattern set $\{A\}$. For the elements of the set $\{k\}$ let us specify the "advantage" of widespread estimates:

1. The percentage – in this case the elements of $\{k\}$ take values $0 \leq k_i \leq 100$; $i = 1$–l where i is the number of patterns in the set $\{A\}$;
2. The *probability* – in this case we have: $0 \leq k_i \leq 1$; $i = 1$–l.

Denote by "BWT" the recognition reliability estimating process. Then like expression (2), the process of obtaining the elements of $\{K\}$ may be described as

$$\text{BWT: } \{C[m, r, l]\} \rightarrow \{k\} \tag{9}$$

Thus, in order to estimate the recognition process reliability, the following initial information is needed: the set of patterns $\{A\}$, the realization study collection $\{A[X]\}$ and the clustering procedure – CLS. The intermediate procedure is the cluster set assignment $\{C[m, r, l]\}$ obtained by the procedure CLS, and the final step is the determination of the values of the

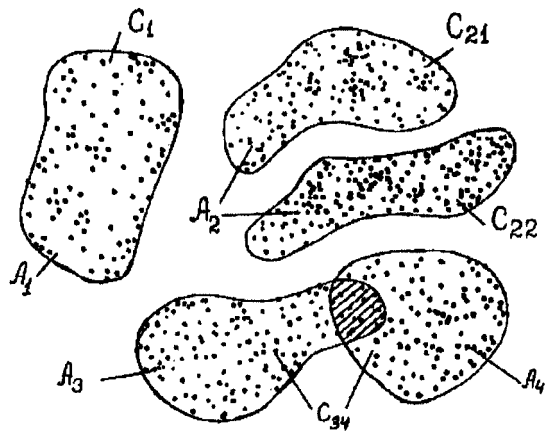


FIGURE 10 The compact patterns A_1 and A_2 and the noncompact patterns A_3 and A_4.

elements from the set $\{k\}$ by means of the BWT procedure (algorithm). The aim of this section is to develop the BWT procedure and to demonstrate its efficiency.

3.2 BWT Procedure

3.2.1 Forecasting the Reliability of Decisions, Obtained by Clustering Process for Compact Patterns

Denote by "ERK" the decision making process in recognition of unknown realization. Then decisions, obtained by clustering, may be expressed by the symbols ERK{CLS}. According to this notation the estimate determination process may be described by the following expression:

$$\text{BWT[ERK\{CLS\}]} = \{k\} \tag{10}$$

In order to formalize the process given by (10) it is necessary to state several axioms.

AXIOM 1 *In given feature space the cluster topology is assumed to be formed and does not depend on the observer's estimate;*

AXIOM 2 *A realization, integrated in any compact pattern cluster, may or may not belong to the given pattern.*

Assume that in pattern set $\{A\}$ there exist two compact patterns A_j and A_j. Each of them creates the clusters $C_j(r_j; l_j; m_j)$ and $C_j(r_j; l_j; m_j)$ respectively, in the feature space (Fig. 10). The unknown realization recognition process, like the clustering process, is carried out by means of the rank links that exist among the points either integrated in the same cluster or belonging to different clusters. Denote by $\{X_j\}$ the set of points integrated in the cluster C_j. Then for any points, integrated in the cluster C_j, the following inequality holds

$$\text{Rang}\{\forall \text{Xj} \in \text{Cj};\ \text{Xjq} \in \text{Cj}\} \leq \text{rj} + \text{lj}, \quad \text{q} = 1 - (\text{rj} + \text{li}) \tag{11}$$

If an unknown realization belongs to cluster, *e.g.*, the point X_1 (Fig. 11) – to the cluster C_j, then instead of expression (11) we will have the following decision making procedure:

$$X' \in A_j \quad \text{if } \text{Rang}\{X'; X_{jq} \in C_j\} \leq r_j + l_j, \quad q = 1 - (r_j + l_i) \tag{12}$$

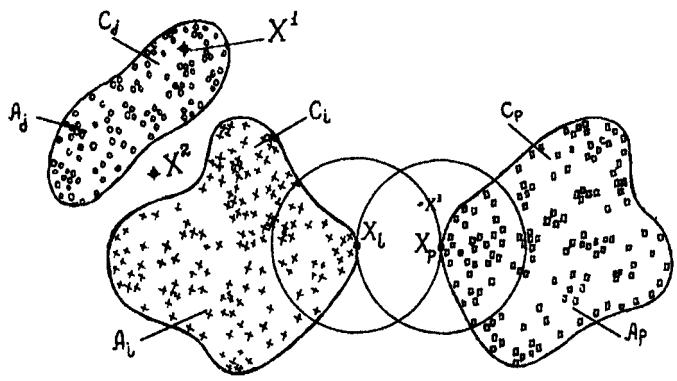


FIGURE 11 The identical links with two or more compact pattern clusters.

According to Axiom 1, if the condition of expression (12) is valid, then the recognition is precise; that enables us, while forecasting the recognition reliability, to give to this situation the highest grade equal to $\max\{k\} = K$, where K is a forming factor.

The most ambiguous situation in the recognition point of view will occur when the unknown realization has identical links with two or more compact pattern clusters, *e.g.* the point X_2 (Fig. 11), in this case we have

$$\mathrm{ERK}\{C_j(r_j; l_j; m_j)\} = \mathrm{ERK}\{C_i(r_i; l_i; m_i)\} \tag{13}$$

In case of rank links expression (13) takes the following form:

$$\mathrm{Rang}\{X^2; X_{jq}\} = \mathrm{Rang}\{X; X_{i,q}\} \leq \min\{(r_j + l_j), (r_i + l_i)\} = Q \tag{14}$$

where $q = 1\text{–}Q$. The realization of expression (14) means that in clusters C_j and C_i points are located symmetrically to the unknown realization X^2. It is natural that in this case the preciseness of recognition or of adopted decision is minimal, that should give the minimal value of its estimation or prognosis equal to $\min\{k\}$. Note that if it is possible to give up the recognition in cases when all patterns are compact and condition (12) is not valid for any pattern, then according to Axiom 2 we will have precise recognition. For cases, when it is necessary to make decision while condition (12) is not valid for any pattern, then it is necessary to make into consideration the cluster interlocation or their parameters.

DEFINITION 8 *The influence zone of any cluster $C_j \in \{C\}$ is the area, where at least one of the points integrated in the cluster establishes closed rank with a point located outside the cluster not exceeding the rank link value $(r + l)$.*

Denote by symbol "$\forall$" the *predicate* "there exists at least one" and by "$\bar{\forall}$" symbol its negation "there does not exist any". Denote by $\mathrm{Ers}(\mathrm{Rang}\{X; X_j \in C_j\})$ the first closed rank link that is established by the unknown realization X with a certain point of cluster C_j. Then the decision making of belonging of unknown realization may be carried out only for the patterns to whose corresponding cluster influence zone belongs the realization. For such clusters the following property is valid

$$\mathrm{Ers}(\mathrm{Rang}\{X; \forall X_i \in C_i\}) \leq r_i + l_i; \quad i = 1\text{–}I_1, \tag{15}$$

where $I_1 < I$ is the number of the patterns to whose corresponding cluster zone belongs the unknown realization. For the decision making rule we have

$$X \in A_j \quad \text{if } \mathrm{Ers}(\mathrm{Rang}\{X; \forall X_j \in C_j\}) = \min\{\mathrm{Ers}(\mathrm{Rang}\{X, \forall X_i \in C_i\})\}, \quad i = 1\text{–}I_1, \tag{16}$$

DEFINITION 9 *The patterns influence zone is the union of their corresponding cluster zones. They are disjoint, if there does not exist a point belonging simultaneously to influence zones of the clusters that correspond to the given patterns.*

The pattern influence zones in Euclidean space may have extremely complicated forms and locations. Therefore in order to formalize. Definition 9, it is necessary to find out the influence zones of clusters for which there exists potential possibility of intersection. In the case of clusters located in the plane, *e.g.*, (Fig. 11), the intersection of influence zones of clusters C_j and C_i may happen in the area where the point X^2 is located. In order to determine

such area, *e.g.*, for clusters C_i and C_p (Fig. 11), let us evaluate the minimal closed rank link by which a point from one cluster is linked to a point from another cluster. There are such points: $X_i \in C_i$ and $X_p \in C_p$ (Fig. 2), for which the following conditions holds

$$\mathrm{Ers}(\mathrm{Rang}\{C_i; C_p\}) = \mathrm{Rang}\{X_i, X_p\} = \min(\mathrm{Rang}\{\forall X_\alpha \in C_i; \quad \forall X_\beta \in C_p\})$$

Here, $\mathrm{Ers}(\cdot)$ is the minimal rank link established by a point of cluster C_i to a point of cluster C_p. Assume that point X^3 is between points X_i and X_p, that means that it belongs to the intersection of the two hyper spheres having radii the distance between points X_i and X_p, and these points as centers. The interpretation of this situation is given in Figure 11, in case of two dimensional Euclidean space, namely, the shaded area, where point X^3 is located, represents the intersection area. Denote this domain by C_{ip}. If for such points the condition from Definition 9 is valid then the following inequality systems hold:

$$\begin{cases} \mathrm{Ers}(\mathrm{Rang}\{\forall X^3 \in C_{ip}; \forall X_i \in C_i\}) \le r_i + l_i \\ \mathrm{Ers}(\mathrm{Rang}\{\forall X^3 \in C_{ip}; \bar{\forall} X_p \in C_p\}) < r_p + l_p \end{cases} \tag{17}$$

$$\begin{cases} \mathrm{Ers}(\mathrm{Rang}\{\forall X^3 \in C_{ip}; \forall X_p \in C_p\}) \le r_p + l_p \\ \mathrm{Ers}(\mathrm{Rang}\{\forall X^3 \in C_{ip}; \bar{\forall} X_i \in C_i\}) < r_i + l_i \end{cases} \tag{18}$$

In the first case the point X^3, according to Definition 1, is in the influence zone of cluster C_i, and in the second case it is in the influence zone of cluster C_p.

Assume that we have the situation described by (17). Then the unknown realization can be conferred on the pattern A_i, but quantitative estimation of the precision of such conferment will be less than that of the decision obtained before. We express the estimation of the precision of the decision obtained the following equality:

$$\mathrm{BWT}(X^3 \in A_i) = ((r_i + l_i)) - \frac{\mathrm{ERS}(\mathrm{Rang}\{X^3; X_i \in C_i\})}{r_j + l_j} K \tag{19}$$

where $i = 1, \ldots, m$, and $K = \max\{k\}$ represents the norming factor and depends on the estimation scale chosen by us. The same can be written when estimating the situation given by expression (17) for the pattern A_p:

$$\mathrm{BWT}(X^3 \in A_p) = (r_p + l_p) - \frac{\mathrm{ERS}(\mathrm{Rang}\{X^3; X_p \in C_p\})}{r_p + l_p} K \tag{20}$$

In the case of the cluster influence zones intersection, *e.g.*, for the patterns A_j and A_i, conditions, given by the following inequality system, are valid:

$$\begin{cases} \mathrm{Ers}(\mathrm{Rang}\{\forall X; X_j \forall X_j \in C_j\}) \le r_j + l_j \\ \mathrm{Ers}(\mathrm{Rang}\{\forall X; X_i \in X_i \in C_i\}) \le r_i + l_i \end{cases} \tag{21}$$

In this case, the unknown realization can be conferred on to the pattern A_j as well as to A_i. According to expression (19) for the estimators of decision making precision we will have: $K_j = \mathrm{BWT}(X \in A_j), K_i = \mathrm{BWT}(X \in A_i)$. Taking into account the decision making rule given by the expression (16) we obtain $X \in A_j$ if $K_j \in K_i$ and $X \in A_i$ if $K_i \in K_j$.

THEOREM 4 *The pattern influence zones do not intersect, if the minimal closed rank link between the clusters corresponding to the patterns is greater than the sum of the cluster construction and rank omissions.*

Proof Denote by $G\{A\}$ the cluster influence zones. Then the condition of this theorem for the clusters A_j and A_i takes the following form

$$G(A_j) \wedge G(A_i) = \emptyset \quad \text{if } \mathrm{Ers}(\mathrm{Rang}\{X_j \in C_j; X_i \in C_i\}) > r_i + l_i + r_j + l_j \tag{22}$$

Assume that we are given the clusters that correspond to the patterns A_j and A_i as well as the points $X_i \in C_i$ and $X_j \in C_j$, that realize the minimal closed rank link between the clusters (Fig. 12).

$$\mathrm{Ers}(\mathrm{Rang}\{X_j; X\}) = R_{ij} \tag{23}$$

According to the expression (23), we get the following form for (22):

$$R_{ij} > (r_i + l_i) + (r_j + l_j) \tag{24}$$

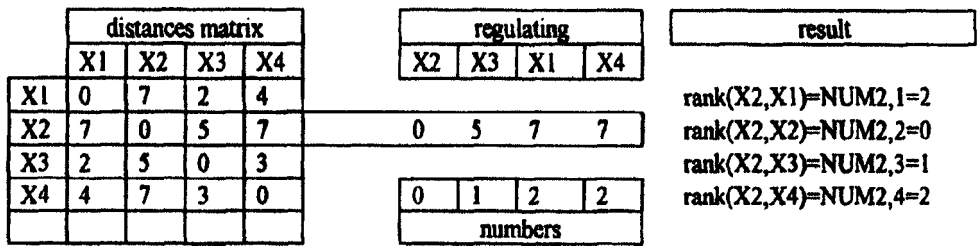
As one rank link establishes relationship between only two points, therefore any point, *e.g.*, an unknown realization, should establish rank link of value $(r + l)$ with $(r + l)$ points. As $\mathrm{Rang}\{X_j, X_i\} = \max\ \mathrm{Rang}\{X_j; X_\alpha \in C_j\}a = 1\text{–}m_j$ (Fig. 12), we have that for the cluster of the pattern A_j the maximum capacity area for establishing link with an unknown realization is equal to the space area confined by the hyper sphere with the radius equal to the distance between the points X_j, and X_e and center X_j, but for the pattern A_j that of the distance between X_i and X_g and center X_i. These hyper spheres intersect only in the case if $R_{ij} < (r_j + l_j) + (r_i + l_i)$, and do not intersect $R_{ij} \geq (r_j + l_j) + (r_i + l_i)$. ■

The greater the intersection area, the greater the error making possibility when recognizing A_j and A_i pattern realizations and vice versa. In this case, for the estimation of recognition reliability we have

$$\mathrm{BWT}(A_j; A_i) = \frac{(r_j + l_j) + (r_i + l_i) - R_{ij}}{(r_j + l_j) + (r_i + l_i)} K \tag{25}$$

If the condition, given by expression (18), is valid, then, according to (25), we obtain the negative values of estimators that are unacceptable for any scale of reliability estimation. Thus, for this situation we can use the following expression

$$\mathrm{BWT}(A_j; A_i) = \left(\frac{R_i - (r_j + l_j + r_i + l_i)}{R_i}\right) \cdot K \tag{26}$$



Expressions (25) and (26) allow us to estimate the reliability of decision making for one pattern in respect with others, in the meaning that for any pattern we have the set of estimators of $(I-1)$ elements. If required, it is possible to average these estimators obtaining one value of an estimator or a scalar.

3.2.2 Prognosis of Recognition Reliability for Noncompact Patterns

Assume that the patterns A_j and A_i intersect, that is, the realizations, corresponding to these patterns, create the same cluster (Fig. 13) denote it by C_{ij}:

Naturally in this case the pattern influence zones intersect significantly, but for their determination the existence of separate isolated clusters is necessary that is not the case in the given situation. Therefore we have to work out another method for the prognosis of the recognition reliability.

DEFINITION 10 *Any pair of points represents neighbors in a cluster if the closed rank link between them does not exceed the rank link value of the cluster formation.*

Let the points given by Definition 10 be $X_j \in C_{ij}$ and $X_i \in C_{ij}$ (Fig. 13). Then according Definition 10, it follows

$$\text{Rang}((X_j \in A_j); (X_i \in A_i)) \leq r_{ij} \tag{27}$$

If the neighboring points $X_j \in A_j$ and $X_i \in A_i$ belong to different patterns, then these points participate (belong) in the patterns intersection.

Denote by m_{ji} the number of points of the pattern A_j that participate in the intersection with the pattern A_i, and that of the pattern A_i by m_{ij}. Reliability of precise recognition of the unknown realization in the intersection area is minimal, but besides the intersection area there may exist in a cluster such an area where the neighboring points belong to only one pattern, that implies that the patterns do not intersect in this area. *Example*, in Figure 13 such are the areas outside the shaded stretch. For the prognosis of the decision making reliability we should take into account the relationship of these areas, that can be realized by the following relationships:

$$\text{BWT}(A; A_i) = \frac{m_{ji}}{m_j}; \quad \text{BWT}(A_i; A_j) = \frac{m_{ij}}{m_i} \tag{28}$$

Note that the values of the estimators obtained by expression (27) belong to the range 0–k. Besides, the form of these expressions will not change by the increase of patterns that participate in the intersection.

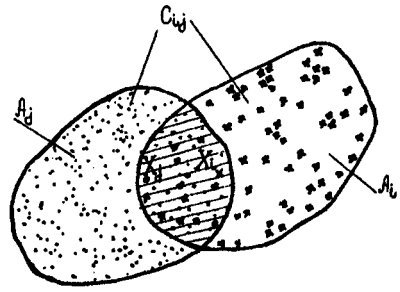


FIGURE 13 The Intersecting patterns.

If the realizations that participate in intersection are taken out of study collection, then noncompact patterns transform into compact ones and the method from the previous paragraph may be applied.

3.3 Algorithmization of Prognosis and Estimation Process

The decision making and estimation problem can be divided into several consecutive procedures, that allows the algorithm realization of this process. Thus, if we are given a pattern set and for each pattern the study collection $\{X_j\} \in \{X\}\ j = 1\text{–}I$ of realizations is known, then for carrying out the outlined problem we have the following sequence of required procedures:

1. Calculation of the distances matrix between any pair of realizations from the study collection;
2. Calculation of rank links matrix using the distances matrix;
3. Formation of the clustering process;
4. Determination of the clusters and their parameters;
5. Determination of intersection areas and of realizations located in these areas for any cluster;
6. Evaluation of the recognition reliability estimators for each pair of patterns that participate in the intersections;
7. Determination of the influence zones for disjoint patterns;
8. Evaluation of the recognition reliability estimators for each pair of disjoint patterns;
9. Calculation of an integrated (scalar) estimator of the recognition reliability for any pattern.

According to the item 9, the obtained estimators are final, but the result depends on the method of working out of an integrated estimator. The most widespread method is the arithmetic mean calculation of the separate estimators. In the other cases, it is possible to take into consideration the most "dangerous" patterns as regards of precise recognition or averaging the estimators obtained concerning several estimators of this type.

The most widespread scale of estimators is the percentage scale, with the norm coefficient $K = 100$. The probabilistic scale is wide spread as well as with the norm coefficient being equal to 1. The choice of the norm coefficient depends on a specific problem and the scale accepted in a given field.

4 CONCLUSIONS

The use of clustering process for the prognosis of decision-making results may be considered as an original method in pattern recognition.

Experimental research has shown the good coincidence of the prognosis with the results obtained by recognition of check realizations that allows us to confirm the effectiveness and good prospects of the given method for cases when the rank link method is used for clustering and decision-making process as well.

The system is implemented in some programming languages. Data-Lab system has been operational for four years and it has been used in the undergraduate image-processing course, and as a platform for different MS and Ph.D dissertation projects. In addition, it is in everyday use within a university signal and image processing research group. The following developments may be related to some nets analysis and their descriptions such as neural and wavelet networks.

References

[1] Bouguettaya, A. and Le Viet, Q. (1998) Data clustering analysis in a multidimensional space, *Information Sciences*, **112**(1–4), 267–295.
[2] Campbell, J., Murtagh, F. and Kokuer, M. (2001) DataLab-J: A signal and image processing laboratory for teaching and research, *IEEE Transactions on Educations*, **44**(4), 329–335.
[3] Diday, E. and Simon, J. C. (1980) In: Fu, K. S. (Ed.), *Clustering Analysis: Digital Pattern Recognition*. Berlin: Springer-Verlag, New York: Heidelberg.
[4] Duda, R. and Hart, P. (1973) *Pattern Recognition and Scene Analysis*. New York: John Willey.
[5] Hartigan, J. A. (1996) Recognition, *Computational Statistics & Data Analysis*, **23**(1), 97–103.
[6] Koilios, G., Tsotras, V. J., Gunopulos, D., Delis, A. and Hadjieleftheriou, M. (2001) Indexing animated objects using spatiotemporal access methods, *IEEE Transactions on Knowledge and Data Engineering*, **13**(5), 758–777.
[7] Prabhakar, S. and Jain, A. K. (2002) Decision-level fusion in fingerprint verification, *Pattern Recognition*, **35**(4), 861–874.
[8] Verulava, O. (1997) Clustering analysis by "Rank of links", *Transactions of Georgian Technical University*, Tbilisi, **3**(414), 288–296.
[9] Verulava, O., Khurodze, R. and Grigalashvili, G. (1997) The estimation and prognosis of the decision making reliability by the "Rank of links" method, *Transactions of Georgian Technical University*, Tbilisi, **3**(414), 277–287.
[10] Verulava, O. (1993) Rank of links method in pattern recognition, *Dissertation Bulletin*, Georgian Technical University, Tbilisi (in Russian).
[11] Weir, N., Fayyad, U. M., Djorgovski, S. G. and Roden, J. (1995) The SKICAT system for processing and analyzing digital imaging sky surveys, *Publications of the Astronomical Society of the Pacific*, **107**(718), 1243–1254.