Synthesis of Automatic Recognition Systems Based on Properties Commonality

Viktor Krasnoproshin
Faculty of Applied Mathematics and Computer Science
Belarusian State University
Minsk, Belarus
krasnoproshin@bsu.by

Vadim Rodchenko
Faculty of Mathematics and Informatics
Yanka Kupala State University of Grodno
Grodno, Belarus
rovar@grsu.by

Anna Karkanitsa
Faculty of Mathematics and Informatics
Yanka Kupala State University of Grodno
Grodno, Belarus
a.karkanica@grsu.by

***Abstract*—The paper explores an actual applied problem related to the synthesis of automatic recognition systems. The conceptual base of synthesis is determined by the methods of describing and separating classes. Three basic principles are known: enumeration of class members, commonality of properties, and clustering. The report proposes an original method for implementing the principle of commonality of properties, based on the search for combinations of features that provide classes distinguishing. The efficiency of the approach is confirmed by the results of a numerical experiment.**

Keywords—pattern recognition system, data mining, instance-based learning

I. Introduction

The intensive development and application of information technology has led to the accumulation of huge amounts of data, which are currently organized into databases and data warehouses. The experience of using the structured query language SQL has shown that its capabilities of identifying the patterns existing inside the data are very limited. Operative analytical processing based on OLAP technology is focused on extracting from data a knowledge, which should be attributed to a "shallow" level of occurrence. The greatest practical interest, however, are the hidden patterns. The Data Mining is focused on the detection of such patterns [1].

In computer science, the problem of pattern recognition is among the fundamental. Its successful solution largely determines the progress in the field of artificial intelligence [2-5]. Pattern recognition is the assignment of initial data to a certain class based on the selection of significant distinguishing features that characterize these data from the general set of homogeneous data [6].

If a class is characterized by some common properties inherent in all its members, then the construction of a recognition system can be based on the principle of properties commonality. The main assumption in this case is that patterns of the same class have a number of common properties that reflect their similarity [7].

The paper proposes an original method for implementing the principle of commonality of properties. The method provides that first, based on the analysis of the data of the training set, combinations of features that ensure the distinction of classes are identified. After that, the construction of a classification algorithm becomes a trivial procedure. The effectiveness of the method is confirmed by the results of a numerical experiment.

II. About The Statement Of Recognition Problem

The basis of the idea of constructing automatic recognition systems is the methods of describing and separating classes (Fig. 1).

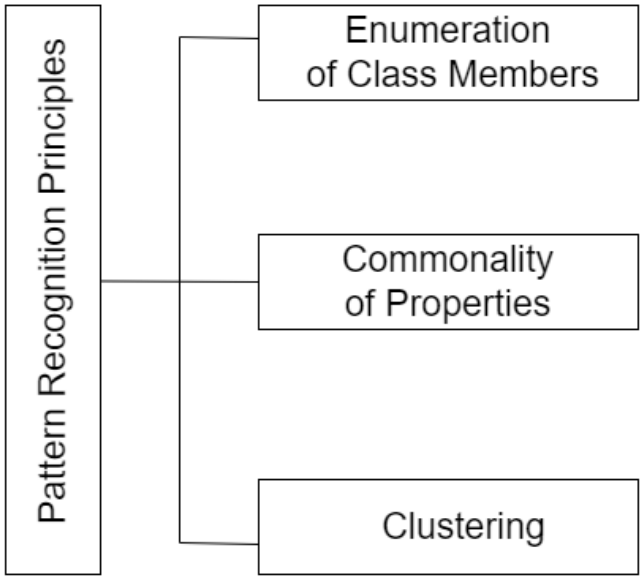

Fig. 1. Pattern recognition principles

When a class is defined by an enumeration of its constituent objects, the construction of a pattern recognition system can be based on the principle of ownership to this enumeration (Fig. 2). The set of class objects is remembered by the recognition system. When a new object is presented to the system, it refers it to the class to which the object located in the system's memory and coincided with the new one, belonged.

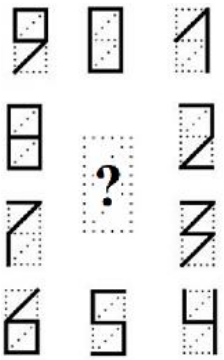

Fig. 2. Enumeration of class members

If all objects of one class have a number of common properties or features that are absent or have different values for all representatives of other classes, then the recognition system can be built on the basis of the principle of *commonality of properties* (Fig. 3).

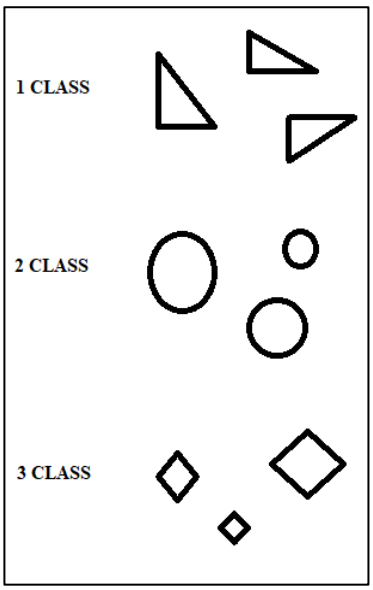

Fig. 3. Commonality of properties

When the objects of a class are vectors in a feature space, then the class can be considered as a cluster. And if clusters of different classes are spaced far enough from each other, then the construction of a recognition system can be carried out using the clustering principle (Fig. 4).

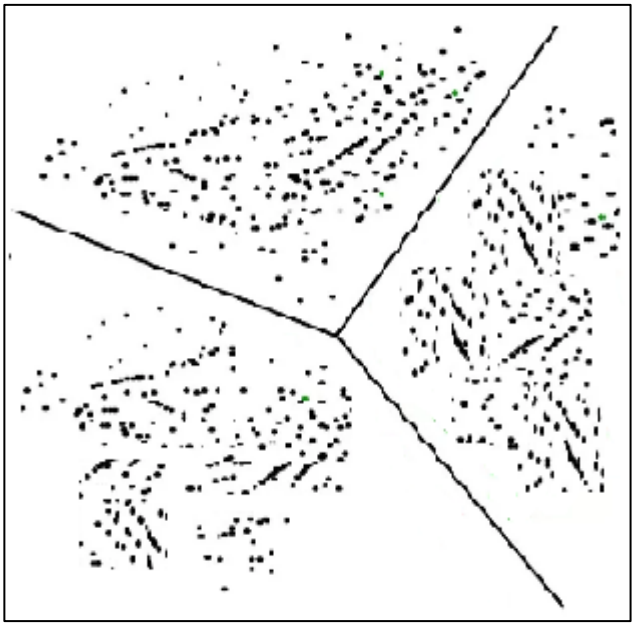

Fig. 4. Clustering

In the classical formulation, the statement of the recognition problem is as follows:

Given a set of objects divided into subsets, called classes. The information about the classes, the description of the entire set, and the description of information about the object who's belonging to a certain class is unknown are given. It is required, based on the available information about the classes and the description of the object, to establish its belonging to one of the classes [8].

As part of the construction of recognition systems based on the principle of clustering, the recognition problem is solved in the following statement:

Let X be the set of objects descriptions, Y – acceptable answers for objects classification. Suppose there is an unknown target dependency $y^* : X \to Y$, *which values* $X^m=\{(x_1, y_1),\ldots, (x_m, y_m)\}$ *are known only for the objects of the final training set. It is necessary to construct an algorithm* $a : X \to Y$, *that would approximate this target dependency not only on objects of a finite training set, but also on the entire set X* [9].

Traditionally, solving a problem is carried out in two stages: first, a model of algorithms is defined to an accuracy of the parameters, and then, in the learning process their values are found that provide the extremum of the selected criterion.

It should be pointed out that there are a number of problematic issues and shortcomings that arise during the implementation of this scenario. First, choosing an algorithm model is a non-trivial problem. The quality of the problem solution results largely depends on the level of training and experience of the data analyst. Secondly, the learning process can only be implemented in an automated mode, and the resulting algorithm $a : X \to Y$ is a "black box" that practically cannot be interpreted in terms of the subject domain. Thirdly, the construction of a classification algorithm based on the data of the training set X^m is carried out only in the original space of objects description, while the question of the existence of subspaces in which the problem is solved more effectively remains open [10].

It is proposed to get rid of the above problematic issues and shortcomings by constructing recognition systems based on the commonality of properties principle [11]. The mathematical problem statement of the recognition problem in this case has the following formulation:

Let X be the set of objects descriptions, Y – acceptable answers for objects classification. There is an unknown target dependency $y^* : X \to Y$, *which values* $X^m=\{(x_1,y_1),\ldots, (x_m,y_m)\}$ *are known only for the objects of the training set. It is required to find feature spaces in which the classes do not intersect, and on their basis to construct an algorithm* $a : X \to Y$, *that would approximate this target dependency not only on the objects of the finite set, but also on the entire set X.*

To solve the recognition problem in such a statement, it is first proposed to find feature spaces where classes do not intersect. After this, constructing a classification algorithm becomes a trivial procedure.

III. Algorithm Of Searching Informative Combinations Of Features

Let the training set $X^m=\{(x_1, y_1),\ldots,(x_m, y_m)\}$ be formed on the basis of an a priory dictionary of features $F=\{f_1, \ldots, f_n\}$.

Let's denote by $V=\{v_1, \ldots, v_q\}$ the set of all possible combinations of features from F. Then V contains $q=2^n-1$ subsets.

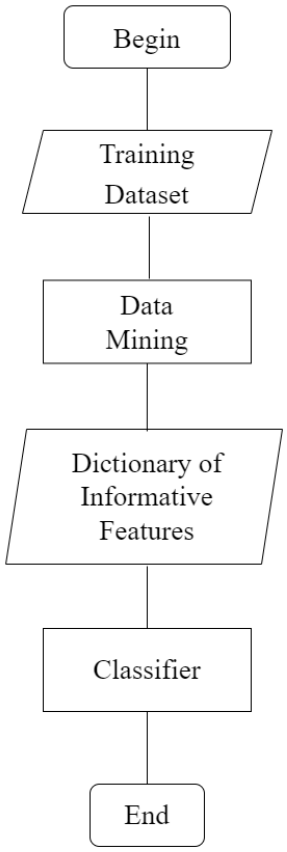

Fig. 5. Algorithm flowchart

The algorithm (see Fig. 5) of searching the combinations of features on the set $V=\{v_1,\ldots,v_q\}$ for which the class patterns do not intersect is as follows.

Step 1. Select from V a subset $V^+=\{v^+_1, \ldots,v^+_n\}$ where v^+_i contains only one feature.

Step 2. For each v^+_i build class patterns and compare their mutual placement.

Step 3. If the class patterns do not intersect, include the feature v^+_i in the set $V^*=\{v^*_1,\ldots,v^*_k\}$.

Step 4. Exclude the subset $V^+=\{v^+_1,\ldots,v^+_n\}$ from the set $V=\{v_1,\ldots,v_q\}$ and get $V^\wedge=\{v^\wedge_1,\ldots,v^\wedge_p\}$.

Step 5. Exclude from $V^\wedge$ all combinations $v^\wedge_i$ that contain any combination from $V^*=\{v^*_1,\ldots,v^*_k\}$.

Step 6. Take the next combination $v^\wedge_i$ from $V^\wedge$ and build a feature subspace on its basis.

Step 7. In this feature subspace we construct class patterns and compare their mutual placement.

Step 8. If the class patterns do not intersect, then include the combination of features $v^\wedge_i$ in the set V^*, and exclude from $V^\wedge$ all combinations that contain $v^\wedge_i$.

Step 9. Repeat the process until $V^\wedge$ becomes empty.

As a result of the algorithm, the set $V^*=\{v^*_1,\ldots,v^*_t\}$, where $0 \le t \le q$, will be constructed. All combinations of features from V* define spaces in which the class patterns do not intersect, and the construction of classification algorithms does not cause any principal difficulties.

IV. Results Of The Numerical Experiment

Let's demonstrate the results of solving the recognition problem based on model data.

Example. Let the following be given:

– classes **NO7 (integer does not contain the digit 7)** and **YES7 (integer contains at least one digit 7)**;

– a priori dictionary of features F = {Units, Tens};

– training set that consists of 220 two-digit integers, among which 180 integers do not contain the digit 7 and 40 integers have at least one digit 7.

Table I shows the values of features *units* and *tens* in the training set used in the numerical experiment.

TABLE I. VALUES OF *UNITS* AND *TENS* IN THE TRAINING SET

Units Tens	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	2	3	2	1	3	3	2	3	2	3
1	3	1	2	2	2	1	3	2	2	2
2	2	3	2	3	1	2	1	3	3	2
3	2	2	1	2	3	3	3	2	2	3
4	2	3	2	3	2	3	2	1	3	2
5	2	2	3	3	3	2	3	2	2	3
6	3	2	3	2	2	3	1	2	2	2
7	2	3	1	2	3	1	2	3	2	3
8	1	2	2	2	2	2	3	2	2	2
9	2	2	2	2	3	3	2	1	1	1

Table II shows the results of a study of the intersection of class patterns based on the features *units* and *tens*, where:

$NO7_i$ = Number of $NO7_i$ for the *i*-th digit;

$YES7_i$ = Number of $YES7_i$ for the *i*-th digit.

TABLE II. RESULTS FOR FEATURES *UNITS*, *TENS*

Digit	**Units**		**Tens**	
	NO7	**YES7**	**NO7**	**YES7**
0	21	3	19	2
1	18	2	20	3
2	19	3	19	1
3	21	2	20	2
4	22	1	21	3
5	23	2	22	1
6	20	2	20	2
7	0	3	0	3
8	18	2	19	2
9	18	1	20	3

Table III shows the results of a study of the intersection of class patterns based on a combination of features (*tens*, *units*).

TABLE III. RESULTS FOR COMBINATION (*TENS*, *UNITS*)

Tens, Units	**NO7**	**YES7**	**Tens, Units**	**NO7**	**YES7**
0, 0	2	0	**5, 0**	2	0
0, 1	3	0	**5, 1**	2	0
0, 2	2	0	**5, 2**	3	0
0, 3	1	0	**5, 3**	3	0
0, 4	3	0	**5, 4**	3	0
0, 5	3	0	**5, 5**	2	0
0, 6	2	0	**5, 6**	3	0
0, 7	0	3	**5, 7**	0	2
0, 8	2	0	**5, 8**	2	0
0, 9	3	0	**5, 9**	3	0
1, 0	3	0	**6, 0**	3	0
1, 1	1	0	**6, 1**	2	0
1, 2	2	0	**6, 2**	3	0
1, 3	2	0	**6, 3**	2	0

Tens, Units	NO7	YES7	Tens, Units	NO7	YES7
1, 4	2	0	**6, 4**	2	0
1, 5	1	0	**6, 5**	3	0
1, 6	3	0	**6, 6**	1	0
1, 7	0	2	**6, 7**	0	2
1, 8	2	0	**6, 8**	2	0
1, 9	2	0	**6, 9**	2	0
2, 0	2	0	**7, 0**	0	2
2, 1	3	0	**7, 1**	0	3
2, 2	2	0	**7, 2**	0	1
2, 3	3	0	**7, 3**	0	2
2, 4	1	0	**7, 4**	0	3
2, 5	2	0	**7, 5**	0	1
2, 6	1	0	**7, 6**	0	2
2, 7	0	3	**7, 7**	0	3
2, 8	3	0	**7, 8**	0	2
2, 9	2	0	**7, 9**	0	3
3, 0	2	0	**8, 0**	1	0
3, 1	2	0	**8, 1**	2	0
3, 2	1	0	**8, 2**	2	0
3, 3	2	0	**8, 3**	2	0
3, 4	3	0	**8, 4**	2	0
3, 5	3	0	**8, 5**	2	0
3, 6	3	0	**8, 6**	3	0
3, 7	0	2	**8, 7**	0	2
3, 8	2	0	**8, 8**	2	0
3, 9	3	0	**8, 9**	2	0
4, 0	2	0	**9, 0**	2	0
4, 1	3	0	**9, 1**	2	0
4, 2	2	0	**9, 2**	2	0
4, 3	3	0	**9, 3**	2	0
4, 4	2	0	**9, 4**	3	0
4, 5	3	0	**9, 5**	3	0
4, 6	2	0	**9, 6**	2	0
4, 7	0	1	**9, 7**	0	1
4, 8	3	0	**9, 8**	1	0
4, 9	2	0	**9, 9**	1	0

From Table III it is clear that:

– all integers of class NO7 do not have the digit 7, and all integers of class YES7 have at least one digit 7;

– a combination of features (*tens, units*) ensures absolute separation of classes NO7 and YES7.

V. CONCLUSION

The paper presents an alternative option for setting and solving the recognition problem. The approach is based on the use of the principle of properties commonality.

In the learning process, based on the contents of an a priori dictionary of features and training set data, the properties of combinations of features are examined and such features are identified that provide class distinguishing.

The results of solving the recognition problem are demonstrated on the example of processing model data.

REFERENCES

[1] Data mining [Electronic resource]. Access mode: https://en.wikipedia.org/wiki/Data_mining. Access date 07/01/2023.

[2] M.M. Bongard, The problem of recognition. Nauka. Moscow, 1967, p. 320, (in Russian).

[3] A.G. Arkad'ev and E.M. Braverman, Training machines for classifying objects. Izdatel'stvo Nauka. Moscow, 1971, p. 192, (in Russian).

[4] N.G. Zagoruiko, Applied methods of data analysis and knowledge. Izdatel'stvo Instituta matematiki SO RAN. Novosibirsk, 1999, p. 268, (in Russian).

[5] V.V. Krasnoproshin and V.G. Rodchenko, "Classification Based on Decision Spaces," Doklady BGUIR №6 (2019), p. 20–25, (in Russian).

[6] (2023, July) Pattern recognition – Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Pattern_recognition.

[7] J.T. Tou and R.C. Gonzalez, Pattern recognition principles. Izdatel'stvo Mir. Moscow, 1978, p. 412, (in Russian).

[8] Yu.I. Zhuravlev, "On an algebraic approach to solving problems of recognition and classification," Problems of Cybernetics, №33 (1978), p. 5–68.

[9] (2023, July) Supervised learning – Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Supervised_learning.

[10] V.V. Krasnoproshin and V.G. Rodchenko, "Learning by Precedents Based on the Analysis of the Features Properties," Doklady BGUIR №6 (2017), p. 35-41, (in Russian).

[11] V.V. Krasnoproshin and V.G. Rodchenko, "Cluster Structures and Their Applications in Data Mining," Informatics №2 (2016), p. 71–77, (in Russian).