New methods of pattern analysis in the study of Iris Anderson-Fisher Data

Alexey Myachin
National Research University Higher School of Economics
NRU HSE
Institute of Control Science of Russian Academy of Science
ICS RAS
Moscow, Russian Federation
amyachin@hse.ru

Abstract: A new method of pattern analysis, based on paired index comparison is introduced. Key properties of the method are described. The effectiveness is demonstrated on the Iris Anderson-Fisher Data.

Index Terms - cluster; pattern analysis; ordinal-invariant pattern-clustering; Iris Anderson-Fisher Data

I. INTRODUCTION

The Iris Anderson-Fisher Data [1] is de-facto classical widely-used for testing different methods of clustering, which makes it useful not only for checking newly-developed methods, but also for comparing newly-obtained results with those of the other well-known algorithms. Let us point out [2] where many possibilities of pattern analysis are demonstrated.

As a development of pattern analysis methods, a new algorithm is proposed, being described as «ordinal-invariant pattern-clustering», the effectiveness of which is demonstrated on the Anderson-Fisher Iris Data.

A. Problem statement

There is a given set of measurement results of linear dimensions of 150 Iris blossoms of three kinds: *Iris Setosa*; *Iris Versicolor*; and *Iris Virginica*, 50 samples of each sorts. Each blossom is characterized by four parameters: *sepal length*; *sepal width*; *petal length*; and *petal width*.

The goal is to develop a method allowing to define the blossom kind by its measurements. In other words – to part blossoms of different kind into separate groups.

II. PATTERN ANALYSIS

A. Description of the method

Method of pattern analysis is based on decomposition of the given set of objects into a number of subsets by a pre-chosen subset metrics. Decomposition condition is significant difference of the subsets objects, while objects of each subset are very close to one another by the selected metrics.

In general a variety X consisting of k objects if being studied. The $x_i \in X$ objects are described as $x_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{im})$ vectors, where x_{ij} – is the j-th parameter of the i-th object. For convenience and easy visual perception, an m-dimensional system of parallel coordinates [3], build of m equidistant vertical lines each j-th of which corresponds to j-th parameter, k piecewise linear functions are build, each corresponding to an objects in study.

Further, using the measure of closeness $d(x_a, x_b)$, the X variety is decomposed to non-intersecting sub-varieties. Let us point out that for different cases the measure of closeness may vary depending on initial data and expected results. Let us show how the method works in practice. The variety in study is X, consisting of three hypothetic banks with three parameters [4].

TABLE I. EXAMPLE WITH HIPOTHETICAL BANKS

Bank	A	B	C
Bank 1	50	20	40
Bank 2	55	10	45
Bank 3	10	60	20

We construct a piecewise linear functions of the researched banks:

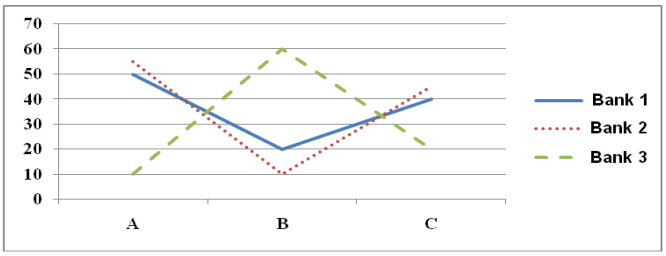

Fig. 1. Picewise linear functions of hipotetical banks

A pattern analysis method has placed the three Banks into 2 different clusters: {Bank 1, Bank 2} and {Bank 3}.

This method has already successfully established itself in solving practical problems in different fields: macroeconomics

[5,6], politics [7], management and personnel management [8].

III. NEW METHOD OF PATTERN ANALYSIS

We propose a new method of pattern analysis to split the original set in disjoint subsets of in this paper: ordinal-invariant pattern clustering. The algorithmic implementation, complexity and some properties are described.

A. *Ordinal-invariant pattern clustering.*

Every object $x_i \in X$ put in one-to-one correspondence vector $r_i = (r_{i1}, r_{i2}, \ldots, r_{ij}, \ldots, r_{im-1})$, where:

$$r_{ij} = 1, \text{ if } x_{ij} < x_{ij+1} \quad (1)$$
$$r_{ij} = 0, \text{ if } x_{ij} = x_{ij+1} \quad (2)$$
$$r_{ij} = 2, \text{ if } x_{ij} > x_{ij+1} \quad (3)$$

Generated positional decimal codes for the studied objects:

$$q_i = \sum_{j=1}^{m-1} 10^{j-1} r_{im-j} \quad (4)$$

As an estimation of the distance d between 2 objects x_a and x_b is used the difference q_a-q_b. We are interested in 2 cases:

1) q_a-$q_b \neq 0$, objects x_a and x_b are divided into different clusters;

2) q_a-q_b=0, form a complete weighted digraph of objects x_a and x_b, where the vertices are the values of the investigated parameters, and the values of edges e_{ij}^s is determined similary (1)-(3):

$$e_{ij}^s = 1, \text{ if } x_{ij} < x_{is} \quad (5)$$
$$e_{ij}^s = 0, \text{ if } x_{ij} = x_{is} \quad (6)$$
$$e_{ij}^s = 2, \text{ if } x_{ij} > x_{is}, \quad (7)$$

where s is the number of studied parameter.

The number of edges is defined as:

$$u = \frac{m(m-1)}{2} \quad (8)$$

For each object additional position code is formed according to the values of edges:

$$q_i^{dop} = \sum_{j=1}^{m(m-1)/2} 10^{j-1} e_{ij}^s \quad (9)$$

Two cases are considered:

1) q_a^{dop}-$q_b^{dop} \neq 0$, objects x_a and x_b are devided into different clusters;

2) q_a^{dop}-q_b^{dop}=0, objects x_a and x_b are combined into one cluster.

The complexity of the initial partitioning (the comparing of q_a and q_b) is defined as:

$$z = \frac{k^2(k-1)(m-1)}{2} \quad (10)$$

The complexity of the new method is determined by the formula (8).

IV. IRIS ANDERSON-FISHER DATA

On the one hand the aim of the experiment using Iris Anderson-Fischer Data was to test the effectiveness of the ideas that will enable more detailed analysis of patterns, and use its results in clustering tasks.

Describe of the findings of the experiment results.

A. *The use of ordinal-invariant pattern clustering.*

The result was the division of the original set into two clusters, containing 50 and 100 flowers. Testing showed that the first cluster contained the flowers of the varieties Iris Setosa, and the second – flowers of Iris Versicolor and Iris Virginica.

Successfully separated the flowers of the varieties Iris Setosa, demonstrated ordinal-invariant pattern-clustering, is not something that is unique and characteristic of the work of a number of known methods. In the literature, in particular, the data highlight this class by the threshold value (24.5 mm) length of petal (fig.2).

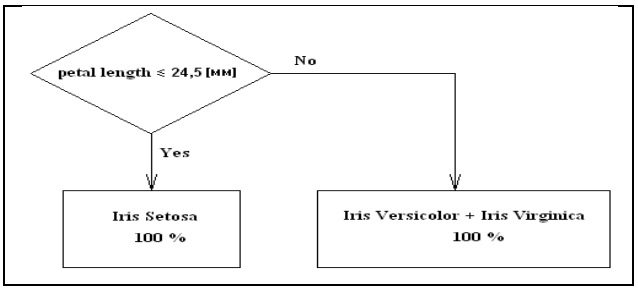

Fig. 2. The Selection of Iris Setosa using the threshold value

This way we can only say that the result of this stage was not worse than others. However, this information content is significantly different. At the first glance, the main difference in the semantic content of these two is the same results.

Threshold method work divided the initial set into two clusters. Splitting criteria was the length of the petal. In the first cluster placed flowers with long petals not exceeding 24.5 mm, and the second – lobe length exceeding this value. Any information on the nature of the clusters themselves, the result is not responsible. In particular, if the original set contained data sizes colors ten (twenty, etc.) the different types and varieties, used method, anyway, would divide them into two clusters by the threshold value of the length of the petal.

In contrast to this result, ordinal-invariant pattern-clustering revealed that the initial set consists of the objects of two types, forming two clusters. In this case, the method did not separate the objects one from the other, but combined in

the separate clusters of the objects having a set of common, that is inherent in this cluster of properties (defined by pattern).

The result of this combination has been the allocation of two clusters. Testing showed that one of them contains the flowers of the varieties Iris Setosa, and the second: varieties of Iris Versicolor and Iris Virginica.

Does any substantial aspect in the fact that the varieties of Iris Versicolor and Iris Virginica are merged in one cluster {Versicolor + Virginica}? A search in literature [9] gives a positive (and very interesting) answer to this question:

"These species are so similar to each other that the monograph of the genus William dykes was recognized only the first (Iris Versicolor) as a separate species. The second is the Virginia iris (Iris Virginica) — was related to synonyms first."

Thus, the second cluster {Versicolor + Virginica}, also, according to W. Dixu, contains flowers of the same species, and demonstrated the algorithm combining these colors within a single cluster is not random.

What is the difference between allocated clusters? The above-mentioned threshold method makes it possible to separate the colors Setosa from Versicolor and Virginica flowers against a threshold of 24.5 mm of petal length. Although this threshold and provides a 100% selection of flowers Setosa, however, it is difficult (though maybe possible) to assume that nature, creating different varieties of flowers, divided them along the length of the petal defined with precision accurate to tenths of a millimeter.

On the other hand, the result of ordinal-invariant pattern-clustering gives another criterion: clusters {Setosa} and {Versicolor + Virginica} differ flower shape, defined by different ratios between its petals that demonstrate a piecewise linear function of the data clusters (fig.3).

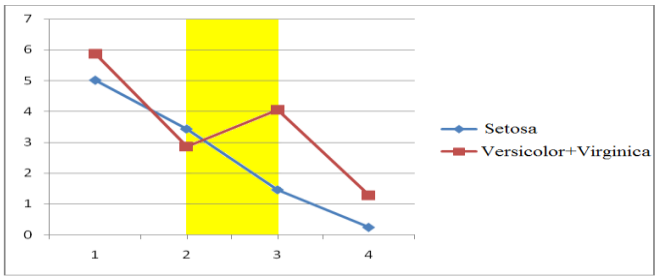

Fig. 3. Piecewise linear function of culester {Setosa} and {Versicolor + Virginica}

This difference is most pronounced on the section 2-3, where piecewise linear functions are of opposite character. In particular you can see that the width of the sepals (parameter 2) flower Setosa significantly greater than the length of the petal (parameter 3), while the class {Versicolor + Virginica} the situation is reversed.

Using this criterion, it is possible to determine which flower in front of you (without resorting to the use of precision measuring technique). Clearly this can be done on the example of fig.4.

Fig. 4. Varieties of Iris flowers

It is obvious that the left flower is much broader petals sepals. Therefore, it is Iris Setosa. To the right is Iris Versicolor or Iris Virginica (which, according to William Dixo same) – Pic.5.

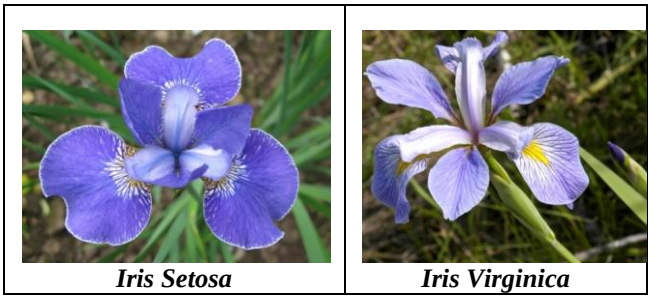

Fig. 5. Varieties of Iris flowers

Preliminary results:

1. The initial set of colors is divided into two subsets {Setosa} and {Versicolor + Virginica}. This division corresponds to the classification of monographia kind of Iris flowers, botany William Diksa;
2. Revealed a clear criterion to divide the selected clusters. Unlike other threshold or statistical criteria, it relies on the shape of the flower determined by the ratios of the dimensions constituting its petals.

B. Pattern analysis for values of the differences

Continue to study. According to the conditions of the problem, we need to break the class {Versicolor + Virginica} in two distinct subclusters of {IrisVersicolor} and {IrisVirginica}. The fact is that [9]:

"In 1927, botanist E. Anderson was recognized the virgin iris as the autonomy species. Both species are members of the North American flora, but the first one northerner and second is southerner. In the us-canadian Botanical literature they are called: first, the colored iris Northern Blue Flag, and second, Iris Virginia – southern Blue Flag."

This task is substantially more complicated, and with varying degrees of success may be achieved by various methods. For example the method of K-means could not only separate the classes {Iris Versicolor} and {Iris Virginica}, but erroneously divided {Iris Setosa} into two clusters.

However, following Anderson, still try to divide {Versicolor + Virginica} into two subclasses: {Versicolor} and {Virginica}, resorting to a more detailed analysis.

C. *The complexity of the task*

Iris Versicolor and Iris Virginica substantially intersect (moreover – "mixed") . This predetermined negative result of K-means. Similar difficulties arise when analyzing the patterns of these varieties. Let us demonstrate it with the following example.

Scale 1 presents the data of the two flowers –Iris Setosa, and figure their patterns. Although these flowers belong to one class -Iris Setosa, however, as can be seen from the figure, the linear patterns of these two flowers, having a similar shape, vary greatly in the values of the parameters.

TABLE II. IRIS SETOSA

Iris	Sepal Length	Petal length	Sepal width	Petal width
Setosa	5,1	1,6	3,8	0,2
Setosa	4,5	1,3	2,3	0,3

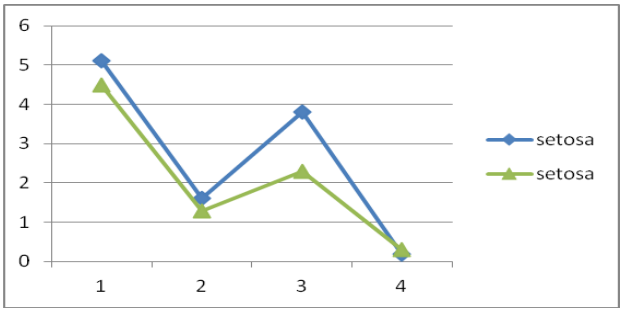

Fig. 6. Piecewise linear functions of two flowers varieties Iris Setosa

On the other hand, scale 2 presents these two different kinds of flowers (Iris Versicolor and Iris Virginica), and in fig.7 – their piecewise linear functions:

TABLE III. IRIS VERSICOLOR AND IRIS VIRGINICA

Iris	Sepal Length	Petal length	Sepal width	Petal width
Versicolor	6	2,7	5,1	1,6
Virginica	6	2,2	5	1,5

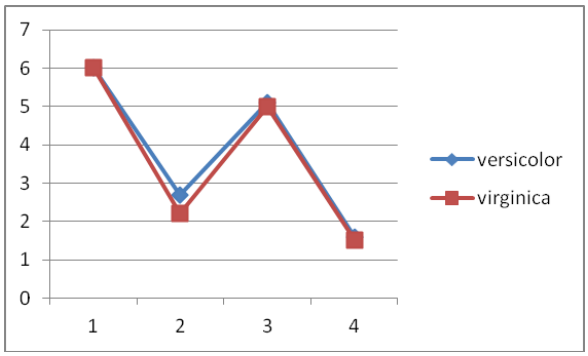

Fig. 7. Case when piecewise linear functions different varieties of flowers (Iris Versicolor and Iris Virginica) are almost identical

Although the parameter values are significantly closer than in the first case, and piecewise linear functions are almost identical, the flowers belong to different species.

D. *Analusis of the nature of the change pattern*

Previously identified nature of the relationship between the analyzed parameters of the object (values of the predicates of the use of mesh edges in a weighted digraph describing the pattern).

The next task is to assess the magnitude and significance of the identified relationships. We will solve this task by using the values of differences of parameters. Explain the proposed procedure with the following example. Consider a dataset of objects, characterized by three parameters:

TABLE IV. DIFFERENCES OF PARAMETERS

Object	Difference 1	Difference 2	Difference 3
Object 1	10	20	70
Object 2	10	40	70
Object 3	10	60	70

Construct piecewise linear functions data object

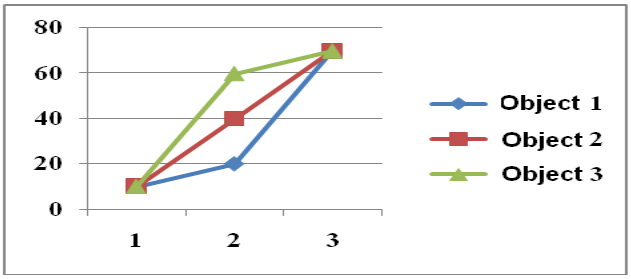

Fig. 8. Piecewise linear functions of the objects

Note that although all three piecewise linear functions have the General character of increasing parameter values (which determined their belonging to the same cluster), the speed of piecewise linear function is substantially different, as the significantly different magnitude relations between the values of the parameters.

So, for example, the third object the second parameter substantially exceeds the value of the first and not much less than the value of the third parameter; while the first object is

the reverse: the value of the second parameter is not much larger than the value of the first, but significantly less than the value of the first.

Set the task to divide these objects, using a more detailed analysis, the essence of which is to assess the significance of the relationship parameters. This will allow you to split the previously obtained cluster.

The idea and the implementation of the method such detailed analysis of patterns is show in the presented example.

1. Write a table of absolute difference between the values of neighboring parameters:

TABLE V. NEW PARAMETERS

Object	1	2
Object 1	10	50
Object 2	30	30
Object 3	50	10

2. Construct a piecewise linear function of the difference:

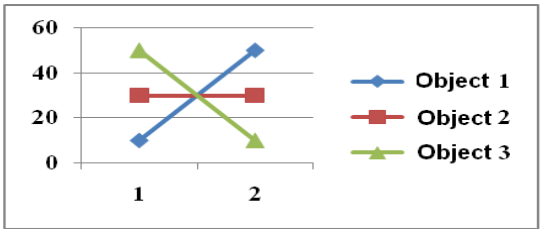

Fig. 9. Piecewise linear functions of the new parameters

3. Fig. 9 demonstrates that the piecewise linear functions of the difference has a trend that can be fixed by classical methods of analysis of patterns (applied to values of the differences). For example, using the encoding:

TABLE VI. FIXED TREND

Object 1	2
Object 2	0
Object 3	1

Thus, all three objects will be placed in the different clusters.

V. RESULTS

The method described above is applicable to selected at the first stage, ordinal-invariant pattern-clusters {Setosa} and {Versicolor + Virginica}. They are divided into a number of subclusters, each of which is characterized by a very close stroke of piecewise linear functions belonging to a specific entity. For convenience of further presentation, we call such "model classes" – "benchmark".

A. *Clustering*

As noted, the result of the second stage is a number of selected "reference" subclasses. For each reference of a subclass will find the average object. This averaged pattern of the object will correspond to the pattern of a subclass. Let's call this the averaged center of the object subclass.

B. *Integration of clusters*

The step of combining the individual small clusters in larger characteristic of many known methods. Important is the choice of the criterion of Association. For these purposes, we use the criterion proposed in the already quoted above work [1]. It noted the following feature of Iris Virginica: the length of its petals, on average, somewhat greater than the length of the petal Iris Versicolor. A similar situation is observed and the width of petal: petal width Iris Virginica, on average, somewhat greater than the width of Iris Versicolor petal.

Based on this fact, the authors suggested using the product of these parameters (the length and width of petal) to amplify the difference, and use this work as one of the clustering options.

We use this criterion, but not for the purpose of separating individual clusters, and for solving the inverse problem is combining them.

NOTE. The considered criterion is, in our opinion, a definite "Botanical" sense. The fact that the multipication of their selected parameters, with accuracy up to the multiplier, gives some estimate of the area of a petal (for example, using the formula for the area of the ellipse $S = \pi ab$, where a and b are the sizes of its semi-axes).

Therefore, this figure (the product of length and width of petal) will be called "the area of the petal."

Based on the noted feature, perform the calculation of the mean evaluation values for the area of a petal. To this end we calculate, for all subsets {Versicolor + Virginica} the value of these estimates (as the product of length and width of petal), put the calculated values in ascending order, and find the averages for each half.

The values obtained are respectively of 5.72 mm^2 and 11.3 mm^2.

Calculate features square petals for the centroids of the clusters.

Next, we merge clusters according to the criterion of proximity of the estimated areal centroids to the computed estimates of mean values for colour varieties {Versicolor} and {Virginica}.

C. *Results of experiment*

The results of the experiment showed a fairly high efficiency of this method:

TABLE VII. RESULTS

Flower	Number	Error
Iris Setosa	50	0
Iris Versicolor	50	1
Iris Virginica	50	4

Thus, the error of clustering is 5/150 = 0,03(3), i.e. about 3%. This is a rather high score, comparable to the best known clustering techniques.

VI. Conclusion

The major advantage of the method of pattern analysis that it directly indicates the distinctive features of the considered clusters. Thus, the analysis of the Iris Anderson-Fisher Data immediately reveals a distinctive feature of the Setosa cluster: this is a substantial difference of this method from other methods of cluster analysis, aimed at assessing the degree of closeness of the parameters of the real object to one of several abstract (in the General case, not existing) objects, referred to as "cluster centroid".

The process of extracting distinguishing features in this method is close to that which a man operates. This feature allows to consider the possibility of using this method, for example, in systems modeling human choice.

The results of clustering the Iris Anderson-Fischer Data demonstrate the effectiveness of this method, however, the issue of merging subclusters requires further study.

In the already cited paper [2] notes: "...the urgent task seems to be the formation of such mathematical models and methods that seek to identify patterns in the data directly, without first identifying the clusters. The need for relevant theoretical, algorithmic and experimental developments are long overdue". One possible use of the results of this work is the use of the procedure for selecting the centers of the reference classes for the initialization of the algorithm K-means.

Acknowledgment

The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program. The work was partially financed by the International Laboratory of Decision Choice and Analysis (DeCAn Lab) of the HSE and by the Laboratory №25 of V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences.

References

[1] Fisher R.A . The use of multiple measurements in taxonomic problems // Annals of Eugenics, 1936, pp. 179-188

[2] Aleskerov F., Belousova V., Egorova L., Mirkin B. Analiz patternov v statistike i dinamike. Chast' 1: Obzor literatury i utochnenie ponjatija [Pattern analysis in the statistics and dynamics. Part 1: Literature Review and refinement of the concept]. Business Informatics, 2013, Vol. 3,pp. 3-18

[3] Few S. Multivariate analysis using parallel coordinates, Perceptual Edge http://www.perceptualedge.com/articles/b-eye/parallel_coordinates.pdf (07.01.2016)

[4] Aleskerov F.T., Solodkov V.M., Chelnokova D.S. Dinamicheskij analiz patternov povedenija kommercheskih bankov Rossii [Dynamic pattern analysis of behavior of commercial banks in Russia]. Economic Journal of the Higher School of Economics, 2006 vol.10, no.1, pp.48-62

[5] Aleskerov F., Alper C.E. Inflation, Money, and Output Growths: Some Observations // Bogazici University Research Paper, No. SBE 96- 06, 1996.

[6] Aleskerov F., Alper C.E. A clustering approach to some monetary facts: a long-run analysis of cross-country data // The Japanese Econo mic Review, 2000, Vol. 51, No. 4, pp. 555–567

[7] Aleskerov F., Ersel H., Yolalan R. Multicriterial Ranking Approach for Evaluating Bank Branch Performance // International Journal of Information Technology and Decision Making, 2004, Vol. 3, No. 2, pp. 321– 335

[8] Aleskerov F., Ersel H., Gundes C., Yolalan R. A Multicriterial Method for Personnel Allocation among Bank Branches // Yapi Kredi Discussion Paper Series, 1998, No. 98–01

[9] Enciclopedia of ornamental garden plants. http://flower.onego.ru/voda/iris_ver.html (07.01.2016)