Classless Logical Regularities and Outliers Detection

Alexander Dokukin [1)]
1) FRC CSC RAS, 119333, Vavilova Str. 40, Moscow, Russian Federation, dalex@ccas.ru

Abstract: *The paper defines classless logical regularities in a similar way to widely known logical regularities of classes. The main reason of their introduction is unsupervised outlier detection in the areas where main part of available data represents normal situation and outliers are rare and different from each other. The substantiation and formal definition is followed by the method for their search and a real world example.*

***Keywords*:** classless logical regularities, unsupervised outlier detection, information value.

1. INTRODUCTION

The basic definition of the logical regularities of classes is given in [1]. They are used in supervised classification and correspond to rectangular areas in feature space that contain precedents of a single class.

Here is the formal definition of a logical regularity of a class. Let's consider a recognition task and especially a training set $S = \{S_1, \ldots, S_m\}$. The set is divided into classes $K_1, \ldots, K_l$ (for simplicity classes don't overlap) and classification of the training objects is known, i.e. $S_i \in K(S_i)$, $i = 1, \ldots, m$. Let's consider also hyperrectangles of the following structure

$$R = r_1 \times \ldots \times r_n, \tag{1}$$

$$r_i = \begin{cases} [a_i, b_i] \\ (-\infty, b_i] \\ [a_i, \infty) \\ (-\infty, \infty) \end{cases}, \ i = 1, \ldots, n.$$

A hyperrectangle R is called a logical regularity of the class K_j if the following conditions hold:

$$1) \ \forall S \notin K_j, S \notin R, \tag{2}$$

$$2) \ |\{S \mid S \in R\}| \to \max.$$

A generalization of the logical regularity that allows entry of other classes' objects to it is called a partial logical regularity. First, a fixed part of such objects was allowed [1]. Later a more generalized version was described that involves information value of the regularity. There are different ways of defining the information value that is a functional that reward presence of the target class precedents and penalize presence of the rest of them in it [2]. The most important of them for the purposes of the present article is statistical information value:

$$I_h(R, S) = -\ln h\begin{pmatrix} p_j(R) & n_j(R) \\ P_j & N_j \end{pmatrix}, \tag{3}$$

where $h\begin{pmatrix} p & n \\ P & N \end{pmatrix}$ is hypergeometric distribution $h\begin{pmatrix} p & n \\ P & N \end{pmatrix} = \frac{C_P^p C_N^n}{C_{P+N}^{p+n}}$, $p_j(R)$ – number of class K_j objects in R, P_j – number of all class K_j objects, $n_j(R)$ – number of objects of other classes in R and N_j – number of all objects of other classes. The greater the $I_h(R, S)$ the lesser the probability to choose such a combination of objects from the set. In addition $p_j(R)$ is required to be greater than $n_j(R)$.

Thus, a generalized regularity is defined as an area in feature space with great enough information value. Rectangular shape is not necessary but it has certain advantages.

Logical regularities based classification is made by voting over the number of regularities of different classes covering the specific object. At that regularities can be weighed according to their quality as a part of class' objects covered by it or in general case according to their information value. A number of other conclusions beside object classification can be made from a set of logical regularities. They are used to acquire simple class descriptions in human readable form, to assess importance of a feature or typicality of the precedent [3]. The latter can be used to detect outliers of a specific class.

The basic idea behind mentioned applications of logical regularities of classes is very simple. The quality of an entity is assessed as a weighed part of logical regularities set that involves the entity and it works well in practical tasks [3]. However, there is one major drawback of the described methodology it requiring information about classification of the studied objects. It is hard to construct nontrivial regularities in a case where most precedents represent normal situations whereas outliers are few and very different. For example an analysis of surgery outcome in a good clinic can be considered [4], where complications are scarce and diverse. Moreover, a situation is feasible in which even the described scarce information is unavailable. For example, in the mentioned medical tasks an outcome that is considered generally well can require more attention. Thus, the analysis requires unmarked data.

To overcome the limitation while preserving the rest of the developments in the area the concept of classless regularities is presented. The formal definition will be given in the following section but the idea is similar to the logical regularities of classes though instead of penalty for the inclusion of objects of other classes the size of the area is penalized.

2. CLASSLESS REGULARITIES

Classless regularity is an area of feature space defined by a specific information value functional.

Let’s consider a training set of precedents $S = \{S_1, \ldots, S_m\}$ of the same structure as in section 1 $\left(S_i \in \Re^n\right)$, but without any classification information. Let’s also define the containing hyperrectangle R_S which is the minimal hyperrectangle containing S. The statistical information value of an arbitrary hyperrectangle $R \subset R_S$ is defined as follows:

$$I_b(R,S) = -\ln b\left(k(R), m, \frac{V(R)}{V(R_S)}\right), \qquad (4)$$

where $b(k,n,q)$ is binomial distribution $b(k,n,q) = C_n^k q^k (1-q)^{n-k}$, $k(R)$ – number of objects in R, $V(R)$– volume of R. Again, the greater the $I_b(R,S)$ the lesser the probability of such a number of uniformly distributed in R_S objects getting into the hyperrectangle. But to prevent rectangles being chosen for the rarity of their complement an additional condition is checked:

$$q^k < (1-q)^{n-k}. \qquad (5)$$

The functional (4) with the condition (5) has all the required properties to serve as an information value. It increases with the number of objects getting into the regularity and decreases with its volume. It is also important that the functional doesn’t depend on the scale of the features, because only the volume ratio is involved in the formula.

Thus, we define the classless logical regularity as a hyperrectangle with great enough information value $I_b(R,S)$.

To illustrate the concept let’s consider a normally distributed two-dimensional set of 100 dots and a set of 100 classless regularities imposed over it (see Fig.1). The regularities for the example are obtained via the random search.

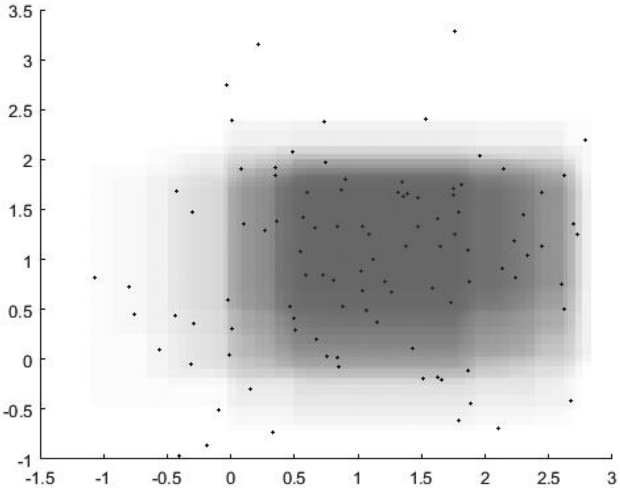

Fig.1 – Classless regularities over a set of random dots.

The intensity of gray corresponds to the number of regularities overlapping the specific area and subsequently the estimated typicality of its precedents. Though the rectangular shape of regularities infer some limitations to the picture the general idea is quite clear. The central area and neighboring dense extensions are considered most typical and the outbound singular objects are estimated outliers.

The limitations will be more obvious on the subsequent figures. The two-cluster structure on the Fig.2 makes the in-between area quite typical, though clusters are formed by the similar distribution as on Fig.1. But if considered separately the same clusters will produce a different picture more resembling the Fig.1.

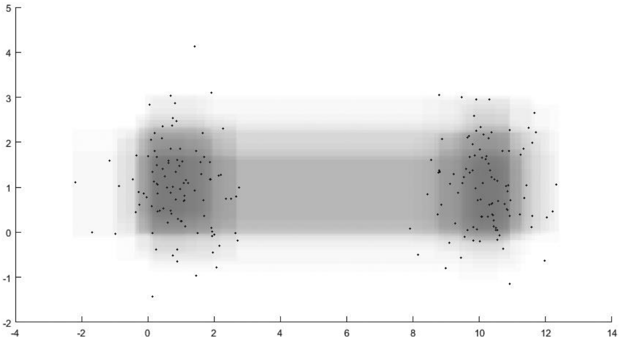

Fig.2 – Classless regularities over two-cluster set.

Quite the opposite can be said about Fig.3. Though the set of dots is circular and quite symmetric, the regularities tend to horizontal and vertical edges making the resulting picture more square like.

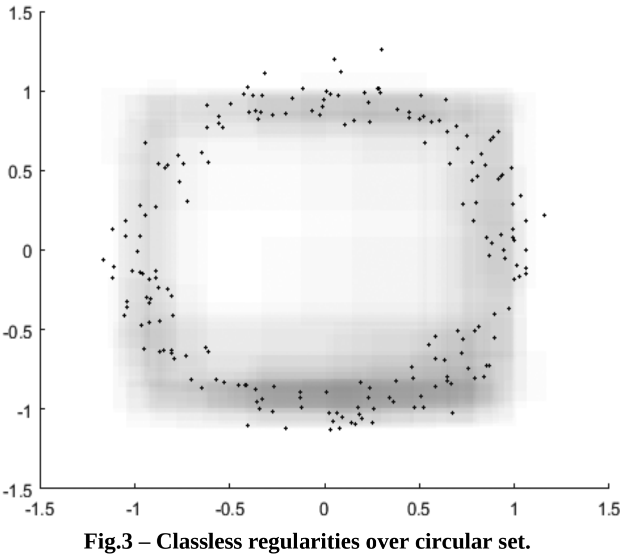

Fig.3 – Classless regularities over circular set.

Everything said it can be mentioned though that the same limitations stand for logical regularities of classes as well. And despite the fact the whole concept works quite well. Thus, the next section will be devoted to the description of the actual algorithm for searching the classless regularities which hopefully will inherit that performance.

3. THE ALGORITHM

In the years since the logical regularities were invented a number of methods for their search were developed. Some of them were integrated into Recognition software system and well tested [3]. For example, in one method the task of searching for regularity is reduced to the task of searching for the maximum consistent subsystem of system of inequalities. Such a system is constructed in correspondence to a particular object that is supposed to be contained by the regularity. The number of considered seed objects is then estimated by a permutation test. Genetic optimization was the basis of another method as well as fastest ascent in a similar approach [5].

In the current paper a simple genetic method will be used to justify the idea with real data. It is described not for its novelty but purely to provide means for reproducing the results. Nevertheless, the method has one peculiarity because it’s aimed for searching a diverse set of regularities in different parts of the considered space instead of one optimal solution and its neighbors. At that the two modifications will be considered. The first one is a straight forward genetic search for the set of best regularities. The second one will inherit the concept of the central object. Thus, it will search for a set of best regularities with respect to a single precedent and then the found sets will be combined.

So, the training sample $S = \{S_1, \ldots, S_m\}$ is considered, each precedent being a real vector $S_i = (a_{i1}, \ldots, a_{in}) \in \Re^n$. A grid of possible regularities’ boundaries is calculated then. Let $\{a_{1j}, \ldots, a_{m_j j}\}$ be the set of unique values of jth feature, $2 < m_j \le m$. The respective set of boundaries will contain the following values:

$$b_{1j} = a_{1j} - \Delta_j ,$$
$$b_{ij} = \frac{(a_{(i-1)j} + a_{ij})}{2}, \; i = 2, \ldots, m_j - 1, \qquad (6)$$
$$b_{m_j j} = a_{m_j j} + \Delta_j .$$

Here Δ_j is half of an average distance between the sample values. When describing the regularity informally the boundary values can be treated as infinity.

The initial population P_0 is defined by selecting left and right boundaries randomly from the respective sets:

$$R_i = r_{i1} \times \ldots \times r_{in} , \; i = 1, \ldots, N_0 , \qquad (7)$$
$$r_{ij} = [b_{q_i j}, b_{t_i j}],$$
$$q_i, t_i \in \{1, \ldots, (m_j + 1)\}, \; q_i < t_i .$$

where N_0 is population size.

Each next generation P_k, $k > 0$ is obtained in following steps.

Step 1. A set of c randomly selected pairs of regularitiess from P_{k-1} is used for producing a posterity C_k. For each pair R_x, R_y a random value $p_{xy} \in [1, \ldots, n-1]$ is generated and C_k is supplemented by a new regularity $R_{xy} = r_{x1} \times \ldots \times r_{xp} \times r_{y(p+1)} \times \ldots \times r_{yn}$.

Step 2. Objects of P_{k-1} undergo mutations to form the set M_k. It means that each interval of each regularity with some probability μ is replaces with a random one.

Step 3. The whole set $[P_{k-1}, C_k, M_k]$ is cleared from duplicates and N_0 of its best objects are transferred to the P_k generation.

Thus, the first algorithm has four parameters: N_0, c, μ, and the number of generations g.

The second algorithm is mostly the same but the initial random population and mutations are performed in respect to a certain object $S_v \in S$. That is the formulae (7) is updated with condition $q_i < a_{vj} < t_i$. The resulting algorithm is executed multiple times in respect to a number of random objects and a part of each resulting set is extracted to form the resulting set of N_0 regularities.

The difference between the algorithms can be clearly seen in the following two images (see Fig.4 and Fig.5). The algorithms are applied to the same random sets of dots with the same set of parameters $N_0 = 1000$, $c = 1000$, $\mu = 0.01$, $g = 10$. At that 20 random seeds are used and 200 best regularities shown.

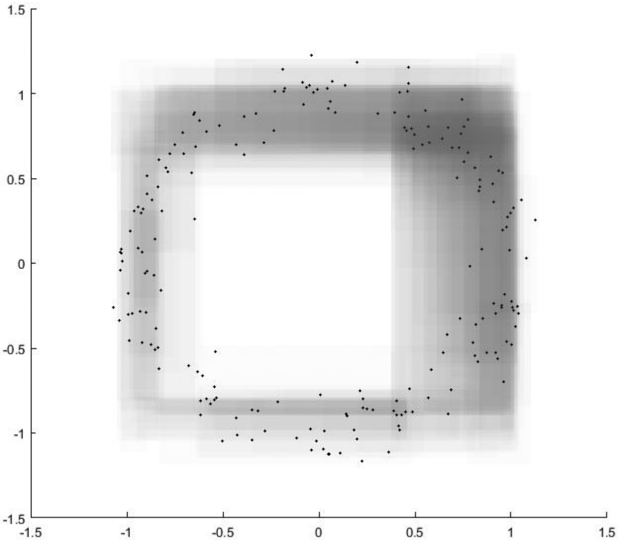

Fig.4 – Results of the first algorithm.

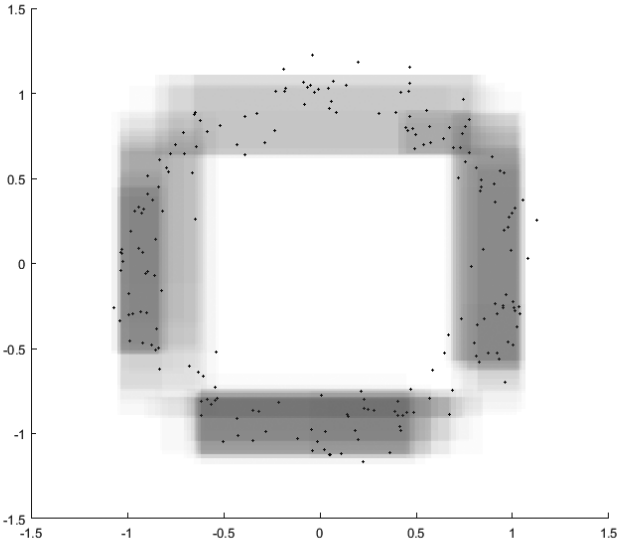

Fig.5 – Results of the second algorithms with 20 seeds.

As can be seen from the figures the second algorithm provides more uniform coverage of dots that better corresponds to expectations. Thus, the second one will be used to analyze data in the following section.

4. REAL WORLD TASKS

Despite artificial examples shown above the proposed method for outlier detection has its drawbacks. The earlier mentioned shape problems are one of them but there are more. For example, missing values or categorical features need special approach. So, a single set of neurological data was chosen to illustrate method's potential, for it has no missing values and all features are binary or quantitative. It is data about outcomes of ischemic stroke [4]. 125 patients are described by 30 features including age, gender, risk factors, anamnesis and so on. Outcomes of disease are also known, bad ones are much rarer as 95 patients survived.

The data was used for unsupervised outliers detection with the proposed method. At that the second modification of the algorithm was applied with the following parameters: $N_0 = 100$, $c = 10000$, $\mu = 0.1$, $g = 10$ and 20 seeds. The resulting numbers of covering regularities were used to estimate patients' typicality and compared to the known outcome of the disease. Correlation coefficient between the estimate and the outcome was 0.3743 that is significant at $p < 2 \cdot 10^{-5}$.

5. CONCLUSION

The concept of classless logical regularities has been proposed. Those are rectangular areas of feature space with statistically justified high density of precedents. For that a special information value functional has been suggested that rewards presence of precedents and penalizes size of the regularity.

The goal of the proposed concept is unsupervised outlier detection that is aimed at finding small number of diverse exceptions in a mostly "normal" data. An algorithm has been proposed to illustrate its potential using artificial and real world tasks. At that simple artificial tasks generate plausible picture and the real one shows significant correlation between the calculated estimate and a real hidden parameter.

The algorithm has its drawbacks though. Namely, it deals poorly with missing values and categorical features. But this is only the first attempt and further research will be aimed at removing those limitations. It is also important to try other approaches beside genetic optimization and to test other applications of logical regularities such as feature selection.

6. ACKNOWLEDGEMENTS

The work was supported by Russian Foundation for Basic Research, grants 13-01-12033, 15-07-00980.

7. REFERENCES

[1] V. Ryazanov. Logical Regularities in Pattern Recognition (Parametric Approach), *Computational Mathematics and Mathematical Physics* 47 (10) (2007), p. 1720-1735.

[2] K. Vorontsov. *Lections on Logical Classification Algorithms (in Russian)*. Moscow, 2007. *http://www.ccas.ru/voron/download/LogicAlgs.pdf.*

[3] Yu. Zhuravlev. V. Ryazanov. O. Senko. *RECOGNITION. Mathematical Methods. Software System. Practical Solutions* (in Russian). Phasis. Moscow, 2006.

[4] Yu. Zhuravlev. G. Nazarenko. A. Vinogradov. A. Dokukin, N. Katerinochkina, E. Kleimenova, M. Konstantinova, V. Ryazanov, O. Sen'ko, and A. Cherkashov. Methods for Discrete Analysis of Medical Data Based on Recognition Theory and Some of Their Applications, *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications.* 26 (3) (2016).

[5] A. Dokukin. Use of Information Value in AVO-Polynomial Method Training, *International Journal "Information Models & Analyses"*. 2 (2) (2013), p. 123-126.