SEARCH FOR NEIGHBORS AND OUTLIERS VIA SMOOTHED LAYOUT

Elena Kleymenova, Elena Nelyubina, Alexander Vinogradov

Abstract: *A new approach to the problem of quick search of the nearest neighbors and outlying objects in the training sample is presented. The approach is based on a special model for the area of mutual attraction of objects whose shape is consistent with the direction of the principal axes, and the form of the attraction dependency is natural. For this type of the zone a smoothed pre-structuring of the sample can be done that allows one to replace laborious procedure for finding nearest neighbors and outliers by simple bit addressing on the set of structure blocks. All actual distances and summed attraction levels for classes are calculated on the final stage for a limited number of objects*

Keywords: *feature space, logical regularity, attraction zone, even distribution, outlier, nearest neighbor, hyper-parallelepiped, bit addressing, dropout threshold*

Introduction

Accumulation and use only reliable information becomes of special importance in tasks of gathering and analyzing big data [Berman, 2003]. But usually such data are recorded just where the risk of error is raised because of intervention of the "human factor" – in medicine, education, environmental monitoring, social surveys and statistics, etc. Thus, often some parts of medical data records significantly differ from the set of average values of parameters for certain kind of patients. This may occur as a result of hardware malfunction, improper use of measuring methods, errors in the recording of results, etc. Similarly the environmental monitoring data are often taken 'in the field', and it's also associated with increased risk of data corruption while its registration and recording into reports. Mathematical treatment of incomplete, inaccurate and partially contradictory data presupposes revealing erroneous objects in order to provide correct application of precise methods of analysis and forecast to the rest of the data. Let R^N be the feature space of a recognition or prediction problem. For small volume of the training sample $X \subset R^N$ each object $x \in X$ makes essential contribution to formation of significant data clusters. It is usually assumed that the object x has its own attraction zone, and it is now known large number of approaches, in which the geometric shape of the attraction zone is modeled in some way - balls, hyper-parallelepipeds, Gaussian "hats", etc. [Tou, 1974]. Such heuristic models allow us to compensate for the deficit of training data at assumption of compactness of classes. In the opposite situation, when the sample has large volume, the use of suitable model helps to optimize the solution, in particular, to reduce the effects of overfitting. Below we consider the problem of exclusion from the training sample the erroneous objects (outliers), which can occur both in small or large-volume sample. The limitation of attraction zone for outliers will serve here as a tool. We are primarily interested in the case of solving problems of recognition and prediction. We describe the data correction method, which is fast at error-detection stage and simultaneously provides efficient addressing for training sample objects and thus the acceleration of algorithms of the type "nearest neighbor".

Model of attraction zone and the generated density

The approach uses a model of attraction area as uniformly filled hyper-parallelepiped with center *x*, volume $\prod_{n=1}^{N}(2a_n+1)$, and density $1/\prod_{n=1}^{N}(2a_n+1)$, where a_n – half of the smoothing interval along the axis *n, n = 1,2, ..., N*. As a result of this smoothing (or spreading) procedure each central object *x* is evenly represented at all points of the hyper-parallelepiped. Location of a new object in zone of attraction of any training object votes for belonging the former to respective class. It is suggested in the approach to consider this impact only in case when the total generated density at a given point exceeds a predetermined threshold.

We will describe one of the reasons for choosing to rectangular zones of attraction. In case of large amounts the problem of analyzing numerous data highlights the priority of processing speed. Under the new conditions simple and well-researched approaches, in particular linear, get rebirth [Berman, 2003]. The most quick are methods in which all calculations can be reduced to comparisons on special linear scales of a particular type. In this series, one of the highly successful approaches turned out to be the one based on the use of Logical Regularities (LR) [Zhuravlev, 2006], [Ryazanov, 2007]. This approach uses data clusters in the form of hyper-parallelepipeds in R^N, each cluster is described by the conjunction of the form $L = \& R_n,\ R_n = (A_n < x_n < B_n)$, and substantially interpreted as recurring joint manifestation of the feature quantities $x = (x_1, x_2, \dots, x_N)$ on intervals $(A_n, B_n), n = 1, 2, \dots, N$. The principle of proximity precedents of the same phenomenon to each other here is embodied in the requirement of filling the interior of the cluster by objects of the same class. Same time, the geometric shape of the cluster represented by the parameters A_n, B_n, becomes of particular importance. Multiple joint appearances of feature values inside this shape are regarded as substantive independent phenomenon that is called Elementary Logical Regularity (ELR).

Thus, in our case the calculation of the predicate of finding new object in the zone of attraction of a training object for the given choice of the form of zone is also reduced to calculation just comparisons of numbers on the main axes omitting more complex operations.

Let's iterate the smoothing process, where each descendant of the central object (i.e. point with non-zero generated density) obtained in the previous steps is considered as a new center of attraction.

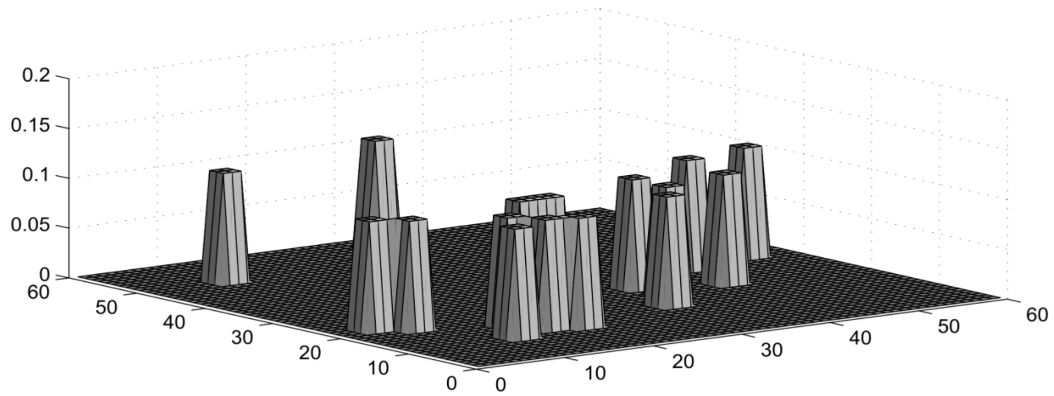

Fig.1. Example of a sample with two outliers after first stages of smoothing.

At s steps of smoothing operation, the attraction zone turns out the hyper-parallelepiped of volume $\prod_{n=1}^{N}(2sa_n+1)$ already unevenly filled with generated density. It is easy to show that the distribution within hyper-parallelepiped rapidly normalized with increasing parameter s, and already for $s > 3$ approximation of the distribution of descendants of a single point via Gaussian $\mu_i \exp\left(-\frac{1}{2}(x_i-x)^T\sigma^{-1}(x_i-x)\right)$ may in some cases be appropriate to construct numerical estimates for classes. With the expansion of volumes $\prod_{n=1}^{N}(2sa_n+1)$ close training objects are beginning to combine their areas of attraction, and this fact results in summation of estimates from neighbors.

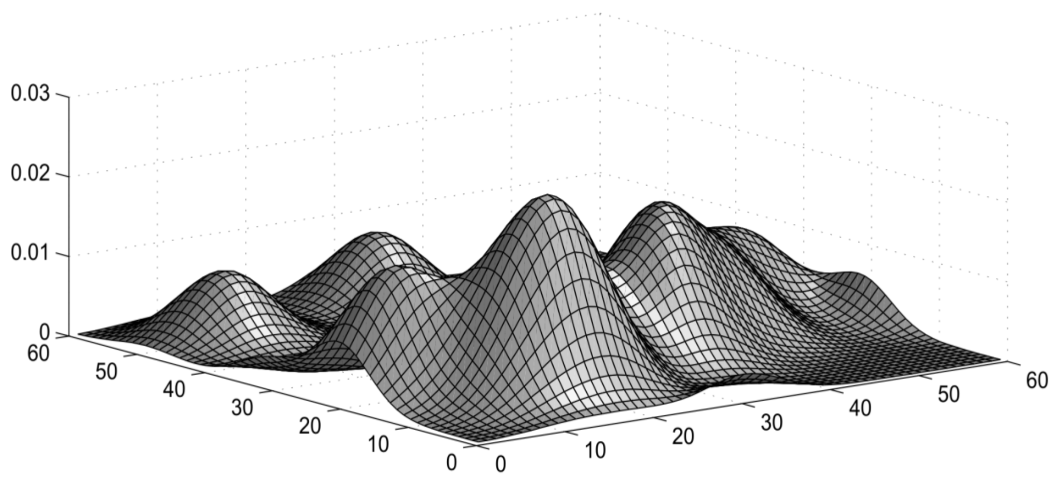

Fig.2. Result of several iterations of the smoothing procedure

The generated density for single objects decreases rapidly, including the maximums in each central point, and at suitable choice of the threshold all isolated objects can be excluded from consideration. The process of normalization generated density for the single point has been well studied, and the desired screening threshold may be calculated in advance with high accuracy. The recognition algorithm proposed in the paper has a structure similar to the standard algorithm of *k* nearest neighbors. It provides advanced possibilities for reconstruction of clusters in generated smoothed densities, and thereby, efficient evaluations for classes. At the same time, controlled expansion of the attraction zones can significantly reduce the amount of search of neighbors at the expense of simple pre-structuring of the sample.

Examples of practical tasks with outliers in data

Below we present two typical practical problems, in which the risk of accidental bias is big in recorded data. The two considered issues are of great importance, and so the quality of such complicated data should be high.

Completion and support of medical registries

In cardiology, neurology, oncology, surgery, including neurosurgery, unified standard forms of reports have been developed on the basis of information from registries. Quality registries are designed for a systematic data acquisition and the application of instruments for improving the quality of healthcare; they can be classified into two categories: disease registries and intervention registries. Quality registries differ from other clinical registries in the existence of special tools that are used in combination with the systematic data acquisition and are aimed at improving the healthcare quality. The tools of support of decision-making analyze the structured data on a patient introduced into the registry and form treatment recommendations on the basis of clinical instructions. During 2013–2014, registries of four directions were introduced with the formation of report templates at the Medical Center of the CB of RF, Moscow. The registry for percutaneous coronary interventions (PCIs, balloon angioplasty and/or stenting of coronary arteries) contains data on 288 patients, of which 138 patients were subjected to planned PCIs and 150 patients were subjected to emergency PCIs. The registry contains 230 indicators, the report on coronary interventions contains 7 sections: implementation of clinical protocols, demographic indicators, characteristics of patients with PCIs, preprocedural state for planned PCIs, preprocedural state for PCIs in case of acute coronary syndrome (ACS) without ST elevation, specific features of the procedure, and postoperative indicators. The registry of the acute cerebrovascular accident (ACVA) is represented by three types: ischemic stroke, hemorrhagic stroke, and transitory ischemic attack (261 patients had ACVA during 2012-2014, of which 197 patients had an ischemic stroke, 25 a hemorrhagic stroke, and 39 a transitory ischemic attack). The registry of ACVA is formed of 240 indicators, and a report on each type of ACVA consists of six sections. By an example of an ischemic stroke, we can represent the contents of the sections: demographic indicators, indicators at prehospital stage, main risk factors, and the estimate of clinical data and the results of examinations (during the first day) during hospitalization and at discharge from the hospital. The registry of general surgery includes 3 nosological forms by which a decision is made on surgery: cholecystitis, appendicitis and inguinal hernia. The registry contains data on 403 patients operated during 2013-2014: cholecystectomy (214), appendectomy (67), and herniotomy (122). The registry contains 260-240 indicators for each patient, taking into account a specific character of pathology. Reports are formed automatically, separately for each nosology according to the following sections: demographic features, estimate of the condition of a patient before operation, the hospital stage, and the audit of the healthcare quality. Oncological registries include prostate cancer (104 patients), gastric cancer (94), renal cancer (64), and pancreas cancer (11).

The scope of indicators ranges within 316-150. The description of sections is made by an example of CPG: demographic data, regular medical check-up, diagnostics, initial treatment, local recurrence, remote metastases, hormone-resistant CPG, and outcomes. The patient's condition and the general and recurrence-free survival rate are evaluated, and the causes of death and the presence of bone fractures are characterized. Electronic forms are developed for all registries: electronic registration form, protocol of observance of clinical recommendations depending on the stage of a disease and individual risk factors, an electronic form for the audit of the results of treatment and clinical outcomes, and an outpatient form of regular medical check-up for assessing remote results. These numerous and complicated data are further used for gathering statistics, taxonomy and classification, recognition and prediction of events in treatment [Zhuravlev, 2016].

Monitoring of water resources ‘in the field’

At present, all over the world a considerable part of the population consumes contaminated water, of poor quality, because many of the local water intakes on the rivers and lakes have lost the quality of drinking water sources by pollution. At the same time there are many man-made factors of changes in the chemical composition of the water of small rivers and lakes: structural changes in aquatic systems, subtraction of river runoff for local economic needs, direct flows of the domestic wastewater into reservoirs, pollution from fertilizers and pesticides, discharge of industrial waters, and others. For these reasons, there is a great need for regular monitoring of small rivers and a comprehensive analysis of the data. For example, such an application for assessment of the quality of water bodies has been directed by the local administration to specialists of Kaliningrad Technical University, Russia. During the work sampling of water was carried out at various water bodies in Zelenogradsky, Nesterovsky, Gusevsky, Krasnoznamensky, Ozersky, Chernyakhovsky, Pravdinsky, Slavsky, Guryevsky, Polessky areas of the Kaliningrad region. All recorded samples showed different exceedances of standards for various types of pollutants. Table 1 shows the comprehensive pollution data from 25 water bodies [Velikanov, 2013].

Table 1. Multiplicity of excess regulations and water pollution index

Object No	Multiplicity of excess regulations						Water pollution index
	Oxygen	BOD_5	Permanganate oxidability	Ammonia nitrogen	Phosphate phosphorus	Ferrous iron	
1	0,78	1,68	2,03	7,88	4,88	1,50	3,12
2	0,71	1,86	2,79	1,76	5,44	3,70	2,71
3	1,04	2,20	3,68	1,70	7,44	2,20	3,04
4	5,36	2,93	5,16	21,1	28,0	3,40	10,9
5	0,63	1,78	3,91	1,88	23,6	0,90	5,45
6	66,7	1,93	4,59	67,8	34,4	3,10	29,7
7	0,77	1,28	4,71	1,86	6,36	1,30	2,71
8	0,69	1,26	2,66	1,66	2,32	1,40	1,66
9	0,76	0,96	2,82	1,18	3,10	1,50	1,72
10	0,89	1,14	2,25	2,18	3,28	1,0	1,79
11	0,88	3,02	2,57	5,43	6,30	0,20	3,06
12	0,85	0,45	3,25	1,32	2,32	0,20	1,40
13	0,66	0,96	7,80	1,32	1,54	2,10	2,40
14	0,77	0,59	3,68	1,79	3,06	0,50	1,73
15	0,79	0,44	2,42	1,52	2,24	1,30	1,45

16	1,58	3,48	3,07	20,4	9,60	1,60	6,62
17	0,62	2,14	3,04	4,22	6,38	1,80	3,03
18	0,75	0,68	3,05	7,88	3,54	2,20	3,02
19	1,48	2,87	7,57	1,76	0,80	3,10	2,93
20	0,82	2,04	4,03	1,70	1,34	2,10	2,01
21	0,86	0,87	5,79	21,1	0,58	1,0	5,03
22	0,75	0,44	2,93	1,88	0,58	0,80	1,23
23	0,96	0,31	2,91	67,8	1,18	0,90	12,3
24	0,96	2,70	10,2	1,86	2,76	3,20	3,61
25	0,81	2,28	2,50	1,66	0,74	0,80	1,46

Samples collected at 3 sites (4, 6, and 23 in the Table 1) were assigned to the 5th class of water quality (extremely dirty). If one is interested in analyses of data for the just ordinary water bodies, then these three precedents should be excluded from consideration or, at least, analyzed separately as representatives of other taxons. The table represents only the most important integrated indicators of pollution, as well as some specific fixed concentration of harmful substances. In fact, for environmental monitoring of water bodies several groups of symptoms is used including organoleptic and sanitary characteristics of the water, indicators of presence of suspensions and emulsions of various substances, objects of micro-flora and other components of biological origin, concentrations of dissolved chemical compounds and individual elements. In total, this list can unite many tens of numerical parameters, and significant part of them has subjectivized expert origin. Of course, for single act of monitoring it is difficult to talk about the use of quite exact methods and techniques of pollutants sampling and recording of the results of their research. But the environmental safety gradually comes to the fore in many different aspects of human life and activity. It should be noted that the North-West Russia is bordered by several EU countries, and for this reason, EU environmental services are very interested in cooperation in matters of protection of water resources and the improvement techniques of monitoring, data storage and analysis. Improving representation of monitoring data recorded 'in the field' can also be a useful factor in ensuring such cooperation.

We will not show here examples of erroneous records found in complex data of this kind, and continue to consider the model example of a sample with outliers from the previous section as an illustration for explaining the algorithm usage in various applications such as the two shown above.

Marking training sample by zones of attraction

In what follows we show the use of controlled expansion of attraction zones of specified kind for arrangement of efficient addressing to the training data. The latter is especially important in the case of large dimensions, as occurs in two practical problems mentioned above.

Let $x_n^m, n = 1, \dots, N, m = 1, \dots, M$, be a training table and $K^l, l = 1, \dots, L$, be its marking by classes. The characteristic function $k(m) = \{l, \ if \ m \in K^l\}$ yields the number of a class by the number *m* of an object in the table. For the object x^0 to be recognized, we will seek a set $T = \{x^p\}, \ p = 1, 2, \dots, P$ of close points of the sample (i.e., nearest neighbors) of the vector x^0, that are located within the hyper-parallelepiped $[x_n^0 - sa_n, x_n^0 + sa_n], \ n = 1, \dots, N$ with volume $\prod_{n=1}^{N}(2sa_n + 1)$, which arises as a result of application of *s* smoothing operations. One should just find all the points of the sample that fall within the hyper-parallelepiped. The fact whether the point falls within the hyper-parallelepiped can be checked independently with respect to each of the axes *n, n=1,2,…,N*, for all the points of the sample $x_n^m, n = 1, \dots, N, m = 1, \dots, M$. For these reasons, one can start the test from any axis, say, from the first, *n=1*, and, on each subsequent axis, check only those points that withstood the closeness test on the previous axes. Having constructed the set $T = \{x^p\}, \ p = 1, 2, \dots, P$, we find all the points of the sample that extend at least minimal attraction to the object x^0.

On each of the main axes $n = 1, \dots, N$ one and the same test on detection sample points $x_n^m, \ n = 1, \dots, N, m = 1, \dots, M$ inside the limits of interval $[x_n^0 - sa_n, x_n^0 + sa_n]$ has to be performed. If the parameters $s, a_n, \ n = 1, \dots, N$ are fixed in advance, this detection is possible only for the points of a certain restricted subset of the training sample. Thus, when structuring the sample into blocks of size $2sa_n + 1$ for each axis *n = 1, ..., N*, the test should only be done to 3^N blocks that are the nearest to the point x^0 in R^N.

We will go further along this path, and choose blocks such that their boundaries are aligned with the binary bit grid. (Without loss of generality, we assume that all of the data are recorded in the fixed-point numbers) Namely, let q_n – the minimum bit such that $2sa_n + 1 \leq 2^{q_n}$, and the sample is structured into blocks with edge length 2^{q_n} for each axis $n = 1, \ldots, N$. Let $w = (w_1, w_2, \ldots, w_N)$ be block indices. Then q_n-th bit of the coordinate x_n^0 of the new object x^0 serves as immediate address w_n of the interval of length 2^{q_n} on the n-th axis, within which we can find the nearest neighbors. Of course, the two other adjacent intervals also should be taken into account, thus 3 intervals in total for each axis.

So, we have replaced the search for the nearest neighbors in the space R^N by the search for the nearest blocks in the space of block indices $W = \{w\}$. Again, such a search can be performed independently on each axis, and we get a subset W^0 of 3^N blocks as a result. After that each selected block is replaced with training objects it contains, and the final direct search is performed to create the set of neighbors $T = \{x^p\}$, $p = 1, 2, \ldots, P$.

A further reduction of the entire volume of the search is possible via organization of hierarchical structuring, when binary (or another but in concordance with the binary) bit mesh is used to construct index, similar to $w = (w_1, w_2, \ldots, w_N)$, as for the entire grid of blocks $W = \{w\}$, as well as within each of the blocks.

Search for the nearest classes and screening outliers

Let $\mu^l(x)$ the total probability density at position x that is generated by attraction zones of points of a class l, $l = 1, \ldots, L$. Using points x^p of the set $T = \{x^p\}$, $p = 1, 2, \ldots, P$, one can construct at the point x^0 a vector of estimates $\mu = (\mu^1, \mu^2, \ldots, \mu^L)$ for summed densities of all classes, $l = 1, \ldots, L$, and use them as votes for respective classes at decision making. In contrast to the ordinary method of k nearest neighbors, one needn't calculate here the distances immediately during the search. Factually, only carrying out the fast hit test for intervals $[x_n^0 - sa_n, x_n^0 + sa_n]$ is enough during the whole search for subset $T \subseteq X$ on the structured sample. Moreover, the first stage of the search consists in simple collecting indices $W^0 \subseteq W$. Distances and exact contributions of neighbors to the total generated densities of classes may be calculated at the final stage after detection of all points $x^p \in T$.

As is known, the evolution of the distribution of multiple smoothing for a single point falls in conditions of the central limit theorem. The deviation of this distribution $F_s(x)$from the multivariate normal

$$\mathcal{N}(x) = \frac{1}{(2\pi)^{\frac{N}{2}}|\sigma|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(x_l - x)^T \sigma^{-1}(x_l - x)\right)}$$

is described by the Berry-Esseen inequality [Berry, 1941], [Esseen, 1942]:

$$|F_s(x) - \mathcal{N}(x)| \leq Const \frac{\rho}{\sigma^3 \sqrt{s}},$$

where functions of the second and third absolute moments of the distribution of single smoothing (and, thus, the values of all variables $a_n,\ n = 1, \ldots, N$) are included as multipliers, and $Const \approx 0.4784$ (the exact value of this constant poses big challenge and continues to be refined in math statistics so far [Shvetsova, 2010]).

Therefore, at strict adherence to the proposed approach the parameter s should always be taken into account, especially in the case of small values. In other cases for larger values of the parameter s, a good estimate can be also obtained easily without reference to this parameter, for example, when used in computing tabulated Gaussian function, which in these cases already is a good approximation for the distribution of multiple smoothing. Same time, the dropout threshold for outliers should be adjusted accordingly, and the reduction of domain of the Gaussian function within boundaries of the block and the necessary renormalization of final distribution should also be considered as additional cost.

One can continue to work with the vector $\mu = (\mu^1,\ \mu^2,\ \ldots, \mu^L)$ in order to decide on the assignment of the object x^0 to one of the classes $l, l = 1, \ldots, L$, such as by using a maximum likelihood criterion, etc.

Thus, in the case of solving problems of recognition or classification the proposed approach leaves a significant range of possibilities. On the contrary, in the case of the problem of sifting alien objects there is an obvious simple way. You can pre-select the parameters s and $a_n,\ n = 1, \ldots, N,$ so that the spread of the region of attraction for s iterations will be consistent with substantive expert views on the dropout of outliers. Then, for strictly single objects of the training sample the maximum of the total generated density $F_z(x^0)$ at the center of the hyper-parallelepiped x^0 can serve as the screening threshold.

Fig.3. presents levels of the total generated density for the training sample of Fig.1. The lowest level serves as a drop-out threshold for two outliers located in the upper part of the figure. At solving the problem of recognition, classification or prediction, the screening of emissions can be carried out in parallel and in coordination with the decision of the main task. The rest of the sample with reliable part of the data can be used with large bases for decisions, identification natural regularities, creation of forecasts, and assessment of various risks.

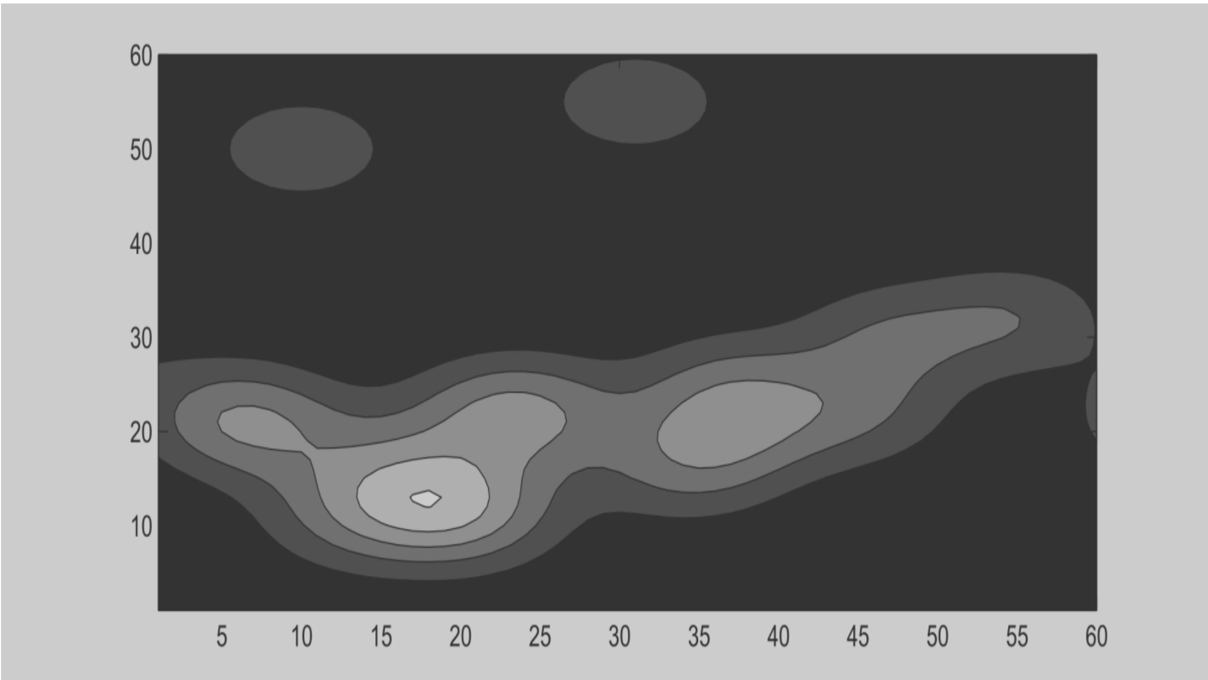

Fig.3. Levels of total generated density built for Fig.1. Senior levels can be used as reliable at construction of a decision rule that cuts off the impact of isolated objects.

Conclusion

This paper presents a fast method of finding the nearest neighbors and separating outlying objects in the training sample. The approach is based on a special model for the zone of mutual attraction of objects and on the pre-structuring of the sample into bit blocks, coordinated with the local geometry of zones. This makes it possible to reduce the search for the set of nearest neighbors to simple choice of a set of relevant blocks, the addresses of which are directly elder bits in the values of parameters of the new object and construction of the final set of neighbors using only comparisons of numbers. The possibilities of the approach analyzed in application to recognition methods of type '*k* nearest neighbors'. In contrast to conventional methods of such kind, there is no need to calculate the distance between the objects immediately. All calculations of distances and local densities of probability distributions for the classes can be made in the final stages and for limited number of objects. Examples are shown of important practical problems in which the decision-making is especially critical with respect to the reliability of the data used in training. In the model example a procedure was presented for removal from the training sample outlying objects that can be executed in parallel and in accordance with the main data processing procedure. The main proposed innovations are related to the model of attraction zone that is compliant to the main axes and uses a natural form of the dependence of attraction on the distance. This model provides efficient bit addressing and fast data processing. In fact, this scheme can be further developed using a variety of other species of such dependence in models of attraction zone but still immersed in hyper-parallelepiped. The approach can be applied in various tasks of information processing and decision making where it is important to monitor the quality of data and perform operational steps to improvement.

Bibliography

[Berry, 1941] Berry, Andrew C. (1941). "The Accuracy of the Gaussian Approximation to the Sum of Independent Variates" // Transactions of the American Mathematical Society 49 (1), pp. 122–136.

[Esseen, 1942] Esseen, Carl-Gustav (1942). "On the Liapunoff limit of error in the theory of probability" // Arkiv för matematik, astronomi och fysik A28: 1–19.

[Tou, 1974] J.T.Tou, R.C.Gonzales Pattern recognition principles, Addison-Wesley, 1977, 390 p.

[Berman, 2003] Berman J. Principles of Big Data, 2003, pp. 1–14.

[Zhuravlev, 2006] Zhuravlev Yu. I., Ryazanov V.V., Senko O.V. "RASPOZNAVANIE. Matematicheskie metody. Programmnaya sistema. Prakticheskie primeneniya", Moscow: Izdatelstvo ”FAZIS” 2006, 168 p. (Russian).

[Ryazanov, 2007] Ryazanov V.V.: Logicheskie zakonomernosti v zadachakh raspoznavaniya (parametricheskiy podkhod) // Zhurnal vychislitelnoy matematiki i matematicheskoy fiziki, T. 47/10 (2007) str.1793 -1809 (Russian).

[Shvetsova, 2010] Shevtsova, I. G. (2010). "An Improvement of Convergence Rate Estimates in the Lyapunov Theorem". Doklady Mathematics 82 (3): 862–864

[Velikanov, 2013] N.L. Velikanov, V.A. Naumov, L.V. Markova, A.A. Smirnova "Results of Natural Researches of Small Water Flows on Reclaimed Soils" // Water: chemistry and ecology No 7 2013 pp.18-26 (Russian).

[Zhuravlev, 2016] Yu. I. Zhuravlev et al. "Methods for discrete analysis of medical information based on recognition theory and some of their applications" // Pattern Recognition and Image Analysis, 2016 (to appear).

Authors' Information

Elena Kleymenova *– Professor, Medical Center of the Bank of Russia, Sevastoposkii pr. 66, 117593 Russia; e-mail: e.kleymenova@gmail.com*

Elena Nelyubina *– Assistant Professor, Kaliningrad State Technical University, Sovietsky prospect 1, 236022 Kaliningrad, Russia; e-mail: e.nelubina@gmail.com*

Alexander Vinogradov *– Senior Researcher, Dorodnicyn Computing Centre, Federal Research Center “Computer Science and Control” of Russian Academy of Sciences; Vavilova 40, 119333 Moscow, Russia; e-mail: vngrccas@mail.ru*