Toward approaches to big data analysis for terroristic behavior identification: child soldiers in illegal armed groups during the conflict in Donbas region (East Ukraine) -

Data classification approach

The accumulated massive of data should be classified. In other worlds, a decision making procedure should be applied to each entry of general community - the multidimensional dataset to extract a class of "infant participant of illegal armed groups". We propose to use for it a data classification approach based on the Bayes rule for minimum classification error in terms of maximum-a-posterior decision task in Markov random field model representation of multi-temporal, multi-source data (Duda, Hart, Stork, 2012).

Let I_j $\{x_1, x_2, ..., x_N\}$ be a given find record describing a community of illegal armed groups combatants, modeled as a set of N identically distributed n-variate random vectors of identifiers. We assume M classes $\omega_1, \omega_2, ..., \omega_M$ to be present among a set of combatants and we denote the resulting set of classes by $\Omega=\{\omega_1, \omega_2, ..., \omega_M\}$ and the class label of the k-th community member ($k=1, 2, ..., N$) by $s_k \in \Omega$. By operating in the context of community classification, we assume that verified events, recognized as a training set to be available, and we denote the index set of the training community member by $T \subset \{1,2,...N\}$ and the corresponding true class label of the k-th training community member ($k \in T$) by s_k*.

When collecting all the feature vectors of the N member of community of combatants in a single ($N \cdot n$)-dimensional column vector $X= col[x_1, x2, ..., x_N]$ and all the community member labels in a discrete random vector $S = (s_1, s_2, ..., s_N) \in \Omega^N$, the Bayes rule approach (Duda, Hart, Stork, 2001) assigns to the community of combatants data X the label vector $\widetilde{S}$, which maximizes the joint posterior probability $P(S \mid X)$:

$$\widetilde{S} = \arg\max_{S \in \Omega^N} P(S \mid X) = \arg\max_{S \in \Omega^N} [p(X \mid S)P(S)] \qquad (1)$$

where $p(X \mid S)$ and $P(S)$ are the joint probability density function (PDF) of the global feature vector X conditioned to the label vector S and the joint probability mass function (PMF) of the label vector itself, respectively. The Markov random field (MRF) approach offers a computationally tractable solution to this maximization problem by passing from a global model for the statistical dependence of the class labels to a model of the local general community properties, defined according to a given neighborhood system (Geman, S., Geman, D., 1993). Specifically, for each k-th member of general community, a neighborhood $N_k \subset \{1,2,...N\}$ is assumed to be defined, such that, for instance, N_k includes the four (first-order neighborhood) or the eight (second-order neighborhood) member surrounding the k-th member ($k=1, 2, ..., N$). More formally, a neighborhood system is a collection $\{N_k\}_{k=1}^N$ of subsets of member such that each member is outside its neighborhood (i.e., $k \notin N_k \forall k = 1,2,...N$) and neighboring member are always mutually neighbors (i.e., $k \in N_h$ if and only if $h \in N_k \forall k, h = 1,2,...N, k \neq h$). This simple discrete topological structure attached to the general community data is exploited in the Markov random field framework to model the statistical relationships between the class labels of spatially distinct member and to provide a computationally affordable solution to the classification problem of (1). Specifically, we assume the feature vectors $x_1, x_2, ..., x_N$ to be conditionally independent and identically distributed with probability density function $p(x \mid s)$ ($x \in \Re^n, s \in \Omega$), that is (Geman, S., Geman, D., 1993):

$$p(X \mid S) = \prod_{k=1}^{N} p(x_k \mid s_k) \qquad (2)$$

and the joint prior probability mass function $P(S)$ to be a Markov random field with respect to the abovementioned neighborhood system. The probability distribution of each *k*-th general community label, conditioned to all the other general community labels, is equivalent to the distribution of the *k*-th label conditioned only to the labels of the neighboring members (*k=1, 2,..., N*):

$$P\{s_k = \blacktriangleleft_i \mid s_h : h \neq k\} = P\{s_k = \blacktriangleleft_i \mid s_k : h \in N_k\}, i = 1,2,...N \qquad (3)$$

The probability mass function of *S* is a strictly positive function on Ω^N, so, $P(S) > 0 \forall S \in \Omega^N$.

The Markov assumption expressed by (3) allows restricting the statistical relationships among the general community labels to the local relationships inside the predefined neighborhood, thus greatly simplifying the contextual model for the label distribution as compared to a generic global model for the joint probability mass function of all the general community labels.

Basing on described approach, the Ho–Kashyap method (Ho, Kashyap, 1965) has been applied to classification of the general community.

As the result of classification we obtain a dataset with all records that meet the specified condition. For example, distribution of members of community "infant participant of illegal armed groups" with age, sex, social status, accessory, spatial and temporal marks inside the general community of members of illegal armed groups' combatants.

Data regularization algorithm

Further, we need regularized spatial-temporal distribution of classified records. It is necessary to avoid duplication and fix falsification, inter-verify data, trace robust trends in the distributions, so get a basis for interpretation of the data. We propose to use a two-stage procedure of data regularization.

The method proposed is based on non-linear kernel-based principal component algorithm (KPCA) modified according to specific of data. Using this method the set of selected records has been analyzed. The robust technique of data regularization for normalization of data reliability is proposed. The technique utilizes data from different sources, different nature, and with different metrics. This approach allows to calculate regularized distributions in units invariant toward data properties and quality. We can analyze simultaneously different types of data, regardless spatial and temporal scales and heterogeneities using this approach.

Correct statistical analysis requires the set of data $\boldsymbol{x}_i$ with controlled reliability (above mentioned "training set"), which reflects distribution of investigated parameters during whole observation period (taking into account variances of reliability of data $\boldsymbol{x}_t$). Set of data $\boldsymbol{x}_t$ ($\mathbf{x}_t \in R^m$) consists of multi-source data, including data with sufficient reliability $\boldsymbol{x}_j$ ($\mathbf{x}_j \in R^m$), where $j = 1, \ldots N$. Problem of determination of controlled quality and reliability spatial-temporal distribution of investigated parameters $\boldsymbol{x}_i$ might be solved in framework of tasks of multivariate random processes analysis and multidimensional processes regularization (Kostyuchenko, Movchan, Kopachevsky, et al, 2015).

Required regularization may be provided by 2-stage procedure. If we able to formulate stable hypothesis on distribution of reliability of data in the framework of defined problem we may to propose relatively simple way to determine investigated parameters distributions $\boldsymbol{x}_t^{(x,y)}$ towards distributions on measured sites $\boldsymbol{x}^m_t$ basing on (Kostyuchenko, Movchan, 2015):

$$x_t^{(x,y)} = \sum_{m=1}^{n} w_{x,y}(\tilde{x}_t^m) x_t^m \qquad (4)$$

where weighting coefficients $w_{x,y}(\tilde{x}_t^m)$ determined as:

$$\min\{\sum_{m=1}^{n}\sum_{x_t^m \in R^m} w_{x,y}(\tilde{x}_t^m)(1-\frac{x_t^m}{\tilde{x}_t^m})^2\} \quad (5)$$

according to (Kostyuchenko, Movchan, 2015). Here m – number of records; n – number of sources/series; $\boldsymbol{x}^m{}_t$ – distribution of data; R^m – set (aggregate collection) of data; $\tilde{x}_t^m$ - mean distribution of searching parameters.

This is the simple way to obtain a regular spatial distribution of analyzed parameters, on which we can apply further analysis, in particular temporal regularization. At the same time, this is the first stage of regularization. This algorithm may be interpreted as the general form of Kolmogorov regularization procedure (Ermoliev, Makowski, Marti, 2012).

Further, second stage of regularization should take into account both data distribution temporal non-linearity (caused by imperfection of available multi-source statistics) and features of temporal-spatial heterogeneity of data distribution caused by systemic complexity of studied community. According to (Kostyuchenko, Movchan, 2015) the kernel based non-linear approaches are quite effective for analysis of such types of distributions.

Proposed method is based on modified kernel principal component analysis (KPCA). In the framework of this approach the algorithm of non-linear regularization might be described as following rule (Kostyuchenko, Movchan, Kopachevsky, et al, 2015):

$$x_i = \sum_{i=1}^{N} \checkmark_i^k \tilde{k}_t(x_i, x_t) \quad (6)$$

In equation (6) the coefficients α selected according to optimal balance of relative validation function and covariance matrix, for example as (Kostyuchenko, 2015):

$$C^F v = \frac{1}{N}\sum_{j=1}^{N}\Phi(x_j)\Phi(x_j)^T \cdot \sum_{i=1}^{N} \checkmark_i \Phi(x_i) \quad (7)$$

Where non-linear mapping function of input data distribution Φ determined as:

$$x_i = \sum_{i=1}^{N} \checkmark_i^k \tilde{k}_t(x_i, x_t) \quad (8)$$

$$\sum_{k=1}^{N}\Phi(x_k) = 0 \quad (9)$$

And $\widetilde{k_t}$ - is mean values of kernel-matrix $\mathbf{K} \in R^N$ ($[\mathbf{K}]_{ij} = [k(\mathbf{x}_i, \mathbf{x}_j)]$). Vector components of matrix determined as $\mathbf{k_t} \in R^N$; $[\mathbf{k_i}]_j = [k_t(\mathbf{x_t}, \mathbf{x_j})]$. Matrix calculated according to modified rule: $\mathbf{k}_t(\mathbf{x}_i, \mathbf{x}_t) = \left\langle \times_{j,t}^{x_j}(1-\times_{j,i})^{x_j} \right\rangle$, where ρ – empirical parameters, selected according to the classification model of study community.

Using described algorithm, it is possible to obtain regularized spatial-temporal distribution of investigating parameters over the whole observation period with rectified reliability and controlled uncertainty (Kostyuchenko, 2015; Kostyuchenko, 2016). As the result we obtain a regular spatial-temporal distribution of child soldiers registered, prepared for the interpretation.