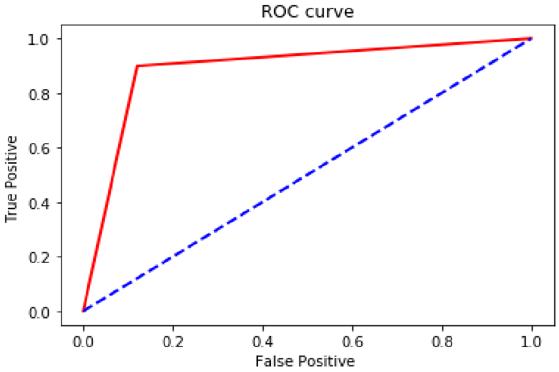

Fig. ROC curve of logistic regression model on “HTTP dataset CSIC 2010” dataset

*This work was supported by the Science Development Foundation under the President of the Republic of Azerbaijan – **Grant No. EIF-KETPL-2-2015-1(25)-56/05/1.***

1. Yu J., Tao D., Lin Z. A hybrid web log based intrusion detection model, Proc. of the IEEE 4th international conference on cloud computing and intelligence systems (CCIS), 2016, Beijing, China, pp. 356-360.

2. Zolotukhin M., Hämäläinen T., Analysis of HTTP requests for anomaly detection of web attacks, IEEE 12th international conference on dependable, autonomic and secure computing, 2014, Dalian, China, pp. 406-411.

UDC 004.056.5

R.M. Alguliyev, R.M. Aliguliyev, L.V. Sukhostat
e-mail: r.alguliev@gmail.com, r.aliguliyev@gmail.com,
lsuhostat@hotmail.com

Institute of Information Technology, Azerbaijan National Academy of Sciences, Baku, Azerbaijan

PURITY-BASED CONSENSUS CLUSTERING FOR ANOMALY DETECTION IN BIG DATA

The paper proposes a weighted consensus clustering for efficient integration of single clustering methods.

A consensus approach is widely used to increase the accuracy and stability of clustering results [1–2]. The proposed method uses a purity-based utility function to aggregate the single clustering methods into a consensus one.

Let us denote the following notations: $X = \{x_1, x_2, ..., x_n\}$ are the points in the dataset, where n is the total number of data points in the dataset. The partition of X into K crisp clusters is represented as a collection of K subsets of objects in $C = \{C_k \mid k = 1, ..., K\}$, $C_k \text{ I } C_{k'} = \varnothing \ \forall k \neq k'$ and $\bigcup_{k=1}^{K} C_k = X$ or as a vector of labels $\pi = \left(L_\pi(x_1), L_\pi(x_2), ..., L_\pi(x_n)\right)^T$, for any i $x_i \xrightarrow{L_\pi} \{1, 2, ..., K\}$. Given the r basic partitions from X to $\Pi = \{\pi_1, \pi_2, ..., \pi_r\}$, $1 \le i \le r$. The goal is to find a consensus partition π by solving the following optimization task:

$$f(x) = (1-\lambda)\sum\nolimits_{i=1}^{r} w_i U(\pi, \pi_i) + \lambda \|w\|^2 \to \max \qquad (1)$$

subject to

$$\pi = w_1\pi_1 + w_2\pi_2 + ... + w_r\pi_r, \qquad (2)$$

$$\sum\nolimits_{i=1}^{r} w_i = 1, \quad w_i \ge 0 \quad \forall i, \qquad (3)$$

where $0 \le \lambda \le 1$ is the regularization parameter [3]. The algorithm of the proposed approach for anomaly detection is as follows:

Input: $\{\pi_1, \pi_2, ..., \pi_r\}$: basic partitions

λ: regularization parameter

$w^{(0)} = \{w_1^{(0)}, w_2^{(0)}, ..., w_r^{(0)}\}$: initial weights of each basic partition

k: number of clusters

Output: Consensus π

Step 1. Obtain a set of basic partitions $\pi_i \ (i = \overline{1, r})$

Step 2. Construct a co-association matrix

Step 3. $s = 0$

Step 4. Calculate the value of the function according to (1) taking into account conditions (2) and (3)

$$f^{(s)} = (1-\lambda)\sum\nolimits_{i=1}^{r} w_i^{(s)} U(\pi, \pi_i) + \lambda \left\| w^{(s)} \right\|^2$$

Step 5. Find $w^{(s)}$.

Step 6. $s = s + 1$

Step 7. Repeat steps 4-6 until the convergence condition is met

Step 8. Return the partition π

End

The experiments were focused on comparing the clustering results of the proposed approach with different distance metrics. The weights of the DBSCAN, OPTICS and k-means are smaller than the weights of the CLARANS and shared nearest neighbor Clustering (SNNC) methods. The best result for $\lambda = 0.2, 0.6, 0.7$ and $\lambda = 0.9$ showed the CLARANS. The best results were obtained at $\lambda = 0.6$. A set of weights obtained with the proposed approach is shown in the following table. The best partition was determined when applying the squared Euclidean and cosine distances. Based on the experimental results, it can be concluded that the proposed approach compensates for the shortcomings of each considered method and increases the efficiency of clustering.

Table.

Experimental results

	DBSCAN	OPTICS	CLARANS	k-means	SNNC
Euclidean distance	0.202	0.197	0.225	0.201	0.175
Cosine distance	0	0	0	1	0
Squared Euclidean distance	0	0	0.999	0	0.001
Minkowski distance (p=3)	0.201	0.199	0.222	0.200	0.178
Minkowski distance (p=4)	0.200	0.200	0.220	0.200	0.180
Chebychev distance	0.185	0.224	0.214	0.184	0.193

This work was supported by the Science Development Foundation under the President of the Republic of Azerbaijan – Grant № EİF-KETPL-2-2015-1(25)-56/05/1.

REFERENCES

1. Wu J., Wu Z., Cao J., Liu H., Chen G., Zhang Y. Fuzzy consensus clustering with applications on Big data // IEEE Transactions on Fuzzy Systems, 2017, vol. 25, no. 6, pp. 1430–1445.

2. Liu H., Wu J., Liu T., Tao D., Fu Y. Spectral ensemble clustering via weighted k-means: theoretical and practical evidence. IEEE Transactions on Knowledge and Data Engineering, 2017, vol. 29, no. 5, pp. 1129–1143.

3. Alguliev R. M., Aliguliyev R. M., Fataliyev T. Kh, and Hasanova R. Sh. Weighted Consensus Index for Assessment of the Scientific Performance of Researchers. COLLNET Journal of Scientometrics and Information Management, 2014, vol. 8, no. 2, pp. 371–400.

UDC 004.93

M.Sh. Hajirahimova, A.S. Aliyeva
e-mail: makrufa@science.az, aliyeva.a.s@mail.ru

Institute of Information Technology of ANAS, Baku, Azerbaijan

METHODS AND MODELS FOR PREDICTING OİL RESERVİOR PVT PROPERTIES: A REVIEW

Knowledge about reservoir fluid properties plays a vital role in improving reliability of oil reservoir simulation. In the absence of laboratory PVT data fluid properties are predicted by empirical correlations. A large number of empirical correlations have been developed for prediction of PVT properties of crude oil.

An adequate knowledge of any PVT (pressure-volume-temperature) properties is very important in reservoir engineering computations such as material balance calculations, reserve estimation, well test analysis, and planning future enhanced oil recovery projects. Determination of such properties in laboratory is too expensive and time consuming. A large number of empirical correlations and models have been developed for prediction of PVT properties of crude oil.

There are two main categories of methods for predicting oil reservoir PVT properties in the literature:empirical correlations and ML techniques.