Methods of Collective Intelligence in Exploratory Data Analysis: A Research Survey

Piotr A. KOWALSKI[1,2,*],Szymon ŁUKASIK[1,2] and Piotr KULCZYCKI[1,2]

[1]Faculty of Physics and Applied Computer Science,
AGH University of Science and Technology,
al. A. Mickiewicza 30, 30-059 Cracow, Poland,

email: {pkowal, slukasik, kulczycki}@agh.edu.pl

[2]Systems Research Institute, Polish Academy of Sciences,
ul. Newelska 6, 01-447 Warsaw, Poland,

email: {pakowal, slukasik, kulczycki}@ibspan.waw.pl

*Corresponding author

Keywords: Computational Intelligence, Collective Intelligence, Exploratory Data Analysis, Data Science, Classification, Clustering, Outlier Detection, Data and Dimensionality Reduction, Metaheuristics.

Abstract. This study contains a brief presentation of the basic tasks for Exploratory Data Analysis (EDA), namely: classification, clustering, reduction of data dimensionality and number of data instances as well as detection of outliers. Herein, solutions to the aforementioned problems incorporating a wide range of computational intelligence algorithms, in particular procedures based on collective intelligence, are under consideration. Furthermore, the combination of metaheuristic algorithms with basic EDA procedures applied and verified within many domains of science, technology and engineering are being presented.

Introduction

Exploratory data analysis is currently one of the most active field of research. Its use is exemplified for instance in the organization of collections of digital information contained in large digital libraries or in image archive management tasks dealing with large sets of illustrations associated with some text and numerical descriptions. It is also employed in the health care domain. There, we can distinguish its use in the interpretation of medical pictures (assembly of EKGs, MRIs, radiographs, etc.), and in the handling of non-image-types of patient information that has been gathered in databases (i.e., the results of morphological tests, blood pressure, temperature, etc.). Another embodiment of this domain is the area of business, manufacturing and marketing. Here, analysed data is utilized for forecasting, planning and management tasks. In factories, huge sets of data can be efficiently analysed with regard to current operating states so as to achieve a natural optimization of the production process. One more very important EDA application lies within fields of telecommunication and computer networks. In this case, an immense data sets containing diverse types of information, very often require analysis and identification in an online regime. Recently, a huge demand for data analysis has come about the World Wide Web arena. Indeed, within the Internet, we are practically surrounded by a vast variety of data that is of distributed type which does not facilitate its analysis and represents a significant challenge for contemporary data mining [26].

From the above examples, we can easily deduct that modern data analysis must be dedicated towards handling large datasets. This applies both to the amount of data instances, as well as to the number of their attributes. For this reason, the classical data science methods fail and there is a natural need to replace them with more advanced procedures that often are based on computation intelligent techniques. While the application of current analytical methods guarantees optimal solutions, yet, in most cases, it does not allow to handle the large number of data. What is more, a problem arises due to the sensitivity of these conventional procedures to initial conditions and to the internal value of algorithm's parameters. That is why in contemporary data science, we often choose algorithms that provide approximate solutions but with reduced computational effort. These procedures are very often based on methodology of computational intelligence.

This paper aims to provide the review of our work carried out on the application of advanced computational intelligence algorithms, especially that of swarm intelligence dedicated to data science tasks. Initially, the basic data analysis tasks will be described, subsequently, the paper will inform the reader about traditional methods of computational intelligence such as Fuzzy Logic (FL), Artificial Neural Networks (ANN) and Evolutionary Algorithms (EA), as well as the more advanced procedures of collective intelligence in the context of their application in selected EDA tasks.

An Overview of Basic Issues of EDA

In this Section, the main tasks of data science are shortly introduced. Figure 1 demonstrates a typical flowchart of data processing activities. In such scheme, we initiate the process of data analysis by acquiring a collection of samples from the object of concern. In this case, the word object can have multiple meanings, on one hand, it may refer to a specific technical device, on the other – data may be obtained from economic process or from other abstract phenomena.

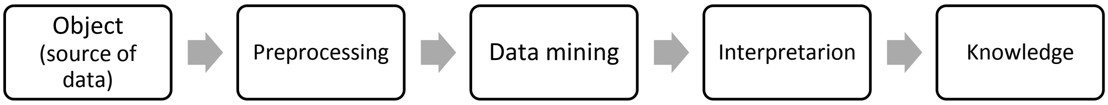

Figure1. Data processing steps

Very often, acquired raw data consists of elements or samples which are difficult to analyse. In addition to that, sometimes they are neutral or even harmful for the efficiency of data mining and its interpretation. Thus, dataset should be normalized, noise removal has to be undertaken and missing data should be additionally considered. All of these are considered as data pre-processing stage. A baseline and very important EDA task, and one considered to be very challenging in data analysis, is outliers detection. It should be highlighted that some of procedures are resistant to outliers, stillat the same time their detection and discovering their sources can lead to additional interesting findings.

Let us denote the data sample as:

$$X = x_1, x_2, \dots\, x_M \qquad (1)$$

where each element of data set (1) could be interpreted as being a point in N dimensional space, i.e. $x_i = x_{i1}, x_{i2}, \dots\, x_{iN}$. In statistics, we can define an outlier as being a data sample element x_i that is exceptionally distant from other samples in X. It is worth to note however that there is no strict mathematical definition of what constitutes an outlier, so determining the outliers subset is ultimately a subjective exercise. Outliers detection methods can be classified as falling within the following categories: Extreme Value Analysis, probabilistic and statistical models, linear models, proximity-based models, information theoretic models and high-dimensional outlier detection [26].

Subsequent and very important class of algorithms employed for data preprocessing are procedures which free data from the redundant or harmful information. Data reduction procedures consist of dimensionality and sample length reduction. The first task, frequently referred to as data compression, assumes a decrease of the data set numerosity, i.e. transformation is undertaken to obtain a representation of the set $X$with $M^{'} < M$ sample examples. The second type of data reduction is that of dimensionality reduction. This is based on the algorithms of feature selection or feature extraction. Within this procedure, the data elements are represented by a smaller number of features, i.e. with$N^{'} < N$. Here, it is worth stating that data extraction from such data sets is a very complicated task. Generally, limitations are encountered due to computer system performance when processing large data sets. Another obstacle is the negative influence of the "curse of dimensionality". After applying aforementioned preprocessing procedures, the transformed data is represented in a compact manner. In the steps following data processing, we have at our disposal for the tasks data analysis, a variety of methods based on traditional, as well as modern statistical techniques. One of the fundamental EDA procedures is clustering. Clustering or cluster analysis corresponds to the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Clustering, as an optimization task, is known to represent a NP-hard complexity, with numerous heuristic approaches being employed to find representative groups. For this task, we can distinguish the following categories of procedures: partitioning methods, hierarchical algorithms, density-based procedures, grid-based methods, and, finally, the computational intelligence-based approaches [26]. In the case of almost all of these categories, we can consider utilizing as well, constraint-based procedures. It should be noted that the clustering algorithm can also be used as an approach for outlier detection.

Another extremely important data analysis task is classification. This considered to be a so-called supervised method. Therefore, in this case, we need to expand the representation of data (1) by way of an additional M-element vector which contains information about the membership of individual elements of $X$to particular classes. Hence, the act of classification consists of assigning the element under consideration to one among a number of previously defined sets of categories called “classes”. These are most often represented by samples constituting sets of elements representative for particular classes. In this case, we should stress the double stage character of the classification task. The first phase it is difficult and often time consuming, as it is the process of constructing the classifier. The second stage involves identifying to which class, a new data item belongs to, and is achieved by the way of applying the ready-to-use algorithm obtained in the first step. The best known and most frequently used algorithms for performing this task are based on the ideas oflinear classifiers, Bayes classifier, Support Vector Machines (SVM), quadratic classifiers, statistical Kernel Density Estimation (KDE), decision trees, Neural Networks and Learning Vector Quantization [26].

In modern data analysis, we quite often expect to use these algorithms to manage diverse data characteristics, i.e. real, integer, categorical or binary, but we also must consider having to cope with information which, in different forms, exhibits an imprecision that is dependent upon the conditions of the problem. Among these are "uncertain" feature values – as seen in statistical methods or "fuzzy" values known in fuzzy logic and, finally, that which arise in handling imprecise information of interval type [13].

Finally, the last steps in the data analysis procedure is the interpretation of results and the application of discovered knowledge.

Optimization Inspired by Nature as a Tool for EDA – a Survey of Research Results

In this part of the paper, selected results obtained by using metaheuristics procedures as a tool for data mining are being presented. Although many methods of collective intelligence can be applied, in this study, we focus mostly upon utilizing the most recent procedures. The first is the Krill Herd Algorithm (KHA) introduced in 2012 [4], the subsequent is the Flower Pollination Algorithm (FPA) invented by Yang in 2012 [30], the third is the Firefly Algorithm (FA), created by Yang, in 2009[29]. We shall also present the applications of older and well-known metaheuristic algorithms for such purpose, namely Particle Swarm Optimization (PSO) [23] and Evolutionary Algorithms (EA) [1]. Furthermore, we reveal the results of using classical computational intelligence type of algorithms like FL, ANN etc. Due to the random nature and novelty of algorithms inspired by nature, in first stage of presented investigation, the basic research had to be performed for KHA [7], FPA[24], FA[21] and PSO [23] procedures.

The modern heuristic algorithm of KHA and FPA are applicable for the purposes of deriving best solution for the clustering task [10, 12, 25]. For these, the comparison with regards to the quality of obtained results has been performed employing selected real and synthetic data sets drawn from the UCI Machine Learning Repository. In particular – in [10] – Celinski-Harabasz, Davies-Bould in and Silhouette Value Indexes as clustering variants has been utilized to validate the cluster division criteria. Furthermore – in [2, 15] – for clustering purposes the Complete Gradient Clustering heuristic algorithm, based on the density of the data is introduced, and in this case is applied to a real-world data set of grains [3], as well as to other benchmark data sets.

The KHA metaheuristic was also successfully applied in the classification task as a tool for the training of neural classifier [9]. Moreover, EA for the Fuzzy Flip-Flop Neural Network [6] is adaptable as a more advanced classifier. The notion of classification has been extended to interval data processing [14], and a new type of ANN, called "Interval Probabilistic Neural Network" introduced in [8]. Additionally, as demonstrated in [5, 11], reducing pattern data, with regards to the classification task involving interval type data is also feasible. Subsequent research considers a classification algorithm dedicated for situations in which samples of data were characterized by nonstationarity for both real [17] and interval [18] types of sample sets.

The concept of reducing the dimension and size of a data set can be based on a linear transformation towards a space of a smaller dimension. It is a case for traditional Principal Component Analysis algorithm. Here, we consider similar strategy assuring preserving distances between data sample elements and the metaheuristics of a Parallel Fast Simulated Annealing [16, 17, 22]. This approach is applicable to all EDA procedures. Another algorithm based on utilizing the sensitivity analysis procedure, as applied in the computation of the weights for the PNN model, can also be dedicated towards pruning the neural network used for the classification task [20].

In the course of research on the issue of outliers detection procedures, certain methodologies have been formulated both in statistical terms [19], as well as being based on the computational intelligence techniques [11]. Finally, recent investigations of our team concern notably the application of modern metaheuristics, in particular, KHA and FPA, for the modelling and optimization of large databases [27, 28].

References

[1] J. Arabas, Evolutionary Computation for Global Optimization – Current Trends, Journal of Telecommunications and Information Technology 4 (2011) 5–10.

[2] M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Łukasik, S. Żak, A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images, Information Technologies in Biomedicine, E. Pietka, J. Kawa (eds.), Springer-Verlag, Advances in Intelligent and Soft Computing, vol. 2, pp. 15-25 (2010).

[3] M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Łukasik, Discrimination of Wheat Grain Varieties Using X-Ray Images, Information Technologies in Medicine: 5th International Conference, ITIB 2016 Proceedings, vol. 1, E. Piętka, P. Badura, J. Kawa, and W. Wieclawek (eds.) Springer, pp. 39-50 (2016).

[4] A.H. Gandomi A.H. Alavi, Krill herd: A new bio-inspired optimization algorithm. Commun Nonlinear Sci Numer Simul, 17(12) (2012) 4831-4845.

[5] P.A. Kowalski, P. Kulczycki, Data Sample Reduction for Classification of Interval Information using Neural Network Sensitivity Analysis, LNAI, 6304 (2010) 271-272.

[6] P.A. Kowalski, Evolutionary Strategy for the Fuzzy Flip-Flop Neural Networks Supervised Learning Procedure, LNAI, 7894 (2013) 294-305.

[7] P.A. Kowalski, S. Łukasik, Experimental Study of Selected Parameters of the Krill Herd Algorithm, Intelligent Systems'14, P. Angelov (eds.), Springer pp 473-485 (2015).

[8] P.A. Kowalski, P. Kulczycki, Interval Probabilistic Neural Network, Neural Computing and Applications, (2015)DOI: 10.1007/s00521-015-2109-3 (in print).

[9] P.A. Kowalski, S. Łukasik, Training Neural Networks with Krill Herd Algorithm, Neural Process Lett 44(1) (2016) 5–17, DOI 10.1007/s11063-015-9463-0.

[10] P.A. Kowalski, S. Łukasik, M. Charytanowicz, P. Kulczycki Clustering based on the Krill Herd Algorithm with Selected Validity Measures, Proceedings of the 2016 FedCSIS, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, vol. 8, pp. 79-87 (2016).

[11] P.A. Kowalski, P. Kulczycki, A Complete Algorithm for the Reduction of Pattern Data in the Classification of Interval Information, International Journal of Computational Methods, 13(3), (2016) 1650018 (26 pages).

[12] P.A. Kowalski, S. Łukasik, M. Charytanowicz, P. Kulczycki, Comparison of KHA with FPA in Clustering Problem, 8th European Symposium on Computational Intelligence and Mathematics(2016).

[13]P. Kulczycki, O. Hryniewicz, J. Kacprzyk (eds.): Techniki informacyjne w badaniach systemowych, WNT, Warszawa 2007.

[14] P. Kulczycki, P.A. Kowalski, Bayes classification of imprecise information of interval type, Control and Cybernetics, 40(1) (2011) 101-123.

[15] P. Kulczycki, M. Charytanowicz, P.A. Kowalski, S. Lukasik, The Complete Gradient Clustering Algorithm: properties in practical applications, Journal of Applied Statistics, 39(6) (2012) 1211-1224.

[16] P. Kulczycki, S. Łukasik, An algorithm for reducing the dimension and size of a sample for data exploration procedures, International Journal of Applied Mathematics and Computer Science, 24(1) (2014) 133-149.

[17] P. Kulczycki, P.A. Kowalski, Bayes Classification for Nonstationary Patterns, International Journal of Computational Methods,12(2) (2015) 1550008-1-19.

[18] P. Kulczycki, P.A. Kowalski "Classification of Interval Information with Data Drift",LNCS, 9405 (2015) 495-500.

[19] P. Kulczycki, M. Charytanowicz, P.A. Kowalski, S. Łukasik, Atypical (Rare) Elements Detection – A Conditional Nonparametric Approach, Computational Modeling of Objects Presented in Images: Fundamentals, Methods, and Applications, (2016).

[20] M. Kusy, P. A. Kowalski, Modification of the Probabilistic Neural Network with the Use of Sensitivity Analysis Procedure, in Proceedings of the 2016 FedCSIS, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, vol. 8, pp. 97-103 (2016).

[21]S. Łukasik, S. Żak: Firefly Algorithm for Continuous Constrained Optimization Tasks, LNAI, 5796 (2009) 97-106.

[22] S. Łukasik, P. Kulczycki,An algorithm for sample and data dimensionality reduction using fast simulated annealing, International Conference on Advanced Data Mining and Applications, Springer Berlin Heidelberg, pp. 152-161 (2011).

[23] S. Łukasik, P.A. Kowalski, Fully Informed Swarm Optimization Algorithms: Basic Concepts, Variants and Experimental Evaluation, in Proceedings of the 2014 FedCSIS, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, vol. 2, pp. 155-161, (2014).

[24] S. Łukasik, P.A. Kowalski, Study of Flower Pollination Algorithm for Continuous Optimization, Intelligent Systems'14, P.Angelov et al. (eds.), Springer, pp 451-459 (2015).

[25] S. Łukasik, P.A. Kowalski, M. Charytanowicz, P. Kulczycki, Clustering using Flower Pollination Algorithm and Calinski-Harabasz Index, 2016 IEEE Congress on Evolutionary Computation, paper E-16413 pp. 2724-2728 (2016).

[26] S. Mitra, T. Acharya, Data Mining: Multimedia, Soft Computing, and Bioinformatics, John Wiley & Sons, 2005.

[27] A. Nowosielski, P. A. Kowalski, and P. Kulczycki, The column-oriented database partitioning optimization based on the natural computing algorithms in Proceedings of the 2015 FedCSIS, M. Ganzha, et al (eds). ACSIS, vol. 5, pp. 1035-1041 (2015).

[28] A. Nowosielski, P. A. Kowalski, and P. Kulczycki, The Column-oriented Data Store Performance Considerations, in Proceedings of the 2016 FedCSIS, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, vol. 8, pp. 877-881 (2016).

[29] X.-S. Yang, Firefly algorithms for multimodal optimization, International Symposium on Stochastic Algorithms. Springer Berlin Heidelberg, pp. 169-178 (2009).

[30] X.-S. Yang, Flower pollination algorithm for global optimization, LNCS, 7445 (2012) 240-249.