A STUDY ON PERFORMANCE ASSESSMENT OF ESSENTIAL CLUSTERING ALGORITHMS FOR THE INTERACTIVE VISUAL ANALYSIS TOOLKIT INVEX

M.A. Titov[1,2 a], M.A. Grigorieva[1,2 b], A.A. Alekseev[1,2], N.A. Belov[1], T.P. Galkin[1,3], D.V. Grin[1,4], T.A. Korchuganova[1], S.A. Zhumatiy[1]

[1] *Lomonosov Moscow State University, Leninskie Gory, 1, Moscow, 119991, Russia*

[2] *Plekhanov Russian University of Economics, Stremyanny lane, 36, Moscow, 117997, Russia*

[3] *National Research Nuclear University "MEPhI", Kashirskoe shosse, 31, Moscow, 115409, Russia*

[4] *National Research Center "Kurchatov Institute", Akademika Kurchatova pl., 1, Moscow, 123182, Russia*

E-mail: [a] mikhail.titov@cern.ch, [b] maria.grigorieva@cern.ch

Interactive visual analysis tools bring the ability of the real-time discovery of knowledge in large and complex datasets using visual analytics. It involves multiple iterations of data processing using various data handling approaches and the efficiency of the whole chain of the analysis process depends on the performance of chosen techniques and related implementations, as well as the quality of applied methods. Stages, where data processing includes intellectual handling (i.e., data mining and machine learning), which are the most resource-intensive, require a distinct attention for evaluation of different approaches. Clustering is one such machine learning technique that is commonly used to discover groups of data objects for further analysis. This work is focused on evaluation of clustering algorithms within the interactive visual analysis toolkit InVEx (Interactive Visual Explorer). InVEx represents a visual analytics approach aimed at cluster analysis and in-depth study of implicit correlations between multidimensional data objects. It is originally designed to enhance the analysis of computing metadata of the ATLAS experiment at the LHC for operational needs, but it also provides the same capabilities for other domains to analyze large amounts of multidimensional data. The experiments and evaluation processes are carried out using operational data from the supercomputer at the Lomonosov Moscow State University. These processes include benchmark tests to assess the relative performance between chosen clustering algorithms and corresponding metrics to assess the quality of produced clusters. Obtained results will be used as guidelines in assisting users in a process of visual analysis using InVEx.

Keywords: visual analytics, clustering, benchmarks, InVEx

Mikhail Titov, Maria Grigorieva, Aleksandr Alekseev, Nikita Belov,
Timofei Galkin, Dmitry Grin, Tatiana Korchuganova, Sergey Zhumatiy

1. Introduction

The interactivity as an integral part of the visual analysis, besides its essential objective of the knowledge discovery in real-time (with a focus on large and complex datasets) while doing the analysis, also brings challenges of keeping up performance and efficiency. Every stage of the process of interactivity should be evaluated to estimate the overall performance metrics. Thus, this work is emphasized to the one of crucial processes of the analysis - intellectual data processing, e.g., application of machine learning (ML) algorithms for clustering.

Clustering algorithms, used in the developed visual analysis toolkit InVEx, require a significant effort to select the most relevant object attributes for different chosen algorithms [as for user/analyst], and to adjust primary parameters of provided algorithms [as for developers]. In addition, the outcome of different clustering algorithms should be compared between each other and being evaluated by the quality of clustering based on expert reviews.

This paper brings the assessment of the clustering algorithms usage in the InVEx toolkit to enhance the user experience and to improve the quality of the analysis process.

1.1 InVEx overview

InVEx stands for the **In**teractive **V**isual **Ex**plorer toolkit [1,2]. It provides advanced interactive data visualization tools, which are used for the analysis of large volumes of multidimensional data with its core process as an interactive visual clustering. Its development has been started for the ATLAS Distributed Computing project (the ATLAS experiment [3] at the Large Hadron Collider) to enhance the analysis of computing metadata [2]. Large amount of ATLAS ProdSys2/PanDA metadata provides means to test and prove the efficiency of applied technologies and methods. The ATLAS **Prod**uction **Sys**tem (ProdSys2) [4], in conjunction with the workload management system namely the **P**roduction **an**d **D**istributed **A**nalysis system (PanDA) [5], represents a complex set of computing components that are responsible for organizing, planning, starting and executing distributed computing tasks and jobs. Initial integration of InVEx with PanDA includes the direct access to the information about computing jobs from PanDA's monitoring system.

The stack of technologies for InVEx includes: Python-based *Django* web framework; cross-browser JavaScript library *Three.js* to create and display animated 3D computer graphics in a web browser (uses WebGL); Python libraries for data handling and analysis such as *Pandas* (data manipulation and analysis), *SciPy* (scientific computing and technical computing, as well includes clustering algorithms: hierarchical clustering, vector quantization, K-means), *Scikit-learn* (ML library, it features various classification, regression and clustering algorithms), *Kmodes* (clustering for categorical data, implementations of k-modes and k-prototypes clustering algorithms), *Intel Data Analytics Acceleration Library / daal4py* (optimized algorithmic building blocks for data analysis stages), *Prince* (factor analysis that aims to find independent latent variables).

1.2 Lomonosov-2 supercomputer overview

The current study uses log data about computing jobs gathered from the Lomonosov-2 supercomputer. This supercomputer is designed by the T-Platforms company and installed at the Lomonosov Moscow State University (MSU) [6,7] (its rank is #93 in the TOP500 list[1]). It is characterized by the Intel Xeon/FDR InfiniBand cluster, accelerated with NVidia Tesla K40s and Tesla P100 GPUs, and with overall 1696 nodes (Intel Haswell-EP E5-2697v3, 2.6GHz, 14 cores and Intel Xeon Gold 6126 2.6GHz, 12 cores) with 64/96 GB of memory per node. Theoretical peak performance is 4.946 petaFLOPS. (For more details please follow the references [6,7].)

[1] Position of the Lomonosov-2 supercomputer within the TOP500 ranking of supercomputers. Available at: https://www.top500.org/system/178444 (accessed on 20.11.2019)

2. Methods and techniques

2.1 Clustering overview

Clustering is a ML technique, that is aimed at grouping similar objects into unlabeled groups called clusters (unsupervised learning). The process of clustering in InVEx is implemented in two stages: i) *Level-of-Detail (LoD) generator* [2] that brings the initial (optional) grouping to reduce the amount of data objects provided to the user (algorithms that are used: scikit-learn/MiniBatchKMeans, daal4py/KMeans, kmodes/KPrototypes); ii) *cluster analysis*, which is a core process to analyze data objects similarities (algorithms that are used: same as for LoD, as well as scikit-learn/KMeans, scikit-learn/DBSCAN, scipy/Hierarchical).

2.2 Clustering validation measures

Validation measures (i.e., quality metrics) are classified as internal and external. Internal measures reflect compactness, connectedness and separation of the cluster partitions. The following metrics were chosen: *Silhouette coefficient* (ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster); *Calinski-Harabaz Index* (the maximum value for index indicates a suitable partition for the data set); *Davies-Bouldin Index* (closer to 0 is better, it computes the ratio between the within cluster distances and the between cluster distances). External measures, which provide comparison of the identified clusters to external preset labels, are represented in this paper by the following metric - *Adjusted Rand Index* - a function that measures the similarity of the two assignments (given the knowledge of the ground truth class assignments and clustering algorithm assignments). (Further extension of using external metrics considers *Fowlkes-Mallows score* that is defined as the geometric mean of the pairwise precision and recall.)

3. Experiments

3.1 Data pre-processing

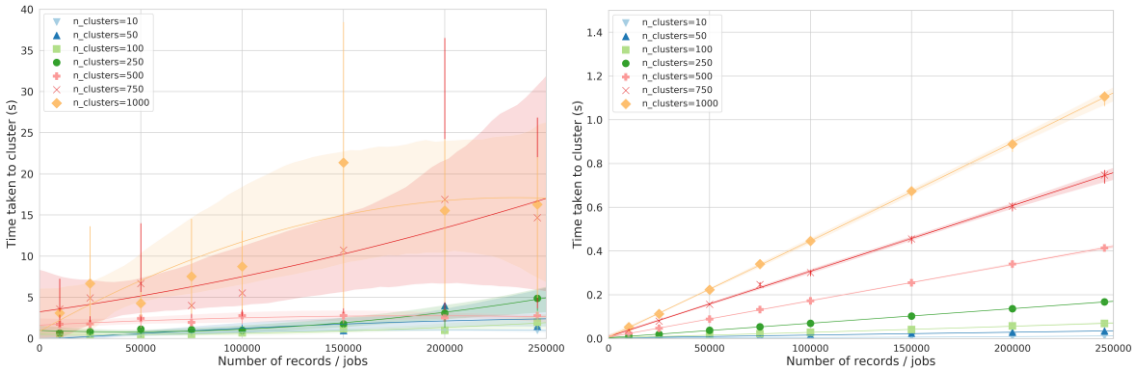
Gathered data for experiments required initial transformations because of the format of data objects attributes. Log data from the Lomonosov-2 supercomputer was collected from the period of 300 days (from June 2018 to March 2019): 245K records with 12 attributes (user ID, execution time duration, number of allocated nodes, CPU load during the job execution per user, GPU load during the job execution per user, number of executed instructions per second, etc.). Almost all of the attributes were of the categorical type (nominal and ordinal data) and most of them are with such nominal values as: "none", "low", "average", "high". Thus, the technique of dimensionality reduction was applied, that is used to map the data record to a lower-dimensionality space.

The process of the data transformation was the following: i) apply multiple correspondence analysis (MCA) to the dataset with all 11 categorical attributes to represent data objects in a multidimensional Euclidean space with 5 dimensions; ii) apply principal component analysis (PCA) to the dataset with 5 attributes from the previous step and 1 non-categorical attribute from the original dataset. The outcome of this transformation is a dataset with 5 attributes (per record), which will be used for clustering in the benchmarks.

3.2 Benchmarks and quality metrics

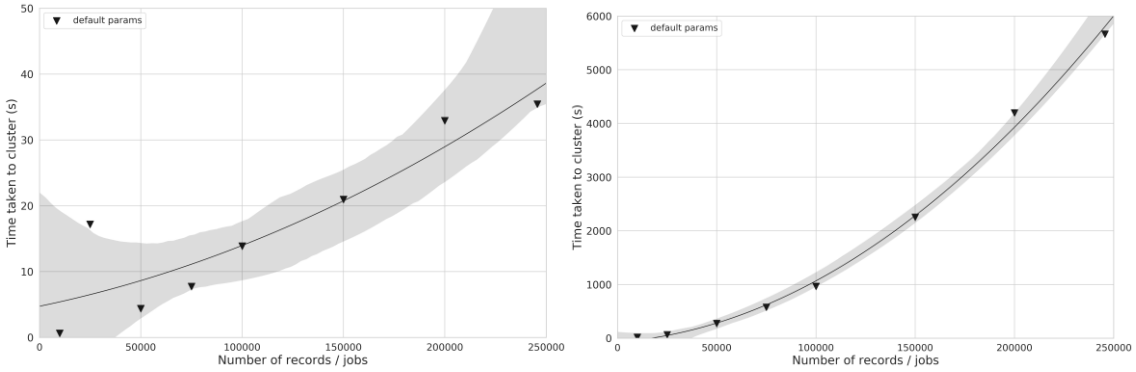
There are several essential algorithms (partitioning-based and density-based methods) that were chosen for performance and quality evaluation. Figure 1 presents benchmarks for KMeans algorithms of different implementations. Scikit-learn/KMeans was significantly inferior to other implementations, and especially for input data over 150K records that took up to hundreds of seconds to be executed, thus its benchmark was not included into the figure. Figure 1 shows that daal4py/KMeans outperforms scikit-learn/MiniBatchKMeans. Quality metrics (internal measures) for KMeans implementations gives better results compared to MiniBatchKMeans (Table 1). Table 2 shows that scikit-learn/KMeans and daal4py/KMeans give very close to each other labeling of clusters, which makes daal4py implementation preferred to be used in the next versions of InVEx for performance improvements and without loss of quality. Benchmarks for two density-based algorithms

are presented in Figure 2, which shows greater performance for HDBSCAN. These algorithms produce very close results in terms of quality metrics (Tables 1,2), thus HDBSCAN is more suitable for use (which is also more robust to parameter selection in comparison to DBSCAN).



(a) (b)

Figure 1. Performance comparison for partitioning-based clustering algorithms with different number of records as input data and different number of outcome clusters: a) scikit-learn/MiniBatchKMeans; b) daal4py/KMeans



(a) (b)

Figure 2. Performance comparison for density-based clustering algorithms with different number of records as input data: a) HDBSCAN; b) OPTICS

Table 1. Internal measures for clustering algorithms

	Silhouette Score	Calinski-Harabaz Index	Davies-Bouldin Index
scikit-learn / KMeans	0.61	92853.7	0.86
scikit-learn / MiniBatchKMeans	0.58	88390.1	0.94
daal4py / KMeans	0.68	32503.6	1.08
HDBSCAN	0.58	211.5	1.37
OPTICS	0.59	73.4	1.36

Table 2. External measures to compare the closeness of results between pair of clustering algorithms

	Adjusted Rand Index
(scikit-learn) KMeans vs. MiniBatchKMeans	0.55
scikit-learn / KMeans vs. daal4py / KMeans	0.95
HDBSCAN vs. OPTICS	0.84

4. Conclusion

The interactive visual analysis toolkit InVEx represents a visual analytics approach aimed at cluster analysis and in-depth study of implicit correlations between multidimensional data objects and object parameters interdependencies. Its capabilities were applied in analysis of log data from the Lomonosov-2 supercomputer, which also were used to conduct experiments on performance estimation for InVEx clustering algorithms.

Experiments outcome is the process of evaluation essential clustering algorithms. Benchmark tests assess the relative performance between chosen algorithms and corresponding metrics, and the quality of produced clusters. Obtained results and the approach itself will be integrated into InVEx and will be used as guidelines in assisting users in a process of visual analysis (as well as a self-adjustment mechanism to configure initial parameters for clustering algorithms).

5. Acknowledgement

Many thanks to all members of the InVEx team and colleagues from the Research Computing Center (RCC) of MSU for providing experimental data and for the continued support. This work was financially supported by the Russian Science Foundation (grant No.18-71-10003).

References

[1] InVEx project, "InVEx" [software], 2019. Available at: https://github.com/PanDAWMS/InVEx (accessed on 20.11.2019)

[2] Grigorieva M.A. et al. Evaluation of the Level-of-Detail Generator for Visual Analysis of the ATLAS Computing Metadata // Lobachevskii Journal of Mathematics, vol.40, no.11, pp.1788--1798 (2019)

[3] ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider // JINST, vol.3, S08003 (2008)

[4] Barreiro F.H. et al. The ATLAS Production System Evolution: New Data Processing and Analysis Paradigm for the LHC Run2 and High-Luminosity // J. Phys.: Conf. Ser., vol.898, no.5, 052016 (2017)

[5] Barreiro F.H. et al. PanDA for ATLAS distributed computing in the next decade // J. Phys.: Conf. Ser., vol.898, no.5, 052002 (2017)

[6] Voevodin V.V. et al. Supercomputer Lomonosov-2: Large Scale, Deep Monitoring and Fine Analytics for the User Community // Supercomputing Frontiers and Innovations, vol.6, no.2, pp.4--11 (2019)

[7] Leonenkov S., Zhumatiy S. Supercomputer Efficiency: Complex Approach Inspired by Lomonosov-2 History Evaluation // RuSCDays 2018: Supercomputing. Communications in Computer and Information Science, vol.965. Springer, Cham (2019)