High-Content Screening images streaming analysis using the STriGen methodology

Laura Melgar-García
Data Science & Big Data Lab, Pablo de Olavide University
Seville, Spain
lmelgar@upo.es

David Gutiérrez-Avilés
Data Science & Big Data Lab, Pablo de Olavide University
Seville, Spain
dgutavi@upo.es

Cristina Rubio-Escudero
Department of Computer Science, University of Seville
Seville, Spain
crubioescudero@us.es

Alicia Troncoso
Data Science & Big Data Lab, Pablo de Olavide University
Seville, Spain
atrolor@upo.es

ABSTRACT

One of the techniques that provides systematic insights into biological processes is High-Content Screening (HCS). It measures cells phenotypes simultaneously. When analysing these images, features like fluorescent colour, shape, spatial distribution and interaction between components can be found. STriGen, which works in the real-time environment, leads to the possibility of studying time evolution of these features in real-time. In addition, data streaming algorithms are able to process flows of data in a fast way. In this article, STriGen (Streaming Triclustering Genetic) algorithm is presented and applied to HCS images. Results have proved that STriGen finds quality triclusters in HCS images, adapts correctly throughout time and is faster than re-computing the triclustering algorithm each time a new data stream image arrives.

CCS CONCEPTS

• **Information systems** → **Clustering**; **Data stream mining**; • **Computing methodologies** → **Genetic algorithms**; • **Applied computing** → *Molecular evolution*; *Imaging*;

KEYWORDS

Real-time, Triclustering, Genetic operators, High-Content Screening

1 INTRODUCTION

Nowadays, one of the biggest challenges in biology is understanding genes and their biological circuits. Due to that, techniques that investigate complete cellular processes are becoming more relevant, as High-Content Screening (HCS). HCS combines an automated imaging and analysis of intact cells exposed to some perturbations (chemical or genomic) that alter their phenotype [11].

HCS is made by many steps that can take too much time if they are not effectively done. On the other hand, these days, stream computing trends are rising, i.e., algorithms that react in the fastest way to provide information from data in real time. Consequently, the processing and analysis of HCS images in a streaming environment could give quick information from them.

In [8] the TriGen algorithm is presented as a Triclustering algorithm that discovers groups of 3D datasets throughout instances, attributes and time. In [12] STriGen algorithm is introduced. STriGen is a new incremental learning method that finds groups of similar behaviour patterns in 3D stream data continuously. In this paper, STriGen algorithm is applied to HCS images to get information from images in real-time.

The article is structured as follows: STriGen and HCS methodologies are presented in Section 2; the experimental setup and the yielded results in Section 3; and finally the conclusions are in Section 4.

2 METHODOLOGY

2.1 STriGen methodology

STriGen is a new incremental learning method that creates triclusters (based on TriGen algorithm [8]) and keeps them updated taking into account the knowledge from previous streams and upgrading its learning method. STriGen meets all 4 Data Streaming requirements, i.e., data has to be processed in the order of its arrival and one by one; the learning model has to be updated incrementally; the model has to deal with small amount of memory referring to the huge quantity of data and it has to be fast.

STriGen algorithm is inspired in the offline/online approach of stream algorithms. Consequently, STriGen starts with an execution of the TriGen modified to treat stream data as static data with W data, where W is the maximum number of data streams that can be used in an iteration of the STriGen algorithm. Afterwards, the algorithm processes each new data stream and the model updates incrementally and quickly basing itself on the new streams in order to provide the upgraded triclusters.

During this second phase of STriGen, it tries to extend the actual triclusters throughout time removing the "oldest" time point to keep always a maximum of W data points, a regular procedure in Data Streaming algorithms [6]. In addition to the extension of the actual triclusters over time, the learning model adjusts itself incrementally depending on the GRQ (*GR*aphical *Q*uality) measure, part of one of the TriGen fitness functions [7].

More specifically, the STriGen learning model makes mutations into triclusters, i.e., adding, deleting and/or changing both instances and/or attributes, to be updated. These operations are quicker than re-training TriGen each time a new stream arrives to keep the learning model updated. Mutations allow to find current and global real solutions as STriGen depends on the results from the first execution of TriGen that can change every time it is executed. In this way, triclusters are mutated until their GRQ values are higher than *minGRQ* or until the number of iterations made is higher than *numIt*. In this way, STriGen tries to include or remove some current tricluster's components in order to keep most accurate components. These 2 parameters and also the "window" W parameter and a minimum GRQ threshold (that can delete the oldest time included in the tricluster when its GRQ is smaller than this) take different values depending on the dataset, to be able to adjust to small or/and abrupt changes in streams.

When the dataset is synthetic or when the resulting triclusters are known in advance, a validation process to compare founded and real triclusters is done with accuracy and F1 Score measures. Datasets that are neither synthetic nor with known triclusters in advance, are evaluated depending on the GRQ value.

2.2 High-Content Screening methodology

High-Content Screening or Analysis (HCS or HCA) is gaining importance. HCS combines automated imaging acquisition and image analysis using, most frequently, automated fluorescence microscopy [4]. During the HCS process intact cells are incubated with substances that alter their phenotype in a desired way [3]. These cells are screened and multiple fluorescence readouts are measured in parallel. This process provides big volume of data with high biological information content. Subcellular locations and fluorescence colour intensity during different complex cellular events, in terms of space and time, are measured with the automated image analysis phase of HCS [5]. HCS images have been useful to detect and study DNA, cytokinesis, cell division, cell migration, apoptosis, mitosis, and more cellular events of target components.

HCS processes involve different tasks as cell preparation and labelling, image acquisition, image analysis and data management [9]. In terms of data challenges, HCS has 2 principal issues: data storage and data processing. One of the main interesting points of HCS is the simultaneous analysis of images that requires a high computing power and quickly computer network connections [1].

2.3 STriGen application to High-Content Screening images

Applying STriGen to HCS images allows to discover the best features to group to get information from cells. Actual numerical features extracted from HCS images are: 1) fluorescent marker colour; 2) cell component's shape; 3) spatial situation; 4) distribution of pixel-colour in a cell region of interest to study interactions or co-occurrences [11]. Moreover, evolution throughout time is added to these 4 features when applying STriGen to HCS.

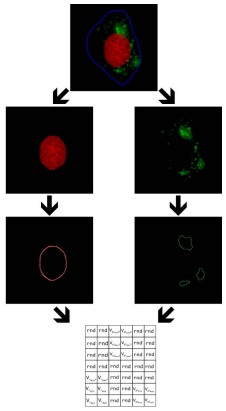

Figure 1: Example of preprocessing phase of HCS images

The type of images used in this experiment are individual intact cells fluorescent microscopy-based HCS images. Images are processed in order to create a dataset that fits STriGen requirements. Firstly, each RGB image is transformed into a 3D matrix representing the coordinates of each image pixel in decimal numbers. In other words, values that represent images colours are the pixel-colour or fluorescent-marker colour features mentioned above. A filter is passed through every image to not include any image that present noise due to optic aberrations, microscopy issues or even bad acquisition of the image. Secondly, data is prepared to fit STriGen dataset specifications, i.e., taking into account: detecting the colours specified to analyse and detecting areas limits (for areas with more than an user fixed value). In addition, random values between 0 and 1000 are added in order to make STriGen able to ignore the background of images. A graphical example of this methodology is in Fig. 1.

3 RESULTS AND DISCUSSION

In this section, the results obtained by the application of the STriGen algorithm to a HCS images dataset are presented.

STriGen has been applied to a set of HCS images from [2] of HeLa cells (cervical cancer cells taken from a woman in the 50s that can divide themselves an unlimited number of times in well preserved laboratory conditions [10]). For this experiment, the dataset selected is the one that presents the reaction of HeLa cells to Transferrin receptors (usually applied as cancer cell target because they enhances site-specific therapies). The quality of the resulting triclusters is evaluated with the GRQ measure. In addition, for this experiment, a dataset with the desired STriGen founded triclusters has been created in order to compute F1 Score and Accuracy measures to check the performance of the algorithm.

460 HCS images similar to the above image in Fig. 1 have been used for this experiment with black for the background image, blue for the cell border, red for the nuclear DNA and green for the endosome compartment. Cell borders (blue colour) has been ignore because they do not provide extra information about HCS features.

Table 1: STriGen configuration parameters

Execution	*minGRQ*	thresholdGRQ	*numIt*
1	0.95	0.80	10
2	0.88	0.70	15
3	0.90	0.75	20

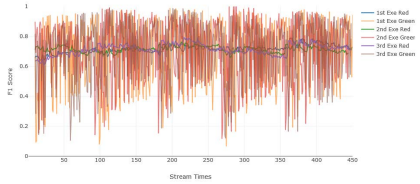

Figure 2: F1 Score results

Figure 3: Accuracy results

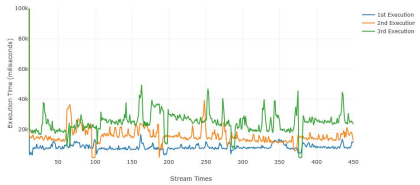

Figure 4: STriGen execution time

STriGen has been executed 3 times due to the fact that the first triclusters results depend on the first triclusters founded by TriGen. In that way, configuration parameters take different values in each execution to see the influence of them in the results (see Table 1) excepting W that has as maximum value 3.

STriGen performance results are in the following figures: Fig. 3 shows accuracy values for all triclusters and Fig. 2 shows F1 score values. Figures show that the algorithm performs in an accurate way and can find components correctly. F1 score of the green triclusters in all 3 executions varies a lot, it is due to the fact that green areas spread through cells and are in continuous movement, however the mean accuracy value in all 3 executions is 0.908. Red areas are more stable in both measures because they are in a similar position in all streams.

Apart from these measures, another important parameter that represents the good performance of the algorithm is time execution. This experiment has been done in a computer with an i7-5820K 3.3GHz processor and 48GB RAM memory. The first phase of STriGen (that is just one execution of TriGen with the first 10 streams) takes a mean of 8.9 minutes to execute completely. Afterwards, the Data Streaming phase starts and each new image stream is processed and provides founded triclusters in a mean of 16 seconds.

In general, the quality measures are mostly equal in the 3 executions. However, it can be seen that the execution time of the 3rd execution of STriGen is much smaller, due mostly to the small value of *numIt*.

4 CONCLUSIONS

The STriGen algorithm performs with good results when dealing with stream data as we have seen in Section 3. It is faster than executing the TriGen algorithm when a new stream arrives and so the evolution of data throughout time can be analysed in real-time. It proves that the learning model updates incrementally making mutations and taking into account the W most recent streams.

The application of HCS images into the Data Streaming environment with STriGen leads the possibility of obtaining information about the evolution of HCS features, i.e. components colour, shape, location and distribution, throughout time in real-time. A possible application of this experiment would be the fact that STriGen could detect when external agents like substances, drugs, antibodies, etc are added to the exposed cell and how the cell reacts to them, e.g., changing triclusters components. It allows to do a continuous analysis of the cell in real-time.

ACKNOWLEDGMENTS

Authors thank the Spanish Ministry of Economy and Competitiveness for the support under the project TIN2017-88209-C2-1-R and TIN2017-88209-C2-2-R.

REFERENCES

[1] M. Bickle. 2008. High-content screening : A new primary screening tool ? 11, 11 (2008).
[2] CellOrgnizer Project [n. d.]. CellOrganizer project from Carnegie Mellon University. http://www.cellorganizer.org/2d-hela/
[3] D. Cronk. 2013. Chapter 8 - High-throughput screening. (2013), 95 – 117. https://doi.org/10.1016/B978-0-7020-4299-7.00008-1
[4] R. Flaumenhaft. 2007. 3.07 - Chemical Biology. (2007), 129 – 149. https://doi.org/10.1016/B0-08-045044-X/00080-8
[5] G. Galea and J. C. Simpson. 2013. Chapter 17 - High-Content Screening and Analysis of the Golgi Complex. 118 (2013), 281 – 295. https://doi.org/10.1016/B978-0-12-417164-0.00017-3
[6] M. Ghesmoune, M. Lebbah, and H. Azzag. 2016. State-of-the-art on clustering data streams. *Big Data Analytics* 1, 1 (2016), 1–27. https://doi.org/10.1186/s41044-016-0011-3
[7] D. Gutiérrez-Avilés, R. Giráldez, F.J. Gil-Cumbreras, and C. Rubio-Escudero. 2018. TRIQ: A new method to evaluate triclusters. *BioData Mining* 11, 1 (2018), 1–29. https://doi.org/10.1186/s13040-018-0177-5
[8] D. Gutiérrez-Avilés, C. Rubio-Escudero, F. Martínez-Álvarez, and J.C. Riquelme. 2014. TriGen: A genetic algorithm to mine triclusters in temporal gene expression data. *Neurocomputing* 132 (2014), 42–53. https://doi.org/10.1016/j.neucom.2013.03.061
[9] S. Lee and B. J. Howell. 2006. [25] - High-Content Screening: Emerging Hardware and Software Technologies. 414 (2006), 468 – 483. https://doi.org/10.1016/S0076-6879(06)14025-2
[10] B.P. Lucey, W.A. Nelson-Rees, and G.M. Hutchins. 2009. Henrietta Lacks, HeLa cells, and cell culture contamination. *Archives of Pathology and Laboratory Medicine* 133, 9 (2009), 1463–1467.
[11] F. Heigwer M. Boutros and C. Laufer. 2015. Microscopy-based High-content Screening. *Cell* 163, 6 (2015), 1314–1325. https://doi.org/10.1016/j.cell.2015.11.007
[12] L. Melgar-García, D. Gutiérrez-Avilés, and C. Rubio-Escudero. 2019. Discovering Behavior Patterns in Big Data Streaming Environments : The STriGen Methodology. (2019). Manuscript submitted for publication.