

# GaSvm Software Manual

Daniel Babiak Daniel.Babiak@polsl.pl  
Dariusz Kuchta Dariusz.Kuchta@polsl.pl  
Grzegorz Mrukwa Grzegorz.Mrukwa@polsl.pl  
Maciej Gamrat Maciej.Gamrat@polsl.pl  
Michal Gallus Michal.Gallus@polsl.pl  
Michal Wolny Michal.Wolny@polsl.pl  
Roman Lisak Roman.Lisak@polsl.pl  
Sebastian Pustelnik Sebastian.Pustelnik@polsl.pl  
Wojciech Wilgierz Wojciech.Wilgierz@polsl.pl

26.10.2017

## Abstract

GaSvm is a software for launching Genetic Algorithm for training set selection in Support Vector Machines classifier training.

## Contents

<b>Installation</b>	<b>1</b>
<b>Usage</b>	<b>2</b>
<b>Input specification</b>	<b>2</b>
<b>Output specification</b>	<b>3</b>
<b>Parameters</b>	<b>3</b>
<b>Final notes</b>	<b>3</b>
<b>References</b>	<b>3</b>

## Installation

GaSvm software itself needs no installation. Once the files are unpacked from the .zip archive, it can be used without further delay.

## Usage

Interface of the application allows to provide all the options that are *available* to be used with GaSvm. Default settings comprise configuration representing pipeline used for MALDI IMS data processing that can potentially serve as a starting point for performing initial experiments.

## Input specification

Input file is a text file constructed as follows:

1. Row with global metadata - it can contain anything (unused for now).
2. Row with global  $m/z$  axis - data has to be resampled before usage.
3. Data of each spectra, each consisting of two lines:
  1. Spatial coordinates of a spectrum (X, Y, and Z, separated with spaces).  
*Please note, that Z coordinate is currently not used. Therefore each Z value can be safely set to zero.*
  2. Intensity values for each  $m/z$  value specified above, separated with spaces. Their number **must** be equal to the number of elements in  $m/z$  axis. This is similar to *imzML* format in *processed* form.

Artificial test data (*only for demonstration of this structure*):

```
in this line are global metadata, which is discarded for now
899.99 902.58 912.04
1 1 0
12 20 0
2 1 0
9 18 13
1 2 0
5 10 20
2 2 0
14 2 19
```

*This data cannot be used for testing the program itself; it is just a reference, how to format data file.*

Sample **real** data file is available [here](#). The same data set was used in G. Mrukwa, G. Drazek, M. Pietrowska, P. Widlak and J. Polanska, “A Novel Divisive iK-Means Algorithm with Region-Driven Feature Selection as a Tool for Automated Detection of Tumour Heterogeneity in MALDI IMS Experiments,” in International Conference on Bioinformatics and Biomedical Engineering, 2016.

## Output specification

This section is under construction.

## Parameters

1. **-Destination** - prefix of the experiment result files.
2. **-Source** - location of the input dataset.
3. **-TrainingSetSplitRate** - (Default: 0.7) training set split rate
4. **-MutationRate** - (Default: 0.1) mutation rate
5. **-BitSwapRate** - (Default: 0.1) rate of bit swaps
6. **-PreservationRate** - (Default: 0.3) percentage of individuals treated as elite
7. **-GenerationsNumber** - (Default: 50) number of generations
8. **-NumberOfRestarts** - (Default: 30) number of time the experiment is repeated
9. **-Seed** - (Default: 0) seed for the RNG
10. **-PopulationSizes** - (Default: 10) population sizes used in the experiment
11. **-InitialFillups** - (Default: 4) number of observations considered at the beginning of the experiment.
12. **-help** - display help with the same informations

## Final notes

In case of any questions, do not hesitate to contact us by [mail](#).

## References

This software is part of contribution made by [Data Mining Group of Silesian University of Technology](#), rest of which is published [here](#).

- Marczyk M, Polanska J, Polanski A: Comparison of Algorithms for Profile-Based Alignment of Low Resolution MALDI-ToF Spectra. In Advances in Intelligent Systems and Computing, Vol. 242 of Man-Machine Interactions 3, Gruca A, Czachorski T, Kozielski S, editors. Springer Berlin Heidelberg 2014, p. 193-201 (ISBN: 978-3-319-02308-3), ICMMI 2013, 22-25.10.2013 Brenna, Poland
- P. Widlak, G. Mrukwa, M. Kalinowska, M. Pietrowska, M. Chekan, J. Wierzgon, M. Gawin, G. Drazek and J. Polanska, "Detection of molecular

- signatures of oral squamous cell carcinoma and normal epithelium - application of a novel methodology for unsupervised segmentation of imaging mass spectrometry data,” *Proteomics*, vol. 16, no. 11-12, pp. 1613-21, 2016
- M. Pietrowska, H. C. Diehl, G. Mrukwa, M. Kalinowska-Herok, M. Gawin, M. Chekan, J. Elm, G. Drazek, A. Krawczyk, D. Lange, H. E. Meyer, J. Polanska, C. Henkel, P. Widlak, “Molecular profiles of thyroid cancer subtypes: Classification based on features of tissue revealed by mass spectrometry imaging,” *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 2016
  - G. Mrukwa, G. Drazek, M. Pietrowska, P. Widlak and J. Polanska, “A Novel Divisive iK-Means Algorithm with Region-Driven Feature Selection as a Tool for Automated Detection of Tumour Heterogeneity in MALDI IMS Experiments,” in *International Conference on Bioinformatics and Biomedical Engineering*, 2016
  - A. Polanski, M. Marczyk, M. Pietrowska, P. Widlak and J. Polanska, “Signal partitioning algorithm for highly efficient Gaussian mixture modeling in mass spectrometry,” *PloS one*, vol. 10, no. 7, p. e0134256, 2015