# Clustering Algorithms Meta Applier (CAMA) Toolbox

Dmitry S. Shalymov
Mathematics & Mechanics department
Saint-Petersburg State University
Email: shalydim@mail.ru

Kirill Skrygan
Mathematics & Mechanics department
Saint-Petersburg State University
Email: kirillskrygan@gmail.com

Dmitry Lyubimov
Mathematics & Mechanics department
Saint-Petersburg State University
Email: ancient.punk@gmail.com

***Abstract*—Clustering is used in many fields, including machine learning, data mining, financial mathematics, etc. There are many different algorithms for cluster analysis. Nevertheless only a few software tools are available to perform these algorithms with user's data sets and to compare results. We propose CAMA software toolbox for the new clustering algorithms research. The main character of this tool is a possibility to load as user's input datasets, as user's algorithms and to compare results with classical methods. Basic principles of CAMA and used technologies are briefly described. Some ideas for new clustering algorithms are also proposed for further development.**

## I. Introduction

Data clustering is the partitioning of a data set into subsets (or clusters), so that the data in each subset share some common trait, often proximity according to some defined distance measure. It is a common technique for statistical data analysis, which is widely used in machine learning, data mining, pattern recognition, image analysis, financial mathematics and bioinformatics.

After thirty years research there were proposed a great amount of approaches. Most of them were developed for a specific problem and are somewhat ad hoc. Those methods that are more generally applicable tend either to be model-based, and hence require strong parametric assumptions, or to be computation-intensive, or both.

Researches face with many difficulties while investigating an efficiency of their new algorithms. Especially it is important when researcher have to prove that his method is the best one for a special kind of data sets.

Unfortunately only a few number of software tools are available to restrict such difficulties. These tools have essential limitations in their functionality. For example it is impossible to load your own algorithms and perform experiments with it. The good example is an open sourced Cluster Validation Toolbox CVAP [1]. ClusterPack [2] is a collection of MATLAB functions for cluster analysis. It consists of the three modules ClusterVisual, ClusterBasics, and ClusterEnsemble. They contain general clustering algorithms as well as special algorithms. But the number of available test data sets is very few. COMPACT [3] is GUI MATLAB tool that enables to compare some clustering methods. Only four clustering algorithms are supported. Data Clustering & Pattern Recognition (DCPR) [4] is MATLAB tool with friendly interfaces. It supports 6 basic clustering algorithms and 4 data sets. The Fuzzy Clustering and Data Analysis Toolbox [5] is a collection of MATLAB functions. It supports 6 clustering algorithms and over 7 clustering validation algorithms. The engine to compare effectivenes of algorithms is also implemented. SOM Toolbox [6] is a software library for MATLAB 5 (version 5.2 at least) implementing Self-Organizing Map (SOM) algorithm. To compare effectiveness of SOM algorithm more than 11 clustering algorithms are supported. The engine for creation artificial data sets is implemented. But there is also no possibility to load and to perform your own clustering algorithms

All these tools are implemented with MATLAB which is the commercial product with quite expensive license. Our toolbox is free software. You can redistribute it and/or modify it under the terms of the GNU General Public License [7].

CAMA toolbox allows to structure the known information about existent clustering algorithms and to research new effective user's methods of cluster analysis.

In Section 2 we introduce basic principles of CAMA and describe main steps of algorithm processing. In Section 3 Kernel module is described, Section 4 is about two modifications of CAMA implementation. Available set of preloaded algorithms and data sets is introduced in Section 5. Main ideas of new clustering algorithm are proposed in Section 6. We conclude in Section 7 by discussing possible extensions of CAMA software toolbox.

## II. Clustering Analysis

The classification of objects according to similarities among them and organizing of data into groups is the objective of cluster analysis. Clustering techniques do not use prior class identifiers. That's why they are among the unsupervised methods. The main potential of clustering is to detect the underlying structure in data, not only for classification and pattern recognition, but also for model reduction and optimization. Different classifications can be related to the algorithmic approach of the clustering techniques: partitioning, hierarchical, graph-theoretic methods and methods based on objective function.

Clustering techniques can be applied to data that is quantitative (numerical), qualitative (categoric), or a mixture of both. In this thesis, the clustering of quantitative data is considered.

Cluster is a group of objects that are more similar to one another than to members of other clusters. In metric spaces,

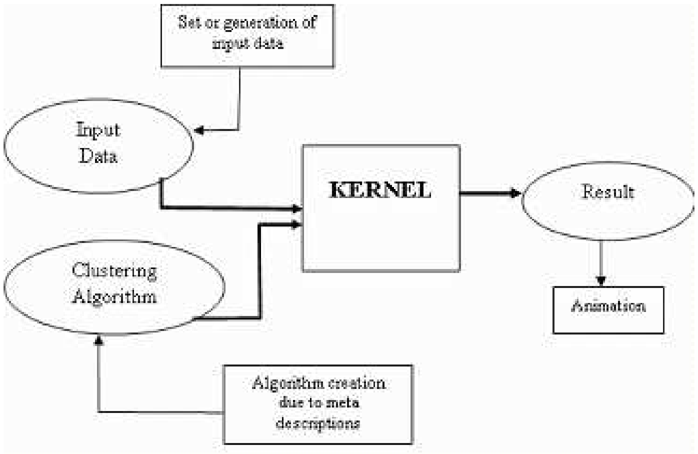
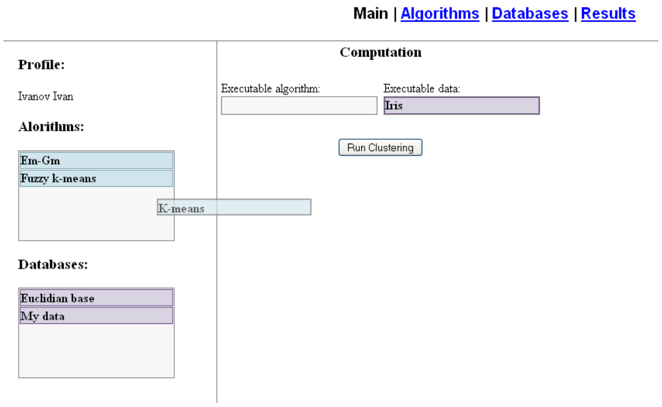

Set or generation of input data
Input Data
KERNEL
Result
Clustering Algorithm
Animation
Algorithm creation due to meta descriptions

Fig. 1. Basic principles of CAMA



Profile:

Ivanov Ivan

Alorithms:

Em-Gm
Fuzzy k-means
K-means

Databases:

Euclidian base
My data

Executable algorithm:

Executable data: Iris

Fig. 2. CAMA Toolbox. Algorithms and data sets drag and drop

similarity is often defined by means of a distance norm. Distance can be measured among the data vectors themselves, or as a distance form a data vector to some prototypical object of the cluster. The prototypes are usually not known beforehand, and are sought by the clustering algorithms simultaneously with the partitioning of the data. The prototypes may be vectors of the same dimension as the data objects, but they can also be defined as linear or nonlinear subspaces or functions.

Since clusters can formally be seen as subsets of the data set, one possible classification of clustering methods can be according to whether the subsets are fuzzy or crisp (hard). Hard clustering methods are based on classical set theory, and require that an object either does or does not belong to a cluster. Fuzzy clustering methods allow objects to belong to several clusters simultaneously, with different degrees of membership.

## III. Clustering Algorithms Meta Applier

The main goal of Clustering Algorithms Meta Applier is the research and approbation of new and existent clustering algorithms on a various number of data sets. CAMA contains an engine to load user's data and user's algorithms, prepared input datasets for experiments and the set of existent algorithms in MATLAB (*.m files) format. CAMA is implemented as a desktop application and as a web-service.

An oracle which knows the correct answer could be used. In that case an accuracy of applied algorithms can be measured due to oracle's knowledge.

Now only hard clustering methods are implemented in CAMA. Other clustering techniques support is a subject for the further development.

Result of algorithm application in CAMA contains the following information: number of clusters in data set, coordinates of cluster centers, number of iterations and character diagrams. It is possible to save extracted results as images and statistic data.

The basic principles of CAMA are shown in Fig.1.

Input data that was generated or was loaded as one of prepared data sets comes to the kernel module with one of the clustering algorithms. Kernel translates algorithm from *.m format and computes the result which is processed with animation module. All GUI interfaces and engines for loading data are implemented with Java. Algorithms processing and calculations are performed with .NET technologies in the kernel which is compiled as DLL.

## IV. Kernel

Kernel is the interpreter of MATLAB language. First *.m files are translated into *.cs files. After that C# compiler compiles it into DLLs. The main problem here is the translation from MATLAB to C# because MATLAB is not the language with pure types. It is partly functional and has a great amount of complex inner functions.

The grammar and semantics are described with ANTLR [8]. The lexical analyzer, parser and additional logics are performed in C# level. Also with C# there were implemented MATLAB functions that are not available in .NET and even in Math.NET Iridium [9].

Interpreter also contains special utility for input data set conversion (usually large number of multidimensional vectors) to the .NET format.

Used technologies: ANTLR and ANLRWorks 1.2.3, MSVS 2008 and .NET 3.5 [10], .NET Reflector [11], Math.NET Iridium.

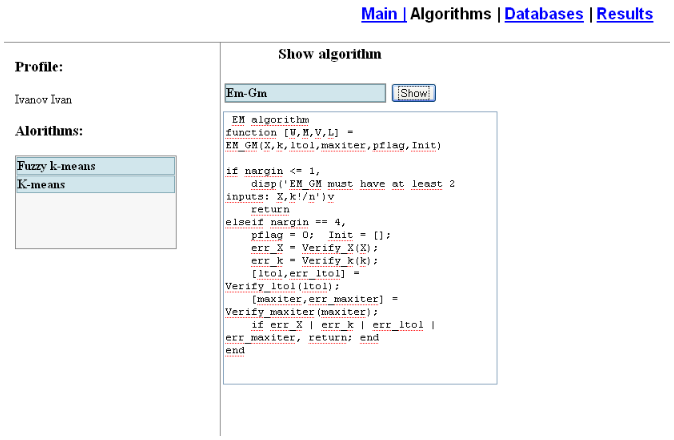
## V. Web-service and desktop application

CAMA is implemented in two modifications: desktop application and web-service application.

All GUI interfaces and modules (except kernel) are implemented with Java technologies.

GUI for desktop and web-service versions are almost the same. Infrastructure of web-service is based on accounts. Each user has its own account for loading and using data sets. After registration user has an access to his personal page which contains personal information, the list of available algorithms and data sets (default data sets and previously loaded by user).

All available algorithms and datasets are visualized so that user can simply drag and drop corresponding items to the evaluation block. It is shown in Fig.2.

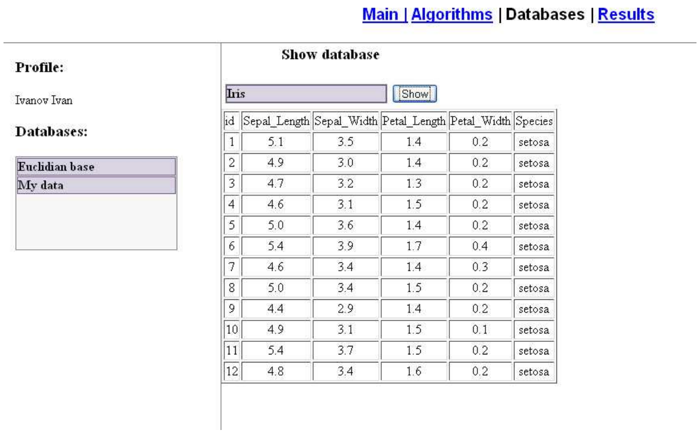


Em-Gm

```
 EM algorithm
function [W,M,V,L] =
EM_GM(X,k,ltol,maxiter,pflag,Init)

if nargin <= 1,
    disp('EM_GM must have at least 2
inputs: X,k!/n')v
    return
elseif nargin == 4,
    pflag = 0;  Init = [];
    err_X = Verify_X(X);
    err_k = Verify_k(k);
    [ltol,err_ltol] =
Verify_ltol(ltol);
    [maxiter,err_maxiter] =
Verify_maxiter(maxiter);
    if err_X | err_k | err_ltol |
err_maxiter, return; end
end
```

Fig. 3. CAMA Toolbox. Clustering algorithm text representation



Iris

| id | Sepal_Length | Sepal_Width | Petal_Length | Petal_Width | Species |
|---|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5 | 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 8 | 5.0 | 3.4 | 1.5 | 0.2 | setosa |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 | setosa |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 | setosa |
| 12 | 4.8 | 3.4 | 1.6 | 0.2 | setosa |

Fig. 4. CAMA Toolbox. Data set representation

User can see the text representation of his algorithms. See Fig.3.

It is possible to load data in csv and text format and check separator between cells in rows. Also user is able to manipulate with data sets by performing SQL queries. Data sets representation is shown in Fig.4.

After execution server will generate a new page with results. The example of clustering artificial data set with EM algorithm [12] can be seen in Fig.5. The new page contains graphics result, number of clusters, coordinates of cluster centers, number of algorithm iterations and other statistics.

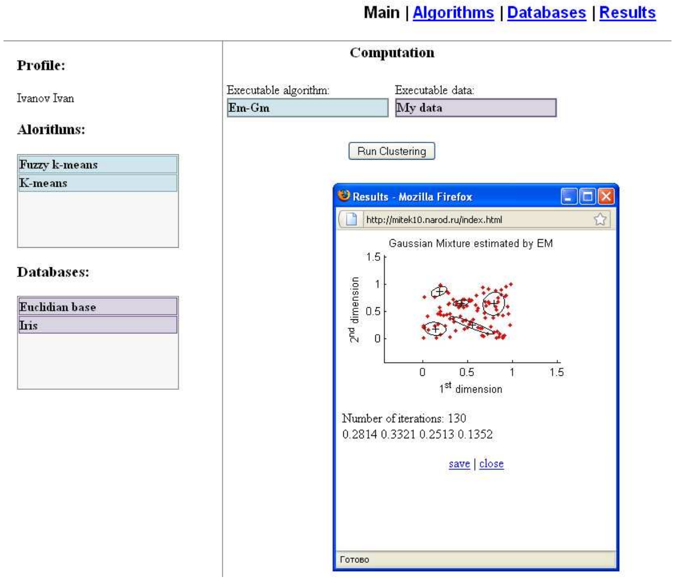
Used technologies: Java Server Pages (JSP) [13], Java Servlets (JDK 1.6) [14] and Apache Tomcat Server [15].

In the desktop application multi user's work is not supported. All web-service pages are replaced with dialogs and no one server is implemented.

Used technologies: JDK 1.6 and Java Swing [16].

## VI. Prepared algorithms and data sets

A fundamental, and largely unsolved, problem in cluster analysis is the determination of the "true" number of groups in a data set. Numerous approaches to this problem have been suggested over the years. We decided to start from such kinds of algorithms. This is Calinski and Harabasz's index (1974), Hartigan's rule (1975), the Kranowski and Lai test (1985), silhouette statistic (Kaufman and Rousseeuw, 1990), Gap statistic (Tibshirani, 2001) and "jump" method (Sugar and James, 2003).



Executable algorithm: Em-Gm

Executable data: My data

Number of iterations: 130
0.2814 0.3321 0.2513 0.1352

Fig. 5. CAMA Toolbox. Result of clustering artificial data with EM algorithm

There are three important problems in clustering algorithms: determination of the "true" number of clusters, initial values of cluster centers and metrics. Now we implemented algorithms that solve only first problem.

Classical data sets that are used to check consistency of clustering algorithms are available by default. This is Iris flower (Fisher, 1936), Wisconsin (Wolberg and Mangasarian, 1990) and Auto Data (Quinlan, 1993). Many data sets are available in UCI Machine Learning Repository [17]. We suppose to support possibility to load and update data from this portal soon.

## VII. New clustering algorithm

The new clustering method is a modification of Sugar and James algorithm [18] that determines the number of clusters. The main procedure in this algorithm, which is called "jump method", is based on "distortion" that determines a measure of within cluster dispersion. It has the following simple steps:

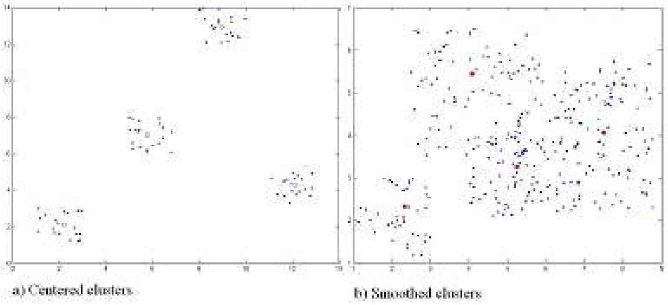
1. Run the k-means [19] algorithm for different numbers of clusters, $K$, and calculate the corresponding distortions

2. Select a transformation power, $Y > 0$ (A typical value is $Y = p/2$)

3. Calculate the "jumps" in transformed distortion $J_k = d_k^{-Y} - d_{k-1}^{-Y}$

4. Estimate the number of clusters in the data set by $K^* = argmax_k J_k$ the value of $K$ associated with the largest jump.

K-means is a typical clustering algorithm but it has two shortcomings in clustering large data sets: number of clusters



a) Centered clusters

b) Smoothed clusters

Fig. 6. SPSA algorithm in Sugar-James method. a) centered clusters; b) smoothed clusters

dependency and degeneracy. Number of clusters dependency is that the value of k is very critical to the clustering result. This shortcoming is not essential in our case. Degeneracy means that the clustering may end with some empty clusters. To avoid this problem and to improve results for noisy data it is proposed to use simultaneous perturbation stochastic approximation (SPSA) [20] algorithms which keep appropriate estimations under almost arbitrary noise [21].

The efficiency of modified Sugar-James method is demonstrated in the Fig.6.

Red circles in the Fig.6 show approximated cluster centers. As we could see they are similar with the ”true” values.

There are the other modifications of Sugar-James method concerned with analysis of ”distortion” curve.

This algorithm is easy to implement and to invoke in CAMA toolbox. Now all results were obtained with MATLAB.

## VIII. Conclusion

Described CAMA software tool for clustering data sets with a number of different algorithms is useful for researchers who want to investigate advantages of their new algorithms and for ordinary users who would like to determine the shape of their data sets, extract atypical objects or reduce amount of stored data.

CAMA also could help to gather and to structure known information about existent algorithms and data sets. Having such information it is possible to tune some clustering algorithms for special cases and even investigate the new ones.

For further development we hope to enrich the number of available algorithms not only for determination of the ”true” number of clusters, but also for determination cluster centers with hierarchical algorithms. It is supposed to implement possibility of working with many metrics at once. The engine to load and to update data from UCI [17] portal will be available soon.

Support of multithreading instructions for simultaneous performing algorithms and integration into the GRID portal is also planned.

## References

[1] Cluster Validation Toolbox CVAP, *http://www.mathworks.com/matlabcentral/fileexchange/14620*

[2] ClusterPack MATLAB Toolbox, *http://www.ideal.ece.utexas.edu/ strehl/soft.html*

[3] COMPACT Toolbox, *http://adios.tau.ac.il/compact*

[4] Data Clustering & Pattern Recognition Toolbox, *http://neural.cs.nthu.edu.tw/jang/matlab/toolbox/DCPR*

[5] Fuzzy Clustering and Data Analysis Toolbox, *http://webscripts.softpedia.com/scriptDownload/Clustering-Toolbox-Download-35404.html*

[6] Self-Organizing Map Toolbox, *http://www.cis.hut.fi/projects/somtoolbox/download*

[7] GNU General Public License, *http://www.gnu.org/copyleft/gpl.html*

[8] ANTLR, *http://www.antlr.org*

[9] Math.NET Iridium, *http://mathnet.opensourcedotnet.info/downloads/IridiumCurrentRelease.ashx*

[10] MSVS 2008 and .NET 3.5, *http://www.microsoft.com/events/series/msdnvs2008.aspx*

[11] .NET Reflector, *http://www.red-gate.com/products/reflector*

[12] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*, Wiley, New York, 1997.

[13] Java Server Pages (JSP), *http://java.sun.com/products/jsp*

[14] Java Servlets, *http://java.sun.com/products/servlet*

[15] Apache Tomcat Server, *http://tomcat.apache.org*

[16] Java Swing, *http://java.sun.com/docs/books/tutorial/uiswing*

[17] UC Irvine Machine Learning Repository, *http://archive.ics.uci.edu/ml*

[18] C. Sugar and G. James, *Finding the number of clusters in a data set : An information theoretic approach*, Journal of the American Statistical Association (98), 2003, pp. 750 - 763.

[19] J. A. Hartigan and M. A Wong, *A K-Means Clustering Algorithm*, Applied Statistics(28), 1979, pp. 100 - 108.

[20] J. C. Spall, *Introduction to Stochastic Search and Optimization. Estimation, Simulation and Control*, Wiley, London, 2003.

[21] O. N. Granichin, B. T. Polyak, *Randomized Algorithms of optimization and Estimation under Almost Arbitrary Noise*, Nauka, Moscow, 2003.