Modern robust fuzzy c-means clustering techniques

This three-year individual postdoc research project aimed at both theoretical and practical advances in the field of clustering algorithms. Results are reported in the following sections.

A UNIFIED THEORY OF FUZZY C-MEANS ALGORITHMS WITH IMPROVED AND SUPPRESSED PARTITION

Suppressed fuzzy c-means clustering (s-FCM) [1] was introduced with the main goal to reduce the execution time of the very popular fuzzy c-means clustering algorithm [2] without significantly damaging the quality of the produced partition. During its iterations, s-FCM manipulates with the partition given by FCM's optimal formula, by proportionally suppressing the lower memberships for each input datum, and giving all suppressed parts to the highest one, while keeping the probability constraint imposed by FCM, thus bringing the partition closer to the hard one. Suppression was found successful in terms of efficiency, but the authors left several doors wide open.

Within the bounds of this research project, we resolved two important problems concerning s-FCM. Firstly, based on a previous own study [3], we introduced a wide series of generalized suppressed fuzzy c-means algorithms (gs-FCM), and showed their advantages in sense of accuracy and efficiency. We showed that gs-FCM clustering models also suppress the multimodality of the probabilistic membership functions produced by FCM, thus enabling the algorithm to find clusters of lower cardinality, without being merged with larger ones.

Secondly, we have found the objective function gs-FCM clustering models minimize, thus proving the optimality of all suppressed c-means clustering algorithms. These clustering models are strongly related to the so-called FCM with improved partition [4], but are definitely distinct ones. Being optimal algorithms and easy to implement, further on efficient and accurate, generalized suppressed fuzzy c-means clustering models will surely have several successful applications [5].

The fuzzy c-means (FCM) algorithm usually produces fuzzy membership functions that are highly multimodal, especially when the number of clusters is high. This adverse effect needs compensation. The easiest way would be to reduce the fuzzy exponent (m) of the algorithm [2], but that eliminates the fuzziness. To maintain the fuzzy nature of the algorithm, a series of modified methods have been introduced, which in each iteration of the alternating optimization scheme manipulate with the partition given by FCM. The so-called FCM with (generalized) improved partition virtually reduces all distances between a given cluster prototype and all input vectors by the same value, causing a rise of the largest membership degree in the detriment of lower ones [4]. The suppressed FCM [1], and our generalizations versions [5], proportionally suppress lower fuzzy membership values and give the suppressed parts to the largest one. In the article [6] we introduced a unified theory of such algorithms and showed the relation among these algorithm families. We have also validated and compared these above mentioned algorithms using an image color reduction framework [7].

The algorithm unification theory presented in [6] is suitable for a more detailed description together with a survey of application papers applying such algorithms that could become a nice journal article in the future.

FUZZY-POSSIBILISTIC PRODUCT PARTITION – THEORETICAL ADVANCES AND APPLICATIONS

The fuzzy-possibilistic product partition (FP3) is a novel way of combining probabilistic and possibilistic factors into the c-means clustering framework, which we first introduced in [8]. At that early stage it confirmed the aim it was designed for: it can fully suppress the effect of outliers, it can treat outlier data the same way as gravity systems, which are not influenced by distant objects. During the first year of the research, the FP3 c-means (FP3CM) clustering model underwent a series of tests using various data and various cluster types. We tested the algorithm's behavior in case of small and large vector datasets, in various circumstances from scalar to multi-dimensional environments, and in case of special shaped clusters as well. We have developed an application of the FPPP c-means algorithm to the detection of clusters of ellipsoidal shape [9]. Numerical tests revealed that the fuzzy-possibilistic product partition is more robust than previous algorithms in this environment, suppressing the effect of distant outlier data. According to the tests, FP3CM is indeed a reliable, robust, accurate clustering model, which produces fine partitions in the presence or absence of outliers as well.

The FP3CM clustering algorithm was employed in a blind speaker recognition problem, and it proved better than its counter-candidates [10].

However, an adverse phenomenon was discovered concerning the FP3CM algorithm, namely its sensibility to initial cluster prototypes. The problem occurs when an input vector coincides with a cluster prototype. In such cases, the vector in question attracts one cluster to itself and no other vectors are accepted there. This problem was recently solved by a slight modification of the objective function. The modification was validated with a long series of tests. A manuscript in this matter is expected to be submitted to IEEE Transactions on Fuzzy Systems by the end of May 2016.

THE CASE OF FUZZY LOCAL INFORMATION C-MEANS CLUSTERING

In the last decade, a large set of automated image segmentation algorithms were published, which integrated local information into the fuzzy c-means clustering model, to enable the basic FCM to deal with several kinds of high frequency noises [11,12]. The fuzzy local information c-means (FLICM) introduced by Krinidis and Chatzis [12] proposed a certain fuzzy local factor added to the objective function of FCM, and an optimization scheme (OS) which led to fine image segmentation. However, after a deeper investigation, we have discovered that the FLICM objective function is not optimized by the OS given in [12]. Consequently, we have elaborated the correct optimization algorithm and a deep study of the FLICM algorithm, finally proposing several ways of improvement [13].

C-MEANS CLUSTERING MODEL IN APPLICATIONS WITH LARGE DATA

Large data requires special processing methods, mostly because of storage space and runtime limitations. The runtime of conventional FCM clustering is directly proportional with the number of input data, as long as it is possible to load the whole dataset. Our solution to reducing execution time is achieved by aggregating identical or similar items in the input dataset before handing them to clustering. In order to show the advantage of such solutions, we have developed a framework for image color reduction that can employ a wide range of clustering algorithms. In paper [14] we have introduced a color aggregation and selection scheme that, combined with fuzzy or hard c-means algorithm, produces fine-quality images with reduced number of colors in very short time. Further on, we have shown the advantages of the above solution using several FCM algorithm versions with improved partition [7]. As different parts of the image are treated independently of each other, the process is highly parallelizable, predicting further improvement in efficiency via GPU implementation.

EFFICIENT MARKOV CLUSTERING – WITH APPLICATION IN BIOINFORMATICS

Markov clustering (MCL) is a useful tool in protein sequence grouping [15] that works on a probabilistic directional graph modeled by a column stochastic matrix. Protein sequence databases have been growing very progressively lately. There is need for efficient clustering methodology which can deal with sequences in order of millions of items, which requires efficient computing with (possibly sparse) matrices with a million rows and columns. Our initial efforts were limited to dealing with the SCOP95 database, which contains 11944 proteins, the processing of which required 2-4 hours depending on the algorithm parameters. At the end of the research project, the Markov clustering of the SCOP data set is performed in 0.4-5 seconds, while a million-protein network is clustered in 100-120 minutes. This improvement was achieved in several steps:

1. In paper [16] we proposed an uncoupling technique to eliminate unnecessary workload of the algorithm. After having the first 5-7 iteration performed, the graph gets fragmented into isolated subgraphs, which can be separated as they have no further effects upon each other. This way the computation in further iterations continues on a large set of very small matrices, thus achieving 1000 times shorter execution times for late iterations, and 20-50 times shorter overall runtime for the whole process, compared to the conventional or naive formulation of MCL. We demonstrated the efficiency of the proposed implementation using the whole SCOP95 database and its carefully selected subsets. This acceleration does not affect the accuracy of the algorithm at all.

2. In paper [17] we proposed a sparse matrix model that implements the columns of the stochastic matrix using lists of records that store only the non-zero elements, and includes a special "push back" feature to insert elements to the end of the list. The latter accelerates the so-called expansion operation within the main loop of the Markov clustering. Eliminating the computation with zeros, we achieve quicker implementation of the first 7-10 loops of the Markov clustering process. The overall runtime is reduced 100-300 times. If we combine this solution with the previous one (first iterations performed by sparse matrix, later ones via uncoupling), 300-1000 times shorter overall runtime is achievable [18].

3. We introduced a so-called sparse supermatrix (SSM) model, which stores nonzero elements of the sparse matrix in arrays, together with its transposed value. This data structure allows us to perform extremely high speed Markov clustering, reducing the overall runtime of processing the whole SCOP95 dataset from over 24 hours to 15-40 seconds [19]. This solution still needed a dense matrix during the expansion operation, which limited the memory-efficiency of the algorithm.

4. We introduced a reformulated MCL solution that does not need a dense matrix when computing the expansion of the similarity matrix. The operations are executed in such an order, that the second power of the huge matrix is computed row by row, thus needing only a buffer that covers a single row of the dense matrix. All other data can be stored in a sparse format, allowing for a much efficient memory management. The limits of processable data sizes (on the same computer) have grown by an order of magnitude, while runtimes (on the same data set) have reduced 2-3 times [20]. A protein network of 250 thousand nodes can fit in the memory of an upper class personal computer.

5. Another memory efficient solution was proposed in [21], which stores only the upper diagonal half of the similarity matrix and executes MCL operations adapted to that data format. This change enabled us to cluster protein sequence networks of a million nodes with a personal computer.

6. Our ultimate version of the efficient MCL algorithm combines the above solution with the matrix splitting technique [16]. This is the algorithm that can cluster a protein sequence networks of a million nodes in 100 minutes. A manuscript describing this solution was recently submitted to Computer Methods and Programs in Biomedicine (Elsevier) journal.

7. Large sized protein sequence data sets, together with ground truth concerning their hierarchical grouping are difficult to find. That is why we have developed a synthetic data generator procedure, to provide large and huge sized test data for highly efficient sparse-matrix based Markov clustering algorithms. Being created according to the structure and properties of the SCOP95 protein sequence data set [22], the synthetic data act as a collection of proteins organized in a four-level hierarchy and a similarity matrix containing pairwise similarity values of the proteins. An ultimate high-speed Markov clustering algorithm was employed to validate the synthetic data. Generated data sets have a healthy amount of variability due to the randomness in the processing, and are suitable for testing graph-based clustering algorithms on large-scale data [23].

APPLICATION IN INFECTION CONTROL

As a member of the internationally recognized Hand-in-Scan team, I was involved in the development of an education system that can help avoid hospital infections. My role was the development of image segmentation methods. Some of the procedures built within the Hand-in-Scan device are based on c-means clustering algorithms developed within this project. A relevant recent result of the Hand-in-Scan team was showing the imperial role of instant visual feedback in the education of correct hand rubbing technique [24].

BRAIN TUMOR DETECTION AND SEGMENTATION

The c-means clustering algorithms developed within this project were involved in brain tumor detection and segmentation based on multispectral magnetic resonance images. We built an environment in which c-means clustering algorithms were employed in semi-supervised learning mode: the best parameters obtained for learning data were applied to test data. The outcome was considerably better than in case of non-supervised learning [25, 26]. To obtain even better recognition, we recently turned to a supervised learning method, namely the random forest technique.

APPLICATION FUZZY LOGIC IN BLOOD GLUCOSE CONTROL

PhD student Péter Szalay was employed in the project during this second year to provide an application of the fuzzy technology in a real-life problem, namely the blood glucose control. The application consisted in handling the drift of available sensors, which bias the signal causing the controller to drive the glucose concentration out of the safe region even in the case of frequent calibration. A linear-quadratic-Gaussian controller was employed on a widely used diabetes model and enhanced with an advanced sparse-grid quadratic filter and a fuzzy inference system-based calibration supervisor. The proposed controller reduced both hypoglycemic and severe hyperglycemic episodes for all virtual patients in the case of extreme meal intake and sensor drift [27].

BONE IMPLANT IMAGING

BSc student Eszter Iklódi was employed in the project to develop another image processing application for the algorithms introduced within this project. This framework handles micro CT images (volumes) of high resolution, created by dentists during experiments with bone implants. The goal of the project was to localize and separate thin and thick bone plates based on geometrical and morphological criteria. Segmentation was based on c-means clustering. Publications in this topic will follow. Eszter Iklódi will keep on working on this matter during her MSc studies.

TEST ENVIRONMENT FOR C-MEANS CLUSTERING

MSc student Bernát Gábor was employed in the project to help the development of a test framework, which will enable us to run enormous amounts of tests in a well-organized manner, using the variety of clustering models created and input data of various kinds and sizes (including huge data). The test framework is under construction, will be finished by the end of the second research year. The test framework is developed in c++ programming language and is planned to be platform independent.

PUBLICATIONS

Results were reported in five peer-reviewed journal articles (PATTERN RECOGNITION LETTERS, NEUROCOMPUTING, COMPUTERS IN MEDICINE AND BIOLOGY, JOURNAL OF HOSPITAL INFECTION, ACTA POLYTECHNICA HUNGARICA), and a couple more follow. Out of the 17 conference papers published, 10 were presented at highly ranked (CORE A) conferences (ICONIP, FUZZ-IEEE, IEEE EMBC, CINC).

REFERENCES

[1] Fan JL, Zhen WZ, Xie WX: Suppressed fuzzy c-means clustering algorithm. Pattern Recognition Letters 24:1607-1612 (2003)

[2] Bezdek JC: Pattern recognition with fuzzy objective function algorithms, Plenum, New York, NY (1981)

[3] Szilágyi, L, Szilágyi SM, Benyó Z: Analytical and numerical evaluation of the suppressed fuzzy c-means algorithm: a study on the competition in c-means clustering models. Soft Computing 14:495-505 (2010)

[4] Höppner F, Klawonn F: Improved fuzzy partitions for fuzzy regression models. Int. J. Approximative Reasoning 32:85-102 (2003)

[5] Szilágyi L, Szilágyi SM: Generalization rules for the suppressed fuzzy c-means clustering algorithms. NEUROCOMPUTING 139:298-309 (2014)

[6] Szilágyi L: A unified theory of fuzzy c-means clustering models with improved partition, Modeling Decisions for Artificial Intelligence, LNCS vol. 9321, pp. 129-140 (MDAI 2015, Skövde, Sweden), 2015

[7] Szilágyi L, Dénesi G, Kovács L, Szilágyi SM: Comparison of various improved-partition fuzzy c-means clustering algorithms in fast color reduction, 12th IEEE International Symposium on Intelligent Systems and Informatics (SISY 2014, Subotica), pp. 197-202 (2014)

[8] Szilágyi L: Fuzzy-Possibilistic Fuzzy Partition: a novel robust approach to c-means clustering. Lect. Notes in Comp. Sci. 6820:150-161 (2011)

[9] Szilágyi L, Varga ZsR, Szilágyi SM: Application of the fuzzy-possibilistic product partition in elliptic shell clustering. In: Torra V, Narukawa Y, et al (Eds.): Modeling Decisions for Artificial Intelligence, Springer, LNCS vol. 8825, pp. 158-169 (2014)

[10] Gosztolya G, Szilágyi L: Application of fuzzy and possibilistic c-means clustering models in blind speaker clustering, ACTA POLYTECHINCA HUNGARICA 12(7):41-56, 2015

[11] Cai W, Chen S, Zhang DQ: Fast and robust fuzzy c-means algorithms incorporating local information for image segmentation. Patt. Recogn. 40:825-838 (2007)

[12] Krinidis S, Chatzis V: A robust fuzzy information c-means clustering algorithm. IEEE T Image Proc. 19:1328-1337 (2010)

[13] Szilágyi L: Lessons to learn from a mistaken optimization. PATTERN RECOGNITION LETTERS 36:29-35 (2014)

[14] Szilágyi L, Dénesi G, Szilágyi SM: Fast color reduction using approximative c-means clustering models. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2014, Beijing), pp. 194-201 (2014)

[15] Enright AJ, van Dongen S, Ouzounis CA: An efficient algorithm for large-scale detection of protein families. Nucl. Acids Res 30:1575-1584 (2002)

[16] Szilágyi L, Szilágyi SM: Efficient Markov clustering algorithm for protein sequence grouping. 35th Ann. Int'l Conf. IEEE Eng. Med. Biol. Soc., Osaka, pp. 639-642 (2013)

[17] Szilágyi L, Szilágyi SM: An efficient Markov clustering approach to protein sequence grouping. J. Patt. Recogn. Imag. Proc. 3:263-272 (2013)

[18] Szilágyi L, Szilágyi SM: Fast implementations of Markov clustering for protein sequence grouping, Modeling Decisions for Artificial Intelligence, LNCS vol. 8234, pp. 214-225 (MDAI 2013, Barcelona), 2013

[19] Szilágyi SM, Szilágyi L: A fast hierarchical clustering algorithm for large-scale protein sequence data sets. COMPUTERS IN BIOLOGY AND MEDICINE 48:94-101 (2014)

[20] Szilágyi L, Szilágyi SM, Hirsbrunner B: A fast and memory-efficient heirarchical graph clustering algorithm. International Conference on Neural Information Processing, LNCS vol. 8834, pp. 247-254 (ICONIP 2014, Kuching, Malaysia), 2014

[21] Szilágyi L, Nagy LL, Szilágyi SM: Recent advances in improving the memory efficiency of the TRIBE MCL algorithm, International Conference on Neural Information Processing, LNCS vol. 9490, pp. 28-35 (ICONIP 2015, Istabul), 2015

[22] Andreeva A, Howorth D, Chadonia JM, Brenner SE, Hubbard TJP, Chothia C, Murzin AG: Data growth and its impact on the SCOP database: new developments. Nucleon Acids Research 36:D419–D425 (2008)

[23] Szilágyi L, Kovács L, Szilágyi SM: Synthetic test data generation for hierarchical graph clustering methods. International Conference on Neural Information Processing, LNCS vol. 8835, pp. 303-310 (ICONIP 2014, Kuching, Malaysia), 2014

[24] Lehotsky Á, Szilágyi L, Ferenci T, Kovács L, Pethes R, Wéber Gy, Haidegger T: Quantitative impact of direct, personal feedback on hand hygiene technique, JOURNAL OF HOSPITAL INFECTION 91:81-84, 2015

[25] Szilágyi L, Lefkovits L, Iantovics BL, Iclanzan D, Benyó B: Automatic brain tumor segmentation in multispectral MRI volumetric records, International Conference on Neural Information Processing, LNCS vol. 9492, pp. 174-181 (ICONIP 2015, Istanbul), 2015

[26] Szilágyi L, Lefkovits L, Benyó B: Automatic brain tumor segmentation in multispectral MRI volumes using a fuzzy c-means cascade algorithm, 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2015, Zhangjiajie, China), pp. 310-316, 2015

[27] Szalay P, Szilágyi L, Benyó Z, Kovács L: Sensor drift compensation using fuzzy interface system and sparse-grid quadrature filter in blood glucose control. International Conference on Neural Information Processing, LNCS vol. 8835, pp. 445-453 (ICONIP 2014, Kuching, Malaysia), 2014