Одеський національний університет імені І. І. Мечникова
Інститут математики, економіки і механіки
Кафедра оптимального керування і економічної кібернетики

Дипломна робота

магістра

на тему: **«Еволюційні алгоритми в задачах кластеризації»**

«Evolutionary algorithms in clustering problems»

Виконала: студентка денної форми навчання
спеціальності 8.04030101 Прикладна математика
Нікітіна Анастасія Олександрівна

Керівник: к. ф.-м. н. Страхов Є. М.
Рецензент: к. ф.-м. н, доц. Яровий А. Т.

Рекомендовано до захисту:
Протокол засідання кафедри
№ ___ від «_____» _____ р.
Завідувач кафедри
_____ _____

Захищено на засіданні ЕК № _____
Протокол № ____ від «____» ____ р.
Оцінка _____ / _____ / _____
Голова ЕК
_____ _____

Одеса — 2017 р.

Odesa I. I. Mechnikov National University
Institute of Mathematics, Economics and Mechanics
Department of Optimal Control and Economic Cybernetics

Diploma thesis

master

«Evolutionary algorithms in clustering problems»

Fulfilled by: full time student
specialty 8.04030101 Applied mathematics
Nikitina Anastasiia Oleksandrivna
Supervisor: PhD. in Ph. and Math. Strakhov Ye. M.
Reviewer: PhD. in Ph. and Math., doc. Yarovyi A. T.

Одеса — 2017 р.

CONTENTS

ВСТУП

Починаючи з 1950-х років кількість і загальний обсяг текстових документів, які використовуються у різних галузях людської діяльності неухильно зростає. Це зростання виражається як у збільшенні загального обсягу текстових колекцій, так і в значному зростанні кількості документів, що вимагають обробки і сортування. Внаслідок такого бурхливого зростання обсягу текстових даних постала необхідність попередньої систематизації текстових масивів в автоматичному режимі: в першу чергу, постала актуальна задача пошуку потрібної інформації у великих колекціях текстових даних (наприклад, що видаються пошуковими системами у відповідь на запити користувачів). Крім того, без систематизації текстів неможливо рішення і ряду інших завдань, зокрема:

- визначення взаємозв'язків між групами документів;
- спрощення візуалізації текстової інформації;
- виявлення дублікатів або близьких за змістом документів;
- контент-аналізу, тощо.

Відмінною особливістю методів кластеризації є здатність автоматично виділяти групи в потоці вхідних даних. Найбільш затребувана на сьогодні і, ймовірно, в найближчому майбутньому є змістова (або, по-іншому, тематична) кластеризація текстових документів. Цей вид кластеризації передбачає поділ текстових колекцій на групи текстів (далі - кластери), такі, що тексти в межах одного і того ж кластера максимально схожі між собою за змістом, в той час як тексти, що відносяться до різних кластерів, мають різний зміст. Саме змістова кластеризація (далі просто кластеризація) текстів розглядається в цій роботі.

Задачу кластеризації набору даних можна розглядати як задачу оптимізації критерію якості отриманого розподілу об'єктів. Одними з методів розв'язання такої задачі можуть служити еволюційні алгоритми, що використовують і моделюють біологічну еволюцію.

Минулого року ми досліджували ефективність застосування еволюційних алгоритмів до задачі кластеризації [1] та розробили програмне забезпечення, що, на основі методу диференціальної еволюції, реалізує автоматичну

кластеризацію об'єктів, представлених числовими даними. Метою цієї роботи було застосування розробленого алгоритму автоматичної кластеризації до практичної проблеми кластреризації колекції текстових документів природною мовою.

Мета роботи

Метою роботи є застосування методу автоматичної кластеризації на основі диференціальної еволюції до колекції текстових документів.

Для досягнення зазначеної вище мети необхідно вирішити такі задачі:

1) провести аналіз методів подання документів у векторному просторі;
2) провести аналіз та удосконалити методи зменшення розмірності простору для векторного подання документів;
3) вибрати ефективну міру подібності для вибраного подання документів у векторному просторі;
4) спроектувати програмне забезпечення для автоматичної кластеризації текстових документів;
5) дослідити ефективність отриманого методу.

INRODUCTION

Since 1950 the number and total amount of text documents that are used in various fields of human activity is steadily increasing. This increase is reflected as an growth of total collections of text, and a significant increase in the number of documents that require processing and sorting. Because of the rapid growth of text data it was necessary to systematize text collections automatically, primarily faced urgent task of finding relevant information in large collections of text data (for example, produced by search engines in response to user requests). Also, without systematization it is impossible to solve a number of other problems, such as:

- define relationships between groups of documents;
- simplify visualization of textual information;
- identify duplicate or content similar documents;
- content analysis, etc.

A distinctive feature of clustering methods is the ability to automatically allocate groups in a stream input. The most in demand today and probably in the near future is the content (or, in other words, thematic) clustering of text documents. This type of clustering involves separation of text collections into groups of texts (hereinafter - clusters) such that the texts within the same cluster most content similar to each other, while texts belonging to different clusters have different content. This content clustering approach (hereinafter simply clustering) of texts considered in this work.

The problem of clustering data set can be viewed as a problem of optimization the quality criterion of the clustering result. To solve this problem we can use evolutionary algorithms that model and use biological evolution.

Last year we investigated the efficacy of evolutionary algorithms in clustering problems [1] and developed software, that, based on differential evolution method, automatically implements clustering of objects presented as numerical data. The purpose of this study was to use automatic clustering algorithm to real-world problem, such as clustering of text documents in natural language.

The purpose of the work is to apply the automatic clustering method based on differential evolution on the collection of text documents.

To achieve the above goal we should solve the following problems:

1) analyze methods of text document representation in vector space;
2) analyze and improve dimensionality reduction methods of the space vector representation of documents;
3) choose effective measure of similarity for the selected representation of documents in a vector space;
4) design software for automatic clustering of the text documents;
5) evaluate the efficiency of the obtained method.

CONCLUSION

The project set out to develop document clustering technique based on evolutionary computing. This has involved a broad investigation from the underlying data models to various algorithms. Our main focus has been investigating different text processing methods in order to enhance the clustering results. Motivation has been on the one hand to reduce dimensionality in order to keep running times low and on the other to enhance clustering results. We have focused on topic modeling methods for dimensionality reduction. We achieved the following results:

1) Analyzed the problem of application area, formalized it and identified the main stages of solving the problem.
2) Reviewed existing methods of text documents representation in vector space and analyzeв their shortcomings.
3) Suggested the use of topic modeling as a method of reducing the dimensionality in order to improve clustering results.
4) Developed a method of finding the optimal number of topics for LDA model.
5) Designed software for automatic clustering of text documents based on the proposed algorithm.
6) Demonstrated efficiency of applying topic modeling to reduce dimensionality of vector space model.

Several most promising directions for future work are briefly described below:

1) Multiobjective clustering: it is important to consider multiple objectives (different clustering validity criteria) when evaluating the fitness of an individual representing a candidate clustering solution.
2) Combine of evolutionary approaches with traditional hierarchical clustering algorithms in order to combination of evolutionary approaches with traditional hierarchical clustering algorithms.
3) Building complex topic models: the LDA model can be considered as a base model, and more complex models can be build on top of it based on the complex needs we have from the data at hand.

BIBLIOGRAPHY

[1] Анастасія Нікітіна. "Про застосування методу диференціальної еволюції до задачі автоматичної кластерізації". In: *VIII Міжнародна школа-семінар Теорія прийняття рішень*. Праці школи семінару. (Sept. 26–Oct. 1, 2016). Ed. by Повідайчик Маляр Млавець. Ужгород: Інвазор, 2016, pp. 195–196.

[2] H. J. Bremermann. *The evolution of intelligence. the nervous system as a model of its environment*. Technical Report No.1, Department of Mathematics, University of Washington, Seattle, 1958.

[3] A. S. Fraser. "The evolution of intelligence. the nervous system as a model of its environment". In: *Australian Journal of Biological Science* 10 (1957), pp. 484–491.

[4] L. J. Fogel, A. J. Owens, and M. J. Walsh. *Artificial Intelligence through Simulated Evolution*. John Wiley and Sons, Inc., New York, 1966.

[5] T. Back. *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, New York, USA, 1996.

[6] J. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, 1992.

[7] R. Storn and K. Price. "Differential Evolution: A Simple and Efficient Heuristic for global Optimization over Continuous Spaces". In: *Journal of Global Optimization* 11 (1997), pp. 341–359. URL: `http://dx.doi.org/10.1023/A:1008202821328`.

[8] R. Gämperle, S. D. Müller, and P. Koumoutsakos. "A parameter study for differential evolution". In: *Advances in Intelligent Systems, Fuzzy Systems, Evolutionary Computation* (2002), 293–298.

[9] J. Ronkkonen, S. Kukkonen, and K. V. Price. "Real parameter optimization with differential evolution". In: *Proc. IEEE CEC* 1 (2005), 506–513.

[10] J. Brest et al. "Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems". In: *IEEE Trans. Evol. Comput* 10 (2006), 646–657.

[11] P. Brucker. "On the complexity of clustering problems". In: *Lecture Notes in Economics and Mathematical Systems* 157 (1978), 45–54.

[12] E. Falkenauer. *Genetic Algorithms and Grouping Problems*. John Wiley and Sons, Inc., New York, 1998.

[13] David L. Davies and Donald W Bouldin. "A cluster separation measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (1979), pp. 224–227.

[14] C.H. Chou, M.C. Su, and E. Lai. "A new cluster validity measure and its application to image compression". In: *Pattern Analysis and Applications* (2004), 205–220.

[15] S. Das, A. Abraham, and A. Konar. "Automatic clustering using an improved differential evolution algorithm". In: *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* 38.1 (2008), 218–237.

[16] G. Salton, A. Wong, and C. S. Yang. "A Vector Space Model for Automatic Indexing". In: *Commun. ACM* 18.11 (Nov. 1975), pp. 613–620. ISSN: 0001-0782. DOI: 10.1145/361219.361220. URL: http://doi.acm.org/10.1145/361219.361220.

[17] Man Lan et al. "A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines". In: *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*. WWW '05. Chiba, Japan: ACM, 2005, pp. 1032–1033. ISBN: 1-59593-051-5. DOI: 10.1145/1062745.1062854. URL: http://doi.acm.org/10.1145/1062745.1062854.

[18] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN: 0521865719, 9780521865715.

[19] Thomas Hofmann. "Probabilistic Latent Semantic Indexing". In: *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '99. Berkeley, California, USA: ACM, 1999, pp. 50–57.

[20] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 993–1022.

[21] M. Steyvers and T. Griffiths. "Latent Semantic Analysis: A Road to Meaning". In: ed. by T. Landauer, S. Dennis McNamara, and W. Kintsch. Laurence Erlbaum, 2007. Chap. Probabilistic topic models.

[22] Paul Jaccard. "The Distribution of the Flora in the Alpine Zone". In: *New Phytologist* 11.2 (Feb. 1912), pp. 37–50. URL: http://www.jstor.org/stable/2427226?seq=3.

[23] Tom De Smedt and Walter Daelemans. "Pattern for Python". In: *J. Mach. Learn. Res.* 13 (June 2012), pp. 2063–2067. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=2188385.2343710.

[24] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[25] A. K. McCallum. *Mallet: a machine learning for language toolkit*. http://mallet.cs.umass.edu.

APPENDIX A

PSEUDO-CODE FOR THE DE ALGORITHM

Algorithm A.1 Pseudo-code for the DE Algorithm

1: Read values of the control parameters of DE: scale factor F, crossover rate Cr, and the population size NP from user.
2: Set the generation number $G = 0$ and randomly initialize a population of NP individuals $P_G = \{\vec{X}_{1,G}, \ldots, \vec{X}_{NP,G}\}$ with $\vec{X}_{i,G} = \{x_{1,i,G}, x_{2,i,G}, \ldots, x_{D,i,G}\}$ and each individual uniformly distributed in the range $[\vec{X}_{min}, \vec{X}_{max}]$, where $\vec{X}_{min} = \{x_{1,min}, x_{2,min}, \ldots, x_{D,min}\}$ and $\vec{X}_{max} = \{x_{1,max}, x_{2,max}, \ldots, x_{D,max}\}$ with $i = [1,2,\ldots,NP]$
3: **while** the stopping criterion is not satisfied **do**
4: **for** $i = 1$ to NP // for each individual sequentially // **do**
5: **Mutation Step**: Generate a donor vector as: $\vec{V}_{i,G} = \vec{X}_{r_1^i,G} + F \cdot (\vec{X}_{r_2^i,G} - \vec{X}_{r_3^i,G})$
6: **Crossover Step**: Generate a trial vector$\vec{U}_{i,G}$ in the following way:

$$u_{j,i,G} = \begin{cases} \vec{v}_{j,i,G}, & if\ rand_{i,j}[0,1] \leq Cr\ or\ j = j_{rand} \\ \vec{x}_{j,i,G}, & otherwise \end{cases}$$

7: **if** $f(\vec{U}_{i,G}) \leq f(\vec{X}_{i,G})$ **then**
8: $\vec{X}_{i,G+1} = \vec{U}_{i,G}$
9: **else**$\vec{X}_{i,G+1} = \vec{X}_{i,G}$
10: **end if**
11: **end for**
12: Increase the Generation Count $G = G + 1$
13: **end while**

APPENDIX B

PSEUDO-CODE FOR THE JDE ALGORITHM

Algorithm B.1 Pseudo-code for the DE Algorithm

Read value the population size NP from user.
Set the generation number $G = 0$ and randomly initialize a population of NP individuals $P_G = \{\vec{X}_{1,G},\ldots,\vec{X}_{NP,G}\}$ with $\vec{X}_{i,G} = \{x_{1,i,G},x_{2,i,G},\ldots,x_{D,i,G}\}$ and each individual uniformly distributed in the range $[\vec{X}_{min},\vec{X}_{max}]$, where $\vec{X}_{min} = \{x_{1,min},x_{2,min},\ldots,x_{D,min}\}$ and $\vec{X}_{max} = \{x_{1,max},x_{2,max},\ldots,x_{D,max}\}$ with $i = [1,2,\ldots,NP]$
Randomly initialize $\vec{F}_G = \{F_{1,G},\cdot,F_{NP,G}\}$ and $\vec{Cr}_G = \{Cr_{1,G},\cdot,Cr_{NP,G}\}$, where each $F_{i,G}$ uniformly distributed in the range [0.1,1] and each $Cr_{i,G}$ uniformly distributed in the range [0,1] with $i = [1,2,\ldots,NP]$
while the stopping criterion is not satisfied **do**
    for $i = 1$ to NP // for each individual sequentially // **do**
        Generate new control parameters $F_{i,G+1}$ and $Cr_{i,G+1}$ in the following way:

$$F_{i,G+1} = \begin{cases} 0.1 + rand_1[0,1],\ if\ rand_2[0,1] < 0.1 \\ F_{i,G},\ otherwise \end{cases}$$

$$Cr_{i,G+1} = \begin{cases} rand_3[0,1],\ if\ rand_4[0,1] < 0.1 \\ Cr_{i,G},\ otherwise \end{cases}$$

        Mutation Step: Generate a donor vector as: $\vec{V}_{i,G} = \vec{X}_{r_1^i,G} + F_{i,G+1} \cdot (\vec{X}_{r_2^i,G} - \vec{X}_{r_3^i,G})$
        Crossover Step: Generate a trial vector $\vec{U}_{i,G}$ in the following way:

$$u_{j,i,G} = \begin{cases} \vec{v}_{j,i,G}, & if\ rand_{i,j}[0,1] \le Cr_{i,G+1}\ or\ j = j_{rand} \\ \vec{x}_{j,i,G}, & otherwise \end{cases}$$

        if $f(\vec{U}_{i,G}) \le f(\vec{X}_{i,G})$ **then**
            $\vec{X}_{i,G+1} = \vec{U}_{i,G}$
        else
            $\vec{X}_{i,G+1} = \vec{X}_{i,G}$
            $F_{i,G+1} = F_{i,G}$
            $Cr_{i,G+1} = Cr_{i,G}$
        end if
    end for
    Increase the Generation Count $G = G + 1$
end while

APPENDIX C

PSEUDO-CODE FOR THE ACDE ALGORITHM

Algorithm C.1 Pseudo-code for the ACDE Algorithm

1: Initialize each search variable vector in jDE to contain K number of randomly selected cluster centers and K (randomly chosen) activation thresholds in [0, 1].
2: Find out the active cluster centers in each chromosome with the help of the rule described in Algorithm 2.1.
3: **for** $iter = 1$ to $MAXITER$ **do**
4: For each data vector $\vec{S}_p$, calculate its distance metric $d(\vec{S}_p, \vec{m}_{i,j})$ from all active cluster centers of the ith DE-vector $\vec{X}_i$
5: Assign $\vec{S}_p$ to that particular cluster center $\vec{m}_{i,j}$ where $d(\vec{S}_p, \vec{m}_{i,j}) = \min_{\forall b \in \{1,2,\ldots,K\}} d(\vec{S}_p, \vec{m}_{i,b})$
6: Change the population members according to the jDE algorithm proposed in section 1.3. Use the fitness of the vectors to guide the evolution of the population.
7: **end for**
8: Report as the final solution the cluster centers and the partition obtained by the globally best vector (one yielding the highest value of the fitness function)