IEEE Catalog Number: CFP17561-PRT
ISBN: 978-1-5386-0500-4

11th IEEE International Conference

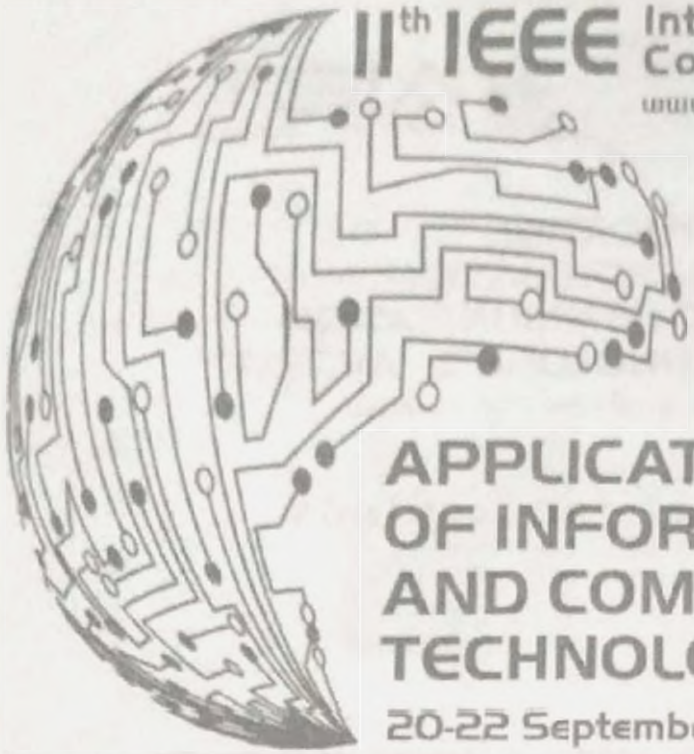
AICT2017
INTERNATIONAL CONFERENCE

APPLICATION of INFORMATION and COMMUNICATION TECHNOLOGIES - AICT2017

Moscow, Russia
20-22 September 2017

CONFERENCE PROCEEDINGS VOL 1.

www.aict.info/2017/



11th IEEE International Conference

www.aict.info/2017

AICT 2017 INTERNATIONAL CONFERENCE

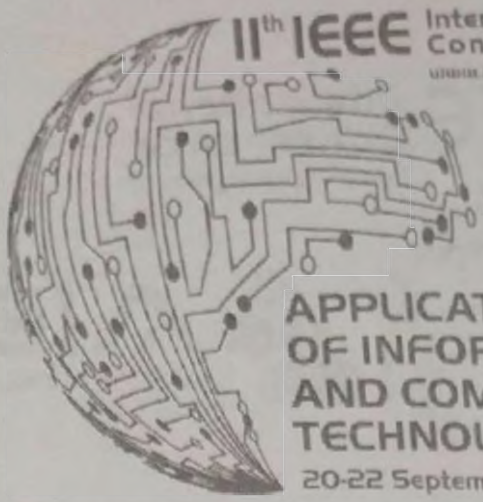
APPLICATION OF INFORMATION AND COMMUNICATION TECHNOLOGIES - AICT2017

20-22 September 2017, Moscow

CONFERENCE PROCEEDINGS

20-22 September 2017, Moscow, Russia

www.aict.info/2017



APPLICATION OF INFORMATION AND COMMUNICATION TECHNOLOGIES - AICT2017

20-22 September 2017, Moscow

CONFERENCE PROCEEDINGS

2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT)

IEEE Catalog Number: CFP1756H-PRT
ISBN: 978-1-5386-0500-4

Additional copies of this publication are available from
Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
+1 845 758 0400
+1 845 758 2633 (FAX)
email: curran@proceedings.com

© Designed by the AICT2017 Publication Team, 2017.

IEEE Catalog Number: CFP1756H-PRT
ISBN: 978-1-5386-0500-4

2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT) is financially supported by the Russian Foundation for the Basic Research, proiect #17-08-20518.

RUSSIAN FOUNDATION FOR BASIC RESEARCH

SESSION 3. CYBER SECURITY ISSUES

Semi-Supervised Evolving Approach for Data Streams Classification Based on Online Gustafson-Kessel Algorithm

I.V. Gorbunov, M.O. Kalmykov, E. V. Rasskazov
Department of Complex Information Security
Tomsk State University of Control Systems and Radioelectronics
Tomsk, Russia
giv@keva.tusur.ru, keso.azure@gmail.com, rev7.azure@gmail.com

A. E. Yankovskaya
Tomsk State University of Architecture and Building, National Research Tomsk State University, National Institute Tomsk Polytechnic University, Tomsk State University of Control Systems and Radioelectronics,
Tomsk, Russia
ayyankov@gmail.com

Abstract— The purpose of the article is to present the approach to data streams qualification in mode that is close to semi-supervised mode. The given approach combines modified online Gustafson-Kessel algorithm for work in mode of data classification. In the approach, there are used such steps as update of clusters, merging of clusters and deleting of unused for a longer time clusters. The benchmark of the accuracy of the suggested approach based on datasets from information repository keel.es is represented. In the conclusion, the recommendations of the suggested approach are given.

Index Terms— Evolving systems, classification, Gustafson-Kessel algorithm, semi-supervised, data streams.

I. Introduction

So far, the data quantity is growing concurrently; data mining is used increasingly frequently to increase living standards and profits. The given facts make the usage of data streams treatment actual. First of all, it happens because of the impossibility of using well-known and well-researched methods of batch computing in conditions of such amount of incoming data. The main reason is that the majority of the methods of batch computing are developed with assumption for quick access to data pre-loaded into RAM. But nowadays data streams are bigger than personal computers are able to embed into the memory. For this reason, the methods of data streams computing are currently important.

Special mention among data streams computing should go to the methods which solve the problem of classification. As a rule, they are based on decision trees [1-4], SVM [5-6], Naïve Bayesian classifier [7], statistical approaches [8-9], and clustering methods [10-12]. The approach suggested is based on the use of online Gustafson-Kessel algorithm of data clustering. The advantages of online Gustafson-Kessel algorithm in combination with such statistical approaches as covariance matrix and heuristic threshold parameters. They make the algorithm fast and flexibly adapted to different types of stream data. The disadvantages of the algorithm are the necessity of store for every cluster, covariance matrix, the number of objects in cluster, the number of object of each class, the total distance from the center of the cluster to all the objects included into it and the indication of recent data income.

We must admit that online Gustafson-Kessel algorithm for the solution of the task of data approximation is used widely [13]. The application of this algorithm to the problem of data approximation is based on the combination of the method of online clusterization for input data set, and for output space the method of approximating line evaluation for each cluster based on recursive least square method. In contrast to the indicated analogue there was a need to solve the problem of calculation of distance in space of class marks and connect the process of input data clusterization with the classification problem of output data.

In the following section the approach to data streams classification and the suggestion of the three methods of class label calculation on the basis of a cluster will be described.

II. Data Stream Classification Approach

Evolving Gustafson-Kessel algorithm is taken as the basis [14], complemented by the procedure of clusters merging suggested in the article [15], and the method of deleting the clusters described in the article [16]. For the description of suggested approach we define the following: α – the speed of cluster center update; β – the boundary probability of layout chi-square; γ – the initial scale for the covariance matrix; n – the spatial dimension of input data; r – the number of classes; $\chi^2_{n,\beta}$ – the boundary probability of layout chi-square; η_1 – the minimum distance to cluster merging; η_2 – the maximum period of cluster update; I – the identity matrix; x_k – the vector of input data which are available at the k time object; z_k – output class label at the k time object; v_i – the vector of the center of i cluster; F^{-1}_i – covariation matrix for i cluster; $d^2_{k,i}$ – the distance between the center of i cluster and k object; $u_{k,i}$ – the weight ratio of the proximity between i cluster and k object; $D^2_{i,k}$ –

closeness estimation between i cluster and k object; w_i – the counter of a non-updatable cluster; p_k – the index of a cluster with the minimum value of $D^2_{k,i}$; M_i – the number of objects in i cluster; $O_{l,i}$ – the number of objects of l class in i cluster; $S_{l,i}$ – the summary estimation of the l class objects closeness to the center of i cluster.

Each time a new object arrives at the time object k, a distance between object x_k and all cluster centers is computed:

$$d^2_{k,i} = (\mathbf{x}_k - \mathbf{v}_i) * (\det(\mathbf{F}_i))^{\frac{1}{n}} * \mathbf{F}_i^{-1} * (\mathbf{x}_k - \mathbf{v}_i)^T, i = [1,c]. \quad (1)$$

and weight ratio of proximity between cluster i and object k is evaluated using the following formula:

$$u_{k,i} = \frac{1}{\sum_{j=1}^{c} \frac{d^2_{k,i}}{d^2_{k,j}}}, i = [1,c]. \quad (2)$$

Next, an estimation of closeness of each cluster to an object x_k is evaluated in the following way:

$$D^2_{k,i} = (x_k - v_i) * F^{-1}{}_i * (x_k - v_i)^T, \quad (3)$$

and an index of the closest cluster is determined as follows:

$$p_k = \arg\min_{i=[1,c]}(D^2_{ik}). \quad (4)$$

If $D^2_{pk} < \chi^2_{nb}$ or z_k is unavailable, object x_k is added to cluster p_k, otherwise a new cluster is created. During the cluster creation, the values describing it are computed according to formulas (5-9):

$$v_{pk} = x_k, \quad (5)$$

$$\mathbf{F}^{-1}{}_{pk} = \gamma * I, \quad (6)$$

$$O_{l,pk} = \begin{cases} 1, \text{if } l = z_k, \\ 0, \text{else} \end{cases}, \quad (7)$$

$$S_{l,pk} = \begin{cases} d^2{}_{k,p_k}, \text{if } l = z_k, \\ 0, \text{else} \end{cases}, \quad (8)$$

$$w_{pk} = 0. \quad (9)$$

While adding an object x_k to a cluster with index p_k, one has to recalculate a covariance matrix determinant using formula (10), covariance matrix itself according to formula (11), the cluster center by formula (12), if z_k value is available, increase an amount of objects of each class in cluster by formula (13),increase total distance between cluster center and objects belonging to each of the classes by formula (14), and update centers of all other classes by formula (15), increasing their counters as follows (16):

$$\det(\mathbf{F}^{-1}_{p_k}) = (1-\alpha)^{n-1} * \det(\mathbf{F}^{-1}_{p_k}) * \left(1 - \alpha + \alpha * D^2_{k,p_k}\right), \quad (10)$$

$$\mathbf{F}^{-1}_{p_k} = \left[\mathbf{I} - \mathbf{G} * \left(\mathbf{x}_k - \mathbf{v}_{p_k}\right)\right] * \mathbf{F}^{-1}_{p_k} * \frac{1}{1-\alpha}, \quad (11)$$

where $\mathbf{G} = \dfrac{\mathbf{F}^{-1}_{p_k}(\mathbf{x}_k - \mathbf{v}_{p_k}) * \alpha}{1 - \alpha + \alpha * D^2_{k,p_k}}$.

$$\mathbf{v}_{p_k} = \mathbf{v}_{p_k} + \alpha * (\mathbf{x}_k - \mathbf{v}_{p_k}), \quad (12)$$

$$O_{l,p_k} = \begin{cases} O_{l,p} + 1, \text{if } l = z_k, \\ O_{l,p}, \text{else} \end{cases}, \quad (13)$$

$$S_{l,p_k} = \begin{cases} S_{l,p_k} + d^2{}_{k,p_k}, \text{if } l = z_k, \\ S_{l,p_k}, \text{else} \end{cases}, \quad (14)$$

$$\mathbf{v}_q = \mathbf{v}_q - \alpha * (\mathbf{x}_k - \mathbf{v}_q), q = [1,c], q \neq p_k. \quad (15)$$

$$w_{q_k} = w_q + 1. \quad (16)$$

If an object x_k with absent z_k was delivered, it is necessary to determine its class (working in classifier mode). It can be achieved by various ways. 3 different options are represented by formulas (17-19):

$$z^1_k = \arg\min_{l=[1,r]}(O_{l,p_k}), \quad (17)$$

$$z^2_k = \arg\min_{l=[1,r]}(S_{l,p_k} / O_{l,p_k}), \quad (18)$$

$$z^3_k = \arg\min_{l=[1,r]}(S_{l,p_k} / M_{p_k}), \quad (19)$$

where $M_{p_k} = \sum_{q=1}^{r} O_{q,p_k}$.

In the following section, we will describe a comparison of these ways of object classification.

During the update of cluster centers a situation can occur when cluster centers are too close to each other and request estimation of data as belonging to the same class. In such situation one has to check following conditions for inputs (20) and outputs (21):

$$\sqrt{D^2_i(\mathbf{v}_j) * D^2_j(\mathbf{v}_i)} < \eta_1, i, j = [1,c], i \neq j, \quad (20)$$

for evaluation of D^2_i a formula similar to (3) with x_k replaced by v_j is used

$$\exists l, \frac{O_{l,i}}{M_i} > 1.2/r \text{ AND } \frac{O_{l,j}}{M_j}, i, j = [1,c], i \neq j. \quad (21)$$

If conditions (20) and (21) are fulfilled for the same cluster pair indexed as i and j, it is necessary to preform cluster merging according to formulas (22-26):

$$\mathbf{v}_q = \frac{M_i * \mathbf{v}_i + M_j * \mathbf{v}_j}{M_i + M_j}, \quad (22)$$

$$F^{-1}_q = \frac{M_i * \mathbf{F}^{-1}_i + M_j \mathbf{F}^{-1}_j + \left(\dfrac{M_i * M_j}{M_i + M_j}\right) * (\mathbf{v}_i - \mathbf{v}_j) * (\mathbf{v}_i - \mathbf{v}_j)^T}{M_i + M_j}, \quad (23)$$

$$O_{l,q} = O_{l,i} + O_{l,j}, l = [1,r], \quad (24)$$

$$S_{l,q} = S_{l,i} + S_{l,j}, l = [1,r]. \quad (25)$$

$$wq_{-} = 0. \quad (26)$$

After performing all the previously described steps one has to check the necessity of cluster deletion: if cluster i was not updated by new values for more than η_2 time objects, a cluster has to be deleted.

Let us show the approach being provided as UML activity diagram on fig 1.

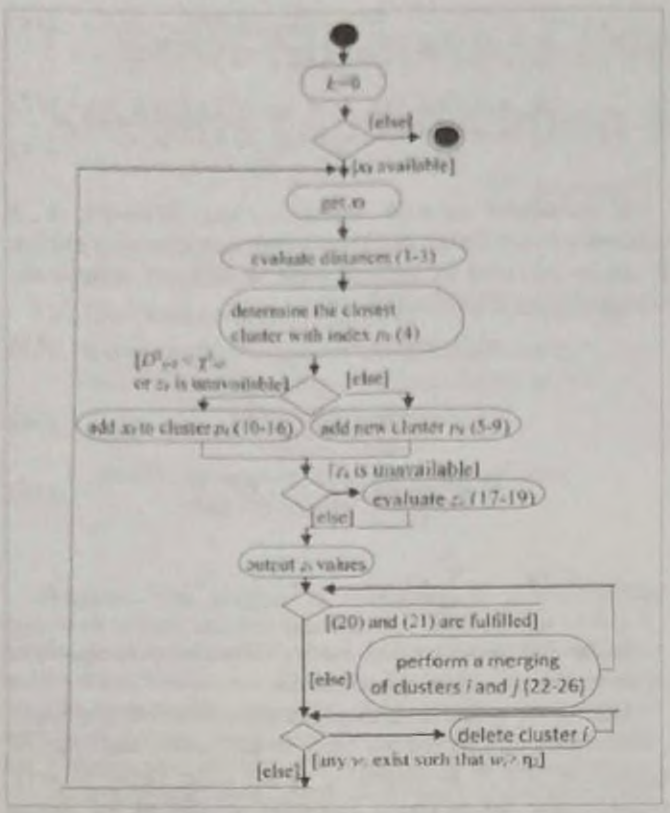


Fig. 1. UML activity diagram of the approach

In the next section, we will move to benchmarking different approaches to determining a class label and their influence on the accuracy.

III. Experiments

The experiments have been performed by considering 4 real-world datasets (http://www.keel.es). To carry the different experiments out, a five-fold cross-validation model has been applied. The characteristics of these datasets for the classification problem are given in the table 1.

TABLE I. Test datasets characteristics

Dataset	Features	Instances	Classes
Hepatits	19	80	2
Wine	13	178	3
Vehicle	18	846	4
Segment	19	2310	7

Given the fact that the data streams computing methods depend greatly on the sequence of the data, each test was carried out 30 times with different data sequences in order to increase the reliability of the results of the experiment.

The results of the experiment are provided in table 2 as average classification accuracy. To enable the possibility to compare with batch mode analogue the last strings are provided containing data for fuzzy classifier built with the usage of algorithms of initialization and optimization [18].

TABLE II. A comparison of the results of the approach with the batch mode analogue

Name	Training set accuracy (%)	Test set accuracy (%)	Evaluation time(s)
Hepatits			
Inference by formula (17)	75,82	65,47	0,34
Inference by formula (18)	78,44	70,7	0,36
Inference by formula (19)	86,47	80,92	0,37
Fuzzy classifier	**94,43**	**88,41**	43
Wine			
Inference by formula (17)	76,13	72,35	0,38
Inference by formula (18)	86,6	82,81	0,40
Inference by formula (19)	88,73	83,38	0,42
Fuzzy classifier	**99,52**	**96,94**	61
Vehicle			
Inference by formula (17)	59,73	48,04	2,92
Inference by formula (18)	57,93	**49,67**	2,99
Inference by formula (19)	**60,57**	49,56	3,05
Fuzzy classifier	53,48	48,87	1287
Segment			
Inference by formula (17)	71,93	71,45	8,72
Inference by formula (18)	73,07	72,08	9,15
Inference by formula (19)	78,52	75,81	9,66
Fuzzy classifier	**89,61**	**85,43**	279683

The conclusions of the experiments' results are provided in the next section.

IV. Conclusion

This paper shows the significance of evolving classifiers for the data streams. General approaches to creating such classifiers are briefly described. The approach is provided, based upon the combination of ideas developed by the authors of online Gustaffson-Kessel method in order to create approximators. That approach is applied to the classification task.

Several ways of evaluating the class label in the systems of that kind are provided. The experiments are carried out in order to determine the influence of an inference type on the accuracy and performance.

In accordance to the experiment the conclusion presented by formula 19 is recommended to use as it provides more accurate results with an insignificant loss of time in comparison to the others. On the majority of sets it is seen that the accuracy of the suggested one is less to a fuzzy classificator, but a fuzzy classificator demands much more time in terms of enlarged range.

From now forward it is planned to combine the suggested approach with the other evolving methods of data streams clasification in case of accuracy improvement. It will be the best option for this approach because in the majority of the operations it may be implemented as a single thread (except some matrix operations), thus the other threads may be uploaded using the other methods.

ACKNOWLEDGMENT

We appreciate the support of the Russian Foundation for Basic Research (16-07-00034, 16-07-00859) for this work.

REFERENCES

[1] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, M. Wozniak, "Ensemble learning for data stream analysis: A survey," Information Fusion 2017, Vol. 37, pp. 132–156.

[2] A. Isazadeh, F. Mahan, W. Pedrycz, "MFlexDT: multi flexible fuzzy decision tree for data stream classification," Methodologies and Application, Soft Computing, 2015, Vol. 20, pp. 1-15

[3] Z. Mirzamomen, M. R. Kangavari, "Evolving Fuzzy Min–Max Neural Network Based Decision Trees for Data Stream Classification," Neural Processing Letters, 2017, Vol. 45, pp. 341-363.

[4] G. Song, Y. Ye, H. Zhang, X. Xu, R. Y.K. Lau, F. Liu, "Dynamic Clustering Forest: An Ensemble Framework to Efficiently Classify Textual Data Stream with Concept Drift", Information Sciences, 2016, Vol. 357, pp. 125–143.

[5] N. Sun, Y. Guo, "A Modified Incremental Learning Approach for Data Stream Classification", Proceeding of the "Sixth International Conference on Internet Computing for Science and Engineering", Henan, 2012, pp. 122-125.

[6] H. M. Gomes,J. P. Barddal,F. Enembreck,A. Bifet, "A survey on ensemble learning for data stream classification. ACM Comput. Surv. 2017, Vol 50, pp. 1-36.

[7] D. K. Babu, · Y. Ramadevi, · K. V. Ramana, "RGNBC: Rough Gaussian Naïve Bayes Classifier for Data Stream Classification with Recurring Concept Drift,"Arabian Journal for Science and Engineering, 2017, Vol. 42, pp 705–714.

[8] B. S. Y. J. Costa, C. G. Bezerra, L. . Guedes, P. P. Angelov, "Unsupervised classification of data streams based on Typicality and Eccentricity Data Analytics," Proceedings of the Fuzzy Systems (FUZZ-IEEE), 2016, pp. pp. 58-63.

[9] E. Lughofer; M. Pratama, "On-line Active Learning in Data Stream Regression using Uncertainty Sampling based on Evolving Generalized Fuzzy Models," in IEEE Transactions on Fuzzy Systems, 2017 vol.PP, no.99, pp.1-1

[10] M. M. Masud. T. M. Al-Khateeb, K. W. Hamlen, J. Gao, L. Khan, J. Han, B. Thuraisingham "Cloud-Based Malware Detection for Evolving Data Streams", ACM Transactions on Management Information Systems, 2011, Vol. 2, pp. 1-27

[11] R. Hyde, P. Angelov, A.R. MacKenzie, "Fully Online Clustering of Evolving Data Streams into Arbitrarily Shaped Clusters," Information Sciences, 2017, Vol. 382–383 382-38. pp. 96-114.

[12] J. L. R. Perez, B. Ribeiro, C. M. Perez, "Mahalanobis Distance Metric Learning Algorithm for Instance-based Data Stream Classification", Proceeding of the International Joint Conference on Neural Networks (IJCNN), 2016, pp. 1857-1862.

[13] S. Shafieezadeh Abadeh, A. Kalhor, "Evolving Takagi–Sugeno model based on online Gustafson–Kessel algorithm and kernel recursive least square method," Proceeding of the IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), 2014, pp. 1-8.

[14] L. Serir, E. Ramasso, P. Nectoux, O. Baurer, N. Zerhouni, "Evidential Evolving Gustafson-Kessel Algorithm (E2GK) and its application to PRONOSTIA's Data Streams Partitioning," Proceeding of the 50th IEEE Conference on Decision and Control and European Control Conference, CDC-ECC'12., 2011, pp. pp. 8273-8278.

[15] I. Skrjanc, D. Dovzan, "Evolving Gustafson-Kessel Possibilistic c-Means Clustering,"Procedia Computer Science, 2015, Vol. 53, pp. 191-198.

[16] D. Filev, O. Georgieva, "An extended version of the Gustafson–Kessel algorithm for evolving data stream clustering", Evolving Intelligent Systems, John Wiley and Sons, 2010, pp. 293-315.

[17] Hodashinsky I.A., Meshcheryakov R.V., Gorbunov I.V., "Designing fuzzy rule-based classifiers using a bee colony algorithm // Informatics," Networking and Intelligent Computing proceedings of the 2014 international conference, 2015, pp 25–34.

Научное издание / Scientific edition

The 11th IEEE International Conference

Application of Information and Communication Technologies

AICT2017

CONFERENCE PROCEEDINGS

20-22 September, Moscow, Russia

Применение информационно-коммуникационных технологий

Труды 11-ой IEEE Международной конференции

20-22 сентября 2017, Москва, Россия

В ДВУХ ТОМАХ

ТОМ 1

Подписано в печать 08.09.2017.
Формат 60×90/8. Усл. печ. л. 57.5
Тираж 150 экз. Заказ 197.

ОТПЕЧАТАНО:
ФГБУН Институт проблем управления им. В.А. Трапезникова
Российской академии наук
117997, Москва,
ул. Профсоюзная, д. 65
http://www.ipu.ru