Position Papers of the of the 17th Conference on Computer Science and Intelligence Systems pp. 3–11
DOI: 10.15439/2022F264
ISSN 2300-5963 ACSIS, Vol. 31

Application of Random Sampling in the Concept-Dependent Granulation Method

Radosław Cybulski
University of Warmia and Mazury,
in Olsztyn
ul. Słoneczna 54, 10-710 Olsztyn, Poland
Email: radoslaw.cybulski@uwm.edu.pl

Piotr Artiemjew
University of Warmia and Mazury,
in Olsztyn
ul. Słoneczna 54, 10-710 Olsztyn, Poland
Email: artem@matman.uwm.edu.pl

***Abstract*—Professor Zadeh in his works proposed the idea of grouping similar objects on the basis of certain similarity measures, thus initiating the paradigm of granular computing. He made the assumption that similar objects may have similar decisions. This natural assumption, operates in other scientific methodologies, e.g. methods based on k nearest neighbours, in reasoning by analogy and in rough set theory. The above assumption implies the existence of grouped information nodes (granules) and has potential applications in reducing the size of decision systems. The hypothesis has guided,, the creation of granulation techniques based on the use of rough inclusions (introduced by Polkowski and Skowron) - according to the scheme proposed by Polkowski. In their work, the possibility of a large reduction of the size of decision systems while maintaining the classification efficiency was verified in experimental works.**

In this paper, we investigate the possibility of using random sampling in the approximation of decision systems - as part of dealing with Big Data sets. We use concept-dependent granulation as a reference approximation method. Experiments on selected real-world data have shown a common regularity that gives a hint on how to apply random sampling for fast and effective size reduction of decision systems.

I. Introduction

IN THIS paper, we employ a granulation technique derived from rough set theory [3]. More specifically, we applied the concept-dependent granulation technique to reduce the size of decision systems [6], a methodology derived from the method proposed by Polkowski in paper [4] and extended in the paper [6]. A comprehensive research in this context is conducted in the monograph [5]. A demonstration of the decision system approximation using the concept-dependent method - showing the use of granulation to reduce the size of decision systems - can be seen in Table I. In this Table, we see how the granulation process allows us to reduce the size of the training systems while retaining the internal knowledge from the original systems. For example, for a radius of 0.682 we have a reduction in the number of objects of almost 98 percent while maintaining the original efficiency. The effectiveness of granulation methods (according to Polkowski's scheme) has been verified in many contexts and works with basically every popular classifier from SVM [8], decision trees [10] to neural networks [9]. The methods have also found applications in the context of steganography [11], preprocessing before feeding data into neural networks [9], in ensemble models [7], in classification processes [12], for absorbing missing values [13], in localization of mobile robots under magnetically variable conditions [14]. Due to the computational complexity of our techniques, an area for exploration that has not yet been adequately explored is their use for with methods for dealing with Big Data. The use of random sampling is our starting point in this area. In this paper, we use examples of relatively small decision systems for a simple illustration of the techniques. The application to big data of our method is to sample up to the size of the data that can be recalculated in the assumed time. Our results tentatively verify this possibility. As a reference classifier, we chose the kNN method, which is not a dedicated choice for our method. Any other classification method adapted to the granular data could be used to verify our assumptions.

The rest of the publication consists of the following sections. In Section II, we have an introduction to the methodology used in the paper, a demonstration of the granulation method and an indication of the classifier. In Section III, we present experimental results divided into two parts. Initial results showing cross-sectional classification performance with different radii, and detailed results with random sampling for selected sensible (giving variable results) granulation radii. We summarise the work in Section IV, where we also present our future research plans.

II. Research Methodology

In this section we will introduce our reference granulation technique and the classifier used.

A. Refence granulatiom method - concept-dependent granulation

Let us illustrate the operation of the concept-dependent granulation technique with an example. The system that is being granulated was generated by the Toy Decision system generator tool [1], [2].

Let us define

$$g^{cd}_{r_{gran}}(u_i) = \{u_j \in U_{trn} : \frac{|IND(u_i, u_j)|}{|A|} \geq r_{gran}$$

TABLE I
EXAMPLE OF CLASSIFICATION USING TOY DATA - MUSHROOM DATA SET. THE RESULTS PRESENTED HERE ARE FOR TWO SUCCESSIVE GRANULATIONS, THE FIRST BEING $layer_1$ THE SECOND $layer_2$.

	$layer_1$		$layer_2$	
r_{gran}	acc	$GranSize$	acc	$GranSize$
0.364	0.887	5.4	0.886	2
0.409	0.884	9.4	0.884	2
0.455	0.891	15.6	0.89	2
0.5	0.915	20.2	0.894	2.8
0.545	0.947	33.8	0.903	4.8
0.591	0.966	40.8	0.887	8.6
0.636	0.983	44.2	0.905	11.2
0.682	0.994	43.8	0.946	15.8
0.727	0.995	48	0.977	21.6
0.773	0.996	58	0.992	27
0.818	1	94.2	0.996	41.6
0.864	1	200.4	1	82.8
0.909	1	504.8	1	226.6
0.955	1	1762.8	1	947.6
1	1	6499.2	1	6499.2

TABLE II
EXEMPLARY DECISION SYSTEM: IRIS-SHORT, 5 ATTRIBUTES, 15 OBJECTS

Day	$a1$	$a2$	$a3$	$a4$	$class$
u_1	4.6	3.1	1.5	0.2	$Iris-setosa$
u_2	4.9	3.1	1.5	0.1	$Iris-setosa$
u_3	5.1	3.3	1.7	0.5	$Iris-setosa$
u_4	4.4	3.0	1.3	0.2	$Iris-setosa$
u_5	5.0	3.6	1.4	0.2	$Iris-setosa$
u_6	6.0	3.4	4.5	1.6	$Iris-versicolor$
u_7	5.9	3.2	4.8	1.8	$Iris-versicolor$
u_8	5.5	2.4	3.8	1.1	$Iris-versicolor$
u_9	6.6	3.0	4.4	1.4	$Iris-versicolor$
u_{10}	5.5	2.6	4.4	1.2	$Iris-versicolor$
u_{11}	6.8	3.2	5.9	2.3	$Iris-virginica$
u_{12}	6.9	3.1	5.1	2.3	$Iris-virginica$
u_{13}	6.5	3.0	5.2	2.0	$Iris-virginica$
u_{14}	6.7	3.0	5.2	2.3	$Iris-virginica$
u_{15}	7.7	2.8	6.7	2.0	$Iris-virginica$

$$and\ d(u_i)=d(u_j)\}$$

$$IND(u_i,u_j)=\{a\in A; a(u_i)=a(u_j)\}$$

U_{trn} is the universe of training objects,

and $|X|$ is the cardinality of set

The sample concept-dependent granules with a 0.25 radius, derived from decision systems from Table II look as follows,

$$g^{cd}_{0.25}(u_1)=\{u_1,u_2,u_4,u_5,\}$$
$$g^{cd}_{0.25}(u_2)=\{u_1,u_2,\}$$
$$g^{cd}_{0.25}(u_3)=\{u_3,\}$$
$$g^{cd}_{0.25}(u_4)=\{u_1,u_4,u_5,\}$$
$$g^{cd}_{0.25}(u_5)=\{u_1,u_4,u_5,\}$$
$$g^{cd}_{0.25}(u_6)=\{u_6,\}$$
$$g^{cd}_{0.25}(u_7)=\{u_7,\}$$
$$g^{cd}_{0.25}(u_8)=\{u_8,u_{10},\}$$
$$g^{cd}_{0.25}(u_9)=\{u_9,u_{10},\}$$
$$g^{cd}_{0.25}(u_{10})=\{u_8,u_9,u_{10},\}$$
$$g^{cd}_{0.25}(u_{11})=\{u_{11},u_{12},u_{14},\}$$
$$g^{cd}_{0.25}(u_{12})=\{u_{11},u_{12},u_{14},\}$$
$$g^{cd}_{0.25}(u_{13})=\{u_{13},u_{14},u_{15},\}$$

TABLE III
PART1 - TRIANGULAR INDISCERNIBILITY MATRIX FOR CONCEPT-DEPENDENT GRANULE GENERATION ($i<j$), DERIVED FROM TABLE II
$c_{ij}=1,\ if\ \frac{|IND(u_i,u_j)|}{|A|}\geq 0.25\ and\ d(u_i)=d(u_j),\ 0,\ otherwise.$

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
u_1	1	1	0	1	1	0	0	0
u_2		1	0	0	0	0	0	0
u_3			1	0	0	0	0	0
u_4				1	1	0	0	0
u_5					1	0	0	0
u_6						1	0	0
u_7							1	0
u_8								1

TABLE IV
PART2 - TRIANGULAR INDISCERNIBILITY MATRIX FOR CONCEPT-DEPENDENT GRANULE GENERATION ($i<j$), DERIVED FROM TABLE II
$c_{ij}=1,\ if\ \frac{|IND(u_i,u_j)|}{|A|}\geq 0.25\ and\ d(u_i)=d(u_j),\ 0,\ otherwise.$

	u_9	u_{10}	u_{11}	u_{12}	u_{13}	u_{14}	u_{15}
u_1	0	0	0	0	0	0	0
u_2	0	0	0	0	0	0	0
u_3	0	0	0	0	0	0	0
u_4	0	0	0	0	0	0	0
u_5	0	0	0	0	0	0	0
u_6	0	0	0	0	0	0	0
u_7	0	0	0	0	0	0	0
u_8	0	1	0	0	0	0	0
u_9	1	1	0	0	0	0	0
u_{10}		1	0	0	0	0	0
u_{11}			1	1	0	1	0
u_{12}				1	0	1	0
u_{13}					1	1	1
u_{14}						1	0
u_{15}							1

$$g^{cd}_{0.25}(u_{14})=\{u_{11},u_{12},u_{13},u_{14},\}$$
$$g^{cd}_{0.25}(u_{15})=\{u_{13},u_{15},\}$$

Random coverage of training systems is as follows,

$$Cover(U_{trn})=\{g^{cd}_{0.25}(u_2),g^{cd}_{0.25}(u_3),g^{cd}_{0.25}(u_4),g^{cd}_{0.25}(u_6),g^{cd}_{0.25}(u_7),g^{cd}_{0.25}(u_{10}),g^{cd}_{0.25}(u_{11}),g^{cd}_{0.25}(u_{15}),\}$$

TABLE V
CONCEPT-DEPENDENT GRANULAR REFLECTION OF THE EXEMPLARY TRAINING SYSTEM FROM TABLE II, IN RADIUS 0.25, 5 ATTRIBUTES, 8 OBJECTS; MV IS MAJORITY VOTING PROCEDURE (THE MOST FREQUENT DESCRIPTORS CREATE A GRANULAR REFLECTION)

Day	$a1$	$a2$	$a3$	$a4$	$class$
$MV(g^{cd}_{0.25}(u_2))$	4.6	3.1	1.5	0.2	$Iris-setosa$
$MV(g^{cd}_{0.25}(u_3))$	5.1	3.3	1.7	0.5	$Iris-setosa$
$MV(g^{cd}_{0.25}(u_4))$	4.6	3.1	1.5	0.2	$Iris-setosa$
$MV(g^{cd}_{0.25}(u_6))$	6.0	3.4	4.5	1.6	$Iris-versicolor$
$MV(g^{cd}_{0.25}(u_7))$	5.9	3.2	4.8	1.8	$Iris-versicolor$
$MV(g^{cd}_{0.25}(u_{10}))$	5.5	2.4	4.4	1.1	$Iris-versicolor$
$MV(g^{cd}_{0.25}(u_{11}))$	6.8	3.2	5.9	2.3	$Iris-virginica$
$MV(g^{cd}_{0.25}(u_{15}))$	6.5	3.0	5.2	2.0	$Iris-virginica$

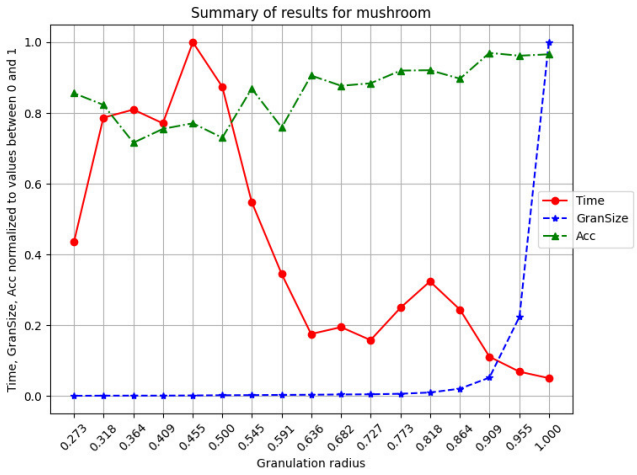

Fig. 1. Mushroom dataset - summary of results

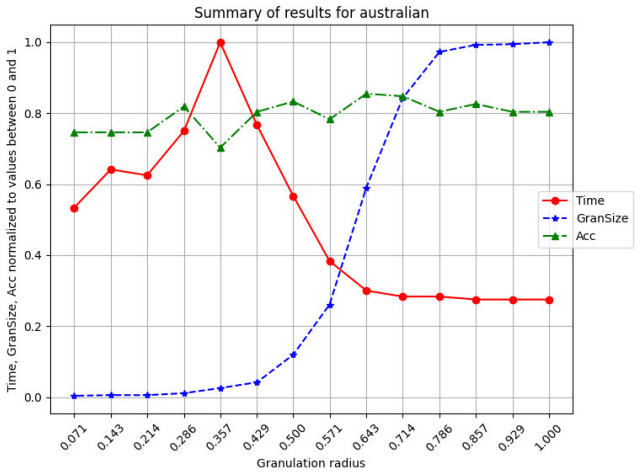

Fig. 2. Australian dataset - summary of results

The granulation process can be supported by using the indiscernibility matrix - see Tables III i IV. A granular reflection of the training system can be seen in Tab V.

B. Reference classifiers

1) Description of k-nearest neighbors algorithm: We use the kNN method from the Scikit-learn package as a reference classifier.

C. Random sampling

We use random sampling in our work on the basis of selecting a fixed percentage of objects - draw with return. In the results showing the use of this method on the x-axis we have the number of objects drawn. The parameters we present in our results are appropriately projected onto the interval [0,1]. We used standard normalization. The implementation was done in python language using standard libraries.

III. Experimental session

In the experimental part, we use three decision systems from the UCI repository [15], including Mushroom, Australian Credit and Heart Disease. In the kNN classifier we use k=1. For Mushroom and Heart we use the Euclidean metric, for Australian Hamming metric. For the initiation experiment, the data are split in a ratio of 0.8 to 0.2 and a cross-classification is performed for the granular systems created for the entire spectrum of granulation radii.

A. Reference results for concept- dependent granulation.

In the following, we will present a reference result for the granulation of the training system and the test classification - where the data are split in a ratio of 0.8 to 0.2. We take these results as a starting point for the analysis of the other results. Let us interpret the results the experiments, which are available in Figures 1, 2 and III-A.

When considering the approximation speed of decision systems, initial radii in the <0,0.5> range require training systems to be covered by a large number of granules which makes granulation slow. Once the threshold of 0.5 is exceeded, the granulation is already less time-consuming and the running time decreases. It is quite easy to find with this result areas where the radius of granulation is optimal, giving the result accuracy classification at a high level with a large reduction in the size of training systems. The optimal radius is a parameter that allows to achieve high classification accuracy (close to efficiency on full data). Our earlier discovery, the determination of optimal granulation radii by applying the layered granulation method, can also be used for this purpose - see [5]. When looking at the accuracy curve, we can see that the level of classification accuracy increases as the radii increase, this is due to the increase in the confidence of determining the classification parameters. At the same time the size of the granular decision systems increases, in the region of radius one, where we use the whole training system the classification level sometimes decreases because the noise existing in the data can be used for classification. We have shown previously that the granularity process for certain radii reduce the noise in the data - which increases the quality of the classification [5]. The last curve shows the percentage of granular systems in relation to the original training systems. It helps to determine in which area the granulation process should be completed. These overall granulation results are our starting point for research into the use of random sampling in tuning our granulation method. We show the results in the next section.

B. Concept-dependent granulation with random sampling.

The results, which are shown in Figures 4 to 21, demonstrate the interesting dependence of the granulation process on random sampling. By drawing a fixed percentage of objects from the original training set, we use a return draw. This causes that

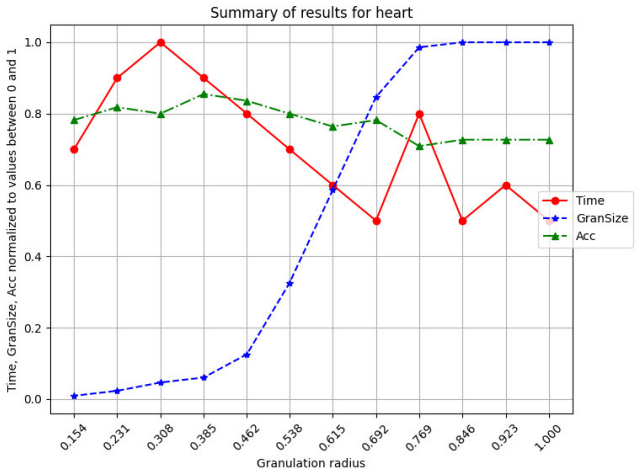

Fig. 3. Heart disease dataset - summary of results

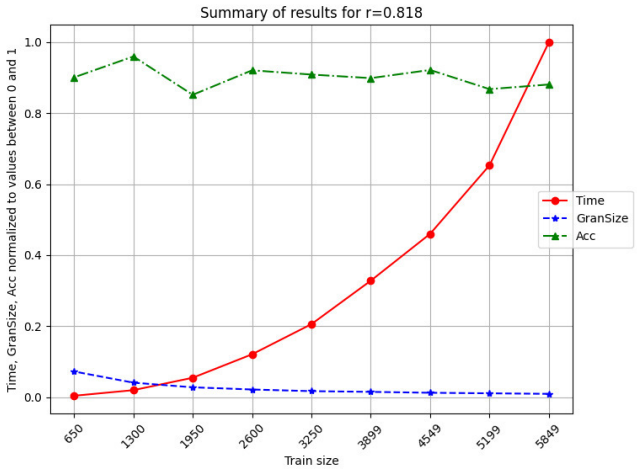

Fig. 4. Mushroom dataset - the result of random sampling for r=0.818; concept-dependent granulation

in the granulation process some objects are absorbed. Hence, the percentage of granular systems in relation to original training systems starts to decrease with increasing random sample size.

First of all, for the individual radii, the decision-making system, regardless of the starting size of the random sample, has a similar final size. This can be observed by looking at the Gransize curve, where with increasing random sample the ratio of granular to pre-granular systems starts to decrease. At the same time, the classification accuracy shows a fairly high stability starting from radii in the region of 0.5. Which gives the conclusion that the use of random sampling significantly reduces the time required for the granulation process and allows the use of a strongly reduced random training sample. The level of reduction is individual to the specific data. In the decision systems studied, we observed that the classification accuracy is at a stable level with a reduction in the running time of the approximation of up to 80 percent over the full data.

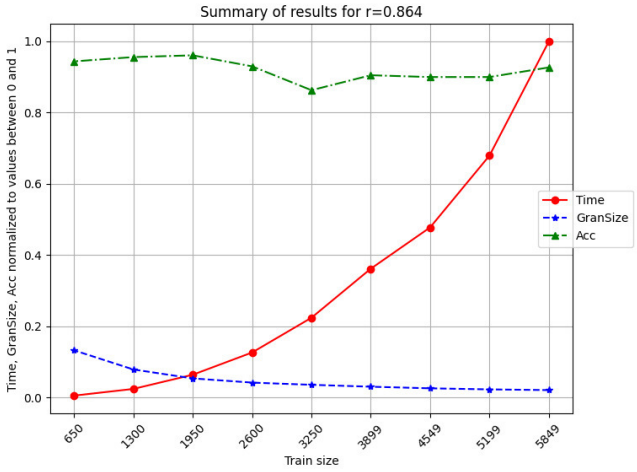

Fig. 5. Mushroom dataset - the result of random sampling for r=0.864; concept-dependent granulation

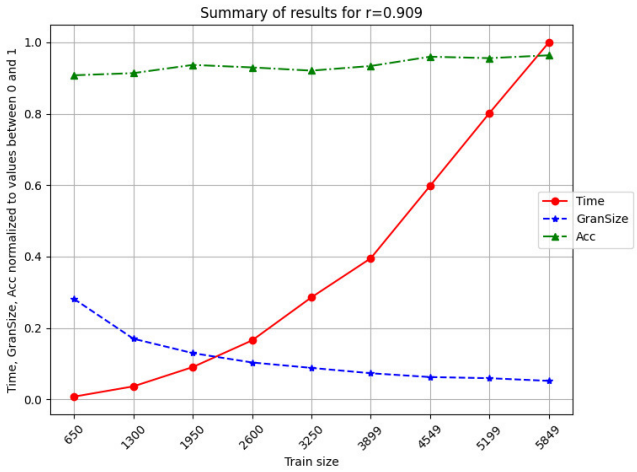

Fig. 6. Mushroom dataset - the result of random sampling for r=0.909; concept-dependent granulation

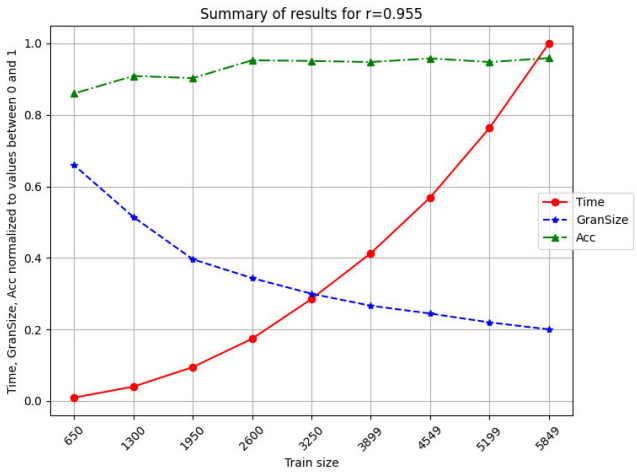

Fig. 7. Mushroom dataset - the result of random sampling for r=0.955; concept-dependent granulation

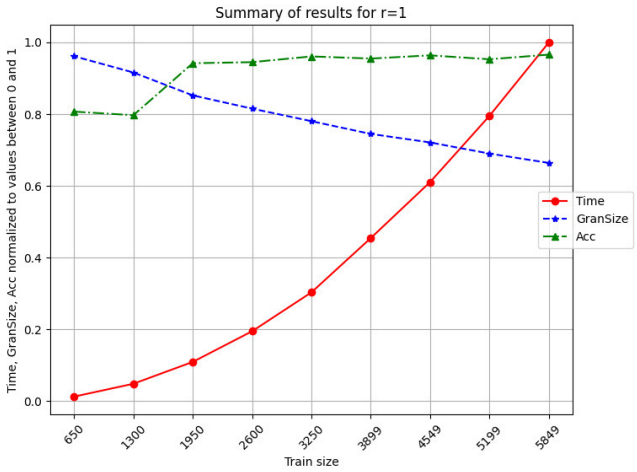

Fig. 8. Mushroom dataset - the result of random sampling for r=1.0; concept-dependent granulation

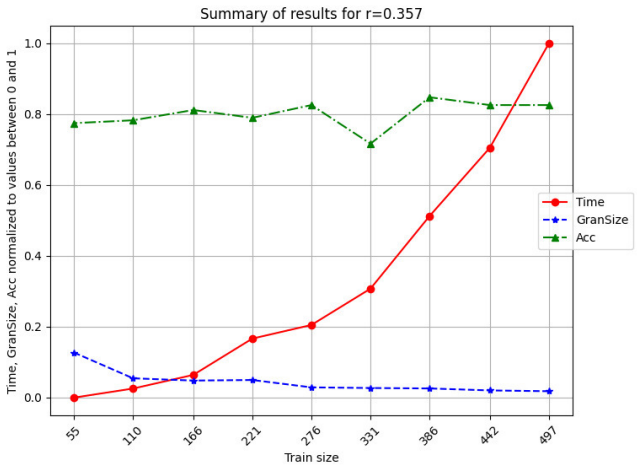

Fig. 9. Australian dataset - the result of random sampling for r=0.357; concept-dependent granulation

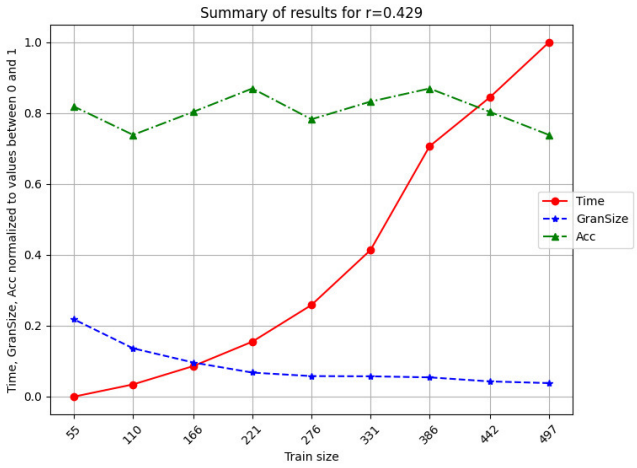

Fig. 10. Australian dataset - the result of random sampling for r=0.429; concept-dependent granulation

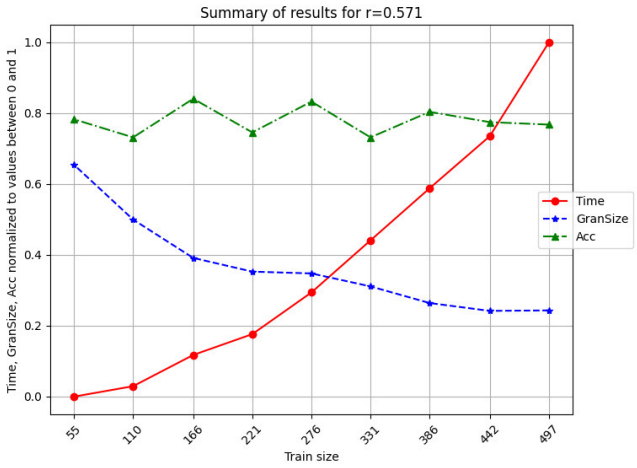

Fig. 12. Australian dataset - the result of random sampling for r=0.571; concept-depnedent granulation

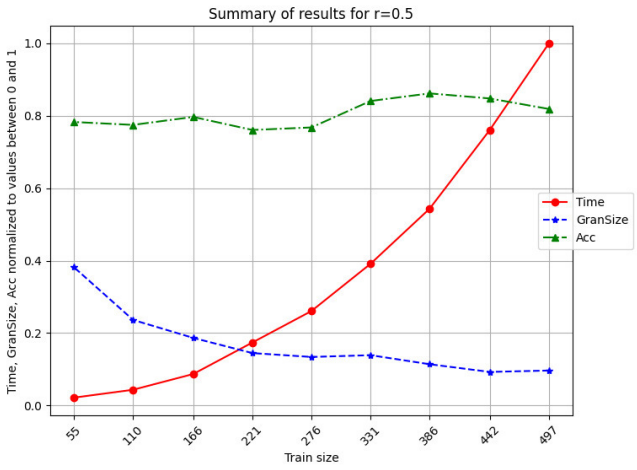

Fig. 11. Australian dataset - the result of random sampling for r=0.5; concept-dependent granulation

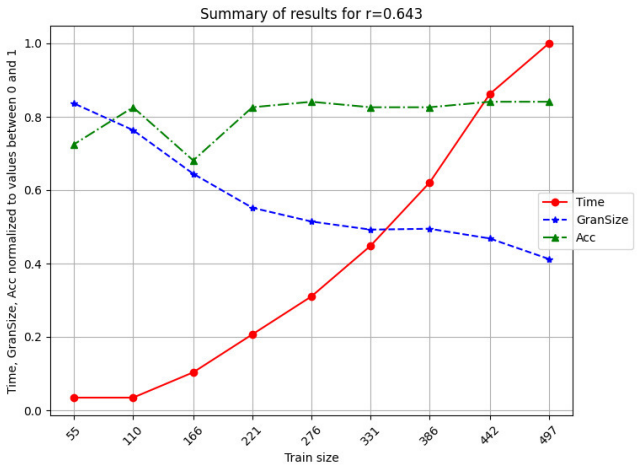

Fig. 13. Australian dataset - the result of random sampling for r=.0643; concept-dependent granulation

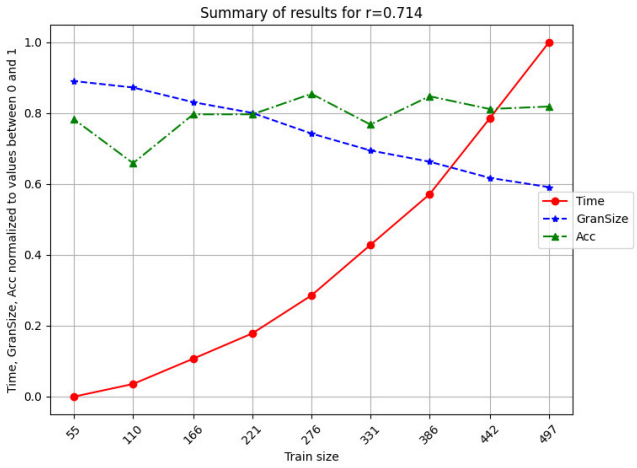

Fig. 14. Australian dataset - the result of random sampling for r=0.714; concept-dependent granulation

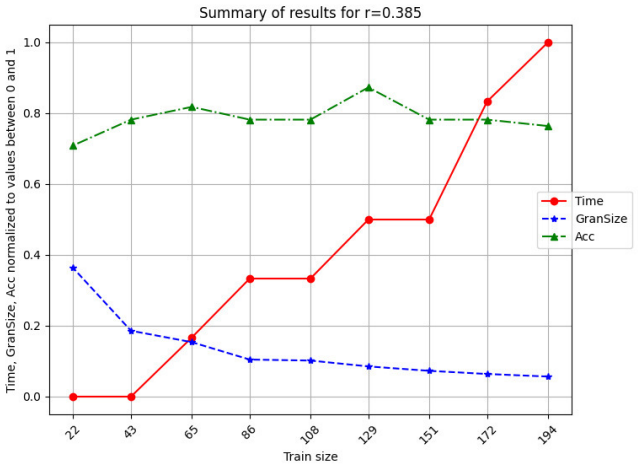

Fig. 16. Heart disease dataset - the result of random sampling for r=0.385; concept-dependent granulation

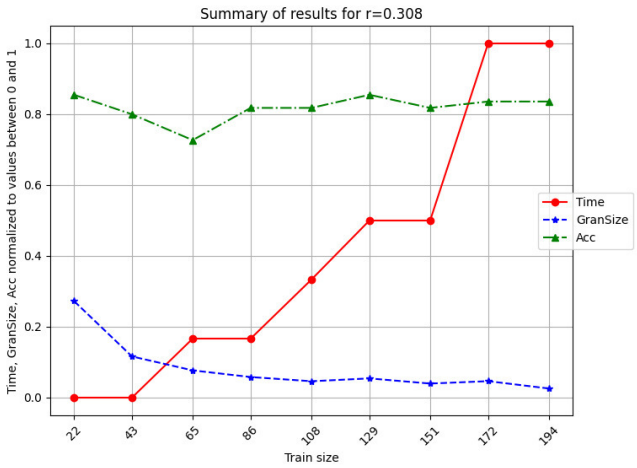

Fig. 15. Heart disease dataset - the result of random sampling for r=0.308; concept-dependent granulation

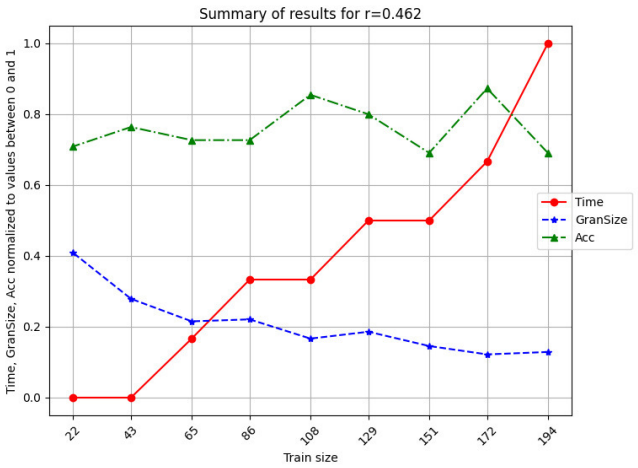

Fig. 17. Heart disease dataset - the result of random sampling for r=0.462; concept-dependent granulation

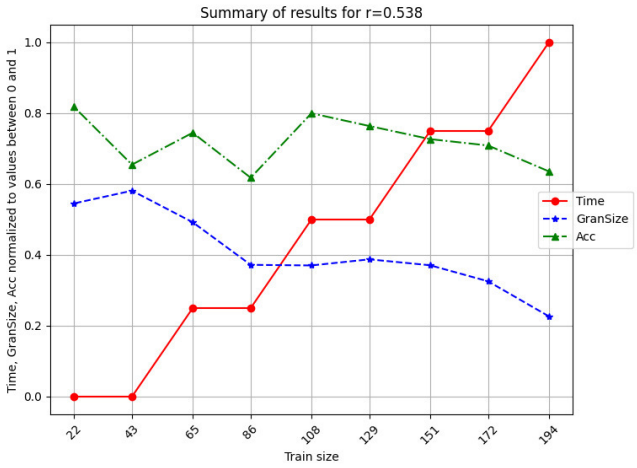

Fig. 18. Heart disease dataset - the result of random sampling for r=0.538; concept-dependent granulation

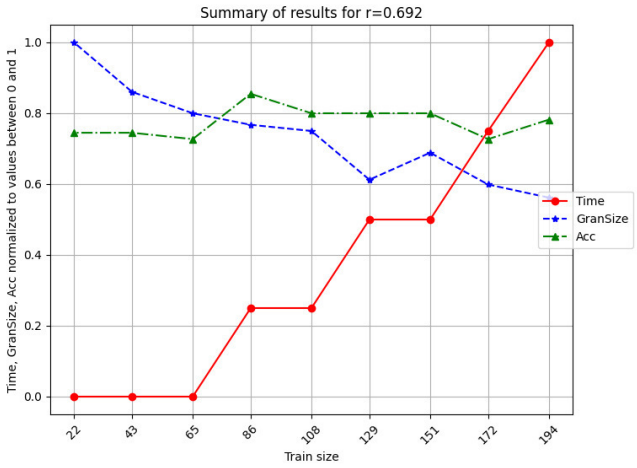

Fig. 20. Heart disease dataset - the result of random sampling for r=0.692; concept-dependent granulation

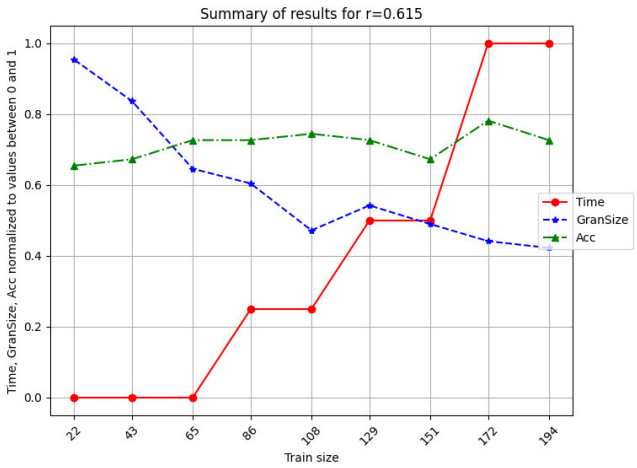

Fig. 19. Heart disease dataset - the result of random sampling for r=0.615; concept-dependent granulation

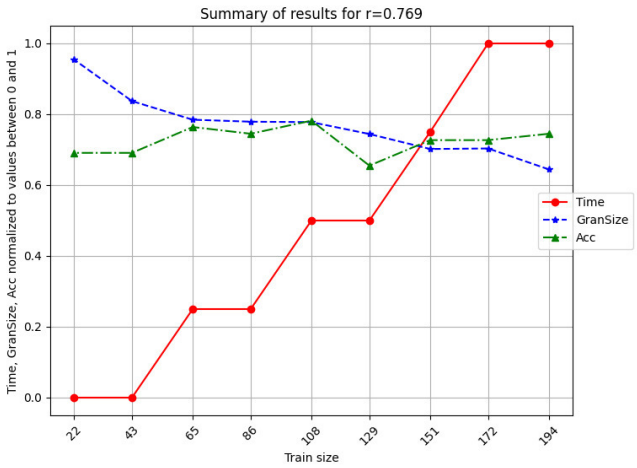

Fig. 21. Heart disease dataset - the result of random sampling for r=0.769; concept-dependent granulation

IV. Conclusion

In the current work, we made an interesting discovery - that random sampling works very well with the concept-dependent granulation method while maintaining the classification quality in reduced systems. For the decision systems studied, stable results, comparable to the performance of the original training systems - in terms of classification accuracy - are obtained with random sampling allowing us to reduce the running time of our method to as much as 20 percent of the original time (time is reduced by 80 percent). The current result was confirmed on three selected systems from the UCI repository, and represents an initial pilot study that opens new research horizons - using granulation methods based on approximate inclusions in the context of Big Data. An interesting observation is that the final granular systems for specific granulation radii (up to 0.5) have a similar size for individual random samples.

The subject of further research will be to look for a way to discover threshold values of random samples that give meaningful granulation results. In addition, we plan to explore in the context of granulation the whole range of possible techniques used for dealing with Big Data sets.

Acknowledgment

This work has been supported by the grant from Ministry of Science and Higher Education of the Republic of Poland under the project number 23.610.007-000

References

[1] Toy decision system generator, http://toyds.herokuapp.com/generator/v1/. Last accessed 12 Apr 2022

[2] Artiemjew, P.: (2022). Rough Inclusion Based Toy Decision Systems Generator For Presenting Data Mining Algorithms. Proceedings of the 3rd Polish Conference on Artificial Intelligence, April 25-27, 2022, Gdynia, Poland, 168-171.

[3] Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, 341—356 (1982). https://doi.org/10.1007/BF01001956

[4] Polkowski, L.: A model of granular computing with applications, In: Proceedings of IEEE 2006 Conference on Granular Computing GrC06, pp. 9–16. IEEE Press, Atlanta, USA (2006)

[5] Polkowski, L., Artiemjew, P.: Granular Computing in Decision Approximation - An Application of Rough Mereology, In: Intelligent Systems Reference Library 77, Springer, ISBN 978-3-319-12879-5, pp. 1–422 (2015).

[6] Artiemjew, P.: Classifiers from Granulated Data Sets: Concept Dependent and Layered Granulation, In: Proceedings RSKD'07. The Workshops at ECML/PKDD'07, pp. 1–9., Warsaw Univ. Press, Warsaw (2007)

[7] Artiemjew, P., Ropiak, K.: 'A Novel Ensemble Model - The Random Granular Reflections', Fundamenta Informaticae, 1 Jan. 2021, vol. 179, no. 2, pp. 183-203, 2021(DOI: 10.3233/FI-2021-2020)

[8] J. Szypulski, P. Artiemjew: The Rough Granular Approach to Classifier Synthesis by Means of SVM, In: Proceedings of International Joint Conference on Rough Sets, IJCRS'15, pp. 256-263, Tianjin, China, Lecture Notes in Computer Science (LNCS), Springer, Heidelberg (2015)

[9] Ropiak, K.; Artiemjew, P. On a Hybridization of Deep Learning and Rough Set Based Granular Computing. Algorithms 2020, 13, 63.

[10] Ropiak K., Artiemjew P. (2020) Random Forests and Homogeneous Granulation. In: Lopata A., Butkiene R., Gudoniene D., Sukacke V. (eds) Information and Software Technologies. ICIST 2020. Communications in Computer and Information Science, vol 1283. Springer, Cham. https://doi.org/10.1007/978-3-030-59506-7_16 (2020)

[11] Artiemjew P., Kislak-Malinowska A. (2019) Using r-indiscernibility Relations to Hide the Presence of Information for the Least Significant Bit Steganography Technique. In: Damaševičius R., Vasiljeviene G. (eds) Information and Software Technologies. ICIST 2019. Communications in Computer and Information Science, vol 1078. Springer

[12] Artiemjew P.: Boosting effect for classifier based on simple granules of knowledge. In: Information Technology And Control (ITC) vol. 47(2), pp. 184-196 (2018)

[13] L. Polkowski, P. Artiemjew: Granular Computing: Classifiers and Missing Values, in Proceedings ICCI'07. 6th IEEE International Conference on Cognitive Informatics, IEEE Computer Society, Los Alamitos, CA, 2007, pp. 186-194.

[14] Artiemjew, P., Ropiak, K.: Robot localization in the magnetic unstable environment, 5th Workshop on Collaboration of Humans, Agents, Robots, Machines and Sensors (CHARMS 2019), The Third IEEE International Conference on Robotic Computing (IRC 2019), Naples, Italy

[15] UCI ML Repository, https://archive.ics.uci.edu/ml/index.php.