A Novel Ensemble Model - The Random Granular Reflections

Piotr Artiemjew, Krzysztof Ropiak

Faculty of Mathematics and Computer Science
University of Warmia and Mazury in Olsztyn
Poland
email:artem@matman.uwm.edu.pl, kropiak@matman.uwm.edu.pl

Abstract. One of the most significant achievements in machine learning is the development of Ensemble techniques, which gave a powerful tool for tuning classifiers. The most popular methods are Random Forests, Bagging and Boosting. In this paper we present a novel ensemble model, named Random Granular Reflections. This algorithm creates an ensemble of homogenous granular decision systems. In each iteration of learning process, the training decision system is covered by random homogenous granules and the granular reflection is created, which takes part in classification process. Seeing the initial results - our approach is promising and seems to be comparable with the selected popular models.

Keywords: Random Granular Reflections, Homogenous Granulation, CSG Classifier, Ensemble Model, Rough Sets, Decision Systems, Classification

1 Introduction

This paper is about the application of granular rough computing in new Ensemble model. The techique that we use to prepare the data for each iteration of learning process was inspired by Polkowski standard granulation - see [16]. This method was the beginning of many new algorithms with diverse applications, for instance in Artiemjew [1]-[3], Polkowski [15]–[20], Polkowski and Artiemjew [21] we have the presentation of standard granulation, concept dependent and layered granulation in the context of training data size reduction, missing values absorbtion and usage in the classification processes.

In our recent works - see [24] and [25] - we have developed a new granulation technique - homogenous granulation (see. detail decription and toy example in Sect. 2). This approximation technique is based on creation of groups of r-indiscernible objects around each training object by lowering the ratio of indiscernibility until the granules contain only homogoneus objects in the sense of their decision class. In this method - what distinguishes it from previously studied - there is no need to estimate optimal parameter of approximation. The r-indiscernibility level for each central training object is formed in automatic way and depends on the homogeneity in decision classes.

The ensemble scheme of classification is really effective in many contexts, for instance in rough set methods the exemplary succesfull applications can be found in [6–8, 26, 29]. The recently developed approximation technique - homogenous granulation - gave us motivation to check it in ensemble model creation. Additionally to Random Forests, Bagging and Boosting we propose a novel algorithm - Ensemble of Random Granular Reflections. The method is based on representation of original training system by its granular reflections formed from random homogenous granules, which covers it in each iteration of learning process. Each granular reflection of training decision system is additionally reduced in size in comparison with original training decision system. The granular reflection of each iteration represents the internal knowledge from original system using the random coverage. The level of data reduction is up to 50 per cent of original data.

In this work we have first sight into this method and for simplicity we treat all attributes as cathegorical. For experiments we performed 50 iterations of learning process with use of CSG classifier - the classifier based on simple granules of knowledge - see [4].

We have compared our new method with selected ensemble models - see Sect. 3.

The rest of the work contains the following content. In Sect. 2 we have introduction to homogenous granulation algorithm. In Sect 3 we have brief introduction to selected Ensemble models. In Sect. 4 we present our novel ensemblme model - The Random Granular Reflections technique. In Sect. 5 we show the results of the experiments, and we conclude the paper in Sect. 6.

2 Homogenous granulation

Detailed theoretical introduction to rough inclusions is available in Polkowski [15] – [20].

For given objects u, v from training decision system, r granulation radius, and A the set of attributes, the standard rough inclusion μ is defined as

$$\mu(v, u, r) \Leftrightarrow \frac{|IND(u, v)|}{|A|} \geq r \tag{1}$$

where

$$IND(u, v) = \{a \in A : a(u) = a(v)\}, \tag{2}$$

The homogenous granules are formed as follows,

$$g_{r_u}^{homogenous} = \{v \in U : |g_{r_u}^{cd}| - |g_{r_u}| == 0,\ for\ minimal\ r_u\ fulfills\ the\ equation\}$$

where

$$g_{r_u}^{cd} = \{v \in U : \frac{IND(u, v)}{|A|} \leq r_u\ AND\ d(u) == d(v)\}$$

and

$$g_{r_u} = \{v \in U : \frac{IND(u,v)}{|A|} \leq r_u\}$$

$$r_u = \{\frac{0}{|A|}, \frac{1}{|A|}, ..., \frac{|A|}{|A|}\}$$

2.1 The process of training system covering

In the process of covering - the objects from training system are covered based on chosen strategy. We use simple random choice because it is the most effective method among studied ones - see [21]).

The last step of the granulation process is shown in the next section.

2.2 Granular reflections

In this step we formed the granular reflections of the original training system based on the granules from the found coverage (the coverage is the set of granules, which cover the universe of traning objects completly). Each granule $g \in COV(U, \mu, r)$ from the coverage is finally represented by single object formed using the Majority Voting (MV) strategy (choice the most common values).

$$\{MV(\{a(u) : u \in g\}) : a \in A \cup \{d\}\} \tag{3}$$

The granular reflection of the decision system $D = (U, A, d)$ is the decision system $(COV(U, \mu, r)$, the set of objects formed from granules.

$$v \in g_r^{cd}(u) \text{ if and only if } \mu(v, u, r) \text{ and } (d(u) = d(v)) \tag{4}$$

for a given rough (weak) inclusion μ.

Toy example of described granulation method is presented in the next section.

2.3 Toy example of homogenous granulation

Considering training decision system from Tab. 1.

Homogenous granules for all training objects:

$$g_0.75(u_1) = (u_1)$$

$$g_1(u_2) = (u_2)$$

$$g_1(u_3) = (u_3)$$

$$g_1(u_4) = (u_4)$$

Table 1. Training data system (U_{trn}, A, d), (a sample from Quinlan data set [23])

	a_1	a_2	a_3	a_4	d
u_1	*sunny*	*hot*	*high*	*strong*	*no*
u_2	*rain*	*cool*	*normal*	*strong*	*no*
u_3	*overcast*	*cool*	*normal*	*strong*	*yes*
u_4	*sunny*	*mild*	*high*	*weak*	*no*
u_5	*sunny*	*cool*	*normal*	*weak*	*yes*
u_6	*rain*	*mild*	*normal*	*weak*	*yes*
u_7	*overcast*	*hot*	*high*	*weak*	*yes*
u_8	*sunny*	*mild*	*normal*	*strong*	*yes*
u_9	*overcast*	*mild*	*high*	*strong*	*yes*
u_{10}	*rain*	*mild*	*high*	*weak*	*yes*
u_{11}	*overcast*	*hot*	*normal*	*weak*	*yes*

$$g_0.75(u_5) = (u_5)$$
$$g_0.75(u_6) = (u_6, u_{10})$$
$$g_0.75(u_7) = (u_7, u_{11})$$
$$g_0.75(u_8) = (u_8)$$
$$g_0.75(u_9) = (u_9)$$
$$g_1(u_{10}) = (u_{10})$$
$$g_0.5(u_{11}) = (u_3, u_5, u_6, u_7, u_{11})$$

Granules covering training system by random choice:

$$g_0.75(u_1) = (u_1)$$
$$g_1(u_2) = (u_2)$$
$$g_1(u_4) = (u_4)$$
$$g_0.75(u_6) = (u_6, u_{10})$$
$$g_0.75(u_7) = (u_7, u_{11})$$
$$g_0.75(u_8) = (u_8)$$
$$g_0.75(u_9) = (u_9)$$
$$g_0.5(u_{11}) = (u_3, u_5, u_6, u_7, u_{11})$$

Granular decision system from above granules is as follows:

Table 2. Granular decision system formed from Covering granules

$g_0.75(u_1)$	*sunny*	*hot*	*high*	*strong*	*no*
$g_1(u_2)$	*rain*	*cool*	*normal*	*strong*	*no*
$g_1(u_4)$	*sunny*	*mild*	*high*	*weak*	*no*
$g_0.75(u_6)$	*rain*	*mild*	*normal*	*weak*	*yes*
$g_0.75(u_7)$	*overcast*	*hot*	*high*	*weak*	*yes*
$g_0.75(u_8)$	*sunny*	*mild*	*normal*	*strong*	*yes*
$g_0.75(u_9)$	*overcast*	*mild*	*high*	*strong*	*yes*
$g_0.5(u_{11})$	*overcast*	*cool*	*normal*	*weak*	*yes*

In the next section there is a brief description of the selected popular Ensemble models.

3 Selected popular ensemble models

There are many techniques in the family of Enssemble models. One of the most popular are Random Forests, Bagging and Boosting - see [31]. Short description of mentioned models is to be found below.

Bootstrap Ensembles - Pure Bagging: It is the random committee of bootstraps [33]. It is a method in which the original decision system - the basic knowledge - is split into (TRN) training data set, and ($TSTvalid$) validation test data set. And from the TRN system, for a fixed number of iterations, we form a new Training systems ($NewTRN$) by random choice with returning of $card\{TRN\}$ objects. In all iterations we classify the TRNvalid system in two ways: the first based on the actual $NewTRN$ system and the second based on the committee of all performed classifications. In the committee majority voting is performed and the ties are resolved randomly.

Bagging based on Arcing - Bagging: The main difference between this method and Bootstrap Ensembles is that the TRN is split into two data sets $NewTRN$ and $NewTST$ - see [5] and [27]. This split is based on Bootstraps where weights determine the probability with which objects are assigned to NewTRN set. Initially weights are equal, but after first classification of the NewTST using NewTRN weights are lowered for well-classified objects. The next step is normalization of weights. This algorithm which shows forming of Bootstraps is called Arcing. Classifying the $TSTvalid$ with $NewTRN$ in a single iteration as the committee of classifiers is the last step of this method. In Arcing weights are modified with the factor equal to $\frac{1-Accuracy}{Accuracy}$.

Boosting based on Ada-Boost with Monte Carlo split: Classification method used in this algorithm is similar to the previously described with the difference that the NewTRN and NewTST are formed in a different way - see [9], [28] and [34].

Objects for NewTRN are chosen based on weights and fixed ratio is used to split the TRN data set. Previous experiments show that split ratio equal to 0.6 is optimal, as it is close to the approximate size of the distinguishable objects in the bootstraps. Other parts of this algorithm works like in the previous one.

Random forests: In this model random trees are created based on randomly chosen attributes and then they take part in the classification process in each iteration. This method can be usefull in other classifiers using the random set of attributes before usage in classification process. The number of attributes, which should be chosen depending on internal data logic, have to be found in an experimental way.

In the following section we present introduction to our new Ensemble method.

4 Ensemble of Random Granular Reflections

In each iteration of our new ensemble model we have used a different homogenous granular decision system formed from random homogenous granules, which covers the original training system. The visualization of the model can be found in Fig. 1.

The time comlexity of this model is quadratic. The most time consuming part is granulation, which main component takes $((no.\_of\_obj.)^2) * (no.\_of\_att.)$ operations.

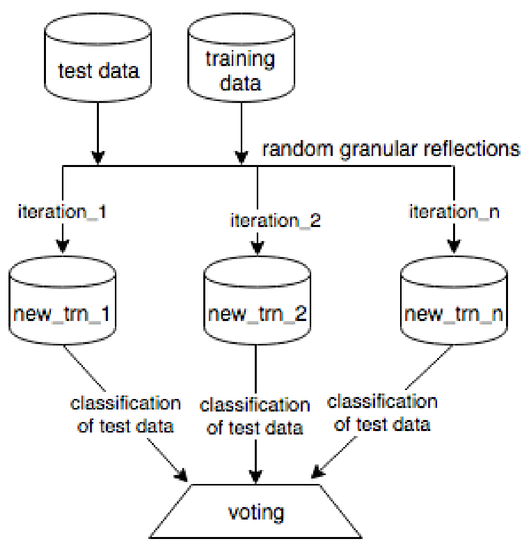

Fig. 1. Ensemble of Random Granular Reflections

5 Experimental Session

To perform initial experiments we used the australian credit data set from UCI Machine Learning Repository [30]. We have run our algorithm with 50 iterations of learning for each tested Ensemble model. As a reference point we have chosen Committee of Bootstraps (Pure Bagging) [33], Boosting based on Arcing (Bagging) [5], [27], and Ada-Boost with Monte Carlo split [9], [28] and [34] - for details see Sect. 3. As a reference classifier we used CSG classifier [4] with radius 0.5. The effectiveness is evaluated by percentage of properly classified objects - the accuracy.

The first result of Random Granular Reflections technique for chosen data set is presented in Fig. 2. The results of the other popular ensemble models are to be found in Figs. 3, 4 and 5. For selected data set our new technique outperformed the other checked methods.

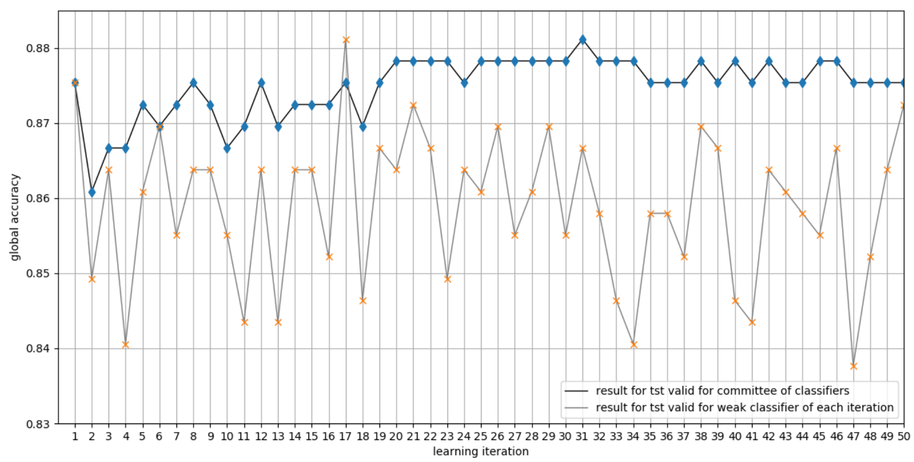

Fig. 2. Ensemble of Random Granular Reflections for australian credit dataset - the accuracy of classification - 5 times 50 iterations of learning

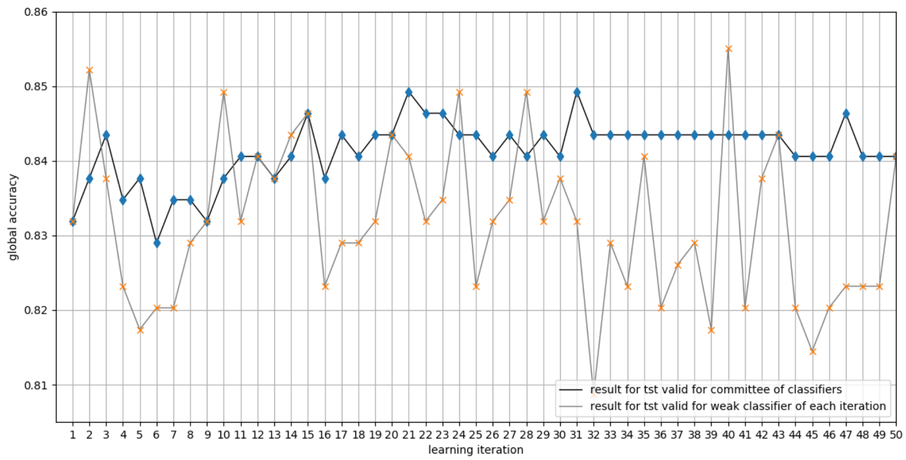

Fig. 3. Bagging ensemble model for australian credit dataset - the accuracy of classification - 5 times 50 iterations of learning

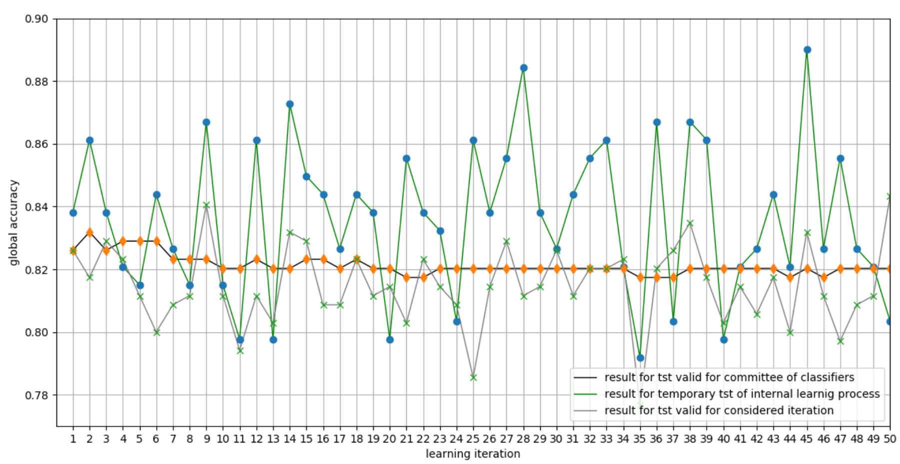

Fig. 4. AdaBoost ensemble model for australian credit dataset - the accuracy of classification - 5 times 50 iterations of learning

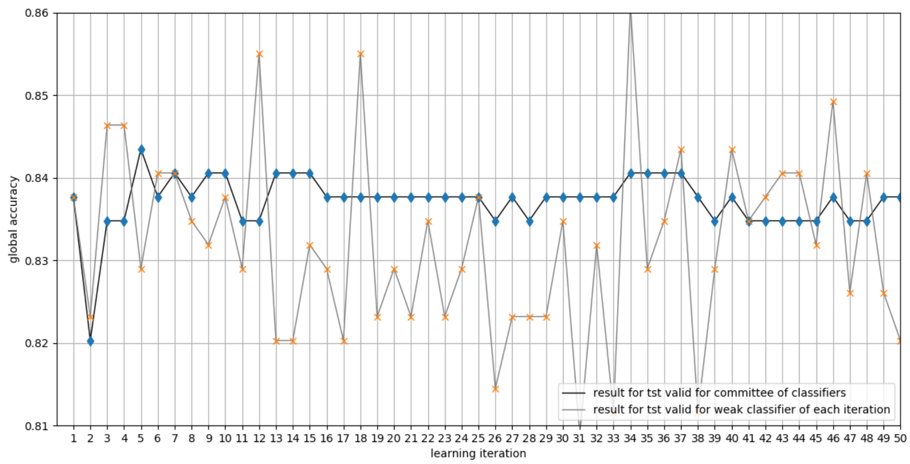

Fig. 5. Pure Bagging ensemble model for australian credit dataset - the accuracy of classification - 5 times 50 iterations of learning

6 Conclusions

The results of the experiments show the effectivenes of our new technique. The Ensemble of Random Granular Reflections turn out to be competitive with other techniqes like Bagging and Boosting. Despite promising initial results, much is left to be done to evaluate the effectiveness and set of applications of this new method.

In the future works we have a plan to extensively check the effectiveness of new model and we are planning to apply the other types of granules in the proposed ensemble model.

7 Acknowledgements

The research has been supported by grant 23.610.007-300 from Ministry of Science and Higher Education of the Republic of Poland.

References

1. Artiemjew, P. : Classifiers from Granulated Data Sets: Concept Dependent and Layered Granulation. In Proceedings RSKD'07. The Workshops at ECML/PKDD'07, Warsaw Univ. Press, Warsaw, 2007, pp 1–9 (2007)
2. Artiemjew, P.: Rough mereological classifiers obtained from weak rough set inclusions. In Proceedings of Int. Conference on Rough Set and Knowledge Technology RSKT'08, Chengdu China, Lecture Notes in Artificial Intelligence, vol. 5009. Springer Verlag, Berlin, 2008, pp 229–236 (2008)

3. Artiemjew, P. (2013): A Review of the Knowledge Granulation Methods: Discrete vs. Continuous Algorithms. In Skowron A., Suraj Z. (eds.)(2013): Rough Sets and Intelligent Systems. ISRL 43, Springer-Verlag, Berlin, 2013, pp 41–59.
4. Artiemjew, P.: Boosting Effect of Classifier Based on Simple Granules of Knowledge, In: Information technolojy and control, Print ISSN: 1392-124X, Vol 47, No 2 (2018)
5. Breiman, L.: Arcing classifier (with discussion and a rejoinder by the author). Ann. Statist. 26 (3): 801849. Retrieved 18 January 2015. Schapire (1990) proved that boosting is possible. (Page 823) (1998)
6. Hu, X., Construction of An Ensemble of Classifiers based on Rough Sets Theory and Database Operations, Proc. of the IEEE International Conference on Data Mining (ICDM2001), (2001)
7. Hu, X.: Ensembles of classifiers based on rough sets theory and set-oriented database operations, Presented at the 2006 IEEE International Conference on Granular Computing, Atlanta, GA (2006)
8. Murthy, C., Saha, S., Pal, S.K.: Rough Set Based Ensemble Classifier, In: Rough Sets, Fuzzy Sets, Data Mining and Granular Computing Lecture Notes in Computer Science Volume 6743, p. 27 (2001)
9. Ohno-Machado, L.: Cross-validation and Bootstrap Ensembles, Bagging, Boosting, Harvard-MIT Division of Health Sciences and Technology, http://ocw.mit.edu/courses/health-sciences-and-technology/hst-951j-medical-decision-support-fall-2005/lecture-notes/hst951_6.pdf HST.951J: Medical Decision Support, Fall (2005)
10. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, pp 341–356 (1982)
11. Polkowski, L.: Rough Sets. Mathematical Foundations. Physica Verlag, Heidelberg (2002)
12. Polkowski, L.: A rough set paradigm for unifying rough set theory and fuzzy set theory. Fundamenta Informaticae 54, pp 67–88; and : In Proceedings RSFDGrC03, Chongqing, China, 2003. Lecture Notes in Artificial Intelligencevol. 2639, Springer Verlag, Berlin, pp 70–78 (2003)
13. Polkowski, L.: Toward rough set foundations. Mereological approach. In Proceedings RSCTC04, Uppsala, Sweden. Lecture Notes in Artificial Intelligence vol. 3066, Springer Verlag, Berlin, pp 8–25 (2004)
14. Polkowski, L.: Granulation of knowledge in decision systems: The approach based on rough inclusions. The method and its applications In Proceedings RSEISP'07, Lecture Notes in Artificial Intelligence vol. 4585. Springer Verlag, Berlin, pp 69– (2004)
15. Polkowski, L.: Formal granular calculi based on rough inclusions. In Proceedings of IEEE 2005 Conference on Granular Computing GrC05, Beijing, China. IEEE Press, pp 57–62 (2005)
16. Polkowski, L.: A model of granular computing with applications. In Proceedings of IEEE 2006 Conference on Granular Computing GrC06, Atlanta, USA. IEEE Press, pp 9–16 (2006)
17. Polkowski, L.: The paradigm of granular rough computing. In Proceedings ICCI'07, Lake Tahoe NV. IEEE Computer Society, Los Alamitos CA, pp 145–163 (2007)
18. Polkowski, L.: A Unified Approach to Granulation of Knowledge and Granular Computing Based on Rough Mereology: A Survey, in: Handbook of Granular Computing, Witold Pedrycz, Andrzej Skowron, Vladik Kreinovich (Eds.), John Wiley & Sons, New York, 375-401 (2008)

19. Polkowski, L.: Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems. In Meyers, R. A.(ed.): Encyclopedia of Complexity and System Sciences. Springer Verlag, Berlin, article 00788 (2009)
20. Polkowski, L.: Approximate Reasoning by Parts. An Introduction to Rough Mereology. Springer Verlag, Berlin, (2011)
21. 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, 1-422 (2015)
22. Poap, D., Woniak, M., Wei, W., Damaeviius, R.: Multi-threaded learning control mechanism for neural networks. Future Generation Computer Systems, Elsevier 2018.
23. Quinlan, J., R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, Kluwer Academic Publishers (1993)
24. Ropiak, K., Artiemjew, P.: On Granular Rough Computing: epsilon homogenous granulation, In: Proceedings of International Joint Conference on Rough Sets, IJCRS'18, Quy Nhon, Vietnam, Lecture Notes in Computer Science (LNCS), vol. 11103, Springer, Heidelberg, pp 546–558 (2018)
25. Ropiak, K., Artiemjew, P.: A Study in Granular Computing: homogenous granulation. In: Dregvaite G., Damasevicius R. (eds) Information and Software Technologies. ICIST 2018. Communications in Computer and Information Science, Springer (2018) in print
26. Saha, S., Murthy, C.A., Pal, S.K.: Rough set based ensemble classifier for web page classification. Fundamenta Informaticae 76(1-2), 171187 (2007)
27. Schapire, R.E.: A Short Introduction to Boosting (1999)
28. Schapire, R.E.: The Boosting Approach to Machine Learning: An Overview, MSRI (Mathematical Sciences Research Institute) Workshop on Nonlinear Estimation and Classification (2003)
29. Shi, L., Weng, M., Ma, X., Xi, L.: Rough Set Based Decision Tree Ensemble Algorithm for Text Classification, In: Journal of Computational Information Systems6:1, 89-95 (2010)
30. University of California, Irvine Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php
31. Yang, P., Yang, Y., H., Zhou, B., B.; Zomaya, A., Y.: A review of ensemble methods in bioinformatics: Including stability of feature selection and ensemble feature selection methods. In Current Bioinformatics, 5, (4):296-308, 2010 (updated on 28 Sep. 2016)
32. Zadeh, L. A.: Fuzzy sets and information granularity. In Gupta, M., Ragade, R., Yager, R.R. (eds.): Advances in Fuzzy Set Theory and Applications. North–Holland, Amsterdam, 1979, pp 3–18 (1979)
33. Zhou, Z.-H.: Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC. p. 23. ISBN 978-1439830031. The term boosting refers to a family of algorithms that are able to convert weak learners to strong learners (2012)
34. Zhou, Z.-H.: Boosting 25 years, CCL 2014 Keynote (2014)