Article | Received 19 February 2022; Accepted 22 April 2023; Published 10 May 2023
https://doi.org/10.55092/aias20230002

Autonomous and sustainable machine learning: pursuing new horizons of intelligent systems

Witold Pedrycz

Department of Electrical & Computer Engineering, University of Alberta, Edmonton, Canada

E-mail: wpedrycz@ualberta.ca

Abstract: The paradigm of Artificial Intelligence and Machine Learning has resulted in an amazingly diverse plethora of models operating in various environments and quite often exhibiting numerous successes. There is a growing spectrum of challenging application areas of high criticality where one has to meet a number of fundamental requirements. Those manifest evidently when Machine Learning constructs have to function autonomously and any decisions being rendered entail far reaching implications. The carefully crafted learning process has to result with advanced models. Along with the developed models, they have to come hand-in-hand with credibility measures that are crucial to assess an extent to which the results generated by such measures are meaningful, trustworthy and credible.

The credibility of the Machine Learning models becomes of paramount importance given the nature of application domains. Autonomous systems including autonomous vehicles, user identification (both using audio and video channels), financial systems (calling for sound mechanisms to quantify risk levels) require the ML system making classification or prediction decisions some level of self-awareness. Among others, this translates to forming sound answers to the following crucial questions emerging within the design process:

How much confidence could be associated with the result?

Could any action /decision be taken on a basis of obtained result?

Given the reported level of credibility, is there any other experimental evidence one could acquire to validate the decision?

In this study, we advocate that a general way to achieve such goals is to engage the mechanism of Granular Computing; subsequently, the granularity endowing the results are sought as a vehicle use to quantify the credibility level. Sustainable (or green) Machine Learning gives rise to the agenda of knowledge reuse, namely exploring possibilities of potential reuse of the already designed models in a spectrum of current environments where

computing overhead as one of the ways to contribute to the agenda of sustainable Machine Learning and discuss a crucial role of information granularity in this context.

Keywords: sustainable machine learning; granular computing; awareness; knowledge transfer; credibility

1. Introduction

In the recent years, we have been witnessing a rapid progress of Machine Learning (ML), bringing a wealth of conceptual developments, impressive learning algorithms and far-reaching applications. In a long run, however, there are some apparent roadblocks that very likely might negatively impact future developments, especially at the application side. Some of them are listed below:

-Enormous computing overhead,
-Limited interpretability and explainability [1],
-Privacy and security issues
-Brittleness of ML solutions
-Credibility of solutions provided by ML models
-Stability of models

All of those challenges are related to each other to some extent and in numerous situations the design criteria are of conflicting character. As an example, one can point at a need to strike a sound trade-off between accuracy and interpretability where such requirements are linked with the aspects of brittleness and privacy. The direction of green AI has started to play a highly visible role [2,3–5].

While the above list is long, there are two items on this agenda that deserve particular attention, namely the credibility measures of ML constructs and ways to curb computing overheads. The credibility of ML models and confidence quantifying their results are also of paramount concern to any critical application especially in situations when dealing with autonomous systems [6,7–9] operating in critical environments where ideally one could anticipate that the constructed system should exhibit some degree of self- awareness. If the credibility of result is low, instead of realising the decision/action, one may anticipate that the system should elicit additional knowledge before making a decision actionable.

The energy consumption associated with the design of complex ML architectures is another highly visible challenge, which is not sustainable in a longer perspective.

In this study, we focus on the two items from the above list that are of paramount relevance. The credibility (confidence) of results produced by ML constructs is inherently expressed in the form of information granules. Several development scenarios are carefully revisited including those involving constructs in statistics (confidence and prediction intervals), and granular parameters (fuzzy sets and interval techniques). We augment the commonly encountered and challenging paradigm of Federated Learning where the aspect of quality of the model and its results calls for a thorough assessment and quantification.

The study is structured into four sections. To make the presentation self-contained, we briefly recall the essentials of GrC. In Section 2, we elaborate on the credibility of models. Section 3 is focused on the ideas of transfer learning. Conclusions are offered in Section 4.

2. Information granularity and the discipline of granular computing

The terms information granules and information granularity themselves have emerged in different contexts and numerous areas of application. Granular Computing is quite often associated with the pioneering studies by Zadeh [10]. He coined an informal, yet highly descriptive and compelling concept of information granules. In general, by information granules one regards a collection of elements drawn together by their closeness (resemblance, proximity, functionality, etc.) articulated in terms of some useful spatial, temporal, or functional dependencies. Subsequently, Granular Computing (GrC) is about representing, constructing, processing, and communicating information granules [11,12]. As a matter of fact, GrC is about realizing mechanisms of abstraction; the required level of abstraction is helpful in coping with complexities of real-world phenomena.

The framework of Granular Computing along with a diversity of its formal settings offers a critically needed conceptual and algorithmic environment. A suitable perspective built with the aid of information granules is advantageous in realizing a suitable level of abstraction. It also becomes instrumental when forming sound and pragmatic problem-oriented trade-off among precision of results, their easiness of interpretation, value, and stability (where all of these aspects contribute vividly to the general notion of actionability).

There are numerous well-established formal frameworks of information granules; the commonly encountered include:

-Sets (intervals) [6,13,14]
-Fuzzy sets [15][16]
-Shadowed sets [17]
-Rough sets [18,19]
-Random sets
-Probabilities
-Hesitant sets
...

There is an important direction of generalizations of information granules, namely information granules of higher type. The essence of information granules of higher type comes with a fact that the characterization (description) of information granules is described in terms of information granules rather than numeric entities. Well-known examples are fuzzy sets of type-2, granular intervals, or imprecise probabilities. For instance, a type-2 fuzzy set [20] is a fuzzy set whose grades of membership are not single numeric values (membership grades in [0,1]) but fuzzy sets, intervals or probability density functions truncated to the unit interval. There is a hierarchy of higher type information granules, which are defined in a recursive manner. Therefore, we talk about type-0, type-1, type-2 fuzzy sets,

etc. In this hierarchy, type-0 information granules are numeric entities, say, numeric measurements. This idea is explored in the construction of granular models.

3. Credibility of models

There are two main challenges when it comes to the construction and an efficient deployment of ML architectures. They have to be carefully addressed:

3.1. Development of ML models by optimizing some loss function

There are a variety of learning schemes aimed at the minimization of the loss function. Typically, structural and parametric optimization tasks are envisioned. Structural optimization in which a number of hyperparameters are optimized focuses in the realm of population-based optimization or a prudent search strategy over a relatively limited search space. The parametric optimization involves some gradient-based optimization.

3.2. Quantification of credibility of the model and its results

This phase, although crucial to any applications addressing the need to express how much confidence could be associated with the constructed ML model, is less visible in comparison to the first one. Yet, the credibility of the ML model and its ensuing parameters is highly relevant implying the usefulness of the developed model and the credibility in the results. The issue of awareness about the quality of the model becomes more central and will play even more visible role given the scope of existing and future applications, especially those concerning critical and autonomous systems.

A numeric result of prediction or classification does not carry any associated credibility measure. We advocate that the credibility can be associated with the numeric results by making its granular description, viz. by forming an information granule formed around the original numeric finding, Figure 1. The information granule delivers a well quantifiable result of credibility of the outcome and makes the user or associated system (e.g., an autonomous vehicle) aware about the quality of the result implying possible activity to be taken, in particular to take some action or rather to collect more experimental evidence.

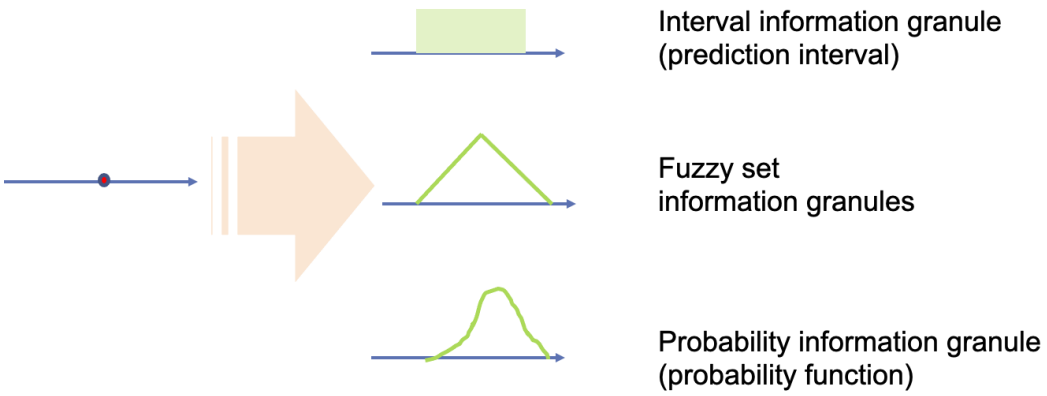


Figure 1. Augmenting numeric results by concept of information granularity.

It is worth noting that this line of thought invoking information granule has been studied in the past under some particular assumptions. For instance, in linear regression analysis, the results are provided in terms of interval information granules guided by some probabilistic evidence and leading to confidence or prediction intervals. In case of nonlinear models, one has to consider more specialized approaches such as a delta method, mean-value estimation (MVE), and bootstrapping.

Another alternative is to resort to Bayesian models and Gaussian processes, in particular. In these cases, the results of the model are probabilistic information granules.

From the architectural perspective, we can think of a granular embedding the original numeric ML model as illustrated in Figure 2. The embedding mechanism is endowed with a level of information granularity ε which can be thought as a design asset whose optimization. From the algorithmic perspective, the embedding is realized by optimizing a certain performance index characterizing the quality of granular results when being confronted with the data.

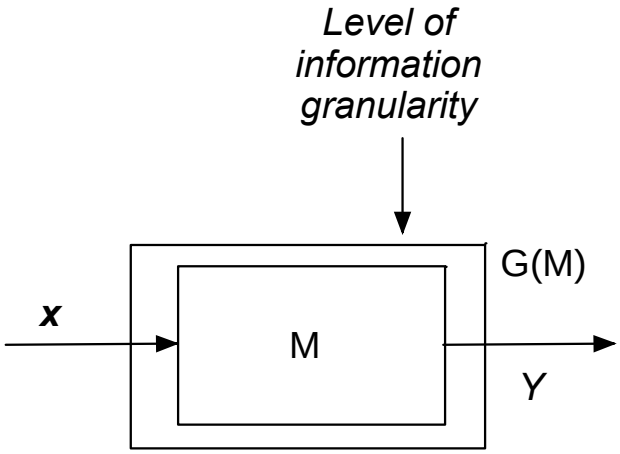


Figure 2. A granular embedding of ML model; emphasized is a level of information granularity treated as a certain design asset.

Let us start with a numeric model M expressed as $y = \mathrm{M}(\boldsymbol{x}; \boldsymbol{w})$ that has been designed in a supervised mode on the basis of pairs of input-output data $(\boldsymbol{x}_k, target_k)$, $k = 1,2, ..., N$. Here $\boldsymbol{x}$ stands for the vector of input variables, $\boldsymbol{w}$, $dim(\boldsymbol{w}) = m$, denotes a vector of estimated parameters of the model, $target_k$ is the output data for the corresponding $\boldsymbol{x}_k$.

The parameters of the model $\boldsymbol{w}$ are elevated to a numeric counterpart in the following fashion

$$\boldsymbol{w} \xrightarrow{G,\varepsilon} \boldsymbol{W} \tag{1}$$

i.e.,

$$\boldsymbol{W} = G(\boldsymbol{w},\varepsilon) \tag{2}$$

where the level of information granularity e gives rise to granular information granules $\boldsymbol{W}$. Namely, if we admit information granules in the form of intervals, we have the following expressions

$$w_i \overset{\epsilon}{\rightarrow} [\min(w_i(1+\varepsilon), w_i(1-\varepsilon)), \max(w_i(1+\varepsilon), w_i(1-\varepsilon))],\ \varepsilon \geq 0 \quad (3)$$

$$w_i \overset{\epsilon}{\rightarrow} [\min(w_i(1+\varepsilon), w_i/(1+\varepsilon)), \max(w_i(1+\varepsilon), w_i/(1-\varepsilon))],\ \varepsilon \geq 0 \quad (4)$$

The higher the value of ε is the broader the interval information granule centered around the original numeric parameter w_i.

The level of information granularity ε is optimized by evaluating the resulting information granule $Y = G(M(\boldsymbol{x}; \boldsymbol{w}))$ in terms of the coverage of data and its specificity. These two criteria are important descriptors of the quality of information granule produced by the model when confronted with numeric data. Coverage is a Boolean (or multivalued predicate in case of fuzzy sets) that returns 1 when the numeric datum is "covered" (included) in the information granule. Let the granule be an interval $[a, b]$ and the numeric datum is y_0. We have

$$\text{cov}(y_0, [a, b]) = 1 \text{ if } y_0 \in [a, b], \text{ otherwise coverage returns zero, } \text{cov}(y_0, [a, b]) = 0 \quad (5)$$

Obviously, we may wish that the coverage requirement is satisfied for all data. The higher the coverage, the better the model in terms of this criterion. Specificity is a measure expressing the precision of the information granule. In general, it can be thought as a decreasing function g of the length of information granule. The length of the interval is computed as $|b - a|$. There are numerous examples of the function g. For instance,

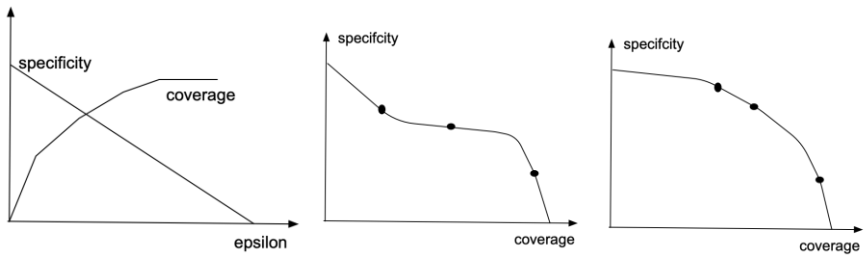
$$g(|b - a|) = 1 - |b - a|/range \quad (6)$$

where *range* is a calibration parameter. Another alternative is

$$g(|b - a| = exp(-\alpha|b - a|) \quad (7)$$

with $\alpha > 0$ serving as scaling coefficient. The higher the specificity of information granule, the more relevant it is. Note that specificity of a single element is the highest, $sp(\{y_0\}) = 1$.

Coverage and specificity are in conflict: to achieve higher level of coverage, one has to reduce specificity and vice versa. Figure 3 illustrates this relationship between coverage and specificity by displaying them for different values of ε.



(a) (b)

Figure 3. Coverage–specificity optimization (a), and coverage–specificity relationship implied by different values of ε (b).

Consider a data set (either training, validation, or testing data). The value of the level of information granularity ε is determined through the optimization of the granular results confronted with the numeric data. In the optimization, a product of coverage *cov* and specifity *sp* (which are essential descriptors of information granules). The optimization yields a certain compromise between these two conflicting criteria of coverage and specificity. With the increase of the values of ε, the coverage increases but results in lower values of specificity.

The pertinent formulas are given as

$$\overline{cov} = \frac{1}{N}\sum_{k=1}^{N} cov(target_k, Y_k) \tag{8}$$

$$\overline{sp} = \frac{1}{N}\sum_{k=1}^{N} sp(Y_k) \tag{9}$$

where the above measures are defined in (5)–(7) and averaged over the corresponding data. We aim to maximize both the measures as in the case of the principle of justifiable granularity. In other words, we have ε_{opt} being a solution to the optimization problem where the product of coverage and specificity is maximized

$$\varepsilon_{opt} = arg\ max_{\varepsilon}(cov(\varepsilon)sp(\varepsilon)) \tag{10}$$

The higher the product of coverage and specificity is, the better the generated granular results are.

The level of information granularity could be more refined by admitting that each parameter of the model, w_1, w_2, ..., w_m can be transformed to its granular counterpart by associating ε_1, ε_2,..., ε_m with the corresponding parameters.

$$w_i \xrightarrow{G,\varepsilon_i} W_i \tag{11}$$

$\boldsymbol{\varepsilon} = [\varepsilon_1\ \varepsilon_2 ... \varepsilon_m]$. This yields the following optimization problem

$$max_{\boldsymbol{\varepsilon}}(cov(\boldsymbol{\varepsilon})sp(\boldsymbol{\varepsilon})) \tag{12}$$

and

$$\boldsymbol{\varepsilon}_{opt} = arg\ max_{\boldsymbol{\varepsilon}}(cov(\boldsymbol{\varepsilon})sp(\boldsymbol{\varepsilon})) \tag{13}$$

The calculus of intervals with the algebraic operations follows the well-known formulas [1].

In case of monotonic functions, we have $f[a, b]) = [f(a), f(b)]$ for increasing functions and $f([a, b]) = [f(b), (a)]$ for decreasing functions. In general case the extension principle is applied [7][12].

In rule-based models the parameters of the local functions in the conclusion are made granular. In the sequel the output is an information granule. Generally, a hierarchy of information granules is build, see Figure 4.

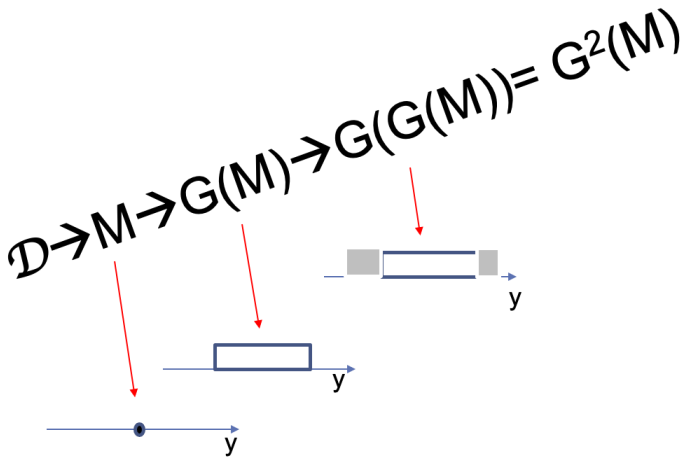


Figure 4. An emerging hierarchy of granular models distributed across information granules of increasing types of information granularity.

4. Transfer learning

The design of ML models calls for substantial computing overhead implying a significant energy consumption; recall that ChatGPT 3 required 936 MWh electricity. To secure further progress it becomes critical to pursue efficient ways to reuse already acquired knowledge (models). Transfer learning is a learning paradigm that is aimed at delivering the badly needed capabilities. While the idea has been around under different names such as learning by analogy, domain adaptation, pretraining..., its role in the current developments is highly relevant. In brief, transfer learning is about an extraction of previously acquired knowledge and applied to a new similar application. There are a number of other advantages motivating the consideration of transfer learning including situations where robustness is required.

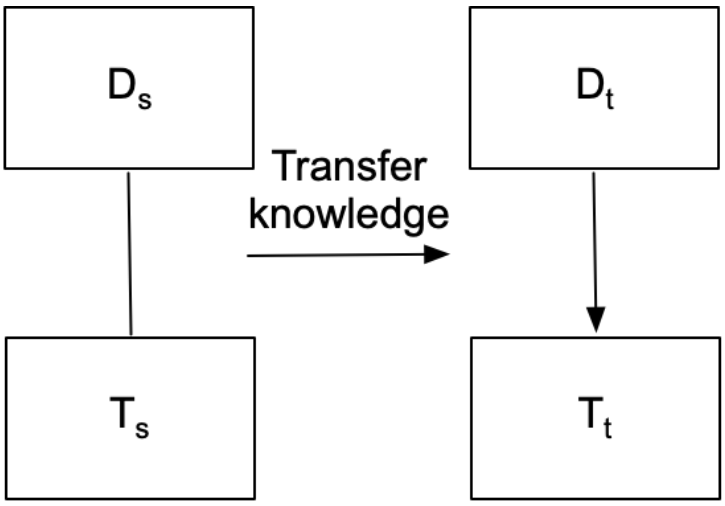


Figure 5. An essence of transfer learning.

What should not be ignored in the overall framework of transfer learning, Figure 5, is a fact that the model (knowledge) transferred from an original environment (referred to as a

source domain D_s) to the new environment (target domain D_t) is no longer of the same quality as it enjoyed in the D_s. Intuitively, the more different D_s is from D_t, this adversely impact the quality of the constructed model located in D_t. This calls for a thorough assessment of M – we argue that M in D_t becomes inherently granular where the concept of information granularity plays a pivotal role.

There are two main approaches in the development of the granular models in the D_t environment.

4.1. Passive approach

As visualized in Figure 6, the model M built in D_s is unchanged and directly positioned in D_t. In light of differences in D_t and D_s, the model M becomes granular $G(M)$. The construction of $G(M)$ follows the scheme discussed in Section 3.

The quality of the resulting granular model can be expressed by computing the product $cov*sp$ obtained for the optimal value of ε, $(cov*sp)_{\varepsilon opt} = \arg\max_{\varepsilon}(cov*sp)$. The lower the value of this product, the lower the quality of the transferred model. This quality is related with the "distance" between D_s and D_t. The more distant the domains are, the lower the quality of the transferred model is up to the point that the passive approach is no longer feasible and one has to explore another alternative such as an active approach.

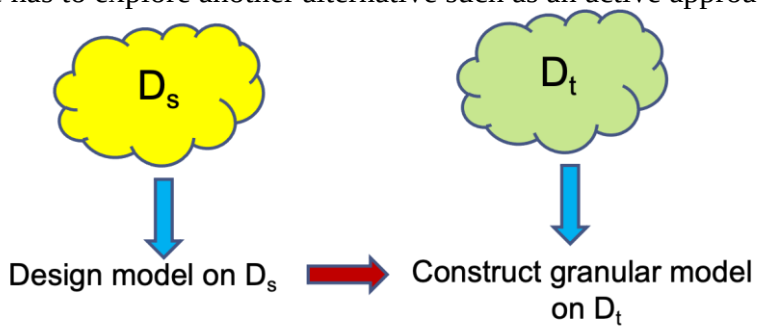


Figure 6. Transfer learning: a passive approach.

4.2. Active approach

As the name stipulates, the model in D_t is constructed by benefiting from the navigation delivered by the granular model $G(M)$. The main idea is displayed in Figure7. The model in D_t is designed on a far smaller data set (hence the reduction in the computing overhead) and its design is guided by the granular results produced by the granular manifestation of the model built on D_s and transferred to D_t.

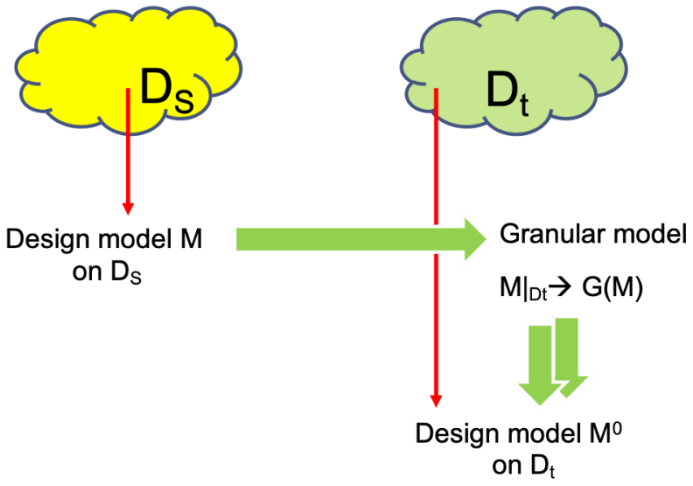


Figure 7. Active transfer learning: a general scheme along with a navigation delivered by the granular model $G(M)$ transferred from D_s.

The minimized loss function is composed of the following two components: (i) the loss computed for M^0 and data coming from D_t, and (ii) granular navigation hints coming from $G(M)$). Formally, it is expressed in the following form

$$Q = \sum_{\boldsymbol{D}_t} ||target_k - M^0(\boldsymbol{x}_k, \boldsymbol{w})|| + \alpha \sum_{\boldsymbol{D}_t} [1 - cov(M^0(\boldsymbol{x}_k, \boldsymbol{w}), G(M(\boldsymbol{x}_k))] * sp(G(M(\boldsymbol{x}_k)) \quad (14)$$

Where α is a hyperparameters controlling an impact coming from the granular model. Higher values of a stress higher reliance of the designed M^0 on the model transferred from D_s. Subsequently the minimization of Q is carried out following the gradient-based optimization scheme

$$\boldsymbol{w}(iter + 1) = \boldsymbol{w}(iter) - \beta \text{grad}_{\boldsymbol{w}} Q \quad (15)$$

The second term of (14) requires some clarification. Refer to Figure 8. The intent is to make the result $M^0(\boldsymbol{x})$ aligned (included in) $G(M(\boldsymbol{x}))$ which implies that $M^0(\boldsymbol{x})$ is covered by $GM(\boldsymbol{x})$) which is expressed by the term $cov(M^0(\boldsymbol{x}), G(M(\boldsymbol{x}))$. Hence in case of full coverage inclusion) the expression $1 - cov(M^0(\boldsymbol{x}_k, \boldsymbol{w}), G(M(\boldsymbol{x}_k))$ attains zero. Specificity measure, $sp(G(M(\boldsymbol{x}_k))$, quantifies the credibility of the granular guidance provided by the transferred model. The lower the specificity, the lower the intensity of support delivered by the granular model. Because of the granular form of the second term in (14), it delivers some regularization mechanism that could be referred to as a granular regularization. If for some data $\boldsymbol{x}_k$, the coverage and specificity are high, the guidance delivered by the granular model to construct the model M^0 becomes higher. If the coverage term is the same for two data $\boldsymbol{x}_k$ and $\boldsymbol{x}_l$, the corresponding regularization term achieves higher value for the data where the specificity of output of the model M is higher.

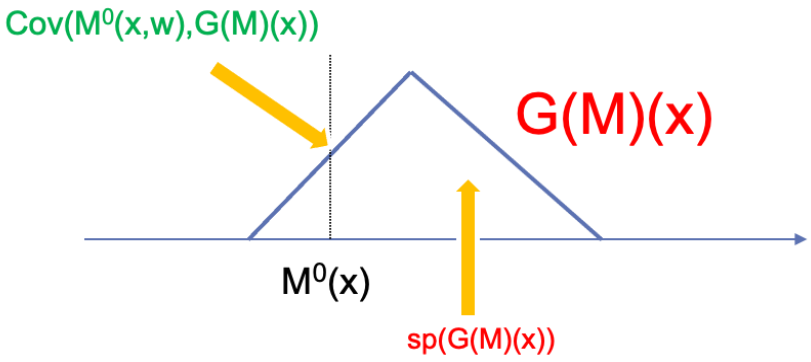


Figure 8. Granular regularization – computational details.

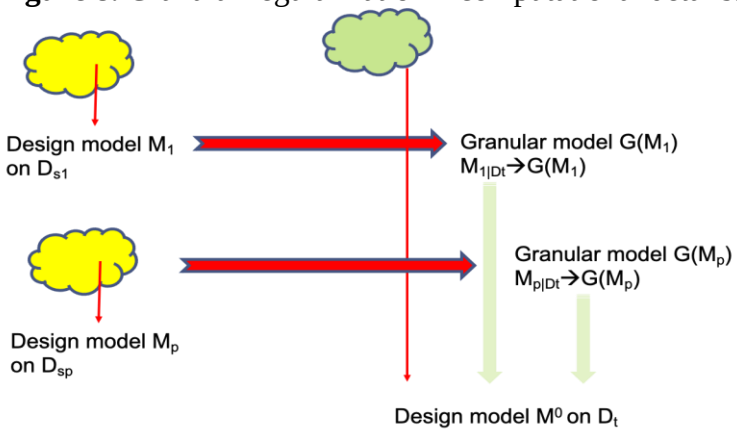


Figure 9. Multi-source transfer learning.

The generalization of the above scheme is the one in which there are several source domains, Figure 9, and from each of them the corresponding models give rise to their granular counterparts. Those is sequel are sought as granular navigation hints that are incorporated in the augmented loss function assuming the following form

$$Q=\sum_{\boldsymbol{D}_t}||target_k - M^0(\boldsymbol{x}_k,\boldsymbol{w})|| + \alpha_1 \sum_{\boldsymbol{D}_t}[1 - cov(M^0(\boldsymbol{x}_k,\boldsymbol{w}), G(M_1(\boldsymbol{x}_k))] * sp(G(M_1(\boldsymbol{x}_k)) + \\ +\alpha_2 \sum_{\boldsymbol{D}_t} \alpha_1 \sum_{\boldsymbol{D}_t}[1 - cov(M^0(\boldsymbol{x}_k,\boldsymbol{w}), G(M_1(\boldsymbol{x}_k))] * sp(G(M_1(\boldsymbol{x}_k)) + \\ +...+ \\ +\alpha_p \sum_{\boldsymbol{D}_t} \sum_{\boldsymbol{D}_t}[1 - cov(M^0(\boldsymbol{x}_k,\boldsymbol{w}), G(M_p(\boldsymbol{x}_k))] * sp(G\left(M_p(\boldsymbol{x}_k)\right) \quad (16)$$

where α_1, α_2, ..., α_p are the hyperparameters associated with the granular models and the overall terms deliver a mechanism of granular regularization.

5. Conclusions

The conceptualization and comprehensive design of intelligent systems involving ML paradigms have been intensively pursued producing a list of successes. With the increasing spectrum of applications, quite often targeting critical domains with far reaching implications,

the development strategies have to include requirements of credibility assessment and ways of learning that are computationally sound. To address these two important issues, we have argued that concepts of abstraction realized in the setting of Granular Computing play is crucial role. The notion of credibility paving a way to realize self-awareness mechanisms in ML architectures and open new directions both in studies on autonomous systems and their applications.

References

[1] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, *et al*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion* 2020(58): 82–115.

[2] Castanyer RC, Martínez-Fernández S, Franch X. Which design decisions in AI-enabled mobile applications contribute to Greener AI? *arXiv* 2021, 2109: 15284.

[3] Pedrycz W. Towards green machine learning: challenges, opportunities, and developments. *J Smart Environ Green Comput* 2022, 2: 163–174.

[4] Schwartz R, Dodge J, Smith NA, Etzioni O. Green AI. *arXiv:*1907.10597v3, 2019.

[5] Tornede T, Tornede A, Hanselle J, Wever M, Mohr F, Hullermeier E. Green Automated Machine Learning: Status Quo and Future Directions. *arXiv*: 2022, 2111: 058550.

[6] Alefeld G, Herzberger J. *Introduction to Interval Computations.* New York: Academic Press, USA, 1983.

[7] Wang Y. On intelligent mathematics (IM): What's missing in general AI and Cognitive Computing? *4th International Conference on Physics, Mathematics and Statistics (ICPMS'21)* 2021, 1.

[8] Wang Y. On abstract intelligence and brain informatics: mapping cognitive functions of the brain onto its neural structures. *Int J Cogn Inform 2*012, 6(4): 54–80.

[9] Wang Y. On cognitive informatics, Brain and Mind: A Transdisciplinary. *J Neurosci* 2003, 4(2): 151–167.

[10] Zadeh LA. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets Syst* 1997, 90: 111–117.

[11] Pedrycz W. *Granular Computing*. Boca Raton: CRC Press, Fl 2013.

[12] Pedrycz W. Granular computing for data analytics: a manifesto of human-centric computing. *IEEE CAA J utom Sin* 018, 5:1025–1034.

[13] Moore R. *Interval Analysis.* Hoboken: Prentice Hall, USA, 1966.

[14] Moore R, Kearfott RB, Cloud MJ. *Introduction to Interval Analysis.* Philadelphia: SIAM, 2009.

[15] Nguyen H, Walker E. *A First Course in Fuzzy Logic.* Chapman Hall: CRC Press, 1999.

[16] Pedrycz W. An Introduction to Computing with Fuzzy Sets—Analysis, Design, and Applications. Berlin: Springer, 2020.

[17] Pedrycz W. Shadowed sets, representing and processing fuzzy sets. *IEEE Trans Syst Man Cybern 1*998, 28: 103–109.

[18] Pawlak Z. Rough sets. *Int J Info Technol Comp Sci* 1982, 15(11): 341–356.

[19] Pawlak Z. *Rough Sets: Theoretical Aspects of Reasoning about Data*. Dordrecht: Kluwer Academic, 1991.

[20] Mendel JM, John RI, Liu F. Interval Type-2 fuzzy logic systems made simple. *IEEE Trans Fuzzy Syst* 2006, 14:808-82.