Mean-field theory of meta-learning

Dariusz Plewczynski

Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Pawinskiego 5a Street, 02-106 Warsaw, Poland

E-mail: darman@icm.edu.pl

Abstract. We discuss here the mean-field theory for a cellular automata model of meta-learning. Meta-learning is the process of combining outcomes of individual learning procedures in order to determine the final decision with higher accuracy than any single learning method. Our method is constructed from an ensemble of interacting, learning agents that acquire and process incoming information using various types, or different versions, of machine learning algorithms. The abstract learning space, where all agents are located, is constructed here using a fully connected model that couples all agents with random strength values. The cellular automata network simulates the higher level integration of information acquired from the independent learning trials. The final classification of incoming input data is therefore defined as the stationary state of the meta-learning system using simple majority rule, yet the minority clusters that share the opposite classification outcome can be observed in the system. Therefore, the probability of selecting a proper class for a given input data, can be estimated even without the prior knowledge of its affiliation. The fuzzy logic can be easily introduced into the system, even if learning agents are built from simple binary classification machine learning algorithms by calculating the percentage of agreeing agents.

Contents

1. Introduction

The *machine learning* (ML) algorithms allow computers to learn based on training data. A major focus of ML is to recognize complex patterns in datasets, or make intelligent decisions based on data. Typically ML algorithms are divided into several classes: (1) supervised learning (generates a function that maps input data into desired outputs); (2) unsupervised learning (models a set of inputs, where no prior classification is given); (3) semi-supervised learning (generates an appropriate function or classifier); (4) reinforcement learning (learns how to act given an observation of the world, where every action has some impart in the environment, with feedback of it back to the algorithm); (5) transduction (predicts new outputs based on training inputs, outputs and test inputs); and (6) learning to learn (learns its own inductive bias based on previous experience) [1]. Different algorithms of ML have been applied successfully to solve real-life problems, for example in the context of bioinformatics [2]–[6] or chemo-informatics problems [7]–[12].

The current trend of machine learning theory and its applications is to develop meta-learning techniques, since no single paradigm is superior to others in all possible situations [13]–[18]. What is the exact definition of *meta-learning*? Most ML researchers do not share any single definition of this term. Therefore it is very important to precisely define what constitutes meta-learning in this paper. Traditionally this term refers to self-adaptation of ML methods, i.e. to readjusting their hypothesis spaces in order to fit better to a changing environment and to allow for higher precision classification of incoming data, or making more accurate predictions about previously unknown cases. Yet, this very general statement cannot be easily described in terms of statistical mechanics and practical applications of such methodology. Therefore, I define here the meta-learning as the process of combining outcomes of individual learning procedures in order to determine the final decision. In that way, the individual learners perform typical machine learning procedures, then their predictions are gathered and integrated. This definition is very similar to other 'meta'-approaches in bio- and chemo-informatics, that are now of great interest in the field of computational biology and biophysics. My approach is currently well supported by recent advances in the protein fold recognition field that is dominated by meta-predictors like 3D-Jury [19, 20], Pcons [21, 22] and Robetta [23]–[25]. Multiple

tests confirmed that consensus methods were more powerful than individual prediction algorithms in sensitivity and specificity even if some meta-predictors used as little as three methods to build a consensus model. Yet, the problem of selecting the proper and most successful procedure of meta-learning is still an ongoing debate.

Here, I provide a simple procedure for the integration of results from different methods into a single prediction that complements previous approaches [14]–[18], [26]. As I mentioned above, in *Meta-Learning* (ML) one trains an ensemble of machine learning algorithms using different types of input training data representations [7], [27]–[29]. All possible solutions are gathered and the consensus is built between them. The final phase of learning, i.e. consensus learning, is trying to balance the generality of solution and the overall performance of the trained model. This approach is similar to other ensemble methods, yet different from bagging (combines many unstable predictors to produce a ensemble stable predictor) or boosting (combines many weak but stable predictors to produce an ensemble strong predictor), it focuses on the use of a heterogeneous set of algorithms in order to capture even a remote, weak similarity of the predicted sample to the training cases. The main problem with such meta-approaches is that they are static and very specific. The meta-approach is optimized for certain combinations of machine learning types of algorithms and selected particular representations of training data. Nevertheless, the model of the consensus should be in principle calculable in a more general way.

Therefore, the goal of this paper is to provide a general, theoretical framework for the general integration of results of individual machine learning algorithms. In order to perform analytical analysis, I assume an infinite, statistical ensemble of different ML methods. The global preference toward a true solution can be described in my approach as the global parameter affecting all learners. Each learner (intelligent agent) performs training on available input data toward classification pressure described by the set of positive and negative cases. When the query testing data is analyzed each agent predicts the query item classification by a 'yes'/'no' decision. The answers of all agents are then gathered and integrated into the single prediction via dynamical evolution described within a cellular automata framework. This dynamical view of the consensus between various machine learning algorithms is especially useful for artificial intelligence or robotics applications, where adaptive behavior is given by the integration of results from an ensemble of ML methods.

2. Brainstorming: cellular automata model of meta-learning

My model of learning is based on a nonlocal *cellular automata* (CA) approach known in physics [30]–[34], with a wide range of applications [35]–[43]. The first statistical mechanics of opinion formation in groups of individuals was proposed by Lewenstein *et al* [44] on the class of models that were based on probabilistic cellular automata and social impact theory introduced by Latane [45, 46]. The mean-field theory with intermittent behavior was observed with a variety of stationary states with a well-localized and dynamically stable clusters (domains) of individuals who share minority opinions [44]. The impact of a group of N agents on a given learner is proportional to three factors: (1) the '*strength*' of the members of the whole ensemble, (2) their '*social*' distance from the individual and (3) their number N. Such a model leads to ferromagnetic and spin-glass phases, when different values of persuasiveness and supportiveness are assumed. Then this approach

was successfully used in a variety of sociological phenomena, with an interesting extension of the model done by Kohring [47, 48], where Latane's theory was extended to include learning. The cellular automata with intrinsic disorder was later solved analytically in the continuous limit by Plewczynski [49] and proved that in the model of Cartesian social space (therefore not fully connected) and containing no learning rules, one can also observe different phases (small clusters in the sparse phase with a large role of strong individuals, and high density phase with almost uniform opinion). The later results of Hołyst *et al*, where numerical simulations and analytical models were tested in simplified geometries, proved the usefulness of mean-field formalism in describing the social impact theory and the presence of the equilibrium states of the system with complex intermittent behavior [50]–[53].

In the present paper I present a novel application of cellular automata models to the meta-learning problem. Each cell in the CA represents a single machine learning algorithm, or a certain combination of parameters, and optimization conditions affecting the classification output of this particular method. I call this single learner the term 'learning agent', each characterized for example by its prediction quality on a selected training dataset. The coupling of individual learners is described by short-, medium- or long-range interaction strength, so-called learning coupling. The actual structure, or topology of coupling between various learners is described using the term 'learning space', and can have different representations, such as Cartesian space, fully connected or hierarchical geometry. The result of the evolution dynamics of such a system given by its stationary state is defined here as the consensus equilibration. The majority of learners define, in the stationary limit, the 'learning **consensus**' outcome of the meta-learning procedure: for example, by the total difference between positive and negative predictions in the binary classification problem.

The information integration, i.e. the consensus building between various machine learning algorithms and various prediction outcomes, is similar to the dynamical changes in cellular automata systems known in physics. The phase transitions can be observed in the system, the global new phase emerging when the system reaches a critical point in terms of its order parameter. Changes between phases of the system are induced by some external factors that can be modeled as a bias added to the local fields. Mean-field theory was successfully used in a number of physical problems, such as a super-fluid effect [54], a weakly interacting Bose gas in an external field [55]–[59], quantum solitons in optical fibers [60, 61] and many others [62]–[68]. The main difference between my approach and those formalisms is the procedure of taking a real function instead of a complex one. The individual learners are modeled here as single machine learning procedures. Therefore each of them is described by a set of calculable real values depending on each ML method's precision, recall and classification error, therefore assuming the input training data. My approach is supervised learning because the description of the quality for each agent depends on the type or representation of the training data. On the other hand, the presented formalism is more general allowing for unsupervised learning, i.e. for searching for unknown patterns in input data, by representing them on the fully connected grid, and allowing for features' similarity exchange. This approach is now studied by Plewczynski *et al* in the context of two wide ranges of applications in bioinformatics: namely protein–protein interaction prediction and protein–ligand docking. In the first case, the whole known proteome of an organism is represented on the two-

dimensional grid, where each node represents a single protein. The coupling between two proteins is given by the structural and sequence similarity focused on likely interaction sites on the surfaces of proteins [69]. This approach allows for unsupervised protein–protein interaction prediction, taking a different route, namely unsupervised meta-learning based on cellular automata and phase transitions, than previous methods [70]–[74].

The cellular automata model of meta-learning is based on several assumptions:

(1) *Binary logic.* I assume the binary logic of individual learners, i.e. we deal with cellular automata consisting of N agents, each holding one of two opposite states ('NO' or 'YES'). These states are binary $\sigma_i = \pm 1$, similar to the Ising model of ferromagnets. In most cases the machine learning algorithms that can model those agents, such as support vector machines, decision trees, trend vectors, artificial neural networks and random forest, predict two classes for incoming data, based on previous experience in the form of trained models. The prediction of an agent answers a single question: is a query data contained in class A ('YES'), or it is different from items gathered in this class ('NO').

(2) *Disorder and random strength parameter.* Each learner is characterized by two random parameters: persuasiveness p_i and supportiveness s_i that describe how individual agents interact with others. Persuasiveness describes how effectively the individual state of an agent is propagated to neighboring agents, whereas supportiveness represents self-supportiveness of a single agent. In the present work I assume that influential agents have high self-esteem $p_i = s_i$, which is supported by the fact that highly effective learners should have a high impact on others in a meta-learning procedure. For example, we can select $p_i = f(\text{precision}, i)$ and $s_i = f(\text{recall}, i)$ in the case where agents are modeled as single machine learning procedures. In general the individual differences between agents are described as random variables with a probability density $\hat{p} = (p_i, s_i)$, with mean values $p = \sum p_i/N$ and $s = \sum s_i/N$. Similar to the social influence theory, the quality of the predictor in some way affects its influence strength, when the final optimization of meta-learning consensus is done. In the case of meta-learning procedure the persuasiveness p_j represents here the ability of learning agent j to persuade agents who hold the opposite state to switch to having the same state as j. The supportiveness s_j represents the ability of learning agent j to support agents who hold the same state, so that it has not only the self-support of an individual agent (of itself), but the support that an agent gives to other agents who share the same state.

(3) *Learning space and learning metric.* Each agent is characterized by a location in the learning space, therefore one can calculate the abstract learning distance $d(i, j)$ of two learners i and j. The strength of coupling between two agents tends to decrease with the learning distance between them. Determination of the learning metric is a separate problem, and the particular form of the metric and the learning distance function should be empirically determined, and in principle can be a very peculiar geometry. In the present paper, I select the fully connected learning space, where all distances between agents are equal $d(i, j) = 1$. This particular geometry is useful, for example, in the case of a simple consensus between different yet not organized machine learning algorithms, where no one group of learners performs significantly better than the others.

(4) *Learning coupling.* Agents exchange their opinions by biasing others toward their own classification outcome. This influence can be described by the total learning impact I_i that the ith agent is experiencing from all other learners. Within the cellular automata approach this impact is the difference between positive coupling of those agents that hold identical classification outcomes, relative to the negative influence of those who share opposite states, and can be formalized as

$$I_i = I_p \left(\sum_j \frac{p_j}{N} \left(1 - \sigma_i \sigma_j \right) \right) - I_s \left(\sum_j \frac{s_j}{N} \left(1 + \sigma_i \sigma_j \right) \right), \tag{1}$$

where $I_p(\cdot)$ and $I_s(\cdot)$ are the functions of the persuasiveness impact and the supportiveness impact of the other agents on the ith agent. It should be noted here that the persuasiveness p_j represents here the ability of agent j to persuade agents who hold the opposite state to switch to having the same state as j. In contrast the supportiveness s_j represents the ability of agent j to support agents who hold the same state, i.e. preventing them from switching to the opposite state. That is, persuasiveness represents the propensity of j to cause other agents to switch to that state, and supportiveness represents its propensity to keep them there.

(5) *The equations of meta-learning.* The equation of dynamics of the learning model defines the state σ_i' of the ith individual at the next time step as follows:

$$\sigma_i' = (-\operatorname{sgn}(\sigma_i I_i)), \tag{2}$$

with rescaled learning influence:

$$I_i = \sum_j \frac{p_j}{N(s+p)}(1 - \sigma_i \sigma_j) - \sum_j \frac{s_j}{N(s+p)}(1 + \sigma_i \sigma_j). \tag{3}$$

I assume a synchronous dynamics, i.e. states of all agents are updated in parallel. In comparison to standard Monte Carlo methods the synchronous dynamics takes a shorter time to equilibrate than serial methods, yet it can be trapped into periodic asymptotic states with oscillations between neighboring agents.

(6) *Presence of noise.* The randomness of state change (phenomenological modeling of various random elements in the learning system and training data) is given by introducing noise into dynamics:

$$\sigma_i' = (-\operatorname{sgn}(\sigma_i I_i + h_i)), \tag{4}$$

where h_i is the site-dependent white noise, or one can select a uniform white noise, where for all agents $h_i = h$. In the first case h_i are random variables independent for different agents and time instants, whereas in the second case h are independent for different time instants. I assume here that the probability distribution of h_i is both site-and time-independent, i.e. it has uniform statistical properties. The uniform white noise simulates the global bias affecting all agents, whereas site-dependent white noise describes local effects, such as prediction quality of individual learners, etc.

The system defined in this way is similar to previously postulated cellular automata models of opinion change in social sciences [44, 49, 50]. The main differences of those approaches from the previously described cellular automata models are given by the infinite-range interactions and fully connected cellular automata, which are better

fitted to the learning context of the problem. In addition, the random strength parameters are introduced, therefore allowing for more complex behavior to be observed. Individual agents are described using probability density $\hat{p} = (p_i, s_i)$, so they differ from each other. The impact function is also included, so learners are able to exchange their states in the form of coupling.

3. Mean-field approximation

The fully connected learning space geometry presents an interesting practical formalization for further analysis of the meta-learning procedure. Here, all agents are coupled with each other with some randomly distributed strength that is independent of the distance between them. The mean-field theory provides a very well-defined and controlled approximation allowing for solving the dynamical equations of such a model. The dynamical 'order' parameter has to be defined, to show the decay of minority groups in the form of 'staircase' dynamics. The fully connected geometry of the learning space is supported by the topology of this problem. We have a set of independent machine learning algorithms, no prior hierarchy or topology for this set is imposed, and we would like to build the consensus between their predictions.

The discrete equation of dynamic is given by

$$\sigma'_i = -\operatorname{sgn}\left(\sum_j \frac{p_j}{(s+p)N}(\sigma_i - \sigma_j) - \sum_j \frac{s_j}{(s+p)N}(\sigma_i + \sigma_j) + h_i\right). \quad (5)$$

Introducing a weighted majority–minority difference for a system:

$$m = \sum_j \frac{(s_j + p_j)\sigma_j}{N(s+p)}, \quad (6)$$

and random parameters to describe effective self-supportiveness of each agent:

$$a_i = \frac{s-p}{s+p} + \frac{\beta}{s+p}s_i, \quad (7)$$

we get the dynamical equation in the noise-absent limit by rewriting equations (5) using $\theta(\cdot)$ as the Heaviside theta function:

$$\sigma'_i = \operatorname{sgn}(m + h_i)\theta(|m + h_i| - |a_i|) + \sigma_i \operatorname{sgn}(a_i)\theta(|a_i| - |m + h_i|). \quad (8)$$

We assume here that $a_i \geq 0$ for any distribution of random variable s_i [44].

The order parameter is defined in physics as a quantity that defines the phase transition between various phases in the physical system, for example defining the evolution of the ordered system toward chaotic behavior. In this model, the order parameter describes the changes between various meta-learning solutions (agreement between learners, uniformity of opinion, minority clusters, chaotic state of not coupled learners) and is given by the formula

$$\pi(\varphi) = \sum_j \frac{(s_j + p_j)\sigma_j}{N(s+p)}\theta(a_j - \varphi), \quad (9)$$

as in the standard mean-field theory of Ising systems [44, 47, 62]. The meaning of this parameter is as follows: π is a positive real number and is equal to the weighted majority–minority difference calculated for those agents that have effective self-supportiveness a_j greater than φ. The order parameter defines different, stationary states of the dynamics. In the noiseless limit it determines uniquely the approach toward a stationary state, and with small noise the only stationary states are close to uniformity with $m \cong \pm 1$. From the practical point of view it is enough to describe the ordering among agents of a given strength in order to completely specify the state of the system.

The derivative $\partial \pi(\varphi)/\partial \varphi$ is related to the weighted majority–minority difference for agents with effective self-supportiveness a_j equal to φ, i.e. for those agents with $s_j = 1/\beta[\varphi(s+p) - (s-p)]$. As in the work of Lewenstein [44] the order parameter π_i in the noiseless limit fulfills the equation

$$\pi'(\varphi) = [\mu(m, \varphi) + \pi(|m|)]\theta(|m| - \varphi) + \pi(\varphi)\theta(\varphi - |m|), \qquad (10)$$

with $\pi(0) = m$, and

$$\mu(m, \varphi) = 1/N \operatorname{sgn}(m) \sum_j \frac{(s_j + p_j)\sigma_j}{N(s+p)}\theta(|m| - |a_j|)\theta(a_j - \varphi). \qquad (11)$$

The mean-field approximation for the system is introduced by replacing the actual variables (like m, $\pi(\varphi)$, $\mu(m, \varphi)$) by their corresponding mean values calculated by averaging over disorder, i.e. random distribution of self-supportiveness s_i and p_i. The averaged equations are then valid for very large N (preferably infinite system), where m is no longer a random variable, because its fluctuations are of the order of $1/\sqrt{N}$. The recurrence equation for m is defined for $\varphi = 0$, i.e. in regions where $\varphi \geq |m| = \pi(0)$:

$$m' = \mu(m, 0) + \pi(|m|). \qquad (12)$$

Reformulating the definition of $\pi(\varphi)$ gives

$$\pi(\varphi) = m - \sum_j \frac{(s_j + p_j)\sigma_j}{N(s+p)}\theta(\varphi - a_j). \qquad (13)$$

Therefore the recurrence equation for m has the following form [44]:

$$m' = m + \sum_j \frac{(s_j + p_j)\sigma_j}{N(s+p)}[\operatorname{sgn}(m) - \sigma_j]\theta(\varphi - a_j). \qquad (14)$$

Let us define the function $f(m)$ as the value of m in consecutive time steps:

$$f(m) = m_0 + \sum_j \frac{(s_j + p_j)\sigma_j}{N(s+p)}[\operatorname{sgn}(m) - \sigma_j(t=0)]\theta(|m| - a_j). \qquad (15)$$

Therefore, we have the mean-field behavior of the recurrence equation described by the general equation

$$f(m) = m_0 + \sum_j \frac{(s_j + p_j)\sigma_j}{N(s+p)}[\operatorname{sgn}(m) - \sigma_j(t=0)]\theta(\varphi - a_j). \qquad (16)$$

In this meta-learning model each agent is described by two random parameters: persuasiveness p_i (the coupling with other agents) and supportiveness s_i (self-supportiveness). The first parameter, persuasiveness, imposes how effectively the individual state of the agent is propagated to neighboring agents, whereas the second parameter, supportiveness, represents self-influence of a single agent. We can select for each type of machine learning algorithm, or instance of a single ML method (depending, for example, on the methods' parameter values), a $p_i = f(\text{precision}, i)$ and $s_i = f(\text{recall}, i)$. In the present paper I assume that influential agents have high self-esteem:

$$p_i = s_i = h(\text{precision}, \text{recall}, i) \sim \tfrac{1}{2}(\text{precision}[i] + \text{recall}[i]). \qquad (17)$$

Here, highly effective learners should have high impact on others in the meta-learning procedure. The individual differences between agents are described as random variables with a probability density $\hat{p} = (p_i, s_i)$, with mean values $p = \sum p_i/N$ and $s = \sum s_i/N$. Therefore the recurrence equation (16) has a simplified form:

$$f(m) = m_0 + \sum_j \frac{s_j}{Ns}[\text{sgn}(m) - \sigma_j(t=0)]\theta(\varphi - a_j), \qquad (18)$$

with

$$a_j = \frac{\beta}{2s}s_j.$$

The procedure of averaging is done here over different random distributions of initial conditions $\sigma_i(t=0)$, and different possible distributions of values of self-supportiveness s_i.

The value of m changes during the evolution of the system:

$$m = \sum_j \frac{s_j\sigma_j}{Ns},$$

and

$$\pi(\varphi) = \sum_j \frac{s_j\sigma_j}{Ns}\theta(a_j - \varphi).$$

The recurrence equation for m' is given by the formula [44]

$$m' = m_0 + \sum_j \frac{s_j}{Ns}[\text{sgn}(m) - \sigma_j(t=0)]\theta(|m| - a_j).$$

And for positive initial value of m we have

$$m' = \mu(m, 0) + \pi_0(m).$$

Reformulating the definition of $\pi(\varphi)$ gives

$$\pi_0(\varphi) = m - \sum_j \frac{s_j\sigma_j(t=0)}{Ns}\theta(\varphi - a_j), \qquad (19)$$

therefore the recurrence map for m has the following form [44]:

$$f(m) = m_0 + \sum_j \frac{s_j}{Ns}[1 - \sigma_j(t=0)]\theta(m - a_j). \qquad (20)$$

The above equations show that $f(m)$ is bounded and an increasing function of m. The m is bounded and an increasing function of the time step, therefore it has at least one stable fixed point. It may have also several fixed stable points separated by unstable ones, fulfilling the equation

$$\sum_j \frac{s_j}{Ns}[1 - \sigma_j(t=0)]\theta(m - a_j) = 0. \tag{21}$$

During time evolution m tends to the nearest stable fixed point from its initial value m_0. Therefore if noise is not present the meta-learning gives the final answer as the local minima in the space of possible solutions. Therefore clusters of minorities appear in the system as generic solutions in the noiseless limit. When the system is forced to make a consensus prediction, for example by majority rule, the final prediction outcome is given simply by the majority result, yet the actual probability of the correct answer can be approximated as the percentage of states in agreement with majority rule, in comparison to the number of minority groups.

For large enough m the system reaches the stationary state close to the uniformity state. For small m the system is unstable and grows quadratically in m. The similar description is valid for negative values of m, therefore $f(m)$ has two stable and one unstable fixed point [44]. In the field theoretical formulation the system has a set of local minima describing clusters of minority states characterized by different s_j. Those groups collapse successively with stronger s_j when a small noise is added to the system [44]. I have to point out that the above model is valid also for different initial conditions, where weaker agents have a random state in opposition to the uniform state of stronger ones. If the equilibrium state of the system exists, and it is stable, the system will also be stable for cases where the initial state for weaker agents will be more diverse (non-uniform).

The above solutions for the system are given by the minority clusters surrounded by the majority agents, and the dynamic is of 'staircase' character in the presence of small noise [44, 49]. In general three initial states of the system can be observed: sparse (no correlation between the agent strength and its state), middle density (a state of a agent starts to be correlated with its strength, lots of interesting metastable global configurations) and large density state (most agents initially have similar states, therefore the role of coupling is not so important). In the first case clusters of both types of states may appear, and when the weak coupling is present there is no bias towards a uniform solution. In the second case a variety of sophisticated geometries and shapes of clusters are present; some are robust and metastable, while others disappear slowly changing their state in agreement with the majority rule. Here, no analytical solutions are easy to find, therefore computer simulations have to be applied. I leave this case for my next paper, where extensive computer modeling of the system will be presented.

4. Concluding remarks

Intelligent agents' theory is a fascinating topic in modern science [75]–[79]. Decision making transitions depend to a high degree on global factors influencing an ensemble of independent learners. On the other hand, those changes are dependent to a high degree on individual decisions (predictions) that are based on agents' attitudes. During consensus, i.e. the final decision making, the reciprocal influence is critical as each learner exchanges

its opinion with others. In my approach, I assume that external factors acting on each learner are present only during the first phase of meta-learning, where initial states for a population of learners are setting up. Yet, both processes, even if acting on different timescales, are important for understanding the computational intelligence process.

In this paper I have presented the statistical theory of meta-learning. In my approach I select long-range coupling between agents, as opposed, for example, to the Euclidean two-dimensional learning space, where only nearest neighbors are coupled. This assumption is well supported by the fact that we are typically focused on only equilibrium, stationary states. The fully connected learning space lets agents evolve faster in comparison to other types of cellular automata. In addition, all agents influence each other, so therefore we avoid local minima traps for the global system.

Each learner is characterized by two random parameters: persuasiveness p_i and supportiveness s_i that describe how individual agents interact with others. The random strength parameters simulate different individual features of learning agents. In principle one can define both parameters in various different ways. In the case of a set of machine learning algorithms, each of them can be described by its intrinsic parameters affecting the precision of a single classification model of training data. In the general case, several different types of machine learning algorithms can be used as individual learners. There, the distribution of the quality of local prediction can be described as random providing that algorithms differ significantly between each other in terms both of the quality of prediction (classification accuracy), recall values (the ability to memorize the positive items in the training dataset) or precision (the ability to precisely predict the classification of training items).

The other definition of those parameters (persuasiveness and supportiveness) can enhance the method's persuasiveness (the value of p_i), if the method has the state $\sigma_i = +1$, and makes its p_i value lower when the opposite state is taken. In this way, it allows us to speed up the consensus process by forcing the system to reach the equilibrium state more rapidly, yet pushing it to the $+1$ decision based on the selected training dataset. This can cause several problems with overtraining, so some limitations of this approach should be taken into account. The actual solutions presented in this paper do not depend strongly on the selected form of those parameters. Anyway we assume that they are some random variables describing the variety of individual decisions in the ensemble of learners.

There are two timescales in the system. The first timescale is related to the fast evolution of individual learners. When input testing data is presented to the system, each learner responds by its own single prediction. This local prediction of each agent is done very rapidly, almost instantly. Then those individual predictions are processed by the cellular automata algorithm in order to find the stationary state of the system. This part is denoted as the integration of information. As was shown above, such a stationary state has the form of minority clusters surrounded by the sea of majority prediction. Therefore, the final consensus prediction given by the majority rule still preserves non-orthodox solutions, allowing for fast adaptivity of the system when the training data pattern is changed. The timescale for this integrative process is relatively long in comparison to individual predictions. Therefore very fast (preferably optimized for parallel processing) cellular automata software implementations have to be prepared in order to apply the formalism described above in real-life problems. In the statistical model presented here, I assume that there is no coupling between those two timescales. Therefore I neglect all

details of individual evolution of learners, focusing our attention for the integration phase of incoming local information into a single, consensus answer.

The core question of this paper is how typical initial distribution of learners' state evolve in time. As was shown above different initial conditions are distinguished by the numbers of agents sharing the opposite opinion $m = \sum_j \sigma_j/N$ into three classes. The first class is close to uniform state $|m| \cong 1$, where almost all learners initially are in agreement. The evolution of the system rapidly collapses into a stationary uniform state. This situation is observed when individual learners share similar machine learning algorithms, or a wide spectrum of parameter values does not change the classification model. Opposite states are sparse, and randomly spread over the learning space. For example, most of single ML algorithms (such as Random Forest, SVM) trained on an easy or moderate difficulty training dataset will give very similar predictions for a test cases. Therefore the initial state of the consensus system is close to uniformity of opinion, and the uniformity state is the most frequent final state. The second initial condition $0 \ll |m| < 1$ describes a much richer solution space. A moderate number of agents share the opposite state and those can be distributed randomly over the learning space, or clustered into well-defined groups. This describes the situation where different values of parameters can cause different classification outcomes, or ensembles of ML algorithms contain significant differences between each other's algorithms, which construct distinct classification models of input training data. The system has its intrinsic preferences (or in other words preferable local classification model)—most of the agents agree with their preferences, yet to some degree the opposite consensus state is possible. Therefore, one can assume that agreement between agents is possible, even if there is a significant proportion of learners that classify input data oppositely. The third type of initial conditions $|m| \cong 0$ contain randomly distributed or clustered different agents' states spread over the learning space. Because the number of opposite states is similar, therefore the system is on the edge of a phase transition between two final consensus answers: 'YES' or 'NO'. Therefore, even a small perturbation of initial state, parameter change, or type and nature of testing examples can in principle guide the system into different, opposite answers. This type of consensus is more fitted to the difficult training cases, where both answers are very probable. Here, the final trained system is very fragile and strongly depends on testing input data. The small change of input testing data can build up very different consensus values. Here, the consensus as the final, stable state of the whole system is not obvious, and it can take a significant amount of time. The final state can be either randomly distributed negative learners in the majority of positive states, or clustered minorities.

Here, I apply the analytical results of the mean-field approximation for cellular automata dynamics with moderate- and long-range interactions known previously from the findings of Lewenstein *et al* [44], Plewczynski [49] and Hołyst *et al* [50, 51, 53] for novel problems of meta-learning. The CA model is characterized by long-range interactions between individual agents, the fully connected learning space (each individual learner exchanges their prediction results with all others) and by an intrinsic disorder that allows for complex learning geometries to appear in the system. Two obvious emergent phenomena were observed immediately in this class of models: polarization and clustering [44]. In this paper I prefer to use different terms for the description of those emerging phenomena, namely *integration* and *adaptivity*. The integration is based on a polarization effect (single majority as the stable stationary state of the system, therefore

allowing for consensus prediction), whereas the adaptivity is given by the clustering of similar outputs of groups of learners (small minority clusters of non-preferred learning outcome are grouped together in the learning space).

The order parameter of the system is given by the variable η that characterizes geometrical and dynamical features of the model. The equilibration of the system in a clustered state is given by intermittent, consecutive steps in the form of 'staircase' dynamics [44, 49]. First the strongest agents change their state, then the weaker rest of the minority cluster collapses. The final solution is given in most cases as a single state with uniform prediction outcome. Therefore it defines the final answer of the consensus system, when input data is presented to the learning network of individual agents. The existence of this uniform solution in the limit of infinite time, when noise is present, is of crucial importance for further analysis of this meta-learning model. The real-life realizations of these algorithms will be presented in forthcoming papers, especially in the field of bio- and chemo-informatics.

Acknowledgments

References

[1] Wikipedia 2009 *Machine Learning* cited; available from http://en.wikipedia.org/wiki/Machine_Learning
[2] Bacardit J *et al*, *Automated alphabet reduction for protein datasets*, 2009 *BMC Bioinform.* **10** 6
[3] Plewczynski D *et al*, *AutoMotif server: prediction of single residue post-translational modifications in proteins*, 2005 *Bioinformatics* **21** 2525
[4] Bhaskar H, Hoyle D C and Singh S, *Machine learning in bioinformatics: a brief survey and recommendations for practitioners*, 2006 *Comput. Biol. Med.* **36** 1104
[5] Larranaga P *et al*, *Machine learning in bioinformatics*, 2006 *Brief Bioinform.* **7** 86
[6] Mjolsness E and DeCoste D, *Machine learning for science: state of the art and future prospects*, 2001 *Science* **293** 2051
[7] Plewczynski D, Spieser S A and Koch U, *Assessing different classification methods for virtual screening*, 2006 *J. Chem. Inform. Modelling* **46** 1098
[8] Azencott C A *et al*, *One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical, and biological properties*, 2007 *J. Chem. Inform. Modelling* **47** 965
[9] Willett P *et al*, *Prediction of ion channel activity using binary kernel discrimination*, 2007 *J. Chem. Inform. Modelling* **47** 1961
[10] Byvatov E and Schneider G, *SVM-based feature selection for characterization of focused compound collections*, 2004 *J. Chem. Inform. Comput. Sci.* **44** 993
[11] Karwath A and De Raedt L, *SMIREP: predicting chemical activity from SMILES*, 2006 *J. Chem. Inform. Modelling* **46** 2432
[12] Melville J L, Burke E K and Hirst J D, *Machine learning in virtual screening*, 2009 *Comb. Chem. High Throughput Screen.* **12** 332
[13] Engelbrecht A P, 2007 *Computational Intelligence* (New York: Wiley)
[14] Burton J *et al*, *Virtual screening for cytochromes p450: successes of machine learning filters*, 2009 *Comb. Chem. High Throughput Screen.* **12** 369
[15] Do C B, Foo C S and Batzoglou S, *A max-margin model for efficient simultaneous alignment and folding of RNA sequences*, 2008 *Bioinformatics* **24** i68
[16] Gesell T and Washietl S, *Dinucleotide controlled null models for comparative RNA gene prediction*, 2008 *BMC Bioinform.* **9** 248

[17] Khandelwal A *et al*, *Computational models to assign biopharmaceutics drug disposition classification from molecular structure*, 2007 *Pharm. Res.* **24** 2249

[18] Ying H *et al*, *A fuzzy discrete event system approach to determining optimal HIV/AIDS treatment regimens*, 2006 *IEEE Trans. Inform. Technol. Biomed.* **10** 663

[19] Ginalski K *et al*, *3D-Jury: a simple approach to improve protein structure predictions*, 2003 *Bioinformatics* **19** 1015

[20] von Grotthuss M *et al*, *Application of 3D-Jury, GRDB, and Verify3D in fold recognition*, 2003 *Proteins* **53** (Suppl. 6) 418

[21] Bujnicki J M *et al*, *Structure prediction meta server*, 2001 *Bioinformatics* **17** 750

[22] Lundstrom J *et al*, *Pcons: a neural-network-based consensus predictor that improves fold recognition*, 2001 *Protein Sci.* **10** 2354

[23] Chivian D *et al*, *Automated prediction of CASP-5 structures using the Robetta server*, 2003 *Proteins* **53** (Suppl. 6) 524

[24] Chivian D *et al*, *Prediction of CASP6 structures using automated Robetta protocols*, 2005 *Proteins* **61** (Suppl. 7) 157

[25] Kim D E, Chivian D and Baker D, *Protein structure prediction and analysis using the Robetta server*, 2004 *Nucleic. Acids Res.* **32** W526-31 Web Server issue

[26] Capobianco E, *Model validation for gene selection and regulation maps*, 2008 *Funct. Integr. Genomics* **8** 87

[27] Hotz C S, Templeton S J and Christopher M M, *Comparative analysis of expert and machine-learning methods for classification of body cavity effusions in companion animals*, 2005 *J. Vet. Diagn. Invest.* **17** 158

[28] Klon A E, Glick M and Davies J W, *Combination of a naive Bayes classifier with consensus scoring improves enrichment of high-throughput docking results*, 2004 *J. Med. Chem.* **47** 4356

[29] Plewczynski D, *Brainstorming: consensus learning in practice*, 2009 *Frontiers in Neuroinformatics: 2nd INCF Congress of Neuroinformatics* (doi:10.3389/conf.neuro.11.2009.08.080)

[30] Goldschmidt Y Y *et al*, *Nonequilibrium critical behavior in unidirectionally coupled stochastic processes*, 1999 *Phys. Rev. E* **59** 6381

[31] Hauert C, *Fundamental clusters in spatial 2 × 2 games*, 2001 *Proc. Biol. Sci.* **268** 761

[32] Dolezal J and Hraba T, *Application of mathematical model of immunological tolerance to HIV infection*, 1988 *Folia Biol (Praha)* **34** 336

[33] Gerhardt M, Schuster H and Tyson J J, *A cellular automation model of excitable media including curvature and dispersion*, 1990 *Science* **247** 1563

[34] Ito K and Gunji Y P, *Self-organization toward criticality in the game of life*, 1992 *Biosystems* **26** 135

[35] Bartocci E *et al*, *CellExcite: an efficient simulation environment for excitable cells*, 2008 *BMC Bioinform.* **9** (Suppl. 2) S3

[36] Chauvet G A, *On the mathematical integration of the nervous tissue based on the S-propagator formalism*, 2002 *J. Integr. Neurosci.* **1** 31

[37] Kanai M *et al*, *Ultradiscrete optimal velocity model: A cellular-automaton model for traffic flow and linear instability of high-flux traffic*, 2009 *Phys. Rev. E* **79** 056108

[38] Su M *et al*, *Spatiotemporal dynamics of the epidemic transmission in a predator-prey system*, 2008 *Bull. Math. Biol.* **70** 2195

[39] Tyson J J and Mackey M C, *Molecular, metabolic, and genetic control: an introduction*, 2001 *Chaos* **11** 81

[40] Conrad M *et al*, *Towards an artificial brain*, 1989 *Biosystems* **23** 175 discussion 216-8

[41] Grim J and Hora J, *Iterative principles of recognition in probabilistic neural networks*, 2008 *Neural Netw.* **21** 838

[42] Kawato M, *From 'understanding the brain by creating the brain' towards manipulative neuroscience*, 2008 *Phil. Trans. R. Soc. Lond. B* **363** 2201

[43] Yu D, Liu F and Lai P Y, *Input reconstruction of chaos sensors*, 2008 *Chaos* **18** 023106

[44] Lewenstein M, Nowak A and Latane B, *Statistical mechanics of social impact*, 1992 *Phys. Rev. A* **45** 763

[45] Latane B, 1981 *Am. Psychol.* **36** 343

[46] Nowak A, Szamrej J and Latane B, 1990 *Psychol. Rev.* **97** 362

[47] Kohring G A, *Ising models of social impact: The role of cumulative advantage*, 1996 *J. Physique* I **6** 301

[48] Kohring G A, 1996 *J. Physique* I **6** 301

[49] Plewczynski D, *Landau theory of social clustering*, 1998 *Physica A* **261** 608

[50] Holyst J A, Kacperski K and Schweitzer F, *Phase transitions in social impact models of opinion formation*, 2000 *Physica A* **285** 199

[51] Kacperski K and Holyst J A, *Phase transitions and hysteresis in a cellular automata-based model of opinion formation*, 1996 *J. Stat. Phys.* **84** 169

[52] Kacperski K and Holyst J A, *Opinion formation model with strong leader and external impact: a mean field approach*, 1999 *Physica A* **269** 511

[53] Kacperski K and Holyst J A, *Phase transitions as a persistent feature of groups with leaders in models of opinion formation*, 2000 *Physica A* **287** 631

[54] Kuklov A, Prokof'ev N and Svistunov B, *Commensurate two-component bosons in an optical lattice: ground state phase diagram*, 2004 *Phys. Rev. Lett.* **92** 050402

[55] Tanaka F, *Thermoreversible gelation is a Bose-Einstein condensation*, 2006 *Phys. Rev. E* **73** 061405

[56] Bronski J C *et al*, *Stability of attractive Bose-Einstein condensates in a periodic potential*, 2001 *Phys. Rev. E* **64** 056615

[57] Hansson T H, Leinaas J M and Viefers S, *Exclusion statistics in a trapped two-dimensional Bose gas*, 2001 *Phys. Rev. Lett.* **86** 2930–3

[58] Hodby E *et al*, *Experimental observation of Beliaev coupling in a Bose–Einstein condensate*, 2001 *Phys. Rev. Lett.* **86** 2196–9

[59] Jackson A D *et al*, *Weakly interacting Bose–Einstein condensates under rotation: mean-field versus exact solutions*, 2001 *Phys. Rev. Lett.* **86** 945–9

[60] Staliunas K, *Midband solitons in nonlinear photonic crystal resonators*, 2004 *Phys. Rev. E* **70** 016602

[61] Ponomarenko S A, Litchinitser N M and Agrawal G P, *Theory of incoherent optical solitons: beyond the mean-field approximation*, 2004 *Phys. Rev. E* **70** 015603

[62] Uzunov D I, *Derivation of effective field theories*, 2008 *Phys. Rev. E* **78** 041122

[63] Ciach A, Gozdz W T and Stell G, *Field theory for size-and charge-asymmetric primitive model of ionic systems: mean-field stability analysis and pretransitional effects*, 2007 *Phys. Rev. E* **75** 051505

[64] Range G M and Klapp S H, *Density functional study of the phase behavior of asymmetric binary dipolar mixtures*, 2004 *Phys. Rev. E* **69** 041201

[65] Buceta J and Lindenberg K, *Comprehensive study of phase transitions in relaxational systems with field-dependent coefficients*, 2004 *Phys. Rev. E* **69** 011102

[66] Chitanvis S M, *Theory of polyelectrolytes in solvents*, 2003 *Phys. Rev. E* **68** 061802

[67] Pastor-Satorras R and Sole R V, *Field theory for a reaction-diffusion model of quasispecies dynamics*, 2001 *Phys. Rev. E* **64** 051909

[68] Drossel B, Bokil H and Moore M A, *Spin glasses without time-reversal symmetry and the absence of a genuine structural glass transition*, 2000 *Phys. Rev. E* **62** 7690–9

[69] Plewczynski D, 2009 unpublished results

[70] Craig R A and Liao L, *Phylogenetic tree information aids supervised learning for predicting protein–protein interaction based on distance matrices*, 2007 *BMC Bioinform.* **8** 6

[71] Del Carpio-Munoz C A *et al*, *MIAX: a new paradigm for modeling biomacromolecular interactions and complex formation in condensed phases*, 2002 *Proteins* **48** 696

[72] Krause R, von Mering C and Bork P, *A comprehensive set of protein complexes in yeast: mining large scale protein–protein interaction screens*, 2003 *Bioinformatics* **19** 1901

[73] Mamitsuka H, *Essential latent knowledge for protein–protein interactions: analysis by an unsupervised learning approach*, 2005 *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2** 119

[74] Tsai R T *et al*, *Exploiting likely-positive and unlabeled data to improve the identification of protein–protein interaction articles*, 2008 *BMC Bioinform.* **9** (Suppl. 1) S3

[75] Conte R, *Agent-based modeling for understanding social intelligence*, 2002 *Proc. Nat. Acad. Sci.* **99** (Suppl. 3) 7189–90

[76] Liu J, Zhong W and Jiao L, *A multiagent evolutionary algorithm for constraint satisfaction problems*, 2006 *IEEE Trans. Syst. Man Cybern. B* **36** 54

[77] Pedrycz W and Rai P, *A multifaceted perspective at data analysis: a study in collaborative intelligent agents*, 2008 *IEEE Trans. Syst. Man Cybern. B* **38** 1062

[78] Pedrycz W and Rai P, *A multifaceted perspective at data analysis: a study in collaborative intelligent agents*, 2009 *IEEE Trans. Syst. Man Cybern. B* **39** 834

[79] Rocha A F, *The brain as a symbol-processing machine*, 1997 *Prog. Neurobiol.* **53** 121