Computing and Informatics, Vol. 35, 2016, 1491–1514

ARTIFICIAL INTELLIGENCE AGGREGATING OPINIONS OF A GROUP OF PEOPLE

Marek Bundzel, Jozef Lacko, Iveta Zolotová

Department of Cybernetics and Artificial Intelligence
Faculty of Electrical Engineering and Informatics
Technical University of Košice
Letná 9
042 00 Košice, Slovakia
e-mail: marek.bundzel@tuke.sk, jozef.lacko@student.tuke.sk
iveta.zolotova@tuke.sk

Tomáš Kasanický, Ján Zelenka

Institute of Informatics
Slovak Academy of Sciences
Dúbravská cesta 9
845 07 Bratislava, Slovakia
e-mail: kasanicky@neuron.tuke.sk, jan.zelenka@savba.sk

Abstract. This study deals with the problems of aggregating the opinions of a group of people in such a way that the quality of the group decision surpasses the quality of the decision of the most experienced individual within the group. The methods we have studied fall in the research domain of the so called *collective intelligence*. We provide an overview of the state-of-the-art in the collective intelligence. We describe the method based on adaptive boosting we have proposed aggregatig the opinions of a group of people. We have implemented a web application to gather opinions of people and used the application to collect data for the experimental analysis. The model problem was to identify whether there is or there is not a tumor present in the series of X-ray images of human lungs. We have compared our proposed method to conventional methods such as majority voting. We have concluded that our proposed method can be successfully used to aggregate opinions

of a group of people to increase their collective intelligence above the level of the most successful individual within the group in many cases. We have observed that the highest increase in the collective intelligence may be achieved for intelligence wise homogeneous groups what confirms the results of previous studies.

Keywords: Collective intelligence, modified adaptive boosting, aggregation of opinions

1 INTRODUCTION

Humans make many individual decisions but often the decisions steering human activities are made as a consensus emerging from the synergies, competitions and collective efforts of a group or a collective. People and in fact also animals have use collective decision making on conscious or subconscious levels. The forms vary, ranging from simple (e.g., the majority voting principle) to the complex and difficult to describe scenarios that we see for example in the politics. By watching an ant colony we come to the obvious fact that a collective of cognitively simple individuals forms by means of communication and interaction an entity the intelligence of which surpasses the intelligence of its comprising elements many fold. Watching a group of humans on the other hand does not necessarily lead to the same conclusions. Sometimes humans in a group produce very bad judgments and sometimes "the many are smarter than the few". The questions among many are:

- Are there algorithms that would make the most of the collective decision making in terms of correctness and satisfaction?
- What is better for the given collective – to leave the decision making to the smartest individual or to consider all the opinions?

Aggregating opinions may be useful when solving many kinds of problems. The model problem that we have used in this study was identification of tumors in X-ray images. Although the reasons for our choice were practical and we will explain them later our aim was not to build a real world solution. But let us imagine a place in the third world where highly trained medical experts are scarce. One may give to several local people a crash course in diagnosing a disease but none of the locals will be very good at it. However, with an appropriate algorithm we may aggregate their opinions and make them perform well as a group. This approach may be also used when solving other well formulated problems with a proper feedback on the quality of the decision, for example when deciding whether to buy or sell commodities on the stock market or whether to provide or decline a loan. But eventually, enhanced collective intelligence may be applied in more complex environments with uncertain and time delayed responses to the decisions made, like in the local and state governments, etc. With the growth of the Internet and related technologies such as social networking large scale collecting of the opinions of people is becoming possible and

even simple. We are now able to express ourselves to very wide audiences. Of course only a small fraction of our opinions is useful and utilized, but nevertheless, these opinions represent a large pool where information and wisdom is scattered.

2 RELATED WORK

A standard dictionary definition of collective intelligence states that it is "a phenomenon in sociology where a shared or group intelligence emerges from the collaboration and competition of many individuals". It is not clear when the concept of collective intelligence emerged for the first time but one of the earliest documented mentions of it comes from the 1785 Condorcet's jury theorem [1]. The theorem states that the more members are in the jury the more probable is that they will decide correctly by majority voting if each member of the jury is independent and more likely than not to make a correct decision. The theorem is proven correct under its assumptions but the assumptions are considered unrealistic. Other notions of collective intelligence come from entomology based on observation that seemingly independent individuals can cooperate so closely as to become indistinguishable from a single organism called superorganism (Wheeler, 1911 [2]). The concept of superorganism has also been revisited [3] because the elements of the colony that was seen as an optimized entity perform a commotion of conflicting cooperative and competitive activities. Other definitions of collective intelligence are:

- Collective intelligence is any intelligence that arises from, or is a capacity or characteristic of, groups and other collective living systems [4].
- The basis and goal of collective intelligence is the mutual recognition and enrichment of individuals rather than the cult of fetishized or hypostatized communities [5].
- Collective intelligence defines the ability to decide collectively that assures capability of the group to perform equally or better than the individuals of the group [6].

Also, many other terms are used to describe the same or closely related concepts: community intelligence, swarm intelligence, collective behavior, crowd IQ, collective decision making, etc. *MIT Center for Collective Intelligence* (`http://cci.mit.edu/`) is one of the most prominent research groups working in the respective field. The center aims to create new examples of collective intelligence, to study the collective intelligence in existing organizations and to define theories on collective intelligence. The center also organizes a conference specialized in collective intelligence. In [7] the authors demonstrated the existence of a measurable collective intelligence in groups that is analogous to general intelligence in individuals. Outtake from the statement on the center's webpage:

> "Our basic research question is: How can people and computers be connected so that – collectively – they act more intelligently than any person, group, or computer has ever done before?"

Thomas W. Malone is the Patrick J. McGovern's Professor of Management at the MIT Sloan School of Management and the founding director of the MIT Center for Collective Intelligence. He is also a co-author of the Handbook on Collective Intelligence [8]. Each essay in the book describes the work on collective intelligence in a particular discipline, for example, economics and the study of markets; biology and research on emergent behavior in ant colonies; human-computer interaction and artificial intelligence; and cognitive psychology and the "wisdom of crowds" effect. Other areas in social science covered include social psychology, organizational theory, law, and communications. The center implemented the Climate CoLab (`http://climatecolab.org/`) crowdsourcing platform. The goal of the Climate CoLab is to harness the collective intelligence of thousands of people from all around the world to address the global climate change. The research of the center produced interesting results in the development of measurement methods of collective intelligence and in providing evidence that there is a collective intelligence factor (Wooley et. al. [9]). The concept of measuring the IQ is based on the assumption that people capable of solving certain type of problems will perform well also with other types of problems. Wooley et al. defined the collective intelligence analogously as the general ability of the group to perform a wide variety of tasks. In two studies with 699 people, working in groups of two to five, Wooley et al. found converging evidence of a general collective intelligence factor that explains a group's performance on a wide variety of tasks.

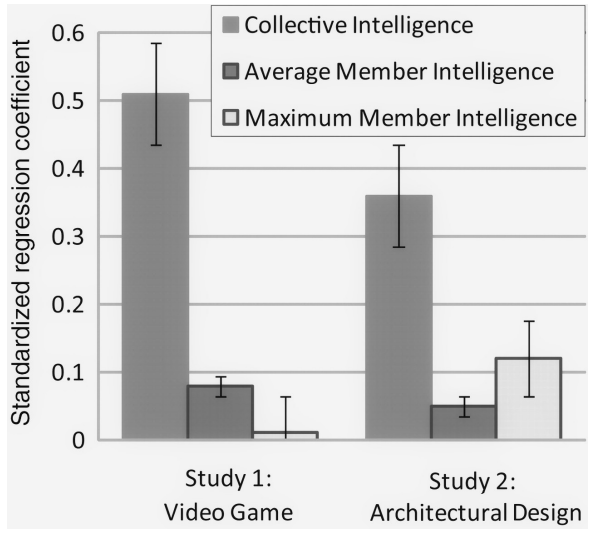

Figure 1. Standardized regression coefficients for collective intelligence and average individual member intelligence when both are regressed together on criterion task performance in Studies 1 and 2 [9]

In the first study, 40 groups of five were working for 5 hours on tasks from all quadrants of the McGrath Task Circumplex [11] with varying difficulty. The tasks comprised visual games, brainstorming, collective creativity tasks, making of moral decisions and negotiating over limited resources. As a criterion task each group played checkers against a standardized computer opponent. The IQ of all participants was also measured. The authors have found that a measurable general collective intelligence factor exists in groups and that it is not strongly correlated with the average or maximum individual intelligence of group members but is correlated with the average social sensitivity of group members, the equality in distribution of conversational turn-taking, and the proportion of females in the group. The average individual intelligence of the group members was not a significant predictor of group performance (Figure 1, Table 1). In the second study the authors have verified the findings for groups of different sizes.

Collective Intelligence	
Brainstorming	0.38*
Group matrix reasoning	0.86**
Group moral reasoning	0.42*
Plan shopping trip	0.66**
Group typing	0.80**
Avg. member intelligence	0.19
Max. member intelligence	0.27
Video game	0.52*

Table 1. Correlations among collective intelligence and group tasks for Study 1, $n = 40$ groups; *P ≤ 0.05; **P ≤ 0.001. Abbreviated from [9]

The experiments Wooley et al. conducted indicate that more than the average intelligence of the group members the performance of the group depends on the average social sensitivity of the group members, the equality in the distribution of conversational turn-taking, and the proportion of females in the group. The authors also state that the data show, the more women are on the team, the better. The extreme is reached when the team has little gender diversity rather than all women. That finding can be explained by differences in social sensitivity among men and women which is important to the group's performance.

Bachrach et al. from Microsoft Research [10] have conducted a thorough study on algorithmic aggregation of the opinions of the groups of people to boost the performance of the group. This study provided a solid base also for our research. The findings imply certain application potential for solving real world problems. Wooley et al. emphasized the role of social skills for the group's performance. When the interaction between the group members is left for an algorithm the problems may be avoided and the group members do not need to meet or communicate. Also, Wooley et al. did not specifically research whether the performance of a group will in general and under what circumstances exceed the performance of the most intelligent individual of the group. Bachrach et al. have focused on a specific task

of solving an IQ test (Raven's Standard Progressive Matrices Test, [12]) rather than on a wide variety of tasks and thus measured the individual and the group IQ conventionally.

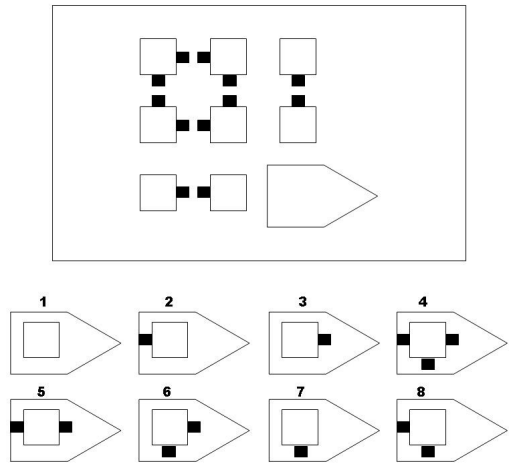

Figure 2. Raven's Standard Progressive Matrices Test is subject to copyright, this item is similar to those used in [10]

The used test takes approx. 30 minutes to accomplish, it does not require literacy of the respondent and it is often used to evaluate the logical thinking of school age children. The test questions type is multiple choice with a single correct answer (Figure 2). The sample consisted of 138 individuals aged 15–17 and it was representative of the British population (Figure 3). The questionnaire was identical for all individuals. Calculating IQ of a respondent based on the individual's questionnaire is straightforward. Different is the situation when IQ of a group of selected individuals is to be calculated. Simple majority aggregator with lexicographical tiebreaking was applied (MAJ). For every given question the answers of the group members are considered and the by the members most frequently selected answer is chosen to represent the aggregated group opinion. Thus a new filled questionnaire is obtained and the group IQ is calculated. Bachrach et al. used also a machine learning (ML) based aggregation method that addresses the limitations of the majority aggregation. The model employing probabilistic graphical models [13] attempts to make better inferences about the correct responses to items by jointly modeling the participants' aptitude and the correct responses. The underlying assumption is that each participant has an associated probability of knowing the correct response to an item, their aptitude, and that they will randomly guess the answer if they do not know the correct response. The ML method learns which participants provide the correct answers more reliably using a naive Bayes classifier.

Bachrach et al. performed a thorough analysis on how the composition of the group influences the aggregate IQ (the "crowd IQ") a relatively simple comparison of the results of the two aggregating methods. The research results show that with the increasing number of the group members (the "crowd size") the aggregate IQ quickly increases but saturates soon (Figure 4). Figure 4 also shows that ML method consistently outperforms MAJ aggregation and that the aggregate IQ is significantly higher than the average IQ of the group members.

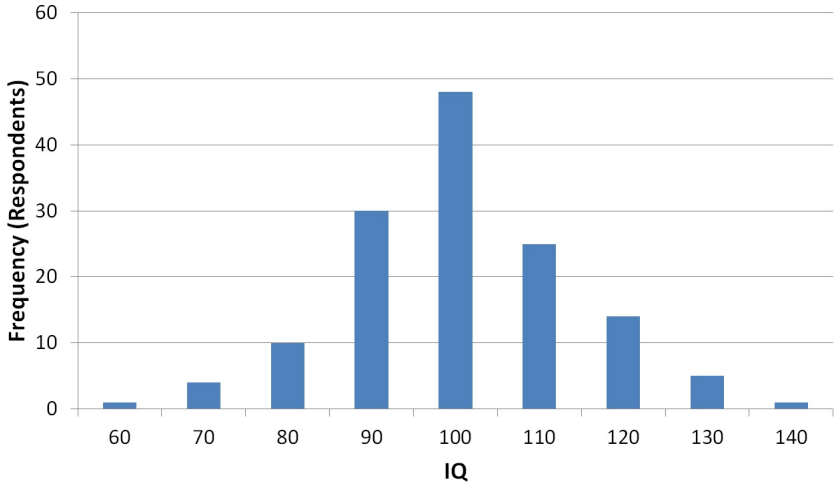

Figure 3. Histogram of IQ scores [10]

Although the aggregate IQ is above the population mean of 115 it is below the group and population maximum. The random crowds used for the analysis in Figure 4 are considered heterogeneous regarding the IQ. Bachrach et al. reported very interesting results for homogeneous crowds. For example if the crowd members fall in the 95 to 105 IQ interval the aggregate IQ surpasses the IQ of the group's smartest member (Figure 5). This phenomenon was observed also by other homogeneous groups of varying IQ.

Next, Bachrach et al. have focused on the relationship between the individual and the contextual IQ of the participants. A participant's contextual IQ is the expected increase in Crowd IQ from adding that participant to a random permutation of the crowd's members. For example, a person possessing a single skill that is rare is more valuable to the team than a person possessing a large set of skills that are common. Figure 6 shows that there is a positive correlation between the individual IQ and the contextual IQ and also a high variance of contextual IQs for participants of equal individual IQ. Individuals of equal IQ will have various effects on the group's performance ranging from beneficial to detrimental. If, for example, there is a person capable to solve the IQ test items that are also many other persons capable to solve, this person will contribute less to the crowd's IQ or even lower it.

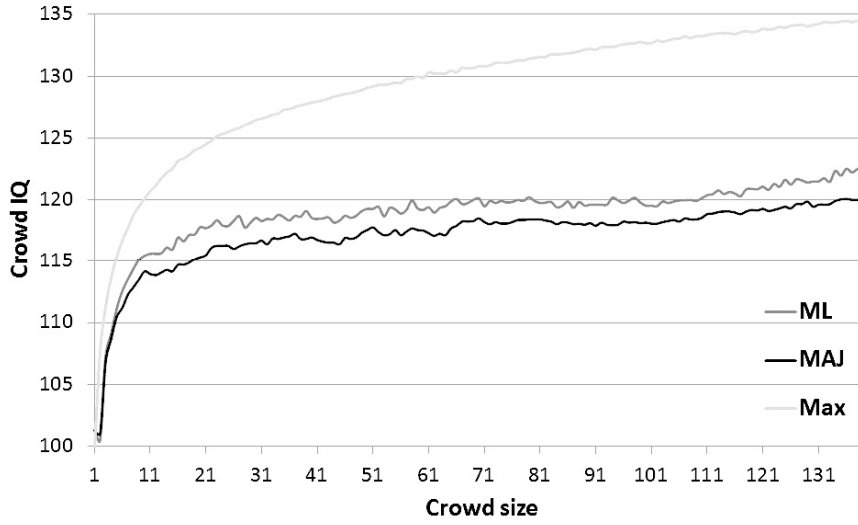

Figure 4. Crowd IQ scores based on the MAJ and ML aggregators for different crowd sizes. The crowds were drawn randomly and repetitively from the sample [10].

On the other hand, a person that is capable to solve the items that the majority of the participants fail to solve but fails in solving the commonly solved items will contribute more to the crowd's IQ.

Although adding the participant with the highest IQ score is a good heuristic, better results can be achieved when the contextual IQ of the participant is considered. The aggregating algorithm influences the contextual IQ. MAJ was used to produce the chart in Figure 6. The crowd IQ of all the participants was slightly higher under the ML aggregator. Bachrach et al. also investigated the dependence of the participants contextual IQ on the aggregator. Figure 7 shows a high correlation between participant's contextual IQ under the MAJ and ML aggregators (correlation coefficient of over 0.95) implying that in this case the key factors affecting contextual IQ are the participant's IQ and the uniqueness of the participant's contribution.

Wolf et al. [14] have investigated the influence of collective decision making in the contexts where dichotomous decisions are to be made. The decision accuracy of a solitary decision maker is fundamentally constrained by the trade-off between true and false positives: a high rate of true positives is possible only at the cost of a high rate of false positives; conversely, a low rate of false positives is possible only at the cost of a low rate of true positives. Wolf et al. used an integrated theoretical and experimental approach to show that a group of decision-makers can overcome this basic limitation. In the generic decision-making context the decision maker has to decide whether or not to take an action A, depending on the state of the environment being 0 or 1. The decision maker does not know the state of

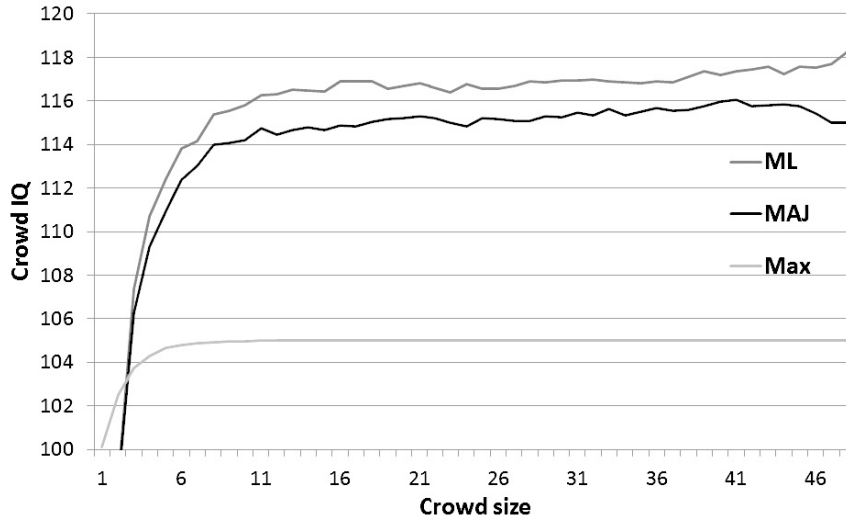

Figure 5. Crowd IQ and maximal IQ for the groups with IQ of the members falling between 95 and 105 (homogeneous) [10]

the environment but perceives a cue of intensity × which provides some information about the state of the environment (Figure 8). An example of such context is medical decision making what is the context of our experiments that we describe below. In that case the doctor (the decision maker) decides whether to apply treatment or not (the action) based on the symptoms, medical images, etc. (the cue), reflecting whether the disease is present or not present (the state of the environment). Wolf et al. have shown that a group of decision-makers can both increase true positives and decrease false positives simultaneously.

Wolf et al. first showed mathematically that, compared with solitary decision-makers, a simple quorum decision rule allows decision-makers in groups to increase true positives and decrease false positives simultaneously. First, every simulated individual in the simulated group takes a decision. Then the individual observes how the other individuals in the group have decided. The quorum decision rule allows the individual to stick to the decision to take action when at least a certain proportion of the group (quorum q, Figure 9) have decided to take action and to change the decision otherwise. Both, the group size and the quorum threshold influence the result. The key assumptions in the model are that the true-positive rate of a solitary decision-maker is higher than its false-positive rate and that the perceived cue intensity by different decision-makers is independent from each other.

The authors then tested the consistency of the predictions experimentally on a total of 436 human participants divided into 24 groups. The model task was predator detection. Each group of individuals was presented an image of a school

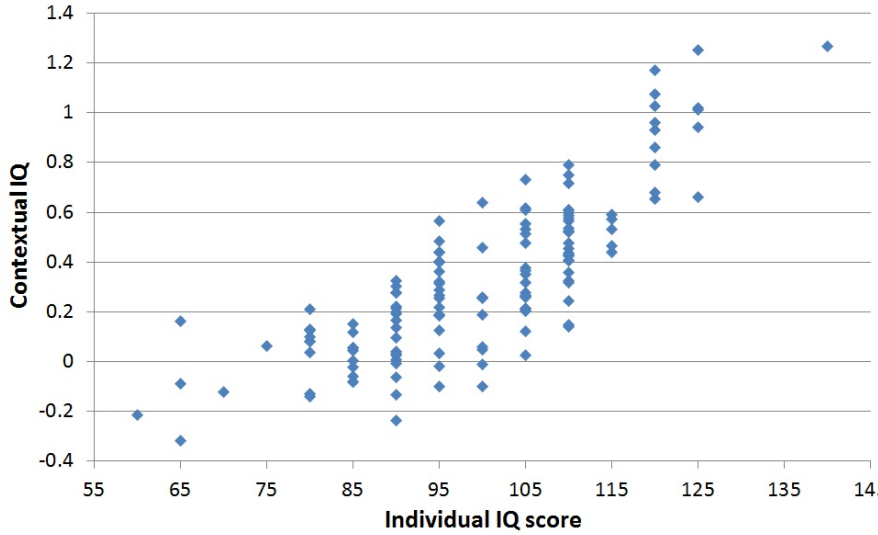

Figure 6. IQ and Contextual IQ [10]

of 144 fish aligned in a 9×16 grid. All fish in this school were identical, except one odd fish, which had either six or seven spines while the remaining fish had no spines. The subjects had to decide whether to "stay" (no odd fish observed or it has 6 spines) or to "escape" (odd fish with 7 spines observed) in a time limit of 5 s after the 2 s of observation have elapsed. The decision was recorded via a keypad (polling 1).

The participants were then presented for 5 s with a bar chart showing the number of individuals that decided to escape. The participants were then asked to decide again (polling 2). Finally the participants were presented with the results of the second polling and the correct answer (to stay or to escape). Consistently with the prediction, the participants achieved higher true positives and lower false positives in the second polling. When comparing polling 2 with polling 1, the average true positive of individuals increased in all of the 24 groups (Figure 10 a)), the average false positive of individuals decreased in 20 of 24 groups (Figure 10 b)). Increasing the group size increased the true positives and decreased the false positives. Wolf et al. provided strong supportive evidence that the participants used quorum responses in their decision to escape in polling 2 based on the social information provided after polling 1 and that individuals adjust their quorum adaptively to the performance of the group. Wolf et al. have confirmed that collective intelligence is possible only when the decision makers evaluate the identical pieces of information differently, dependent on their experience and cognitive style. The decision accuracy in contexts such as medical decision-making may be improved by group decisions in which decision-makers make decisions without a prior ex-

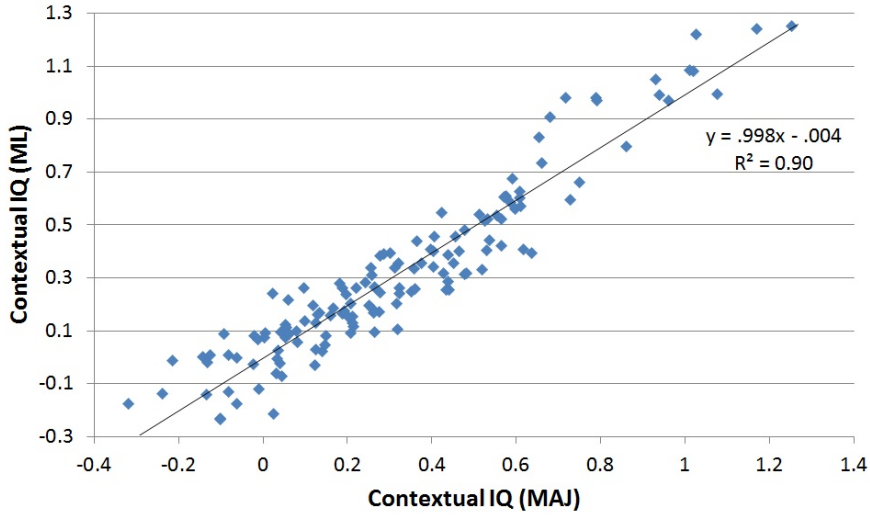

Figure 7. Contextual IQ under the majority and machine learning aggregator [10]

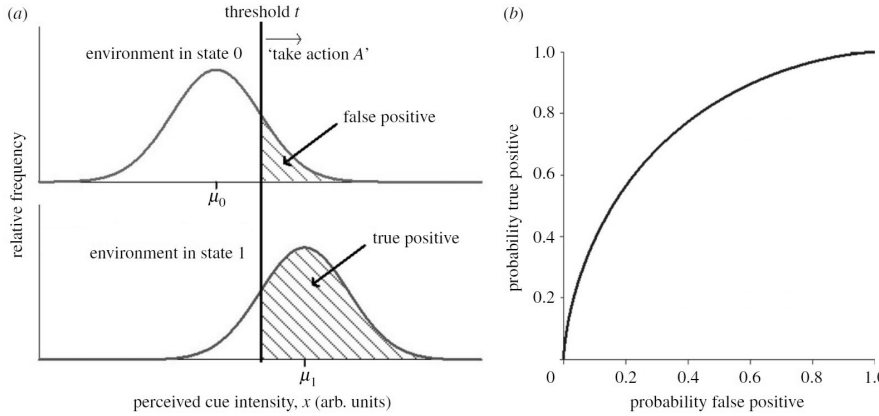

Figure 8. Generic dichotomous decision-making context [14]

change of opinions and differ in their experience, cognitive style and/or personality.

The concept of collective intelligence is not reserved to the systems containing humans. Also systems comprising artificially intelligent entities (or mixed systems) may act collectively, sharing information and responsibilities. Kvasnička and Pospíchal [17] have investigated an application of artificial intelligence techniques to get a better and deeper understanding of Halbwachs concept of "collective memory" [18]. They found that the concept of collective memory forms a base for pro-

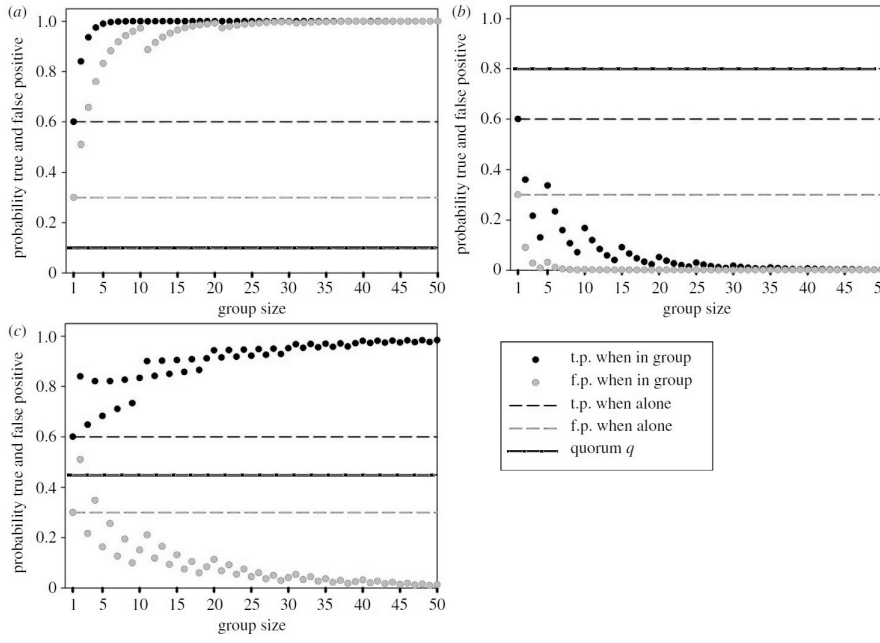

Figure 9. Intermediate quorum thresholds allow decision-makers in groups to overcome the trade-off between true and false positives. [14]

duction of an effective tool for acceleration of adaptive and evolutionary problems in multiagent systems. The concept of collective memory as used in social sciences and in artificial intelligence has many common properties.

3 EXPERIMENTAL DATASET

Our aim was to build an experimental setup that would allow us to collect data and evaluate various methods for aggregating opinions of groups of people. With regard to the technical education of the volunteering participants who were the students of the Faculty of Electrical Engineering and Informatics at the Technical University in Košice, Slovakia, we have chosen a non technical model problem we have assumed the most of the participants will have very little experience with. The model problem was lung nodules detection in X-ray images. For this study the participants had to state whether there is or there is not a lung nodule present in the image. Our assumption was that the participants will initially perform slightly better than a random classifier and after a short education improve their accuracy. Thus we were able to evaluate the impact of collective decision making on groups composed of individuals with various success rates. The X-ray images used were taken from the standard digital image database with and without chest lung nodules [15], cre-

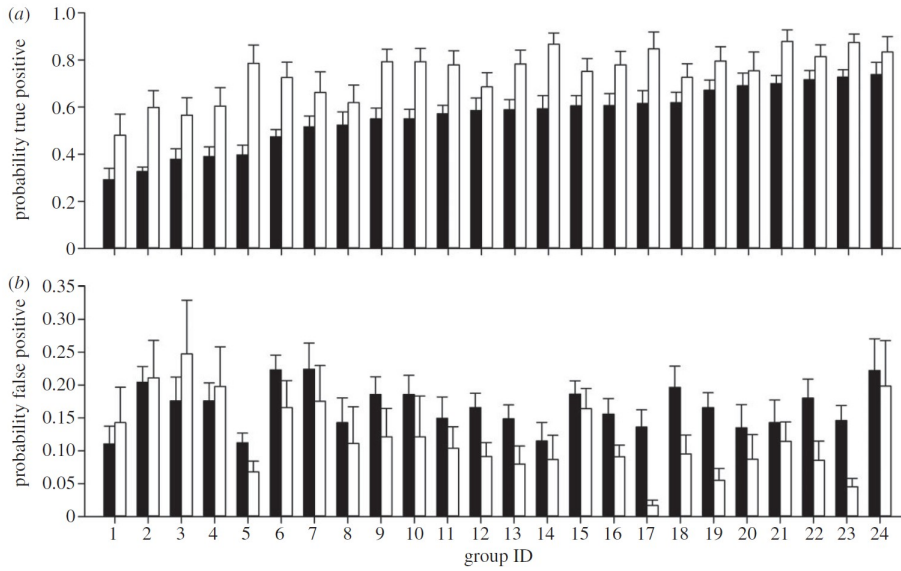

Figure 10. After taking the decision of other group members into account (polling 2, white bars), individuals both increase their true positives and decrease their false positives. Polling 1 results are show by the black bars [14].

ated by the Japanese Society of Radiological Technology in cooperation with the Japanese Radiological Society in 1998. The database consists of 154 nodule and 93 non-nodule high resolution 12 bit grayscale images (2 048 × 2 048 matrix size, 0.175 mm pixel size). Every image is provided with additional information on patient's age, gender, diagnosis (malignant or benign), X and Y coordinates of the nodule, degree of subtlety in visual detection of the nodule (1–5, 0 means there is no nodule). Each image contained zero or one lung nodule. We have built a web based application to collect data independently from the participants. Screenshot is in Figure 11. Data were then stored in a database and analyzed.

4 THE AGGREGATION METHODS USED

We have used three aggregation methods. First we will present the artificial intelligence based method derived from the adaptive boosting (AdaBoost) approach [15]. Yoav Freund and Robert Schapire won the Gödel Prize in 2003 for their work on this machine learning meta-algorithm. Adaptive boosting is used with many types of learning algorithms. The underlying idea is that the output of the learning algorithms that are assumed to perform at least slightly better than random guessing and are called "the weak learners" is combined into a weighted sum representing the final output. The original implementation of AdaBoost generates the weak learn-

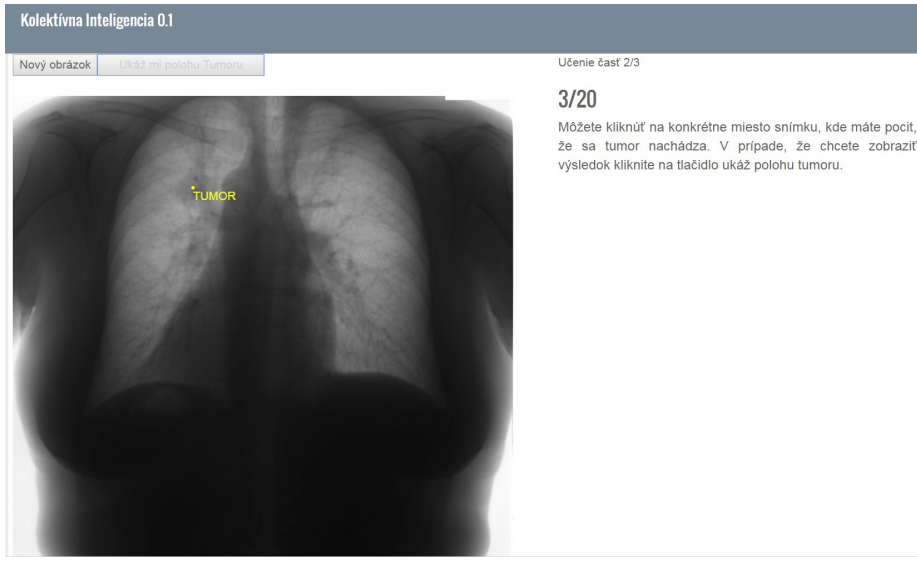

Figure 11. Screenshot of the web application used to collect the participants' responses. Instructions are in Slovak language.

ers gradually so that the training of the subsequent learners is altered in favor of the learning instances that were misclassified by the previous learners. We see this meta-algorithm as a form of collective intelligence where the opinions of the weak learners are aggregated. Each learner processes the information independently and has slightly different "cognitive abilities" (as a result of the altered training conditions). These features comply with the predispositions for forming a collective intelligence behavior as formulated above [9, 10, 14]. In our case we consider the participants to be the weak learners. We do not train the participants on particular and specific examples but we use the confidence measures that are calculated by the AdaBoost.

The modified adaptive boosting (MAB) algorithm that we have proposed for the aggregation of opinions considers the participants to be the weak hypotheses $h_t(x)$. Every participant provides his/her predictions for all presented images. The participant's predictions on a subset of m of the presented images $x_1, \ldots, x_m$ are used to set the parameters of the MAB model. We call this subset of the images the *training* set because it is used to train the MAB model. The participants were first responding to the images without being educated on the topic. Selected participants were then provided some education on the nodule detection and if a subset of the images was used in this process we call this subset the *educational* subset to avoid confusion. The training and the educational subsets are mutually exclusive. The correct responses for the training images are designated $y_1, \ldots, y_m$ and have values of 1 or -1, (nodule or no nodule present). The MAB meta-algorithm is built with

a group of T participants and the selected training images. For any given group of T members and for any given image x MAB combines the weak hypotheses $h_t(x)$:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{1}$$

where:

- h_t is the weak hypothesis (the participant is providing it),
- x is the input (the presented image in our case),
- $\alpha_t \in R$ are the parameters that have to be found.

The machine learning AdaBoost trains the weak hypotheses on different subsets of the examples available. We did not educate the participants this way although it is possible, more complicated and the resulting MAB is limited to use the same group of the participants. It is the subject of the future work. The participants naturally vary in their performance. We have decided to build the MAB models with the assumption that the inherent variability will at least partially replace the different education. The goal is now to assign the appropriate parameters to the weak hypotheses. For the given training set: (x_1, y_1), ..., (x_m, y_m) where y is the expected output the weights $D_1(j)$ for $j = 1, \ldots, m$ are initialized:

$$D_1(i) = 1/m. \tag{2}$$

Let us create a list $\boldsymbol{h}$ containing the items initialized as $[h_t, \alpha_t = 0, \varepsilon_t = -1]$. Let us create an empty list $\boldsymbol{h'}$. MAB continues in a loop of the Algorithm 1.

for $t \leftarrow 1$ **to** T **do**
  for $i \leftarrow 0$ **to** *number of elements in* $\boldsymbol{h}$ **do**
    Calculate ϵ as the sum of $D_t(j)$
    for all $j = 1, \ldots, m$, where $h(j)$ was incorrect
  end
  find the item in the list $\boldsymbol{h}$ with the ε_{min} = minimal ε
  calculate α of the item: $\alpha = 0.5 \log((1 - \varepsilon_{min}/\varepsilon_{min})$
  transfer the item from the list $\boldsymbol{h}$ to the list $\boldsymbol{h'}$
  recalculate D where Z_t is a normalizing factor so that D_{t+1} is a probability distribution:

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t} \tag{3}$$

end

Algorithm 1: Preparing for the aggregation

Now the list $\boldsymbol{h}$ is empty and the list $\boldsymbol{h'}$ contains the parameters of the MAB that are used to calculate the aggregate output of the group (Equation (1)). An important difference to the machine learning AdaBoost is that the participants provide crisp predictions (a nodule or no nodule) instead of real valued output reflecting (at least theoretically) the probability of the correct answer (what would be a confidence in one's decision in our case).

We have used two other methods to compare the proposed MAB method to. One was a simple majority voting aggregation (MV) that considers opinion of the every member of the group to be equal and the most frequent prediction is considered the aggregate prediction. If there is even count of positive and negative predictions the aggregate prediction is chosen randomly. This may occur in the groups with even number of the members only. MV does not take into account the probability of the correct prediction of the individual members. The last aggregation method we have used is weighted voting (WV) that takes the reliability of the individual members into account. The reliability R_t of a group member t is a function of the individual's probability of a correct answer P_t calculated statistically based on the participants predictions made on the training images:

$$R_t = 2\left(P_t - 0,5\right). \tag{4}$$

And the aggregate output is:

$$H\left(x\right) = sign \sum_{t=1}^{T} R_t h_t\left(x\right). \tag{5}$$

5 EXPERIMENTAL ANALYSIS

We have collected 45 voluntary participants in the initial experiments. The participants were the students of the Technical University in Košice, 37 males and 8 females aged 19 to 24 years. The participants had no previous experience with lung nodule identification. The presenter explained to the participants what are they about to do and how do they log in into the web application that we have implemented to collect their responses. The participants were then presented with 30 X-ray images (15 contained a lung nodule) and they decided for each of them whether they see or they do not see a lung nodule in it. The participants also designated the position of the nodule when they saw it. All participants responded to the same set of images but presented in different, random order. The participants were then educated on lung nodule detection. The application presented them several images with lung nodule position shown. After this education the participants responded to another set of 30 X-ray images (15 of them again contained a lung nodule). The image sets used in the pre-education, education and post-education phases were mutually exclusive. The participants were also (falsely) informed that the images are randomly taken from a database and the number of positive and negative images varies

to avoid them having assumptions on the number of positive samples in the image sets. We have performed an initial test on 10 randomly chosen participants to see how the aggregating algorithms perform on a heterogeneous group (performance-wise) and see whether education improves their performance. The accuracy of the participants was calculated as the percentage of correctly classified images from all images in the set. Because there was 1:1 ratio of positive and negative images, the accuracy of 50 % corresponds to random guessing. The accuracies of the participants varied greatly. The results are in Table. 2. We have observed that the process of education improves the performance of most participants.

	Accuracy before Education in %	Accuracy after Education in %
Max. accuracy in the group	70.00	80.00
Min. accuracy in the group	52.22	52.41
Average accuracy	59.48	67.83
MAB aggregate	70.00	77.96
MV aggregate	68.33	76.11
WV aggregate	65.00	80.37

Table 2. Comparing classification accuracies before and after education of the participants

The proposed MAB aggregator performed almost as good as the best performing member of the group. The findings published in [10] indicate that the presently used aggregation methods do not outperform the best individual of a heterogeneous group. Next we have investigated the collective intelligence of the groups composed of the similarly performing participants. We have considered the percentage accuracy of a participant achieved on a subset of the presented images to be the only performance measure. The goal of the experiment was to observe the relationship between the collective intelligence and the composition a group. For this experiment we have formed several groups of the participants based on their performance falling in the given range. Randomly selected subset of the responses of each participant was used to set the parameters of the MAB and WV models and collective intelligence was calculated. This process was repeated 30 times and the collective intelligence of the group was calculated as the average of all values. The results were compared to two heterogeneous groups of 12 and 44 participants (HTR 1 and 2). Table 3 summarizes the experimental results. Please note that the participant's accuracy used to select him/her for a given group was calculated using all responses of the participant and the accuracy of the best performing participant in group was calculated over a randomly chosen subset of the responses. Therefore the listed maximum accuracy of the group's top performing member may exceed the group's stated homogeneity interval's top boundary. The experimental results showed that the more complicated MAB aggregator does not generally outperform the simple MV aggregator. The WV aggregator performed the worst, although it takes into account the group members' reliability in similar way as MAB does. For all testing groups, the accuracy of either MAB or MV aggregators exceeded the accuracy of

the group's top performing member. MAB consistently outperformed the group's top performing member for homogeneous and heterogeneous groups.

	Accuracy in %								
	Accuracy ranges for the homogenous groups							HTR 1.	HTR 2.
	50–55	50–60	60–70	65–70	65–75	70–83	75–83		
Group's top performing member (GTM)	56.1	60.6	70.5	71.5	74.8	81.4	82.0	79.4	81.5
Group's average accuracy (GAA)	53.2	55.5	65.7	67.9	68.6	75.1	78.4	69.5	65.4
MAB aggregate average	64.4	67.1	82.5	84.7	83.2	89.4	89.4	83.9	86.9
MV aggregate average	54.9	70.3	84.6	89.4	82.9	86.9	84.8	85.4	79.6
WV aggregate average	46.1	50.6	79.5	83.0	81.4	86.9	86.3	83.7	83.6
100*(MAB-GTM)/GTM	14.8	10.7	17.0	18.5	11.2	9.8	9.0	5.7	6.6
100*(MAB-GAA)/GAA	21.1	20.9	25.6	24.7	2.13	19.0	14.0	20.7	32.9
100*(MV-GTM)/GTM	-2.1	16.0	20.0	25.0	10.8	6.8	3.4	7.6	-2.3
100*(MV-GAA)/GAA	3.2	26.7	28.8	31.7	20.8	15.7	8.2	22.9	21.7

Table 3. Comparison of the aggregating methods performances on homogenous and heterogeneous groups. The top performing method is highlighted in every column.

We have calculated the percentage increase of the aggregate accuracy over the group's best and the group's average. The values are in Table 3. below the double line. We have used these values to construct the graph in Figure 12. The collective intelligence exceeded the group's top performing member and the group's average the most in the groups composed of the members performing close to the general average of approx. 67 % accuracy. In the contrary to the findings of Bachrach et al. [10] the MAB and MV aggregators outperformed the top performing members also for the heterogeneous groups. We are in the process of collecting a much larger dataset and these findings need to be verified. We have also investigated the influence of the group size on the quality of the aggregate output. We have randomly generated groups of various sizes, calculated the aggregate output and averaged the values for the given group size. We have considered the participants before and after education to represent distinct hypotheses thus virtually doubling the number of the participants. This is not ideal but justifiable because all participants were educated on different image sets. Certainly, the results must be verified using a larger number of participants. Figure 12 shows the experimental results. All aggregation methods

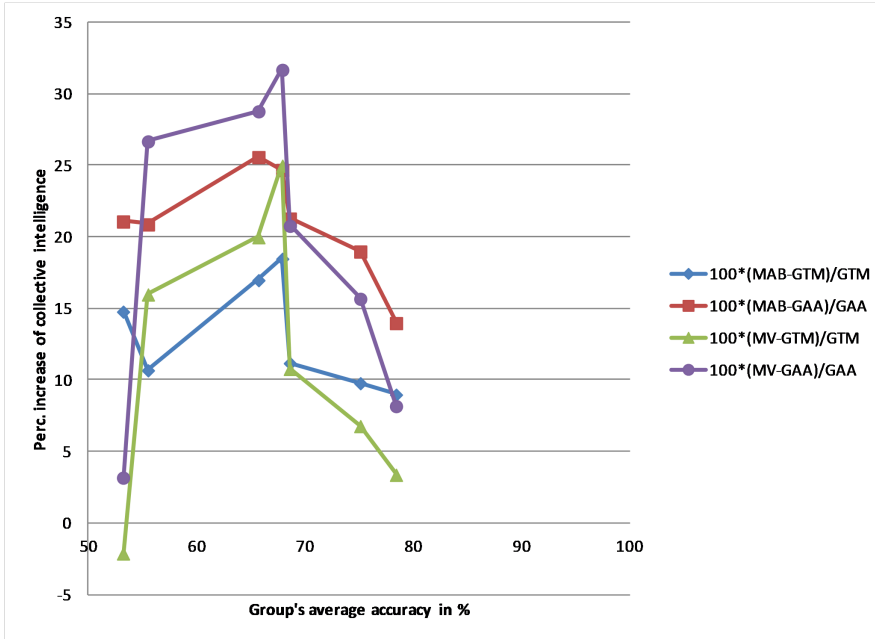

Figure 12. The increase of the collective intelligence (aggregate accuracy) over the group's top performing member and the group's average accuracy measured for homogeneous groups as of Table 3

outperformed the best performing member of the group with no significant differences among each other. At first there was an increase of the collective intelligence with the group size but a plateau was soon reached at the group size of 17 people. This is consistent with the findings of Bachrach et al. [10]. MV and WV methods did not outperform the GTM of the maximum size group. MAB continued to perform better than GTM. If this trend would continue it must be tested in a larger experiment. Theoretically, the principles of adaptive boosting give us the potential to fully utilize the potential of arbitrarily large groups of people. We have mentioned that when the participant answered "yes, there is a nodule" he/she had to mark the location of the nodule by a mouse click. So far we have considered the participants to be dichotomous classifiers and did not address the question what are their answers based on before and after the education. We have investigated the cases when the participant answered positively and there was in fact a nodule in the X-ray image (true positives). We have compared the location of the nodule marked by the participant with the ground truth and if the distance of the coordinates was less than the average nodule radius plus 20 % we have considered the positive answer to be founded, otherwise to be unfounded. Before the education, an average of 31 % of the true positives was founded. This means that the remaining 69 % of

unfounded decisions was based either on random choice or on other features of the X-ray image than the nodule presence. We did not investigate the reasons further at this point. After the education the overall accuracy of the participants has increased. The percentage of the founded true positives increased from the average of 31 % to the average of 52 %.

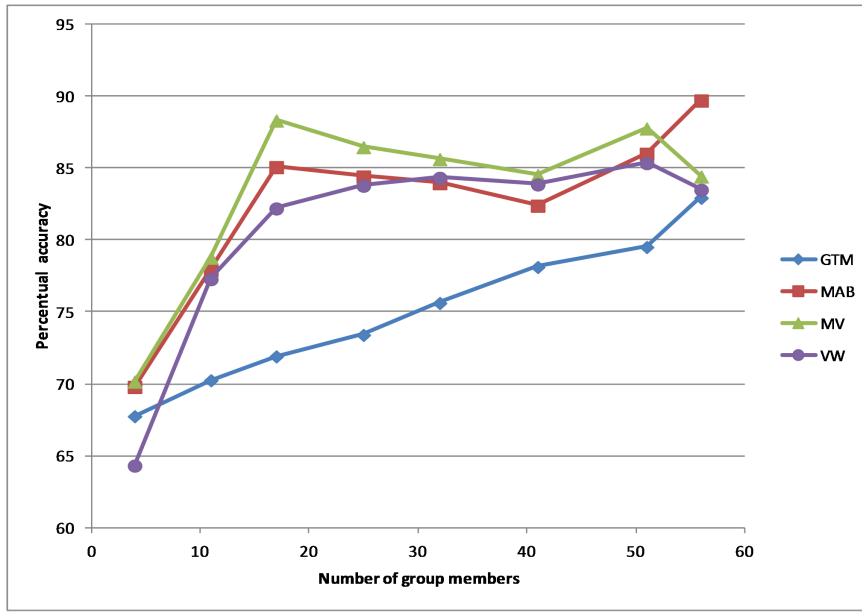

Figure 13. Collective intelligence vs. the number of the group members

6 CONCLUSION AND FURTHER RESEARCH

We have focused on the development and testing of the methods for aggregation of opinions of a group of people. First, we have provided an overview of the state-of-the-art of the related research. The widespread term "collective intelligence" is used in several contexts. We have proposed an opinion aggregation method based on machine learning meta-algorithm AdaBoost and we use it to make a set of intelligent entities (people, in our case) behave and act as a single entity thus creating a collective intelligence. The model problem was lung nodule identification in digital X-ray images of the chest. The problem enabled us to investigate not only the correctness of the individual answers but also whether the answers were founded or unfounded. We have recruited 45 voluntary participants for the experiment. We have compared the proposed MAB method to simple aggregation methods. The experimental results showed that none of the used aggregators performed consistently better than another. We have investigated the aggregate out-

put of performance-wise homogeneous and heterogeneous groups. The aggregate output of a group was consistently better than the output of the best performing member of a group. This contradicts the findings of Bachrach et al. [10] who found that aggregate output outperforms in general the group's top member for homogeneous groups only. We have confirmed that increasing the group size increases the collective intelligence up to a certain limit. After this limit is reached increasing the group size does not increase the collective intelligence. The results indicate that MAB aggregation may overcome this limit and utilize an arbitrarily sized group to the full potential but this hypothesis must be verified in a larger experiment. We have also found that it is possible to build a relatively reliable collective X-ray image classifier with uneducated participants although the majority of their correct answers were unfounded. We have presented what we consider a base for our future research. The adaptive boosting principle may be implemented in several ways for the aggregation of the opinions. We have used a simple and straightforward way. The participant's confidence into his/her answer was not taken into account but we will collect and utilize this information in the future. The group members were not trained on specifically selected images as it is done by the machine learning AdaBoost. This is more complicated with the human participants because the group members must be trained consecutively but this option is also to be tested. We have not aggregated the opinions on the location of the lung nodule yet. We prepare an implementation of the AdaBoost based aggregation that provides a real valued output to do this. We have also upgraded the web application and we are in the process of collecting a larger set of answers.

Acknowledgement

This publication is the result of the implementation of the following projects: Grant KEGA – 001TUKE-4/2015 (60 %), University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, the second phase of project USP TECHNICOM, supported by the Research & Development Operational Programme funded by the ERDF (20 %) and VEGA 2/0154/16 Networked control of heterogenous multi-agent systems (20 %).

REFERENCES

[1] Marquis de Condorcet: Essai sur l'Application de l'Analyse a la Probabilité des Décisions Rendues a la Pluralité des Voix. 1785, PNG Scan (in French). Available at: http://gallica.bnf.fr/ark:/12148/bpt6k417181, retrieved 2016-05-10.

[2] Wheeler, W. M.: The Ant-Colony as an Organism. Journal of Morphology, Vol. 22, 1911, pp. 307–325.

[3] Wenseleers, T.: The Superorganism Revisited. BioScience, Vol. 59, 2009, No. 8, pp. 702–705. Available at: https://doi.org/10.1525/bio.2009.59.8.12.

[4] Atlee, T.—Por, G.: Collective Intelligence as a Field of Multi-Disciplinary Study and Practice. 2000, retrieved from http://www.community-intelligence.com/files/Atlee, 2016-05-10.

[5] Lévy, P.: Collective Intelligence: Mankind's Emerging World in Cyberspace. Perseus Books Cambridge, MA, USA, 1997, ISBN 0306456354.

[6] Hiltz, S. R.—Turoff, M.: The Network Nation: Human Communication via Computer. MIT Press Cambridge, MA, USA, 1993, ISBN 0262581205.

[7] Woolley, A. W.—Aggarwal, I.—Malone, T. W.: Collective Intelligence and Group Performance. Current Directions in Psychological Science, Vol. 24, 2015, pp. 420–424, DOI: 10.1177/0963721415599543.

[8] Malone, T. W.—Bernstein, M. S.: Handbook of Collective Intelligence. MIT Press, 2015, ISBN 0262029812.

[9] Woolley, A. W.—Chabris, Ch. F.—Pentland, A.—Hashmi, N.—Malone, T. W.: Evidence for a Collective Intelligence Factor in the Performance of Human Groups. Science, Vol. 330, 2010, No. 6004, pp. 686–688, DOI: 10.1126/science.1193147.

[10] Bachrach, Y.—Graepel, T.—Kasneci, G.—Kosinski, M.—Gael J. V.: Crowd IQ – Aggregating Opinions to Boost Performance. 2012, retrieved from http://research.microsoft.com/pubs/159452/AIQ.pdf, 2016-05-10.

[11] McGrath, J. E.: Groups: Interaction and Performance. Prentice-Hall, Englewood Cliffs, NJ, 1984, ISBN 0133657000.

[12] Quinn, K.: Ravens Progressive Matrices. Retrieved from http://www.testprepforravens.com/, 2016-05-10.

[13] Koller, D.—Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009, ISBN 9780262013192.

[14] Wolf, M.—Kurvers, R. H. J. M.—Ward, A. J. W.—Krause, S.—Krause, J.: Accurate Decisions in an Uncertain World: Collective Cognition Increases True Positives While Decreasing False Positives. Proceedings of the Royal Society B, 2013, Issue 280. http://dx.doi.org/10.1098/rspb.2012.2777.

[15] Shiraishi, J.—Katsuragawa, S.—Ikezoe, J. et al.: Development of a Digital Image Database for Chest Radiographs with and without a Lung Nodule: Receiver Operating Characteristic Analysis of Radiologists' Detection of Pulmonary Nodules. American Journal of Roentgenology, Vol. 174, 2000, No. 1, pp. 71–74.

[16] Freund. Y.—Schapire, R.—Abe N.: A Short Introduction to Boosting. Journal of Japanese Society for Artificial Intelligence, Vol. 14, 1999, pp. 771–780.

[17] Kvasnička, V.—Pospíchal, J.: Artificial Intelligence and Collective Memory. Emergent Trends in Robotics and Intelligent Systems. Advances in Intelligent Systems and Computing, Vol. 316, 2015, pp. 283–291, Print ISBN 9783319107820.

[18] Halbwachs, M.: On Collective Memory. The University of Chicago Press, Chicago, 1992, a translation of original French edition, La Mémoire Collective, 1950.

Marek Bundzel is with the Technical University of Košice, Department of Cybernetics and Artificial Intelligence, Slovakia. He is active in teaching and research. His expertise includes mainly methods of computational intelligence like artificial neural networks, support vector machines and evolutionary algorithms and applications of the above methods in pattern recognition, forecast and robotics. He has spent two years at Waseda University, Tokyo where he was developing a model based on the memory-prediction framework, a theory of brain function, for the purposes of object recognition in mobile robots.

Jozef Lacko has studied bussiness informatics at the Technical University of Košice. His master thesis focused on application of artificial intelligence to enhance collective intelligence. During his studies he has participated in live projects with USS Košice as a software developer. He was at the internship at The University of Tokyo, researching the efficiency of parallel computation.

Iveta Zolotová graduated at the Department of Technical Cybernetics of the Faculty of Electrical Engineering, Technical University of Košice, Slovakia in 1983. She defended her C.Sc. in the field of hierarchical representation of digital image in 1987. Since 2010 she has been working as Professor at the Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia. Her scientific research is focused on networked control and information systems, supervisory control, data acquisition, human machine interface and web labs. She also investigates issues related to digital image processing.

Tomáš Kasanický is a researcher at Institute of Informatics of the Slovak Academy of Sciences and a Ph.D. student at Faculty of Electrical Engineering and Information Technology, Slovak University of Technology, Slovakia. His main research areas are artificial neural networks, artificial immune systems and bioinspired algorithms. He has participated in several national and European research projects.

Ján Zelenka received his M.Sc. and Ph.D. degrees from Faculty of Electrical Engineering, Department of Control and Information Systems, Zilina University in 2006 and 2009, respectively. He is currently a researcher at the II SAS. His research interests include stochastic search techniques, biologically inspired methods for control and coordination of a group of mobile robots and scheduling process optimisation. He has published several papers in various international conference proceedings and a chapter in a book in the mentioned areas.