Proceedings of the 2013 International Conference on Social Computing,
Behavioral-Cultural Modeling, & Prediction (SBP 2013) , 513-522, Washington, DC, April 2013

Exploiting User Model Diversity in Forecast Aggregation

H. Van Dyke Parunak, Sven A. Brueckner, and Elizabeth Downs

Soar Technology
3600 Green Court, Suite 600
Ann Arbor, MI 48105 USA
van.parunak@soartech.com

Abstract. In many contexts, people generate forecasts about events of interest, and decision-makers wish to aggregate these forecasts to improve their accuracy. These forecasts differ from signals in the physical sciences. In particular, sensor signals are noisy samples from a common underlying distribution, while human-generated forecasts are based on cognitive models that vary from one informant to another. As a result, human forecasts, unlike physical signals, are *not* guaranteed to be statistically independent conditioned on the true outcome. These differences both provide new opportunities for aggregation, and impose restrictions that do not apply to physical signals. This paper describes the difference between forecasts and physical signals, outlines a strategy for exploiting these differences in aggregation, and demonstrates modest but statistically significant gains in the accuracy of aggregated forecasts using data from a large ongoing experiment in forecasting world events.

Keywords: Mental models, generated and interpreted signals, forecast aggregation

1 Introduction

In many contexts, people forecast events, and decision-makers wish to aggregate these forecasts in the hopes of improving their accuracy. An example question is, "Will Bashar al-Assad be removed from power before 31 December 2013?" Forecaster i's answer is a binomial distribution θ_i over outcomes {Yes, No}. Forecasts to questions with more than two outcomes are multinomial distributions.

The intuitive approach is to average the forecasts. With constant weighting, such an approach is called an Unweighted Linear Opinion Pool (ULinOP). One might vary the weights by a range of factors, including the confidence of individual forecasters [3] or the expected information gain that they have to offer [11].

This intuition, which assumes independence among individual forecasts, is misleading. Information from cognitive processes is likely to be very different from that from physical sensors, both qualitatively and statistically. In particular, it depends on idiosyncratic mental models through which people view the world. Appropriate aggregation across such data should take into account the degree to which informants are working with the same or different mental models.

Previous work shows that distinctive characteristics of forecasts derived from cognitive processes (Section 2) urge aggregation methods other than averaging (e.g., voting), and motivate estimations of differences between informants' mental models (Section 3). This paper demonstrates modest but statistically significant gains in forecast aggregation using such estimates in a voting framework (Section 4). Section 5 discusses main lessons.

2 Generated and Interpreted Signals

In this section, we distinguish two main categories of data sources [5] and summarize how differences in their statistical properties affect aggregation [8].

Most analysis methods assume information sources that perform like a physical sensor: given a characteristic of the environment (e.g., temperature t), the sensor "generates" a value $x = t + \varepsilon$, where the error ε is drawn from some distribution. A conventional decision-making process with such "generated" signals compares the signal with a threshold T, and answers "yes" or "no" depending on whether the signal exceeds the threshold. If queried repeatedly, such a source yields answers that are independent of one another, conditioned on the true value of the condition $t > T$.

Some information elicited from people may satisfy this model. For example, in magnitude estimation in psychophysics [4], people confronted with external stimuli (such as sounds of different loudness, or light of different brightness) function as a transducer, converting a simple stimulus into a correlated number.

This model is less satisfying in explaining how people assign probabilities to complex world events. Whether a dictator will leave office depends on a complex web of events, including the country's economy, the level of civil unrest, dissension within the government, and relations with other nations. An intelligence report for decision-makers consists not just of a probability estimate associated with the final event, but also a discussion of the various factors that influence this outcome. Trained analysts describe their work as weighing these factors and the relations among them. In other words, they claim to be *interpreting* the world through an internal cognitive model. The same kind of processing is performed by many AI ("artificial intelligence") systems, which manipulate symbolic information that is mapped onto statements about the world. We call such a signal, an "interpreted signal."

Each source of an interpreted signal may have a different cognitive model, attending to different features of the world. Typically, no single model includes all relevant features, and some features may be invisible to all informants. The responses generated by such a system are not independent, but are guaranteed to be negatively correlated, conditioned on at least one outcome [5].

Interpreted data differ from generated data not only in their correlations, but also in the relation of the optimal aggregation to individual forecasts. Each forecast is a vector of probabilities across the possible outcomes. The space of all such probabilities for a given problem is a "simplex," which is the line segment [0, 1] for a binary problem, an equilateral triangle for a problem with three outcomes, and so forth. For example, Fig. 1 shows three forecasts (*a, b, c*) against a three-outcome question. The

dashed line is the simplex, the space of all triples (p_1, p_2, p_3) such that $p_i \in [0, 1]$ and $\sum p_i = 1$. The corners of the simplex correspond to outcomes. *a* assigns 100% to the first outcome and nothing to the other two. *b* assigns 50% each to the first and third outcomes, and nothing to the second. *c* is the uniform forecast, assigning 33.3% to each outcome.

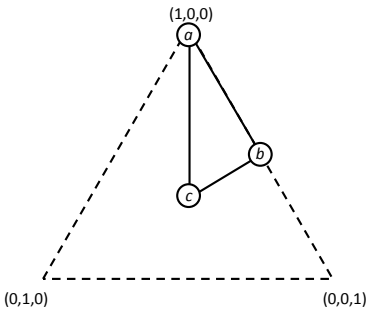

Fig. 1. Three forecasts on the trinomial simplex.

Any set of forecasts on the simplex has a convex hull (subject to certain information-geometric refinements [8]). In Fig. 1 the convex hull of (*a, b, c*) is a solid line.

Under benign constraints, the best aggregation of a set of generated estimates lies within their convex hull, while the best aggregation of a set of interpreted estimates can lie outside the convex hull [8]. Common aggregation methods (such as taking a weighted average with non-negative weights) are constrained to the convex hull, and cannot adequately process interpreted estimates. We explore an alternative approach, voting, which can leave the hull.

These differences between interpreted and generated signals highlight the importance of characterizing forecasters in terms of their individual cognitive models.

A caveat is in order. People can produce generated signals from low-level stimuli, and a response to a forecasting question may include both generated and interpreted components. Such a mixture might result, for example, from emotional stress in the respondent, which in turn might come from a question's subject matter, the complexity of the elicitation environment, or exogenous factors in the respondent's personal life. From our perspective, such a generated component is noise in the signal.

3 Estimating Model Differences

A cognitive model behind an interpreted signal describe the world in terms of *state variables*, mutually exclusive and collectively exhaustive *states*, or non-independent *statements* about the world [7]. Whatever the form, the underlying reasoning deals with a discrete set of features, and in realistic settings, different informants may not attend to the same subset of features.

Estimating a forecaster's actual model would be valuable. For example, a decision-maker who relies on an aggregated forecast is likely to be very interested in understanding the various mental models that support and oppose the conclusion that is presented. However, in this paper we focus simply on the *difference* between pairs of models, to guide aggregation of the resulting forecasts.

We distinguish three broad approaches to estimating model differences [9].

Some approaches to estimating model differences make assumptions about the internal structure of the models. For example, given a graph-structured model with statements as nodes and either conditional probabilities (a Bayesian belief network

[10]) or transition weights (a Narrative Space Model [7]) on the edges, we might ask forecasters what news items they found helpful in responding to a question, and use these to infer what statements a forecaster's model might contain.

At the other extreme, if forecasters are allowed to choose the questions that they address, it is reasonable to assume that forecasters will prefer questions on which their models can shed light. In our experiments, some forecasters are determined to address every question, so their selection does not give a distinct signature. However, for selective forecasters, we can use the overlap in questions between two forecasters to estimate the similarity of their models.

In this paper, we use an intermediate class of measure that requires only the actual forecasts θ_i, θ_j from two forecasters *i, j*. Each time we aggregate two forecasts from these forecasters, we compute a notional "distance" between their forecasts. One would like such a measure to be invariant with respect to invertible differentiable transformations of the set over which probabilities are defined. This condition allows only a 1-parameter family of geometries [2], characterized by the formula [13]:

$$D_\delta(P,Q) = \frac{1-\sum_i p_i^\delta q_i^{1-\delta}}{\delta(1-\delta)} \quad (1)$$

where δ is the parameter identifying the geometry. Well-known examples are the Hellinger divergence (for δ = 0.5) and the Kullback-Liebler divergence (or relative entropy, for δ = 1). We use the δ = 0.95 divergence, which we symmetrize (by averaging with the δ = 0.05 divergence) and normalize to yield a true distance in [0, 1].

Let us denote this difference for aggregation episode *k* as $w_{ij|k}$, or (where there is no risk of confusion) as w_{ij}. w_{ij} is the *within*-event distance between *i* and *j* at a single aggregation event. If forecasters *i* and *j* participate in multiple such events, we can compute a *cross*-event divergence,

$$c_{ij} = \frac{\sum_k w_{ij|k}}{N} \quad (2)$$

where *N* is the number of shared aggregation events. c_{ij} estimates the overall tendency of *i* and *j* to produce different forecasts on different questions, or on different responses to the same questions (which might reflect differential attention to news events relevant to their respective models). For two forecasters with identical models, $c_{ij} \approx 0$, recognizing the potential noise discussed at the end of Section 2.

Both w_{ij} and c_{ij} are in [0, 1]. How might they compare with each other?

- If both are high, we have different forecasts (high w_{ij}) from forecasters with different models (high c_{ij}), an unremarkable situation.
- We are also not surprised if both are low (similar forecasts from similar models).
- If w_{ij} is high and c_{ij} is low, two forecasters give different results even though their underlying models are similar. This circumstance raises questions about whether their models have purchase on this particular question (or perhaps whether one or the other of them has a high noise component in the forecast in question). In either case, we should discount their divergent contributions to the aggregate.

- If w_{ij} is low and c_{ij} is high, different models lead to a similar answer. Such a circumstance encourages us to pay more attention to their joint opinion.

4 Exploiting Model Diversity in Aggregation

This section outlines how to apply the insights from the previous sections, using diversity measures to modulate voting. We outline our methodology, then define the two dimensions of experimental space: how to select the outcome for which a vote is cast, and how to compute the size of the vote.

4.1 Evaluation Methodology

We analyze responses to 98 forecasting questions from various subsets of 169 forecasters, some responding multiple times to a single question, over time periods ranging from 2 to 245 days. Responses are collected through an on-line interface at https://ace-informed.net. Each question is aggregated once a day while it is active, using the most recent forecast from each forecaster.

We measure forecast accuracy with the Brier score [1], a quadratic error score in [0, 2]. We average the scores from daily aggregation events on a question to give an overall "mean daily error" (MDE) for that question, then average the MDE across all questions to give an overall score, the "mean mean daily error"(MMDE). We measure the performance of an algorithm as the percentage improvement of its overall score (the MMDE) compared to that for ULinOP.

Small improvements may result from a method's exceptional success on a few questions. To estimate the significance of improvements between two methods, we compare the question-by-question MDEs for each method to see whether the method with better MMDE also improves more individual questions. We use the nonparametric Wilcoxon signed rank test [12]. Each question contributes two scores, one from each method being compared. We rank the pairs by the magnitude of their difference. The test statistic W is the difference between the sum of the ranks for which the first method is better and the sum of the ranks for which the second is better. If the two methods are comparable, W should approach 0. The distribution of W approaches normality for $N > 10$, a condition clearly satisfied by our 98 questions, allowing us to estimate the probability of the null hypothesis (that the median of the differences between each pair of results is 0). Small values of $p(H_0)$ increase our confidence that one method is superior to the other.

4.2 Averaging vs. Voting

The ideal aggregate for interpreted signals can leave the convex hull of the individual forecasts [8]. Weighted averages with non-zero weights are confined to the hull, but voting methods can leave the hull. Can voting do better than weighted averages?

Let F_i be the set of forecasters who assign their largest probability to outcome j,

$$F_j = \left\{ i: \underset{k}{argmax}\left(\theta_i(k)\right) = j \right\} \tag{3}$$

Assign vote v_i to forecaster i. Then the voting aggregate is

$$\hat{f}_j = \frac{\sum_{i \in F_j} v_i}{\sum_i v_i} \tag{4}$$

With unit votes, $\forall i\colon v_i = 1$, voting outperforms ULinOP by 2.8%, with $p(H_0) = 0.006$. Substantially greater gains (~30%) are achievable by modulating the value of the v_i by various factors, such as a forecaster's historical accuracy, but we have found no benefit to using any diversity measure to change a forecaster-based vote. In the nature of the case, such a vote must be an average across the forecaster's diversity with respect to all other forecasters. We hypothesize that taking such an average discards important information, and we explore here *pairwise* voting methods in which each *pair* of forecasters casts a vote for an outcome derived from their individual forecasts. We explore three ways of selecting the outcome for a given pair of forecasters.

Average Pairwise Voting (APV) votes for the outcome favored by the average of the forecasts. If the forecasts differ, this approach selects the outcome favored by the more extreme forecaster. With unit votes, this approach is 7.4% better than ULinOP with $p(H_0) = 0.002$. To understand this gain, note that the more extreme forecast (the forecast with lower entropy) determines the outcome that receives the vote. Other experiments show that in general, more extreme forecasts (reflecting higher forecaster confidence [3]) are more likely to be correct. To remove this entropy effect, we define two alternative pairwise voting methods.

Split Pairwise Voting (SPV) assigns a vote of 0.5 to the outcome favored by each forecaster. When they agree on the outcome, the entire vote goes to their consensus, but when they do not, each outcome gets half of the credit.

Consensus Pairwise Voting (CPV) gives a vote to a pair only when they agree on the outcome, and registers no vote for pairs who disagree.

4.3 Using Estimates of Model Diversity

We experiment with four vote values. Three use a gain γ reflecting diversity.

The **unit vote** gives each pair has one vote. Combined with SPV, this approach yields the proportion of the forecasters who favor each outcome.

The **ratio γ vote** uses the ratio of c_{ij} and w_{ij}. To avoid singularities, we add a constant ζ (e.g., 0.1) to the numerator and denominator:

$$\gamma_{a|ij} = \frac{c_{ij}+\zeta}{w_{ij}+\zeta} \tag{5}$$

High values of γ_a indicate that the average forecast of i and j should be given higher credence, while low values indicate that the (divergent) forecasts merit less attention. The unit vote represents the limit of ratio γ when ζ → ∞.

The **difference γ vote** transforms (5) so it is bounded to [0, 2] and $c = w \rightarrow \gamma = 1$:

$$\gamma_{ij} = (c_{ij} - w_{ij} + 1) \tag{6}$$

To derive (6), since c and w are in [0, 1], γ is bounded by

$$\gamma_{max} = \frac{1+\zeta}{\zeta}, \gamma_{min} = \frac{\zeta}{1+\zeta} \tag{7}$$

First, we scale γ to [0, 1]:

$$\gamma' = \frac{\gamma - \gamma_{min}}{\gamma_{max} - \gamma_{min}} = \frac{\zeta}{1+2\zeta}\left(\frac{c+c\zeta+\zeta-w\zeta}{w+\zeta}\right) \tag{8}$$

Next, when $c = w$, the parenthesized term in the last element of Equation 4 becomes unity, so we reach γ' = 0.5 at the midpoint by allowing $\zeta \rightarrow \infty$. Taking this limit in (8) and multiplying by 2 so that γ = 1 at the midpoint, we obtain (6). Scaling before taking the limit retains the influence of c and w, avoiding a unit vote.

The **velocity γ vote** looks at how two forecasters are moving relative to one another on the simplex in their successive forecasts against a single IFP. This vote requires multiple forecasts from at least one forecaster. Consider three conditions:

1. If two forecasters move in synchrony with one another, they probably have very similar models. We will give them a unit vote (comparable to ratio γ when w and c are the same size).
2. If they move apart over time, they are responding differently to events they observe in the world, and we should downweight their vote (a value between 0 and 1).
3. If they move together over time, they started in different positions on the simplex (reflecting different models), but now are coming closer together, suggesting that we should give them a greater vote (say, between 1 and 2).

Let $w_{ij|k}$ be the distance between forecasters i and j at aggregation event k, where k is augmented each time one or the other of them updates a forecast to the IFP in question. When we are considering a single pair, we write simply w_k for their separation at event k. (9) has the required properties:

$$\gamma = 1 + \frac{w_{k-1} - w_k}{w_{k-1} + w_k} \tag{9}$$

When two forecasters start with the same forecast ($w_{k-1} = 0$) and move apart, (9) takes the value 0. When they start with different forecasts and move together ($w_k = 0$), it has the value 2. When they move in synchrony ($w_{k-1} = w_k$), it has the value 1.

The value varies reasonably with distance moved in all cases except when w_{k-1} or w_k is 0. In this case, regardless of the value of the other (non-zero) w, γ will be 0 or 2, depending on the direction of the move. To avoid this, we constrain w to be no less than a minimum value w_0 (0.01 in the experiments reported here). In addition, at the first aggregation event involving a pair of forecasters, we use γ = 1.

These diversity measures can be applied both to averaging and to voting aggregation. We have explored two averaging approaches. Naively, one might form all possible pairs of forecasts, and weight the average within each pair by γ:

$$\bar{\theta} = \frac{\sum_{ij} \gamma_{ij} (\theta_i + \theta_j)}{2 \sum_{ij} \gamma_{ij}} \tag{10}$$

Since γ_{ij} is symmetrical, we can define $\gamma_i \equiv \sum_j \gamma_{ij}$, and (10) becomes

$$\bar{\theta} = \frac{\sum_i \gamma_i \theta_i}{\sum_i \gamma_i} \tag{11}$$

This approach gives no benefit over ULinOP. Assigning a single aggregate γ to each forecaster throws away information about the value of specific pairs of forecasts.

An alternative approach, Cluster-Weighted Aggregation (CWA) [6], agglomeratively clusters individual forecasts using a weighted sum of c_{ij} and w_{ij} as the distance measure. It then aggregates forecasts following the structure of the dendrogram, starting at the leaves, and weighting each cluster by γ before aggregating it into higher-level clusters. This approach offers on the order of 5% improvement over ULinOP with either ratio or velocity γ. Difference γ offers no improvement.

5 Experimental Results

We have ten conditions to compare: ULinOP and 3x3 experimental conditions. Table 1 reports two numbers for each pair of conditions: the lift of the column method over the row method, and the statistical significance ($p(H_0)$ from the signed rank test over individual IFPs). The table is symmetrical, so we report only the upper half.

The various voting schemes with APV dominate everything else, because of the entropy factor, reflecting forecaster confidence. Even with this factor removed, unit voting dominates averaging, decisively for SPV ($p(H_0) = 0.018$), and suggestively for CPV ($p(H_0) = 0.053$). All three forms of γ give positive lift over ULinOP, significantly for difference γ in SPV, for ratio γ in CPV, and for velocity γ in both. Restricting our attention to SPV and CPV, velocity γ dominates the other two forms, and also

Table 1. Lift and Significance of Column over Row (only upper half-table reported). Green bold values have $p(H_0) \leq 0.05$, red values have $p(H_0) > 0.1$.

	APV, Unit	APV, Ratio	APV, Diff.	APV, Velocity	SPV, Unit	SPV, Ratio	SPV, Diff.	SPV, Velocity	CPV, Unit	CPV, Ratio	CPV, Diff.	CPV, Velocity	
ULinOP	**.074** **.002**	**.056** **.002**	**.072** **.005**	**.015** **.048**	**.030** **.018**	.013 .118	**.032** **.018**	**.072** **.003**	.015 .053	**.026** **.035**	.011 .061	**.034** **.013**	ULinOP
APV, Unit		-.018 .285	-.002 .463	-.059 .294	**-.044** **.002**	**-.062** **.001**	**-.043** **.001**	-002 .778	-.059 .353	-.049 .105	-.063 .411	**-.040** **.003**	APV, Unit
APV, Ratio			.015 .300	-.041 .257	**-.026** **.048**	**-.044** **.007**	-.025 .086	.015 .261	-.042 .269	-.031 .193	-.046 .317	-.022 .064	APV, Ratio
APV, Diff.				-.057 .269	**-.041** **.002**	**-.059** **.002**	**-.040** **.002**	.000 .526	-.057 .319	-.046 .096	-.061 .410	**-.038** **.004**	APV, Diff.
APV, Velocity					.015 .071	-.003 .061	.016 .071	.057 .378	.000 .851	**.010** **.023**	-.005 .155	.019 .085	APV, Velocity
SPV, Unit						**-.012** **.008**	**.001** **.025**	**.041** **.003**	-.016 .089	-.005 .065	-.020 .112	**.004** **.0001**	SPV, Unit
SPV, Ratio							**.019** **.010**	**.059** **.002**	.002 .074	**.013** **.049**	-.002 .074	**.022** **.003**	SPV, Ratio
SPV, Diff.								**.041** **.003**	-.017 .092	-.006 .087	-.021 .120	.003 .564	SPV, Diff.
SPV, Velocity									-.057 .355	-.046 .117	-.061 .442	**-.038** **.004**	SPV, Velocity
CPV, Unit										**.011** **.028**	**-.004** **.036**	.019 .109	CPV, Unit
CPV, Ratio											**-.015** **.015**	.008 .089	CPV, Ratio
CPV, Diff.												**-.029** **.038**	CPV, Diff.
	APV, Unit	APV, Ratio	APV, Diff.	APV, Velocity	SPV, Unit	SPV, Ratio	SPV, Diff.	SPV, Velocity	CPV, Unit	CPV, Ratio	CPV, Diff.	CPV, Velocity	

gives the highest lift against ULinOP (statistically indistinguishable from the impact of entropy in APV with unit voting).

These results invite three observations.

First, the signal in a forecast has an interpreted component that can be exploited in aggregation. Both voting (which can leave the convex hull of individual forecasts) and measures of diversity between forecasters' mental models give statistically significant improvements in forecast aggregation.

Second, our gains over ULinOP are modest. Our limited gains may reflect a generated-signal component in forecasts (e.g., an emotional response to questions). Also, we are hopeful that refinements can increase the contribution of our methods. For example, estimating c_{ij} from multiple instances of w_{ij} is sensitive to small-number effects when forecasters share only a few questions in common. We are exploring techniques for testing whether c_{ij} is converged, and using other estimates of cross-question diversity that are not subject to the convergence challenge.

Third, benefits gained from diversity effects are not always orthogonal to other benefits. There is a suggestion (with very low significance) that APV with any γ vote is worse than with unit voting. All forms of γ dominate ULinOP. Why don't they help with APV? The answer appears to be that the two effects fight against each other. Entropy (certainty) is only exploited when the members of a pair prefer opposite outcomes, while γ is high only when the members of a pair are close together, which usually means they favor the same outcome. Thus the two effects boost different pairs of forecasts, cancelling out the discriminatory information that each of them has to offer. We have observed such interference with other forecast features as well. An important focus of ongoing research is to understand these interaction effects and how to exploit combinations of features for greater performance improvement.

6 Acknowledgments

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20060. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, Dol/NBC, or the U.S. Government.

7 References

1. Brier, G.W.: Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review, 78(2) (1950)
2. Cencov, N.N.: Statistical Decision Rules and Optimal Inference. Rhode Island, American Mathematical Society (1982)

3. Forlines, C., Miller, S., Prakash, S., Irvine, J.: Heuristics for Improving Forecast Aggregation. In Sun, W., editor, AAAI Fall Symposium 2012: Machine Aggregation of Human Judgment, Arlington, VA (2012)
4. Gescheider, G.A.: Psychophysics: The Fundamentals. Third ed. Psychology Press (1997)
5. Hong, L., Page, S.E.: Interpreted and Generated Signals. Journal of Economic Theory, 144:2174-2196 (2009)
6. Parunak, H.V.D.: Cluster-Weighted Aggregation. In Sun, W., editor, AAAI Fall Symposium 2012: Machine Aggregation of Human Judgment, Arlington, Virginia (2012)
7. Parunak, H.V.D., Brueckner, S., Downs, L., Sappelsa, L.: Swarming Estimation of Realistic Mental Models. Thirteenth Workshop on Multi-Agent Based Simulation (MABS 2012, at AAMAS 2012), pages (forthcoming), Springer, Valencia, Spain (2012)
8. Parunak, H.V.D., Brueckner, S., Hong, L., Page, S.E., Rohwer, R.: Characterizing and Aggregating Agent Estimates. In Ito, T., Jonker, C., Gini, M., Shehory, O., editors, Twelfth International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2013), pages (forthcoming), IFAAMAS, Minneapolis, MN (2013)
9. Parunak, H.V.D., Downs, E.: Estimating Diversity among Forecaster Models In Sun, W., editor, AAAI Fall Symposium 2012: Machine Aggregation of Human Judgment, Arlington, Virginia (2012)
10. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. San Francisco, CA, Morgan-Kaufmann (1988)
11. Waterhouse, T.: Pay by the Bit: An Information-Theoretic Metric for Collective Human Judgment. In Sun, W., editor, AAAI Fall Symposium 2012: Machine Aggregation of Human Judgment, Arlington, Virginia (2012)
12. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin, 1(6):80-83 (1945)
13. Zhu, H., Rohwer, R.: Measurements of Generalisation based on Information Geometry. In Ellacott, S.W., Mason, J.C., Anderson, I.J. (eds.) Mathematics of Neural Networks: Models, Algorithms and Applications, Kluwer, 1997)