



HAL
open science

Combining Users Feedback as a Source of Knowledge for Feature Selection in Regression

Loujain Liekah, Emilie Morvant, Marta Soare, Amaury Habrard

► **To cite this version:**

Loujain Liekah, Emilie Morvant, Marta Soare, Amaury Habrard. Combining Users Feedback as a Source of Knowledge for Feature Selection in Regression. Hubert Curien Laboratory, UMR 5516 CNRS. 2018. hal-04277019

HAL Id: hal-04277019

<https://hal.science/hal-04277019>

Submitted on 13 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Combining Users Feedback as a Source of Knowledge for Feature Selection in Regression

Loujain Liekah

Advisors: Dr. Emilie Morvant Dr. Marta Soare Prof. Amaury Habrard

Abstract

In many machine learning applications, when we face the curse of dimensionality where the number of features is too large while the number of training examples is too small; relevant features selection is a critical need before applying any prediction models on our small data. To extract the maximum information out of the data in this setting; a possible and interesting source of useful information would be the experts of the field. The previous works introduced the benefits of querying one expert on the relevance and weights of the coefficients in sparse linear regression and how this feedback improves the prediction. But this method falls under the assumption that the voter has sufficient knowledge. Naturally, it is very hard to ensure the level of accuracy, which exposes the feedback to the risk of being biased. We define a way to calculate the accuracy of a user feedback and propose two methods to combine votes of multiple users on the relevance of features, calculate the accuracy of the final votes and study the reduction in the prediction error in a synthetic setting.

1 Introduction

The most essential factor in machine learning is data, gathered in different forms and combinations, organized or messy, structured or noisy. Therefore the larger the training sample is; the higher the possibility to learn an efficient model that will perform as expected on test data. Not having adequate training examples is a common problem in many text mining and bioinformatic problems where our small training set has thousands of features; that means high feature to sample ratio, and a high dimension search space for the learning algorithm. To challenge the curse of dimensionality and be able to generalize authentically; we need to reduce the dimension space by methods like principal component analysis or variables selection. Selecting variables in different ways like variable ranking using correlation or with nested subset selection methods; is to improve models runtime and get higher efficiency for the prediction on test data[4].

This work aims to develop an algorithm to combine the feedback of various users on the relevance of features and see the improvement of the transferred knowledge on the accuracy of the predictions by comparing MSE averaged values for different methods. This work is based on the previous research on the improvement of expert feedback on feature relevancy and value of coefficients in sparse linear regression [2], where the authors proposed a model that improved the prediction accuracy by querying one expert on the most informative features formulated as a probabilistic inference process.

Our study aims to extend this work by incorporating and combining feedback from several experts with different levels of accuracy in simulated settings. Our goal is to study which is the best method to merge expert opinions to gain additional needed information in "small n large p" scenarios, depending on variables such as the tuned levels of accuracy of the feedbacks, the number of voters feedback used to formulate the combined Majority Vote feedback and Field Focused Feedback.

Outline: The rest of the report is organized as follows. Section 2 explains some relevant previous works, Section 3 introduces the setting and notations used in this paper and present a model to calculate the accuracy of the combined feedback, Section 4 presents the contributions and explains the two methods proposed to combine multiple users feedbacks, Majority Vote and Field Focused Feedbacks, Section 5 details the experimental parameters of the synthetic setting used in this paper. Section 5.1 presents the results of the combined vote feedback for 3 different cases of accuracy levels and different number of users' feedbacks used to form the combination.

2 Related Works

This research goes under various subjects in the machine learning field; by concept, approach or objective. Many studies focus on the advantage of the user in the interactive machine learning process, where we involve the domain experts who are the expected future users of the applications in the development. They participate in feature selection, tuning parameters or defining new constraints resulting in a faster and more accurate learning [1]. Knowledge Elicitation via Sequential Probabilistic Inference for High-Dimensional Prediction [2] is a recent study to exploit experts' domain knowledge to help determine relevant features efficiently by formulating knowledge elicitation in sparse linear regression as a probabilistic inference process. The authors intend to find a solution to the situation where the number of training examples is inadequate in a particular instance of sparse linear regression. The essence of the method is that it adjusts to the feedback gained from the field professional and consecutively integrates the feedback before choosing the next query that maximizes the information gain. This study continues previous approaches for formulating experts knowledge efficiently like in prior elicitation where the aim is to create a prior distribution for Bayesian data analysis and limit the number of variables relevant to learn efficiently [5]. The method in [2] encodes the knowledge of the

expert as a feedback model and identifies a query algorithm that ranks the features according to their information gain, with a user interface to promote an efficient interaction. The feedback is modeled as additional data. The overall interaction scheme is to compute posterior distribution over the training data; the expert is queried sequentially for the feedback. The feedback is used to update posterior accordingly. They consider two types of feedback: about the relevance of the features, and the value of the regression coefficients. The query algorithm arranges the features to query about the feature that maximize the information gain first. The bigger the information gain, the higher the influence of the new feedback. This work is closely related to ensemble methods in machine learning [3] where we combine individual decisions of multiple classifiers under the conditions that the individual classifiers are 1- accurate (better than random guessing) and 2- diverse (make different errors on new data points) we use an average vote to classify test data and avoid the risk overfitting the model. Ensemble classifiers are good for statistical, computational and representational reasons The author mentions manipulating the input features as a technique to deal with highly redundant features; the main idea in this method is to build various classifiers by selecting unique subsets of the input features, stating that the ensemble classifier was able to match human experts [6].

3 Setting and Notations

We take the notion of expert feedback on relevance of features and the linear regression function to approximate the coefficients in spike and slab linear regression from [2].

In the setting we consider several experts that give feedbacks on the relevance of features (1 means relevant, 0 means not relevant), as we study ways to combine these feedbacks we can consider the feedback of each expert as their **vote** for the relevance of a feature. We introduce here the methods of calculating accuracy and combining voters, we create a synthetic setting explained below.

Setting let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denotes training data generated from standard normal distribution $\mathbf{X} \sim N(0, 1)$, the number of examples n . Let p be the total number of features, and $w_1, \dots, w_p \in \mathbb{R}$ are the regression coefficients, we denote p^* as the number of relevant features, we have w_1, \dots, w_{p^*} unknown true coefficient values are generate from $w_i \sim N(0, 1)$ the rest are set to zero. The observed target values are $\mathbf{Y} \in \mathbb{R}^n$, the experts number e_n (number of users' feedbacks); we constrain e_n to be odd ($e_n \bmod 2 = 1$) to facilitate the calculation of the majority vote. The predefined levels of experts accuracy vector is denoted \mathbf{L} .

We define terms like b the budget which is the number of questions we can ask to each expert; the default of our setting having budget equals to the number of relevant features $b = p^*$. We also create and normalize some test data under the same conditions.

We define \mathbf{F} to be the **feedback** from the voters about the relevance of the

features as a boolean matrix $\mathbf{F}(b, e_n)$ such that ($\mathbf{F}[i, j] = 1$ if the expert j thinks that the feature i is relevant and 0 if it is irrelevant). Individual feedback for voter j is presented by a binary array \mathbf{F}_j .

We define the **accuracy of combined feedback**, as the percentage of 1's in the feedback, denoted \mathbf{A} by querying the voter only on the relevant features (which will all have true coefficients). We calculate the percentage of ones' (relevant) over the total number of relevant features:

$$\mathbf{A}(\mathbf{F}_j) = \frac{1}{p^*} \sum_{i=1}^{p^*} \mathbf{F}_j[i]. \quad (1)$$

We make use of deterministic posterior approximations for fast computation, used to approximate spike-and-slab prior and the feedback models [2] this function takes each feedback individually.

We supply the feedback from each voter to the linear regression model that returns the weights of the voted-as-relevant coefficients. We use the mean squared error as our loss function (the mean of the squared difference between the predicted and actual target value), we calculate the Mean Squared Error (MSE) as the risk function to compare the effect of the feedbacks on the predicted regression model $MSE = \frac{1}{N} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$

In the following we will introduce our two methods to combine the feedbacks from multiple experts with various levels of accuracy on the relevance of features in "small n large p" applications and we will discuss the effect of the combined vote in the reduction of the prediction error.

4 Ensemble Feedback

Following the work on Knowledge Elicitation Via Probabilistic Inference [2], this model suggests combining the knowledge of multiple experts on high dimensional data. We present our data to the voters and we gather their feedback on the relevance of the p^* features.

Ensemble experts provides two methods to merge or assemble those feedbacks provided by the users into a single reliable final vote used to select the relevant features.

4.1 Majority Vote Feedback

Supposing we have a set of features $P = \{p_1, p_2 \dots p_{p^*}\}$, a set of odd number of Experts $E = \{e_1, e_2, \dots, e_{en}\}$ with varying accuracy level for each expert. We want to combine the feedbacks of our users and evaluate the improvement in the results and compare with the score of individual feedback. The combination takes the form of a single feedback composed of the majority vote over the features; and since the feedback is boolean (1 if the expert thinks the feature is relevant and 0 if the expert believes it is irrelevant) the majority vote is assigned

1 if the average of votes > 0.5 indicating that more than half of the number of experts agree that this feature is relevant, and 0 otherwise.

This method suggests averaging the feeds from the users over each feature to generate a generalized vote based on summing the number of positive votes for each feature from the experts over the number of experts. having b the assigned budget of queries to each expert, \mathbf{F} the feedback matrix of all experts, e_n an odd number of experts. we combine the feedbacks to generate the majority vote as follows:

Algorithm 1 Majority Vote for Combing Feedbacks

for $i = 1$ to b

 for $j = 1$ to e_n

$$\mathbf{F}_{MV}[i] = \frac{1}{e_n} \sum_{j=1}^{e_n} \mathbf{F}[i, j]$$

$$\mathbf{F}_{MV}[i] = \begin{cases} 1 & \text{if } \mathbf{F}_{MV}[i] > 0.5; \\ 0 & \text{if } \mathbf{F}_{MV}[i] < 0.5. \end{cases}$$

The **majority-vote feedback** is the boolean vector \mathbf{F}_{MV} obtained from averaging the votes of all users on the relevance of the features.

4.2 Field-Focused-Feedback (\mathbf{F}_{FF})

In applications with high number of features, like medical applications with various data about the patients (general information, blood analysis, genetic background, immune system status, medical history . . .); it is very likely that we can divide those features into subsets or fields; where each of those subsets belong to a specialized field and having a field expert feedback would be more reliable than a general one.

The model assumes in our setting that the number of fields n_f equals the number of experts: $n_f = e_n$, we also take into consideration ($b \bmod n_f = 0$). We define the range of field as budget over number of fields: $r_f = \frac{b}{n_f}$. **field-accuracy**(n_f, e_n) here is a square matrix indicating the accuracy for each expert over each field. We illustrate the steps in a simple example showing the feedback matrix \mathbf{F} for three experts $e_n = n_f = 3, p^* = b = 9$.

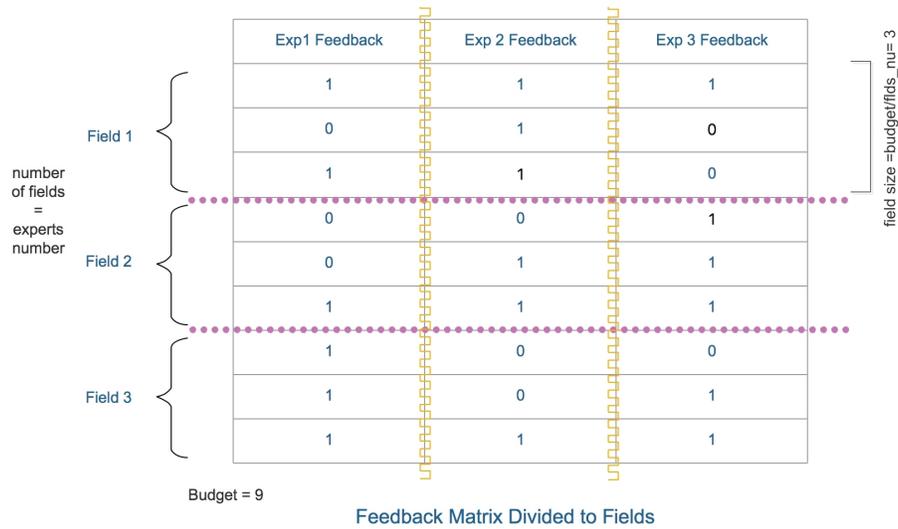


Figure 1: Example of 3 Experts Feedback Matrix divided into 3 Fields.

We take each feedback individually and calculate the accuracy for each expert over the fields as shown in Figure 2:

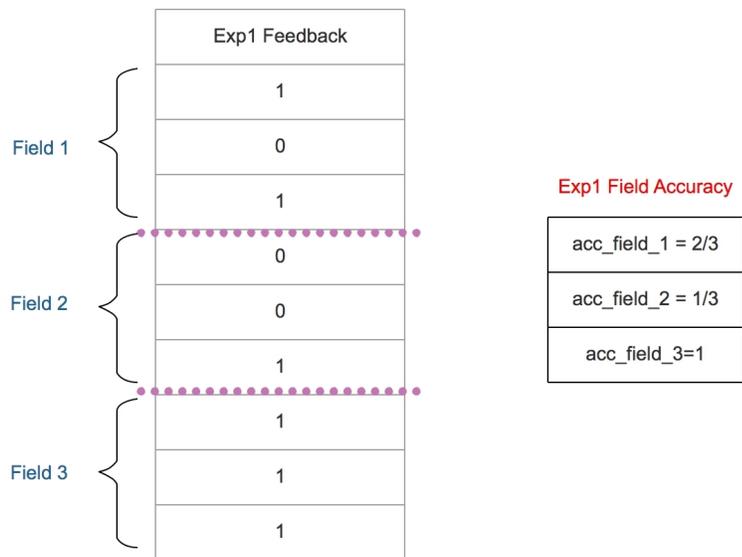


Figure 2: Expert 1 Field Accuracy

We obtain accuracy matrix for the experts over each field. Here we select

the maximum accuracy over each field, and we keep the indices of the experts that maximize the accuracy over the fields in a vector M_A .

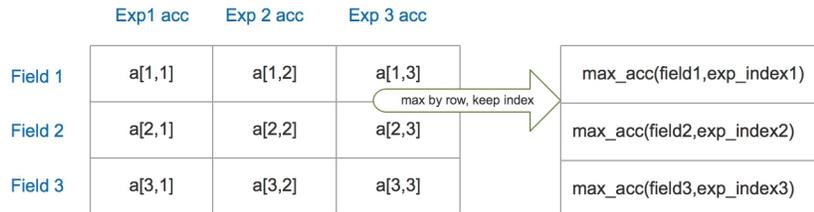


Figure 3: Accuracy Matrix for Experts - Max Accuracy

We retrieve the feedback $\mathbf{F}_{\mathbf{F}\mathbf{F}}$ given by the expert with the maximum accuracy over each field from the feedback Matrix \mathbf{F} by:

$$\text{for } j = 1 \text{ to } n_f : \\ \mathbf{F}_{\mathbf{F}\mathbf{F}}[((j-1) \times r_f) + 1 : i \times r_f] = \mathbf{F}[((j-1) \times r_f) + 1 : i \times r_f, M_A[j]].$$

5 Experiments

To study and evaluate the performance of our proposed methods we run some experiments of a regression problem on synthetic data under controlled settings: our parameters mainly include: n number of training samples $n \in \{10, 15, 20\}$ p is the number of features $p = 100$ for *setting*₁, and $p = 1000$ for *setting*₂, p^* number of relevant features $p^* \in \{20, 39, 70\}$ for *setting*₁, and $p^* \in \{119, 120, 150, 154, 180, 210\}$ for *setting*₂. We generate our training data $\mathbf{X} \in \mathbb{R}^{n \times p}$ from standard normal distribution $\mathbf{X} \sim N(0, 1)$. We generate unknown, true coefficient values \mathbf{w} (only first w_{p^*} are non-zero). We generate the output data $\mathbf{Y} \sim N(\mathbf{X}w_{p^*}, 1)$. We normalize our training and test data for better scaling. We use an odd number of experts $e_n \in \{3, 5, 7\}$ with varying accuracy levels to be detailed in the result section; we assign the level of accuracy for experts in a vector \mathbf{L} . We create the feedback of experts as a matrix \mathbf{F} with number of rows equals to b , and number of columns equals to e_n ; we constraint the budget (number of questions to ask each expert) to be equal to the number of relevant features p^* in both settings. \mathbf{F} is generated by setting $b \times e_n$ to 1 the amount corresponding to the level of accuracy in \mathbf{L} for the expert over the budget, then we make a random permutations of the 1s and 0s to make sure that not only the first features are each time the ones with the correct feedback. We copy the environment setting in [2] for the inputs of the linear regression spike and slab function, prior and other fixed model parameters.

We run our code¹ and repeat the data generation and experiments over 100 times and 500 times for other settings to have confidence under the law

¹Code available on github <https://github.com/Loujainl/ExpertsCombination>

of large numbers. We measure and average the accuracy of the ensembled Majority Vote and Field Focused Feedbacks as in equation 1. We use the Mean Squared Error (MSE) as the risk function to compare the effect of the feedbacks on the predicted regression model $MSE = \frac{1}{N} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$. We measure and average the MSE for individual experts feedbacks, Majority Vote, Field-Focused Feedback and no feedback.

5.1 Results

We tested the model on different settings and parameters, we detail the effect of the users' levels of accuracy on the combined vote accuracy as the most interesting influence on our objective.

5.1.1 Voters with Uncertain Levels of Accuracy

We test the effect of having fraud votes with accuracy level less than random guessing (< 0.5) on the performance of the Majority Vote and Field Focused Feedback.

Number of Experts $e_n = 3$: The figure 4a shows the experts levels initiated $\mathbf{L} = [0.75 \ 0.60 \ 0.45]$. The accuracy of Majority Vote $\mathbf{A}_{MV} = 0.67$, Field Focused Feedback accuracy level $\mathbf{A}_{FFF} = 0.75$. Figure 4b shows the MSE scored by each expert feedback, by our two combination and the MSE for the regression with no feedback.

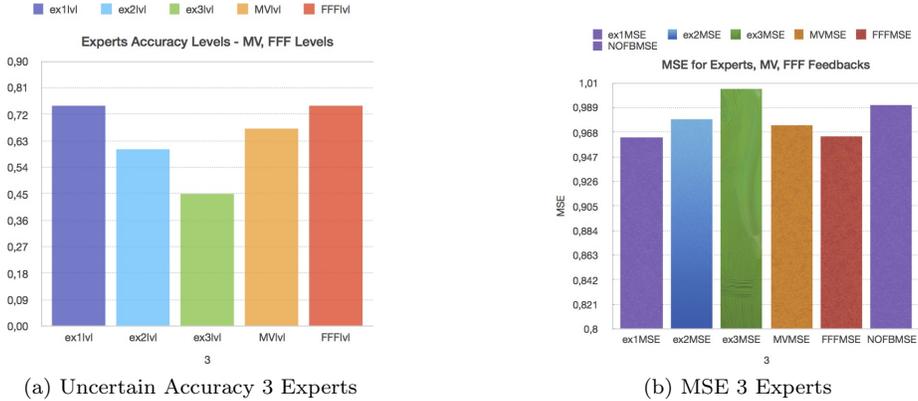


Figure 4: Voters with Uncertain Accuracy - MSE - Three Experts Combination

Number of Experts $e_n = 5$: We try the model with 5 voters on the relevance of features for a linear regression model, we combine the votes of the 5 feedbacks and calculate the accuracy of the Majority Vote and Field Focused Feedback, we measure and average the Mean Squared Error scored by the regression model for every voter.

The figure 5a shows the 5 experts levels initiated $\mathbf{L} = [0.70 \ 0.60 \ 0.55 \ 0.50 \ 0.40]$. The accuracy of Majority Vote $\mathbf{A}_{MV} = 0.59$, Field Focused Feedback

accuracy level $A_{FFF} = 0,63$. Figure 5b shows the MSE scored by each expert feedback, by our two combination and the MSE for the regression with no feedback.

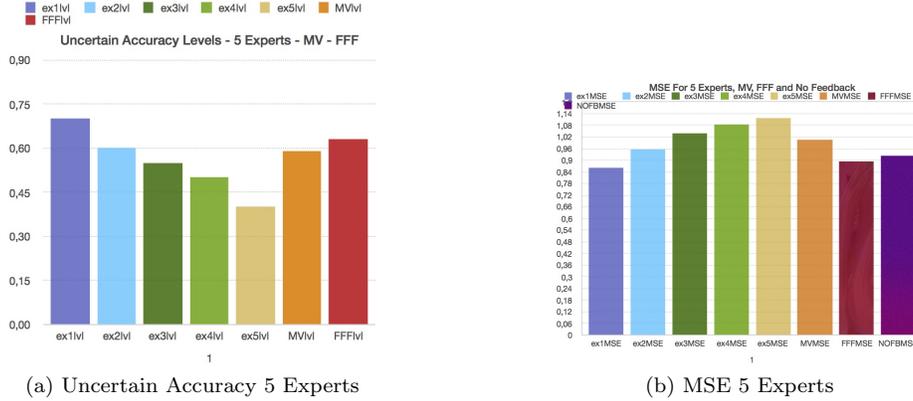


Figure 5: Voters with Uncertain Accuracy- MSE - Five Experts Combination

Number of Experts $e_n = 7$: We test the combination of the feedbacks having 7 voters levels initiated $L = [0,75 \ 0,65 \ 0,60 \ 0,55 \ 0,50 \ 0,45 \ 0,40]$ shown in 6a. With the accuracy of Majority Vote $A_{MV} = 0,613$, Field Focused Feedback accuracy level $A_{FFF} = 0,689$. Figure 6b shows the MSE scored by each expert feedback, by our two combination and the MSE for the regression with no feedback.

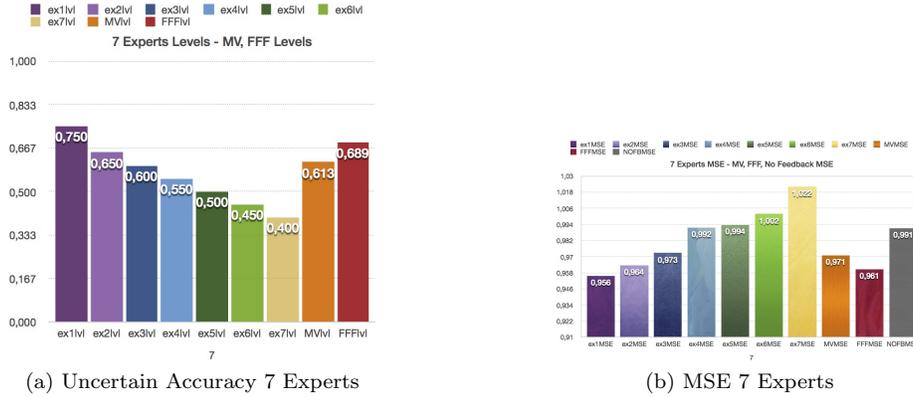


Figure 6: Voters with Uncertain Accuracy- MSE - Seven Experts Combination

After studying the results presented above of the calculated accuracy levels of the majority vote feedback and field focused feedback and the MSE averaged values for the individual voters and the combined votes, also to the model with no feedback, we notice that following the field focused feedback shown in the

last figure for example, we obtain a more important reduction in MSE value $MSE_{FFF} = 0.961$ for the combined vote than the majority vote $MSE_{MV} = 0.971$ shown in orange. Field focused feedback accuracy and MSE values fall between the highest two experts accuracies and MSE values respectively. The model with no feedback MSE = 0.991 and the voter with the highest accuracy $L[1] = 0.75$ and $MSE_{exp1} = 0.956$.

These results leads us to prefer the Field Focused Feedback combination method in the setting where we are uncertain of the voters' accuracy levels, FFF scored lower error resulting in a fairly trustable final feedback than the Majority Vote combination .

5.1.2 Voters Accuracy Levels Better Than Random Guessing

We test the effect of having voters with accuracy levels better than random guessing (> 0.5) on the performance of the Majority Vote and Field Focused Feedback.

Similar Levels of Accuracy: We try different number of experts with similar accuracy levels:

Number of Experts $e_n = 3$: The figure 7a shows the experts levels initiated $L_1 = L_2 = L_3 = 0.7$. The accuracy of Majority Vote $A_{MV} = 0.79$, Field Focused Feedback accuracy level $A_{FFF} = 0.72$. Figure 7b shows the MSE scored by each expert feedback, by Majority Vote of similar levels feedbacks and the field focused feedback, and the MSE for the regression with no feedback.

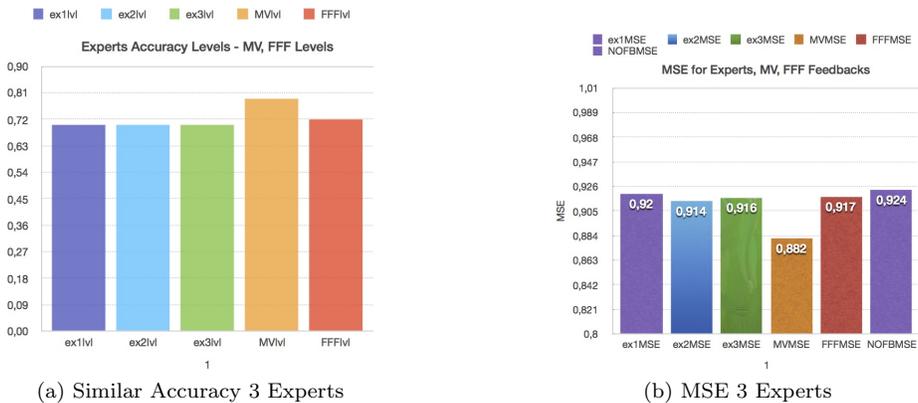


Figure 7: Voters with Similar Accuracy Levels- MSE - Three Experts Combination

Number of Experts $e_n = 5$:

We run the model with five experts having approximately similar levels. The figure 8a shows the experts levels initiated $L_1 = L_2 = 0.75, L_3 = L_4 = 0.7, L_5 = 0.6$. The accuracy of Majority Vote $A_{MV} = 0.82$, Field Focused Feedback accuracy level $A_{FFF} = 0.72$. Figure 8b shows the MSE scored by each expert feedback, by Majority Vote of five experts with similar levels

feedbacks , the MSE of field focused feedback, and the MSE for the regression with no feedback.

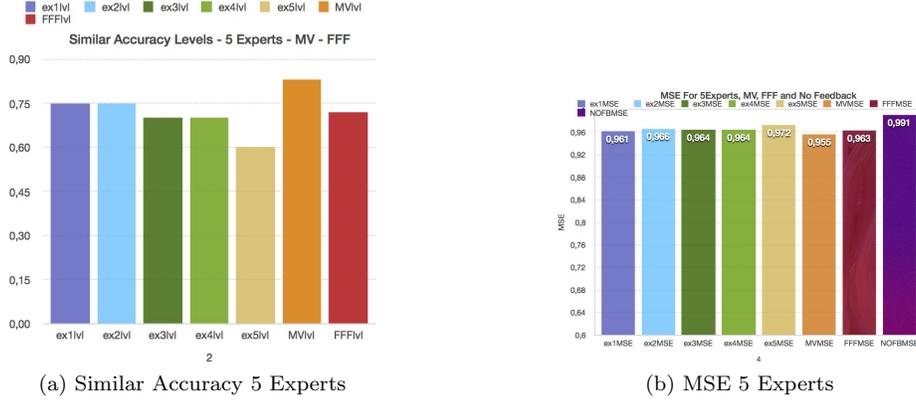


Figure 8: Voters with Similar Accuracy Levels- MSE - Five Experts Combination

Number of Experts $e_n = 7$:

We try combining seven voters with quiet similar accuracy levels, the figure 9a shows the initiated experts levels, $L_1 = L_2 = 0.70, L_3 = L_4 = L_5 = 0.65, L_6 = L_7 = 0.60$ the level of accuracy for the Majority Vote $A_{MV} = 0.792$ and Field Focused Feedback $A_{FFF} = 0.69$.

Figure 9b shows the MSE scored by each expert feedback, in orange the Majority Vote of five experts with similar levels feedbacks , the MSE of field focused feedback in red, and the MSE for the regression with no feedback.

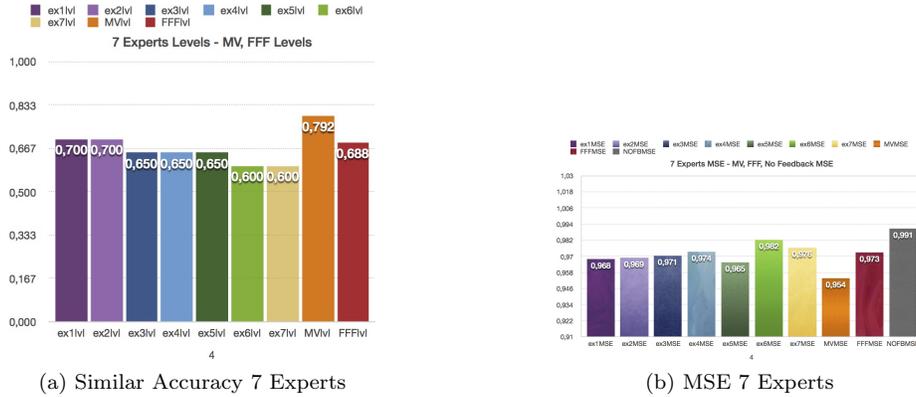


Figure 9: Voters with Similar Accuracy Levels- MSE - Seven Experts Combination

These results presented in the previous section imply that in the case of

voters with sufficient similar degree of knowledge, more precisely similar accuracy levels better than random guessing, the Majority Vote combination returns an interesting reduction in the MSE value than the single voters or the Field Focused Feedback method.

5.1.3 Voters with Ascending Levels of Accuracy:

We here study the case with experts' accuracy levels > 0.5 and increasing degree of credibility for the voters.

Number of Experts $e_n = 3$: We run the model for three experts with close high accuracy levels and illustrate the level of integrity of the combined votes. Figure 10a shows the primarily levels of three experts $\mathbf{L} = [0.75 \ 0.70 \ 0.65]$. The accuracy of Majority Vote $\mathbf{A}_{MV} = 0.85$, Field Focused Feedback accuracy level $\mathbf{A}_{FFF} = 0.73$.

Figure 10b shows the MSE scored by each expert feedback, the MSE for the Majority Vote, Field Focused Feedback error and the MSE for the regression with no feedback.

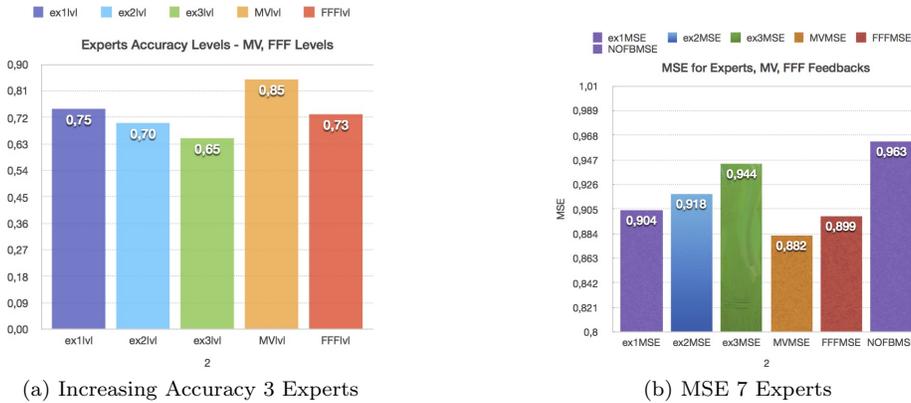


Figure 10: Voters with Increasing Accuracy Levels - Three Experts Combination

Number of Experts $e_n = 5$: We present the result of combining five voters with increasing accuracy better than random guessing using Majority Vote and Field Focused Feedback.

Figure 11a shows the primarily levels of five experts $\mathbf{L} = [0.80 \ 0.75 \ 0.70 \ 0.65 \ 0.60]$. The accuracy of Majority Vote $\mathbf{A}_{MV} = 0.80$, Field Focused Feedback accuracy level $\mathbf{A}_{FFF} = 0.77$.

Figure 11b reflects the averaged error scored by each expert feedback, the MSE for the Majority Vote feedback in orange, Field Focused Feedback error in red and the error for the regression with no feedback.

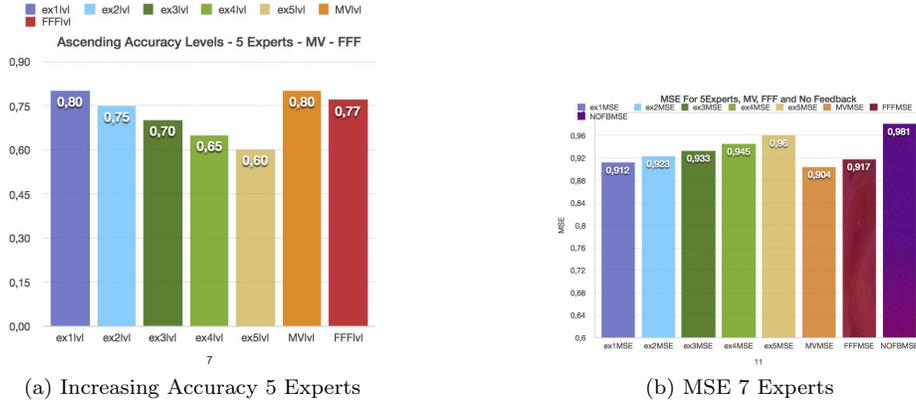


Figure 11: Voters with Increasing Accuracy Levels - Five Experts Combination

Number of Experts $e_n = 7$:

We repeat the experiment of combining accurate voters with increasing credibility levels, here we present seven experts feedback with best voter level $L_1 = 0.85$ and least accuracy level $L_7 = 0.55$. Figure 12a shows that Majority Vote feedback accuracy $A_{MV} = 0.905$, Field Focused Feedback accuracy level $A_{FFF} = 0.771$.

The bar plot in Figure 12b shows MSE scores for the voters and the ensemble votes.

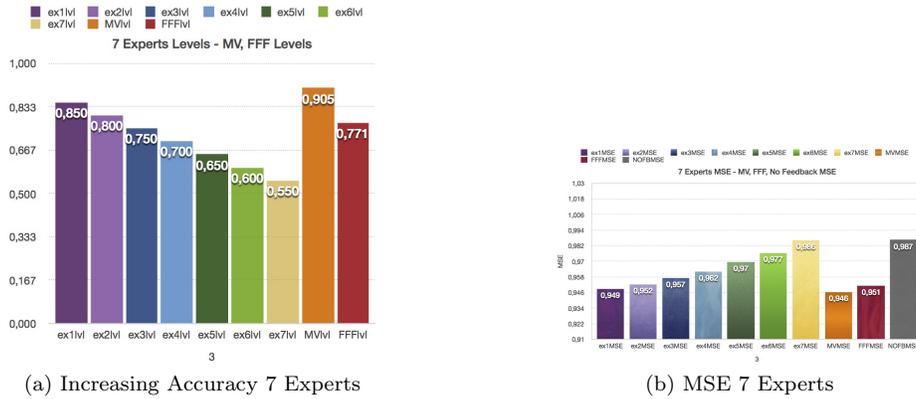


Figure 12: Voters with Increasing Accuracy Levels - Seven Experts Combination

Studying the last three results of combining accurate feedbacks from different number of voters with incremental credibility levels using Majority Vote method shown in orange outcomes a notable decrease in the averaged value of the Mean Squared Error than the error of the user with the best high accuracy level. Field Focused Feedback shown in red resulted in a relatively good combined vote close to the performance of experts between the highest two accuracy levels. This

leads to believe that combining voters with ascending levels of accuracy using the majority vote feedback results in an important decrease of the averaged Mean Squared Error for the linear regression than the error of the voter with the best accuracy level.

6 Conclusion

We proposed two techniques to combine the feedbacks of multiple users on the relevance of coefficients in "small n , large p " linear regression in a synthetic setting, and a model to calculate the accuracy of the combined vote. We illustrated and discussed the effects of the number of voters and their levels of credibility on the accuracy and performance of the Majority Vote feedback and Field Focused Feedback and proposed the preferred model to use when combining the feedbacks according to the environment settings. It is important to note that any single feedback with accuracy $\mathbf{A} > 0.5$ results in a reduction of the averaged MSE value of the model than no feedback. Combining feedbacks is also beneficial in preventing overfitting in machine learning applications where the number of training samples is too small and the number of dimensions is too large.

7 Future Works

It could be an interesting thing to find out when is it necessary to combine votes? is there a threshold of accuracy levels for enhanced combination? How trustable is this accuracy measure of the user knowledge? Will we get better results if we ensure diversity in the generated feedbacks? How will our model behave on different datasets. These could be outlines for developing this work in the future.

References

- [1] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, December 2014.
- [2] P. Daee, T. Peltola, M. Soare, and S. Kaski. Knowledge Elicitation via Sequential Probabilistic Inference for High-Dimensional Prediction. *Machine Learning (ML), Special Issue of the ECML-PKDD 2017 Journal Track*, 106(9-10):1599–1620, 2017.
- [3] Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.
- [4] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003.
- [5] A. O’Hagan, C.E. Buck, A. Daneshkhah, J.R. Eiser, P.H. Garthwaite, D.J. Jenkinson, J.E. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Experts’ Probabilities*. Statistics in Practice. Wiley, 2006.
- [6] KAGAN TUMER and JOYDEEP GHOSH. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8(3-4):385–404, 1996.