Master Thesis

A Human-Machine Ensemble Method for Economic Forecasts

Supervisor Associate Professor Shigeo MATSUBARA

Department of Social Informatics
Graduate School of Informatics
Kyoto University

Takahiro MIYOSHI

February 7, 2017

A Human-Machine Ensemble Method for Economic Forecasts

Takahiro MIYOSHI

Abstract

Forecasts of economic indices such as GDP, inflation, and unemployment rate are important for policymakers, companies, and investors. However, short-term forecasts of them are difficult since they are brought by complex interactions of various factors such as policies and economic activities by companies and individuals. Macroeconomic models and statistical methods have been used so far, but they have limitations.

For economic forecasts, “wisdom of the crowd,” which aggregates forecasts of different humans, is successful. Machine learning has also attracted attention. Forecasts by a human group and machine learning are studied independently, and both of which perform better than traditional statistical methods.

The forecasting method is greatly different between humans and machines. While humans consider information about political and economic circumstances, machines build statistical models from past time series. For example, in an unprecedented financial crisis such as 2007–2008, machines cannot forecast the future well, while humans can forecast it flexibly. In this manner, either one is not always more accurate. Sometimes humans are more accurate, and in other times machines are.

This thesis aims to make more accurate forecasts by combining forecasts by humans and a machine. Combining forecasts has been studied as consensus forecasts in the field of econometrics and as ensemble methods in the field of machine learning. However, there is no way to change the combination of humans and a machine depending on the situation. This thesis introduces such a human-machine ensemble method.

The key challenges in this thesis are as follows:

Modeling forecasts by humans and a machine and inventing an ensemble method using them The characteristic of forecasts by a group of humans is that the expected errors can be reduced by increasing diversity, and the characteristic of forecasts by a machine is that the expected errors can be

quantified based on the prediction model. Models that reflect each feature are necessary to make a human-machine ensemble method. The ensemble method also must make use of these characteristics.

Implementation and evaluation of economic forecasts using the proposed ensemble method It is not clear whether the assumptions of the proposed models are satisfied in real problems. It is also necessary to confirm that the proposed method improves the forecast accuracy for actual problems.

This thesis addresses the first problem as follows. First, we models outputs of machines as posterior distributions. The model can express the changes of the state according to inputs by regarding variance of the distribution as expected squared error. Second, we propose an ensemble method that uses the optimal number of humans depending on the expected error of a machine forecast. If the expected error of a machine forecast is sufficiently small, the method uses a machine forecast only, and if it is large, the method combines the optimal number of humans according to its value.

For the second problem, we apply the proposed method to inflation forecasts in the U.S. to verify the models and evaluate the performance. A time-series analysis model, ARMA(1,1) was used as a forecast accuracy benchmark. We created Recurrent Neural Networks models for machine forecasts, and we used survey data of economic forecasts for humans forecasts. By combining these forecasts, we confirm the behavior of the proposed method and discuss it.

The main contributions of this thesis are as follows:

Modeling forecasts by humans and a machine and inventing an ensemble method using them We modeled the case where the expected error changes with each forecast by assuming posterior distributions to outputs of machines. Additionally, we proposed a method that adopts the optimal number of humans dynamically depending on the expected error of a machine forecast.

Implementation and evaluation of economic forecasts using the proposed ensemble method We applied the human-machine nsemble method to inflation forecasts. As a result, we confirmed the proposed model can be applied to real inflation forecasts and forecast accuracy improves with 4 out of 7 data.

経済予測に関する人間と機械のアンサンブル法の考案

三吉 貴大

内容梗概

GDP やインフレーション，失業率などの経済指標の予測は，政策立案者や企業，投資家にとって重要である．しかし，これらの経済指標の変動は，企業や個人の経済活動，政策など様々な要因の複雑な相互作用の結果生じるため，短期的な変化を正確に予測することは難しい．これまでマクロ経済学によるモデルや統計的な手法が用いられてきたが，その予測能力には限界がある．

これに関して，複数の人間による予測を組み合わせる「群衆の叡智 (wisdom of the crowd)」が有効であることが知られている．また，近年，機械学習による予測も注目されている．人間の集団による予測と機械学習を用いた予測は別々に研究され，どちらも従来の統計的手法を超える予測精度を示している．

しかし，人間と機械の予測の仕方は大きく異なる．人間は，景気や政策などの情報を総合的に考慮して予測を行うのに対し，機械は，過去の時系列のみから統計的なモデルを構築して予測を行う．たとえば，リーマン・ショックのような過去に例のない金融危機の際，機械はうまく予測を行うことができないのに対し，人間はある程度柔軟に予測を行うことができる．このように，どちらか一方が常により正確であるということはなく，人間の方が正確な場合もあれば，機械の方が正確な場合もある．

本研究は，人間と機械の予測をうまく組み合わせることで，より正確な予測を行うことを目的とする．予測の組み合わせは，金融の分野では専門家の予測を総合するコンセンサス予測として，機械学習の分野では複数の予測モデルを合成するアンサンブル法として研究されてきた．しかし，人間と機械の予測方法の違いを踏まえ，状況に応じて使い分けるような方法は明らかではない．本研究では，そのような人間と機械のアンサンブル法の考案を試みる．

本研究における課題として，以下の 2 つが挙げられる．

人間と機械の予測方法のモデル化と，それを用いたアンサンブル法の考案　人間による予測の特徴は，人数を大きくすることで多様性が増し，誤差の期待値を小さくできることであり，機械による予測の特徴は，出力に対する誤差の期待値を定量化できることである．人間と機械のアンサンブル法を考案するにあたって，それぞれの特徴を捉えたモデルが必要である．そして，人間と機械の

アンサンブル法は，これらの特徴を活かしたものでなければならない．

考案したアンサンブル法を用いた経済予測の実装と評価　考案したモデルにおける仮定が実問題で満たされるか明らかでない．また，提案手法が実際の問題に対して，機械や人間のみの場合よりも予測精度を改善するか確認する必要がある．

一つ目の課題に対して，まず，機械による予測の出力を事後分布とするモデル化を行う．出力された分布の分散を二乗誤差の期待値とみなすことで，入力に応じた状態の変化を表現できる．次に，機械による予測の誤差の期待値に応じて，誤差の期待値を最小化する人数の人間による予測を組み合わせる手法を提案する．つまり，機械による予測の誤差の期待値が十分に小さければ，機械のみで予測を行い，誤差の期待値が大きければ，その値に応じた人数の人間による予測を組み合わせる．

二つ目の課題に対しては，以上に提案した手法をアメリカのインフレ予測に適用し，モデルの検証と性能の確認を行う．予測精度の比較には，ベンチマークとして統計的時系列解析の手法を用いた．また，機械による予測として，Recurrent Neural Networks (RNN) による予測モデルを作成した．人間による予測には，長期にわたって継続している経済予測の調査データを用いた．これらの予測を組み合わせて提案手法の振る舞いを確認し，考察を行う．

本研究の貢献は以下のとおりである．

人間と機械の予測方法のモデル化と，それを用いたアンサンブル法の考案　機械の出力に事後分布を仮定することで，予測のたびに誤差の期待値が変化する場合をモデル化した．そして，機械による予測の誤差の期待値に応じて，組み合わせる人間の予測の数を動的に決定できるようにした．

考案したアンサンブル法を用いた経済予測の実装と評価　提案した Human-Machine Ensemble Method を用いてインフレ予測を行なった．その結果，考案したモデルがインフレ予測に適用可能であること，また，7 つのデータ中 4 つで予測精度が向上することを確認した．

A Human-Machine Ensemble Method for Economic Forecasts

Contents

Chapter 1 Introduction

Forecasts of economic indicators such as GDP, inflation, and unemployment rate are important for policymakers, companies and investors. However, accurate forecasts of these indicators are difficult since fluctuations in them are caused by complex interactions of various factors such as policies and economic activities. The theory in the field of macroeconomics mainly deals with long-term fluctuations. Therefore, the theory for short-term fluctuations has not been established yet. Statistical time-series models have been used for short-term forecasts, but they are not sufficient.

The *wisdom of the crowd*, which combines forecasts by many humans, is effective for short-term economic forecasts. According to Ang, Bekaert and Wei [1], the mean of expert forecasts collected by surveys is more accurate than macroeconomic models and time-series models. Surowiecki [13] has summarized many cases where the result of aggregating opinions of many people was accurate as the *wisdom of the crowd*.

Economic forecasts by *machine learning* have also attracted attention. In recent years, the growth of computers and the Internet enabled us to handle large amounts of economic data. By machine learning using these big data, it is possible to create complicated prediction models. Several studies have reported that models based on *artificial neural networks (ANN)* predicted inflation more accurately than the traditional time-series models [3, 9, 10].

Both of a group of humans and machine make more accurate forecasts than the traditional methods. However, the forecasting procedure is largely different in humans and machines. While humans consider information such as economic conditions and policies when forecasting, machines make forecasts based on a statistical model constructed from past time series only. For example, when an unprecedented financial crisis such as 2007–2008 occurred, machines can not forecasts the future well, while humans can forecast it flexibly. In this manner, either one is not always more accurate. Sometimes humans are more accurate, and in other times machines are.

This thesis aims to make more accurate forecasts by combining a group of

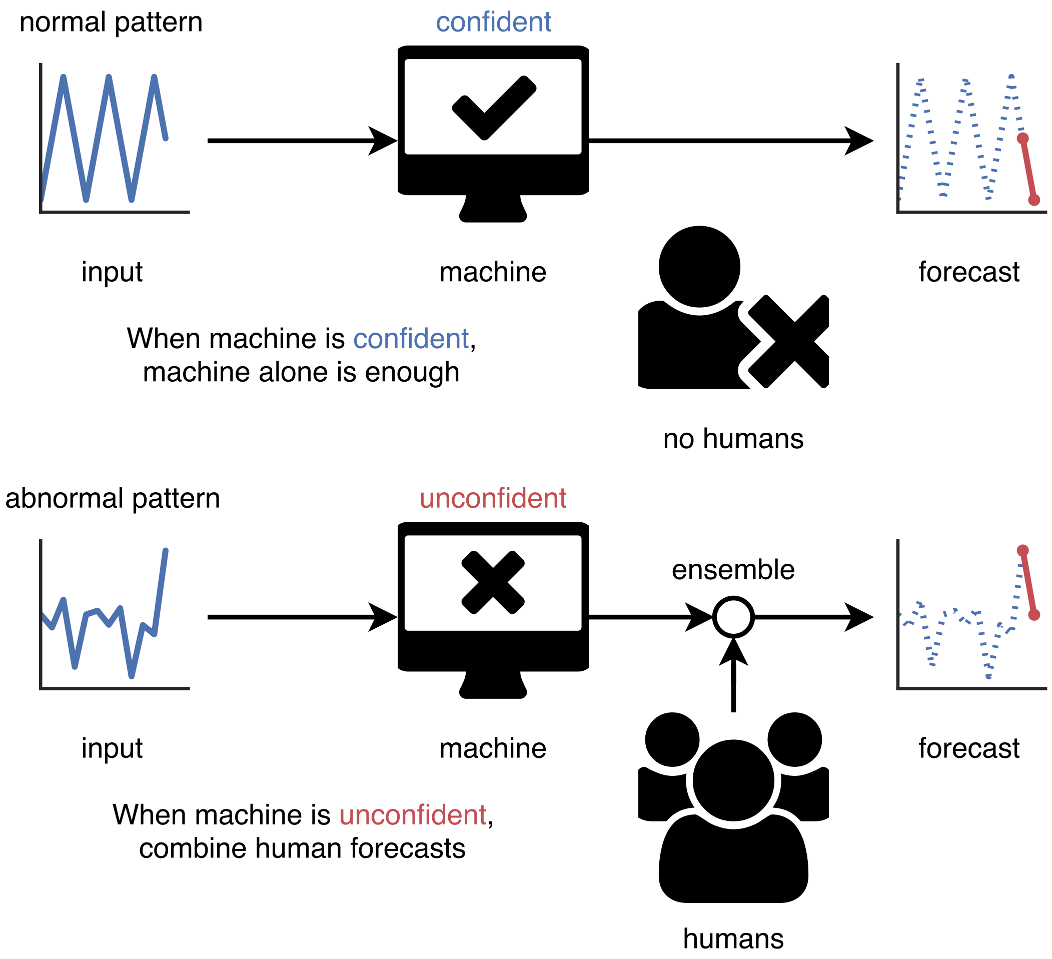


Figure 1.1: Schematic of a human-machine ensemble method

humans and machine with harnessing their characteristics. Combination methods for forecasts have been studied as *ensemble methods* in the field of machine learning [15] and as *consensus forecasts* in the field of finance [2]. In ensemble methods, although many of them include the training phase, combining methods of trained models can be applied to combinations with human forecasts. Consensus forecasts aggregates forecast values by using simple average, weighted average and trimmed average. In either method, however, there is no combining method that harnesses the good points of each forecaster. It is impossible to emphasize human forecasts over the machine at financial crisis since traditional methods always use the same combination regardless of the situation (input).

Therefore, this thesis proposes an ensemble method that emphasizes a machine if the machine is likely to forecast accurately and emphasizes humans otherwise. Figure 1.1 shows a schematic of a human-machine ensemble method this thesis proposes. Machines are good at forecasting that follows patterns since they build a prediction model based on past data. The upper part of the figure shows that when such a pattern is given, a machine forecast alone is enough since it can forecast the future well. On the other hand, when a pattern not in the past is given, machines can not make a forecast well. In such a case, the proposed method combines humans, who can forecast patterns not in the past, as shown in the lower part of the figure. Through the above method, the proposed ensemble method take advantages of a machine and group of humans forecasts. Specifically, we expand the existing model to express confidence in a forecast, which is represented by the expected error. Based on the models, we propose an ensemble method that changes the combination of a machine and humans dynamically according to input. It minimizes the expected errors of the forecasts by the ensemble.

We also conducted an experiment to apply the proposed method to actual inflation forecasts. We made ANN prediction models using *Long short-term memory (LSTM)* in the hidden layer as a machine. And we used two survey data for economic forecasts as humans. We also used traditional time-series models as benchmarks to compare them. We prepared test set separately from the training set for evaluation. We constructed seven ensembles and conducted verification of the proposed models, assessing forecast accuracy and confirmation of the behavior of the ensemble.

We obtained three empirical results as follows. First, the proposed model, which express the expected error according to input, was valid on actual inflation forecasts. Second, the proposed human-machine ensemble method made more accurate forecasts than each of a machine and group of humans in 4 out of 7. Finally, the human-machine ensemble method emphasized human forecasts when a pattern not in the past was given as shown in Figure 1.1.

However, the human-machine ensemble method does not always behave as expected. Therefore, we discuss the conditions to make use of the proposed ensemble method. We also consider applications of the proposed method.

The rest of this thesis is organized as follows. Chapter 2 describes inflation forecasts, the wisdom of the crowd and ensemble methods as related works. In Chapter 3, we modeling forecasts of machines and humans, and propose a human-machine ensemble method. Chapter 4 describes the data set, forecasting models and evaluation methods used in the experiment. Chapter 5 contains the empirical results. We discuss the scope and applications of the proposed method in Chapter 6. Finally, Chapter 7 concludes.

Chapter 2 Related Works

2.1 Inflation Forecasts

Accurate forecasts of future inflation is important for policymakers conducting monetary and fiscal policy. Inflation is also significant when investors hedge the risk of nominal assets; when companies make investment decisions or determine the prices; and when labor and management negotiate wages.

Short-term forecasts of inflation is difficult, so various methods have been devised. This section introduces the study of Ang et al., which comprehensively compares econometric methods, and describes *artificial neural networks (ANN).*

2.1.1 Comparison of econometric methods and surveys

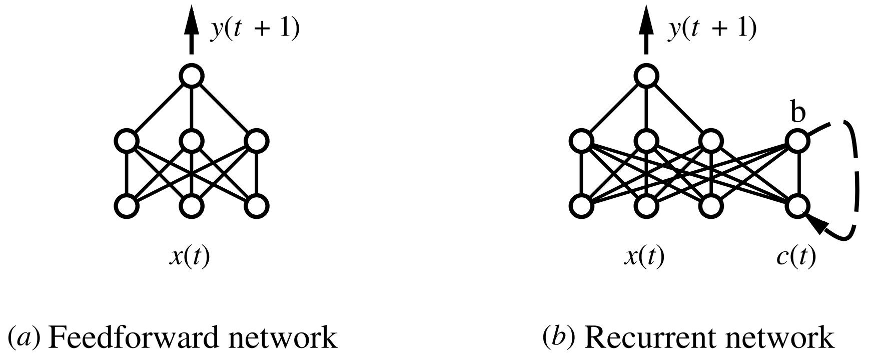
Ang et al. [1] applied four methods to U.S. inflation forecasts and compared their forecasting accuracy. They used time-series models, regressions using real activity measures motivated from the Phillips curve, term structure models that include linear and non-linear specifications, and survey-based measures. As a result, the surveys were the most accurate. Additionally they examined combining methods. However, the combinations of these forecasting methods were not able to achieve higher accuracy than the survey alone.

The forecasting methods they used are roughly four, but each method has several variations depending on the parameters of time-series models, economic indicators included in the regressions or term structure models, and the aggregating methods of forecasting values collected by surveys. They also used three surveys: the Livingston Survey (LIV), the Survey of Professional Forecasters (SPF), and the Michigan survey (MICH). We describe the Livingston Survey and the SPF in Section 4.1.2.

The target of forecasts are annual change rates of four inflation indicators: Consumer Price Index for All Urban Consumers, All Items (CPI); CPI for All Urban Consumers, All Items Less Shelter (CPI-XS); CPI for All Urban Consumers, All Items Less Food and Energy (CoreCPI); and Personal Consumption Expenditure (PCE). Ang et al. set two out-of-sample periods, after 1985 and 1995. And they evaluated the forecasting methods by Root Mean Squared Errors (RMSE) during that periods.

Table 2.1: Empirical results by Ang et al. [1]. Each entry represents the ratio of the RMSE to a benchmark, ARMA(1,1). Bold entries are the smallest RMSEs in each column. The rows represents forecasting methods: the best time-series model (not necessarily ARMA(1,1)), the best Phillips Curve models, the best term structure model, and three surveys, LIV, SPF and MICH. The bottom row is the ensemble method by weighted averaging all models. Each method is applied to two out-of-sample periods, post-1985 and post-1995.

		CPI	CPI-XS	CoreCPI	PCE
Post-1985	Time-Series	1.000	1.000	1.000	1.000
	Phillips-Curve	0.979	1.000	0.862	1.027
	Term-Structure	1.091	1.047	0.945	1.018
	LIV	0.789	0.844	**0.655**	1.082
	SPF	**0.779**	**0.819**	0.691	1.199
	MICH	0.902	0.881	1.185	1.217
	Ensemble	0.873	0.864	0.819	**0.962**
Post-1995	Time-Series	0.764	0.833	0.915	**1.000**
	Phillips-Curve	0.977	0.992	0.767	1.020
	Term-Structure	0.913	0.973	0.866	1.025
	LIV	0.792	0.856	0.557	1.202
	SPF	0.861	0.914	**0.699**	1.250
	MICH	0.862	0.937	0.822	1.338
	Ensemble	**0.722**	**0.735**	0.702	1.005



(*a*) Feedforward network (*b*) Recurrent network

Figure 2.1: Architecture of a feedforward network and recurrent network [8].

Table 2.1 shows the empirical results by Ang et al. Each value reports the relative RMSE ratio to a time-series model, ARMA(1,1). In time-series models, ARMA(1,1) generally showed good accuracy, so it was set as the benchmark. The Phillips Curve models and the term structure models were sometimes worse than the benchmark and they never achieve the best score in each column. Overall, the surveys outperformed the other forecasting methods.

Ang et al. investigated ensemble methods such as mean, median and weighted average. Weighted average made the most accurate forecasts among them. The table shows the RMSEs of the weighted average as an ensemble method. The ensemble method had the smallest RMSEs in only three of the eight. Combining forecasts did not always make more accurate forecasts.

2.1.2 Artificial neural networks

Inflation forecasting using an ANN model has been studied recently[3, 9, 10]. These studies have reported that ANN models can forecast as well as traditional time-series models and sometimes outperform them. This section describes ANN models.

ANN is a method of machine learning that is inspired biological learning system [8]. It has attracted much attention with the development of deep learning [16].

Figure 2.1 (a) shows the basic architecture of an ANN. An ANN consists of layers that include some units called *neuron*. A unit i receives an input vector

$\boldsymbol{x}_i$ and returns an output $o_i(\boldsymbol{x})$. The output $o_i(\boldsymbol{x})$ is

$$o_i(\boldsymbol{x}_i) = f(\boldsymbol{w}_i \cdot \boldsymbol{x}_i + b_i), \tag{2.1}$$

where $\boldsymbol{w_i}$ is weights, b_i is a bias, and f is an *activation function.* Each unit multiplies the inputs by the weights, adds a bias, and returns the output of the activation function with it. The logistic function or rectified linear function is used for activation functions. And an activation function suitable for the problem is used as the output layer of the network. In a feedforward network, a unit receives the outputs of units in the previous layer and its output is received by units in the next layer. By repeating this, the network can express non-linear function $y(\boldsymbol{x})$. For example, it is possible to create a classifier that estimates the age and sex from a face photo.

An ANN learns by adjusting the weights $\boldsymbol{w}$ and biases b so as to approach the input and output of the training set. Regard parameters $\boldsymbol{w}$ include weights and biases. The training set consists of pairs of inputs $\boldsymbol{x}$ and target outputs $\boldsymbol{d}$, $\{(\boldsymbol{x}_1, \boldsymbol{d}_1), (\boldsymbol{x}_2, \boldsymbol{d}_2), \ldots, (\boldsymbol{x}_N, \boldsymbol{d}_N)\}$. The *Gradient Descent* is used for the parameter adjustment. It updates the parameters $\boldsymbol{w}$ gradually so as to reduce the error function $E(\boldsymbol{w})$, which represents the error between the output of the network y and target output $\boldsymbol{d}$. That is,

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \nabla E, \tag{2.2}$$

where η is a learning rate and ∇E is a gradient defined as

$$\nabla E \equiv \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \cdots \frac{\partial E}{\partial w_n} \right]. \tag{2.3}$$

The *Stochastic Gradient Descent*, which updates parameters using not all samples but a batch of a plurality of samples, is often used. It has been proved that learning converges if learning rate is sufficiently small. For efficient gradient calculation, the *Backpropagation* algorithm is generally used.

Recurrent neural networks (RNN) are neural networks including circular structures as shown in Figure 2.1 (b), and are applied to time series data. For example, A RNN model can predict the stock price of the next day based on the fluctuations of the price so far. Learning of RNN also uses the Gradient Descent, and the learning algorithm is the extended Backpropagation.

2.2 Wisdom of the Crowd

A journalist Surowiecki [13] has summarized phenomena where groups of humans made more accurate forecast than an expert. He called it the *wisdom of the crowd*. The reason why the wisdom of the crowd works well is that individuals in the group have diverse information. For instance, while several people forecast that the GDP will increase due to a policy, others forecast that it will decrease due to recession in a country, so the aggregation of them can make more informative forecasts. The results of Ang et al., where the surveys outperform various forecasting methods, can be explained by the wisdom of the crowd.

Page [11] gave the a rationale for the wisdom of the crowd. *Diversity prediction theorem* states "a crowd's collective accuracy equals the average individual accuracy minus their collective predictive diversity." Let h_i denote forecasted value by individual i, and let $\bar{h}$ denote the average of N forecasts. V is a correct value. Then, the following equation always holds [6, 11].

$$(V - \bar{h})^2 = \frac{1}{N}\sum_{i=1}^{N}(V - h_i)^2 - \frac{1}{N}\sum_{i=1}^{N}(\bar{h} - h_i)^2 \tag{2.4}$$

The left side is the squared error of the group's forecast $\bar{h}$. The first term on the right side is the average of the squared error of each individual forecast. The second term is the variance of the forecasts within the group, which represents the diversity of the forecasts within the group. This value decreases as the individuals in the group make similar forecasts, and increases as they make diverse forecasts. The larger this values, the smaller the error of the group forecast (the left side) compared with the forecast error of average individual (the first term on the right side). This is because the averaging the forecasts with positive error and ones with negative error offset the errors.

2.3 Ensemble Methods

This section describes *weighted averaging* and a study on the optimal composition of a group in the case where individual forecasters have types such as professional and non-professional.

In weighted averaging, the optimal weights can be obtained when given a covariance matrix of forecasters' errors [15]. Let h_i denote a forecasted value by a forecaster i, and let w_i denote i's weight. When a vector of forecasts by N forecasters $\boldsymbol{h} = (h_1, \ldots, h_N)$ is given, the weighted average $G(\boldsymbol{h})$ is

$$G(\boldsymbol{h}) = \sum_{i=1}^{N} w_i \cdot h_i, \tag{2.5}$$

where $w_i \geq 0$ and $\sum_i w_i = 1$. The weights $\boldsymbol{w} = (w_1, \ldots, w_N)$ are determined to minimize G's expected squared error $\mathrm{MSE}(G)$:

$$\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \mathrm{MSE}(G). \tag{2.6}$$

Let $\mathrm{cov}(\varepsilon_{h_i}, \varepsilon_{h_j})$ denote the covariance in the errors of the forecasters i and j, and let $\mathrm{var}(\varepsilon_{h_i})(= \mathrm{cov}(\varepsilon_{h_i}, \varepsilon_{h_i}))$ denote the variance of forecaster i. Then w_i that minimizes the expected squared error is obtained by

$$w_i = \frac{\sum_{j=1}^{N} \mathrm{cov}(\varepsilon_{h_i}, \varepsilon_{h_j})^{-1}}{\sum_{k=1}^{N} \sum_{j=1}^{N} \mathrm{cov}(\varepsilon_{h_k}, \varepsilon_{h_j})^{-1}}. \tag{2.7}$$

The covariance matrix can be estimated from past forecasts. Basically, weighted averaging gives big weights to accurate forecasters, and small ones to inaccurate forecasters. However, if the covariance is small, that is, forecasters make errors differently from each other, the weights of those are set not to be too biased to either one.

Weighted averaging assumes that the number of forecasters N is constant. And it cannot make use of the good points of each forecaster as mentioned in Chapter 1 since the weights $\boldsymbol{w}$ are fixed regardless of the input $\boldsymbol{x}$. In other words, even if forecaster i can make an accurate forecast for input $\boldsymbol{x}_1$ but can not do so for input $\boldsymbol{x}_2$, the weights are always same.

Lamberson and Page [7] proposed to use a ratio of forecaster types instead of weights when the forecasters have types. The forecasters are divided into types by a statistical criterion called *type coherence*. Here we assume there are two types of forecasters. If the number of forecasters is small, the ratio of a type having high accuracy should be increased, and if the number of forecasters is big, the ratio of a type having high diversity should be increased.

There are two types of forecasters, a and b. The number of each is A and B

respectively, and the sum is $N = A + B$. The forecast values of each forecaster is modeled as random variables, and follows a probability distribution according to the type of the forecaster. Assume that the forecasters are unbiased, that is, forecasts are not always greater or smaller than the real value. Let $\mathrm{var}(\varepsilon_a)$ and $\mathrm{var}(\varepsilon_b)$ denote the variance of the type a and b respectively. The variance equals the expected squared error if forecasts are unbiased. Let $\mathrm{cov}(\varepsilon_a)$ denote the covariance in the errors of the forecasts made by two different type a forecasters, and assume that this is the same for any pair of type a forecasters. Define $\mathrm{cov}(\varepsilon_b)$ similarly. The smaller the within-type covariance, the more diverse the group of forecasters. Finally, let $\mathrm{cov}(\varepsilon_a, \varepsilon_b)$ denote the covariance in the errors of any two forecasters, one of which is of type a and the other of which is of type b. Let us find the ratio of the number of each type A and B that minimizes the expected squared error of the ensemble.

Type coherence between two types a and b is defined as

$$\mathrm{TC}(a, b) = \mathrm{cov}(\varepsilon_a) + \mathrm{cov}(\varepsilon_b) - 2\,\mathrm{cov}(\varepsilon_a, \varepsilon_b). \tag{2.8}$$

If $\mathrm{TC}(a, b) > 0$, type a and b satisfy type coherence. This means the across-type covariance is less than the average of the within-type covariances.

The forecasts of an ensemble is simple averaging. Then, the expected squared error of the ensemble is

$$\frac{A\,\mathrm{var}(\varepsilon_a) + B\,\mathrm{var}(\varepsilon_b) + 2\binom{A}{2}\mathrm{cov}(\varepsilon_a) + 2\binom{B}{2}\mathrm{cov}(\varepsilon_b) + 2AB\,\mathrm{cov}(\varepsilon_a, \varepsilon_b)}{N^2}. \tag{2.9}$$

Given the size of the group N, the optimal fraction of type a forecasters, which minimizes that expected squared error, is approximated by

$$A^* = \frac{[\mathrm{var}(\varepsilon_b) - \mathrm{var}(\varepsilon_a)] - [\mathrm{cov}(\varepsilon_b) - \mathrm{cov}(\varepsilon_a)]}{2N \cdot \mathrm{TC}(a, b)} + \frac{\mathrm{cov}(\varepsilon_b) - \mathrm{cov}(\varepsilon_a, \varepsilon_b)}{\mathrm{TC}(a, b)}, \tag{2.10}$$

where the approximation error is less than $1/M^2$.

Focusing on $\mathrm{var}(\varepsilon_a)$ and $\mathrm{var}(\varepsilon_b)$, decreasing $\mathrm{var}(\varepsilon_a)$ or increasing $\mathrm{var}(\varepsilon_b)$ make A^* greater. The more accurate the forecasts of type a or the less accurate the ones of type b, the greater the optimal ratio of type a forecasters in the group.

However, if the group size N is sufficiently large, the first term in equation (2.10) can be ignored. Then, the optimal number of type a forecasters, A^* is independent from $\mathrm{var}(\varepsilon_a)$ and $\mathrm{var}(\varepsilon_b)$, and depends on the diversity of each type, not the forecast accuracy of each type.

Lamberson and Page have summaried the above results as follows:

> In a sufficiently small group, the lowest variance type should be in the majority. In a sufficiently large group, the forecaster type with the lowest within-type covariance should be in the majority [7].

For example, consider conducting economic forecasts by employing a group of forecasters. If you want reduce the group size, it is better to increase the ratio of forecasters with high capability such as experts, and otherwise, it is better to increase the ratio of ordinary people.

However, the model of Lamberson and Page also can not utilize the characteristics that each type of forecasters has. For instance, assume that there are two types of forecasters. Type c forecasters use time-series models for economic forecasts, so their forecasts are accurate when the economy is stable; and type d forecasters make forecasts based on economic information, so they can make flexible forecasts even when the economy is unstable. As well as weighted averaging, the method of Lamberson and Page can not change the combination according to the situation. The next chapter proposes a human-machine ensemble method that harnesses each advantege, assuming a prediction model obtained by machine learning as type c and humans as type d.

Chapter 3 Human-Machine Ensemble Method

In this chapter, Section 3.1 describes the characteristics of forecasts by humans and a machine, and overview the proposed *human-machine ensemble method*. Then, I modeling the forecasts of humans and ones of machines in Section 3.2. Finally, Section 3.3 formulates the human-machine ensemble method as a problem to minimize the expected squared error of the ensemble, and solves it.

3.1 Overview

Traditional ensemble methods assumes that the error of each forecaster always follows a fixed probability distribution as mentioned in Section 2.3. However, in fact, since each forecaster has good and weak points, the expected error of each changes depending on the situation. Let us consider economic forecasts. Since machines learn past patterns, they are good at a forecasting when the target index follows a pattern, but they are bad when given a pattern not in the past. On the other hand, humans can consider information such as economy and policies that machines can not lean well, so they are adaptable to patterns not in the past. Therefore, it is necessary to understand the characteristics of humans and machines for developing a human-machine ensemble method.

The characteristic of forecasts by machines is that their expected error can be quantified since they make forecasts statistically based on patterns they leaned. Let us explain by using an illustrative example. Figure 3.1 shows 200 data points belonging to one of two classes A and B. The classes A and B are sampled from the Gaussians with standard deviation $\sigma = 1$ centered at $(-2, 0)$ and $(2, 0)$, respectively. Assume that there is a prediction model that learned those samples. For instance, if this model is given an input $p = (-2, 0)$, which has the typical feature of class A, it returns that the probability of belonging to A is 0.95. Furthermore, given $q = (0.1, 2)$, which is located near the boundary of the two classes, it returns the probability is 0.45. Of course, humans also can answer the probability that p and q belong to class A based on the figure, but since they answer the probability intuitively, the grounds for the value is weak and the reliability is poor. Meanwhile, machines can return values with

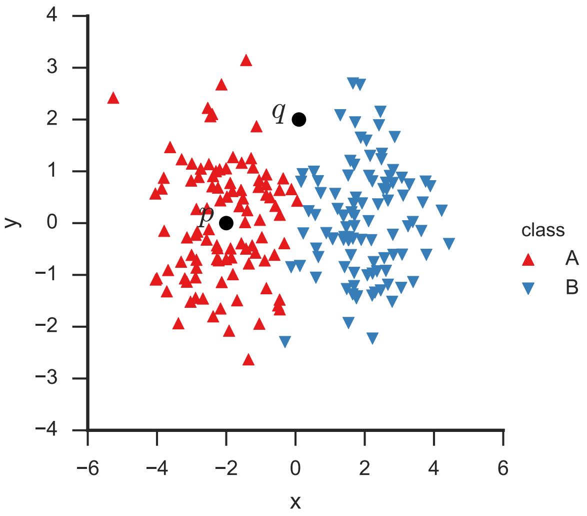


Figure 3.1: An illustrative example of a classification problem. There are 200 samples represented as points in a 2D feature space. Each of them belongs to a class A or B. Given a sample p as input to a prediction model that learned those samples as training set, the model outputs that the probability of belonging to A is high, while given a sample q, the probabilities of belonging to A and B are about the same.

grounds based on the concrete model.

On the other hand, the characteristic of forecasts by humans is that they can make forecasts considering information such as the current economy and new policies, which is difficult for machines to learn. In addition, forecasts by a group of humans can harness the *wisdom of the crowd* as mentioned in Section 2.2. In economic forecasts, opinions are often different among experts, so it is easy to gather diverse forecasts by increasing the number of people. Conversely, it is difficult to construct diverse prediction models by machine learning. Even if the learning algorithm is different, the predicted values will be similar since the training set is the same. Algorithms such as *Random Forest* divide the training set and construct multiple weak learners, but they sacrifices the accuracy of each model.

The human-machine ensemble method combines forecasts of a machine and group of humans according to the expected error of the machine forecast. In other words, if the expected error of machine is small, the ensemble judge that the machine forecast alone is enough, otherwise it combines forecasts of humans. This procedure makes it possible to harness the characteristics of machines that can quantify the expected errors and the characteristics of humans that can compensate for the poor part of machines and increase diversity easily.

3.2 Model

In this section, I modeling forecasts of prediction models obtained by machine learning and forecasts of individual humans. They forecasts continuous values such as annual inflation rate.

3.2.1 Humans model

Forecasts of individual humans are modeled as random variables like Lamberson and Page [7] introduced in Section 2.3. Let random variables h_i and ε_{h_i} denote the forecast and error of a human i.

Assume that the error ε_{h_i} follows a distribution with mean $\mu = 0$ and variance $\sigma^2 = \mathrm{var}(\varepsilon_h)$. This means that forecasts by humans are unbiased. That is, although a forecast can be smaller or greater than the actual value, the positive errors and the negative errors are the same amount as a whole, and the errors are not positively biased or negatively biased.

Moreover, let $H(n)$ denote the average of forecasts by n humans, and let $\mathrm{cov}(\varepsilon_h)$ denote the average covariance in the errors of two different humans as

$$\mathrm{cov}(\varepsilon_h) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \mathrm{cov}(\varepsilon_{h_i}, \varepsilon_{h_j}). \tag{3.1}$$

According to [14], the expected squared error of $H(n)$ is

$$\mathrm{MSE}(H(n)) = \frac{1}{n} \mathrm{var}(\varepsilon_h) + (1 - \frac{1}{n}) \mathrm{cov}(\varepsilon_h). \tag{3.2}$$

3.2.2 Machine model

Forecasts of machines are modeled as that the expected error varies according to the input. The inputs are, for example, inflation fluctuations over the past 12 months. Assume that a prediction model θ obtained by machine learning outputs a probability distribution to an input vector $\boldsymbol{x}$. This is regarded as the posterior distribution $f_\theta(y|\boldsymbol{x})$ for the target value y when given the input $\boldsymbol{x}$.

The forecast value $y_\theta(\boldsymbol{x})$ is the mean of the posterior distribution outputted by model θ given input $\boldsymbol{x}$:

$$y_\theta(\boldsymbol{x}) = \int_{-\infty}^{\infty} y f_\theta(y|\boldsymbol{x}) dy. \tag{3.3}$$

Let $\mathrm{var}(\varepsilon_\theta|\boldsymbol{x})$ denote the variance of the distribution. This represents the expected squared error of the forecast value $y_\theta(\boldsymbol{x})$ from the definition.

Finally, let $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$ denote the average covariance in errors between the machine and an arbitrary human. That is,

$$\mathrm{cov}(\varepsilon_\theta, \varepsilon_h) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{cov}(\varepsilon_\theta, \varepsilon_{h_i}). \tag{3.4}$$

Section 5.1 verify whether these models are applicable to actual inflation forecasts.

Implementation A prediction model that outputs probability distributions can be implemented by using *artificial neural networks (ANN)* approximately. Separate the possible target values into K discrete values $y_1, \ldots, y_K$, and regard the problem as *classification* to the closest value. Then, if the *softmax function* is used for the output layer of ANN, the prediction model can output approximate discrete probability distributions. For example, when forecasting a percentage of annual inflation change, an output of the ANN model is that the probability of 0% is 0.02, 0.5% is 0.05, 1% is 0.33, . . . , and 4% is 0.03. In this case, $y_\theta(\boldsymbol{x})$ is obtained by

$$y_\theta(\boldsymbol{x}) = \sum_{i=0}^{K} y_i f(y_i|\boldsymbol{x}). \tag{3.5}$$

Section 4.2.2 describes more detailed implementation.

For regression problems such as economic forecasts, a model that outputs a single value is often used. However, Rothe, Timofte and Gool [12] have reported a model that outputs probability distributions made more accurate predictions than a model that outputs single value in age estimation.

Probability distributions are also easier to interpret for humans than single values. Even if a single forecast value is outputted, its grounds and how reliable it is are unknown. Meanwhile, humans can interpret the expected error of the forecast from the distribution, and estimate how reliable the forecast is even if the grounds of the distribution is unknown.

3.3 Optimal Composition of Humans and a Machine

This section formulates a human-machine ensemble method as a problem to find the number of humans n that minimizes the expected error of the ensemble, and analyse the solution.

3.3.1 Problem formulation

The forecast value of a human-machine ensemble, $Y_{\theta,h}(n|\boldsymbol{x})$ is the average of a forecast by a machine $y_\theta(\boldsymbol{x})$ and forecasts by n humans $\boldsymbol{h} = (h_1, \ldots, h_n)$:

$$Y_{\theta,h}(n|\boldsymbol{x}) = \begin{cases} \dfrac{y_\theta(\boldsymbol{x}) + \sum_{i=1}^{n} h_i}{n+1} & (n \geq 1) \\ y_\theta(\boldsymbol{x}) & (n = 0). \end{cases} \tag{3.6}$$

The problem is obtaining n that minimizes the expected squared error of the ensemble $\mathrm{MSE}(Y_{\theta,h}(n|\boldsymbol{x}))$. This is formulated as an optimization problem as follows:

$$\begin{aligned} \text{minimize} \quad & \mathrm{MSE}(Y_{\theta,h}(n|\boldsymbol{x})) \\ \text{subject to} \quad & n \geq 0. \end{aligned} \tag{3.7}$$

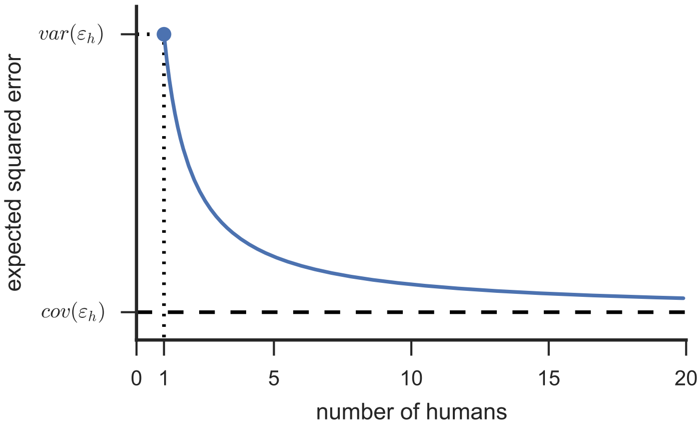


Figure 3.2: Graph of the expected squared error of the average of forecasts by n humans.

The expected squared error $\mathrm{MSE}(Y_{\theta,h}(n|\boldsymbol{x}))$ can be derived from the results of Lamberson and Page [7] introduced in Section 2.3. In equation (2.9), assumed that type a is human and type b is machine, $\mathrm{MSE}(Y_{\theta,h}(n|\boldsymbol{x}))$ is

$$\mathrm{MSE}(Y_{\theta,h}(n|\boldsymbol{x})) = \frac{n\,\mathrm{var}(\varepsilon_h) + \mathrm{var}(\varepsilon_\theta|\boldsymbol{x}) + n(n-1)\,\mathrm{cov}(\varepsilon_h) + 2n\,\mathrm{cov}(\varepsilon_\theta,\varepsilon_h)}{(n+1)^2}. \tag{3.8}$$

It is same as Lamberson and Page to obtain the optimal ratio of forecasters that have types. However, in this study, the expected error of a machine $\mathrm{var}(\varepsilon_\theta|\boldsymbol{x})$ varies due to the input $\boldsymbol{x}$. The next section analyses how the optimal number of humans n is determined according to each parameter.

3.3.2 Theoretical results

First, let us examine the relationship between the expected squared errors of forecasts by a group of humans $H(n)$ and the number of humans n. From (3.2), the expected squared error takes $\mathrm{var}(\varepsilon_h)$ when $n = 1$ and approaches $\mathrm{cov}(\varepsilon_h)$ when n increases. From the definition, $\mathrm{var}(\varepsilon_h) \geq \mathrm{cov}(\varepsilon_h)$ holds, so (3.2) decreases monotonically with $n \geq 1$. Figure 3.2 shows the graph of $\mathrm{MSE}(H(n))$ when n is a variable.

The expected squared error of a human-machine ensemble also approaches $\mathrm{cov}(\varepsilon_h)$ when $n \to \infty$ from (3.8). And it takes $\mathrm{var}(\varepsilon_\theta|\boldsymbol{x})$ when $n = 0$. However, whether $\mathrm{var}(\varepsilon_\theta|\boldsymbol{x})$ or $\mathrm{cov}(\varepsilon_h)$ is greater is different depending on $\boldsymbol{x}$, and sometimes (3.8) has a local minimum.

If (3.8) has a local minimum, the optimal number of humans n^* always exists. The condition to have a local minimum is

$$\operatorname{cov}(\varepsilon_\theta, \varepsilon_h) < \frac{3\operatorname{cov}(\varepsilon_h) - \operatorname{var}(\varepsilon_h)}{2}. \tag{3.9}$$

When taking the local minimum, n is

$$N^*(\boldsymbol{x}) = \frac{2var(\varepsilon_\theta|\boldsymbol{x}) - var(\varepsilon_h) + cov(\varepsilon_h) - 2cov(\varepsilon_\theta, \varepsilon_h)}{3cov(\varepsilon_h) - var(\varepsilon_h) - 2cov(\varepsilon_\theta, \varepsilon_h)}. \tag{3.10}$$

If $N^*(\boldsymbol{x})$ is greater than 0, the optimal n is $N^*(\boldsymbol{x})$, and if $N^*(\boldsymbol{x})$ is less than 0, the optimal n is 0. That is, n^* that minimizes the expected squared error when (3.8) has a local minimum is

$$n^* = \begin{cases} 0 & (N^*(\boldsymbol{x}) <= 0) \\ N^*(\boldsymbol{x}) & (N^*(\boldsymbol{x}) > 0). \end{cases} \tag{3.11}$$

If (3.9) does not hold, the optimal n does not always exist. In this case, the optimal n is 0 when $\operatorname{var}(\varepsilon_\theta|\boldsymbol{x}) \leq \operatorname{cov}(\varepsilon_h)$, and the optimal n does not exist when $\operatorname{var}(\varepsilon_\theta|\boldsymbol{x}) > \operatorname{cov}(\varepsilon_h)$. When n^* does not exist, you can make the expected squared error close to $\operatorname{cov}(\varepsilon_h)$ by increasing n.

Figure 3.3 shows four graphs that have different optimal n. These are grasped by adding a machine forecast to Figure 3.2 as the case of $n = 0$. The parameters related to machine forecasts $\operatorname{var}(\varepsilon_\theta|\boldsymbol{x})$ and $\operatorname{cov}(\varepsilon_\theta, \varepsilon_h)$ influence the values near $n = 0$. In particular, due to (3.9), if $\operatorname{cov}(\varepsilon_\theta, \varepsilon_h)$ is sufficient smaller than the parameters related only to humans, that is, if the forecasts between a machine and humans are sufficiently different, (3.8) has local minimum.

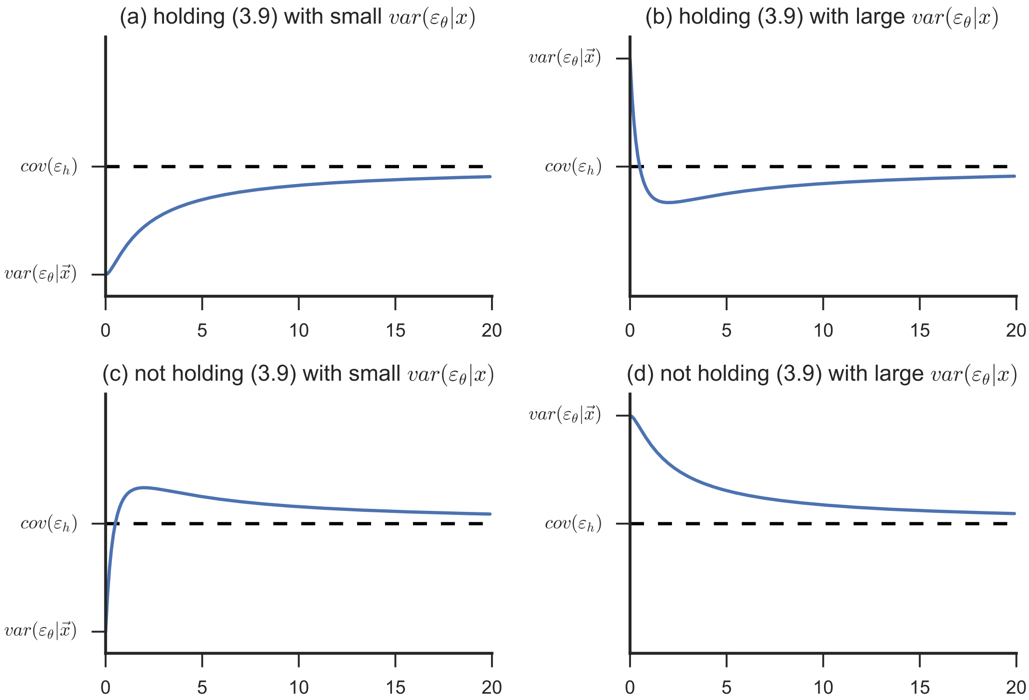


Figure 3.3: Approximate form of graphs about the expected squared errors of ensembles with n as a variable. The upper two are cases where the condition (3.9) holds, and $n^* = 0$ in (a) and $n^* = N^*(\boldsymbol{x})$ in (b). The lower two are cases where the condition (3.9) does not hold, and $n^* = 0$ in (c) and n^* does not exist in (d).

Chapter 4 Experiment

Chapter 3 showed the number of humans that minimizes the expected squared error of the ensemble in the case where a machine and humans follow the proposed models. However, it is unclear whether the proposed models is appropriate for real problems. Furthermore, even if the proposed model is appropriate, it is impossible to know the true values of parameters $\mathrm{var}(\varepsilon_h)$, $\mathrm{cov}(\varepsilon_h)$ and $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$. These can only be *estimated* from past forecasts. The distributions outputted by a machine $f_\theta(y|\boldsymbol{x})$ are also based only on the training set, so they are not always adaptable to unknown inputs. Therefore, we apply the proposed human-machine ensemble method to actual economic forecasts and investigate its performance by comparing with other forecasting methods.

In this chapter, Section 4.1 and Section 4.2 describe data and forecasting methods, respectively. The results are in Chapter 5.

4.1 Data

We used inflation in the U.S. as targets of economic forecasts. This section describes four inflation indicators and two surveys for economic forecasts.

4.1.1 Inflation

We consider four different measures of inflation: CPI-U for All Urban, All Items (*CPI*); CPI-U for All Urban, All items less food and energy (*CoreCPI*); Personal Consumption Expenditures (*PCE*); and PCE excluding food and energy (*CorePCE*). All measures are seasonally adjusted. CPI and CoreCPI are obtained from the Bureau of Labor Statistics[1)] and the sample period is from Jan. 1957 to Oct. 2016. PCE and CorePCE are obtained from the Bureau of Economic Analysis[2)] and the sample period is from Jan. 1959 to Oct. 2016.

CPI measures the total cost of goods and services purchased by urban consumers. It is commonly used as an indicator of inflation. CoreCPI excludes food and energy price from CPI. Short-term fluctuations of CoreCPI is smaller than CPI because the prices of food and energy change easily.

[1)] https://www.bls.gov/

[2)] https://www.bea.gov/

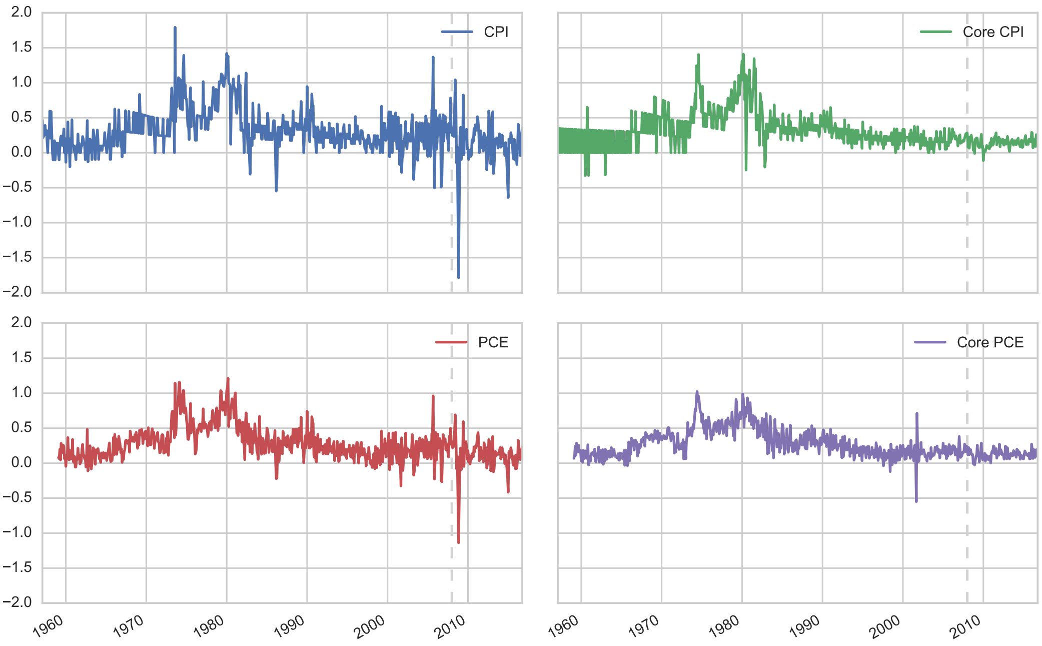


Figure 4.1: Monthly inflation rate measured by CPI, CoreCPI, PCE, and CorePCE. The dotted line represents 2008, which is the boundary between the traing set and test set.

PCE also measures the cost of goods and services purchased by consumers, but includes costs that CPI does not include such as medical expenses, and the data source is different. In addition, calculation procedures are different such that PCE uses chained weights while CPI uses fixed weights. CorePCE exclude food and energy price from PCE.

This thesis defines inflation rate from $t-1$ to t as:

$$\pi_t = \log\left(\frac{P_t}{P_{t-1}}\right) \times 100, \tag{4.1}$$

where P_t is the level of an index at time t. This thesis uses the terms "inflation" and "inflation rate" as defined in (4.1). π_t can be regarded as an approximation of the normal change rate by the following first order Taylor expansion approximation, assuming that the change is sufficiently small,

$$\log\left(\frac{P_t}{P_{t-1}}\right) = \log\left(1 + \frac{P_t - P_{t-1}}{P_{t-1}}\right) \approx \frac{P_t - P_{t-1}}{P_{t-1}}.$$

Figure 4.1 shows monthly inflation of the four indices.

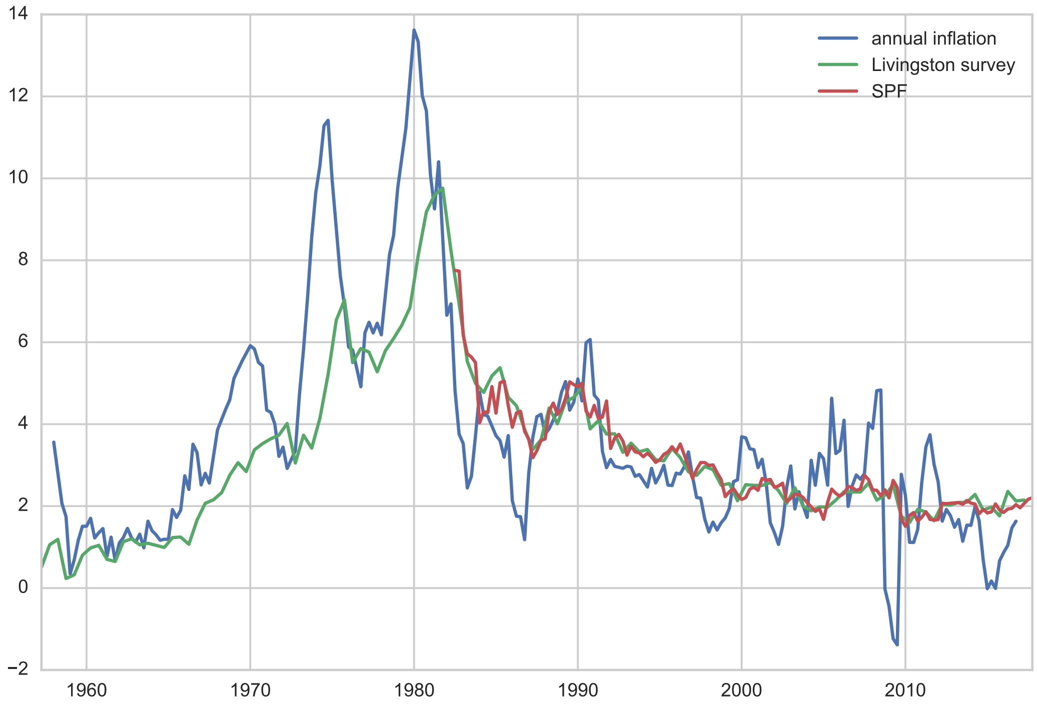


Figure 4.2: Annual CPI change rate and averaged forecasts by the Livingston Survey and the SPF.

4.1.2 Surveys

We used two surveys for economic forecasts: the Livingston Survey[1] and the Survey of Professional Forecasters[2] (SPF) as forecasts by humans. The Federal Reserve Bank of Philadelphia has been taking these surveys to economic experts. While the Livingston Survey includes the economists from government and academia, the SPF mainly covers industry. Figure 4.2 shows forecasts by each survey.

Livingston Survey The Livingston Survey is conducted twice a year, in June and in December, from 1946. This study treats from Jun. 1957 to Jun. 2016. The Livingston Survey includes only CPI as a survey item. It was reorganized in 2004, and started to forecast seasonally adjusted series from not seasonally

[1] https://www.philadelphiafed.org/research-and-data/real-time-center/livingston-survey

[2] https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/

adjusted series. It forecasts inflation 6 months later and 12 months later as the short term forecasts. The average number of respondents to 12 months forecast was 45.6 and the standard deviation was 11.6.

SPF The SPF began in 1968 and is conducted quarterly. It covers all four indicators as survey items. However, the survey of CPI started in 1981Q3, and the other three indicators started in 2007Q1. In this thesis, we used all the survey data up to 2015Q4. The average number of respondents to CPI was 34.4 and the standard deviation was 9.3.

4.2 Forecasting Models and Methodology

This section describes four forecasting methods and evaluation method for the experiment. The forecasting methods are a time-series model *ARMA(1,1)* as benchmark, a *recurrent neural networks (RNN)* as machine forecasts, survey forecasts as human forecasts, and the human-machine ensemble method proposed in Chapter 3. We created models that predict annual change rate of the four indices using ARMA(1,1) and RNN, and also created models that predict 6 months change rate of CPI. Due to the restrictions on the survey data, 12 months and 6 months forecasts for CPI are made from the Livingston Survey, and annual forecasts for all indices are made from the SPF. The human-machine ensemble method follows the restrictions of surveys. The data set is divided into the training set and test set, and creation of prediction models does not include the test set. Forecasting accuracy is evaluated by *Root Mean Squared Error (RMSE)* for the test set.

4.2.1 Time-series model

We used an *autoregressive moving average (ARMA) model* for benchmark forecasts as well as Ang et al. [1]. ARMA(1,1) is a traditional model for inflation forecasts. Ang et al. made a model based on quarterly inflation rate, and we also did so.

Assume that inflation is stationary to apply an ARMA model. A process is stationary means that the following equations hold for arbitrary t and k:

$$\begin{aligned} E[\pi_t] &= \mu \\ \mathrm{cov}(\pi_t, \pi_{t-k}) &= E[(y_t - \mu)(y_{t-k} - \mu)] = \lambda_k, \end{aligned} \tag{4.2}$$

where μ and λ_k are constants. This assumption is totally different from the model of Chapter 3, which assumed that the expected error varies depending on the situation t.

The specifications of the ARMA(1,1) model is

$$\pi_{t+1} = \mu + \phi\pi_t + \psi\epsilon_t + \epsilon_{t+1}, \tag{4.3}$$

where ϵ_t is white noise with variance σ^2 [17]. The parameters are μ, ϕ, ψ, σ^2, and they are estimated by maximum likelihood conditional on a zero initial residual, assumed the Gaussian white noise.

Applying this model, when an inflation rate at time t, π_t, is given, the forecast value of inflation rate after one period is

$$\hat{\pi}_{t+1|t} = E[\pi_{t+1}|\pi_t] = \mu + \phi\pi_t + \psi\hat{\epsilon}_t. \tag{4.4}$$

$\hat{\epsilon}_t$ is obtained by sequentially approximating like $\hat{\epsilon}_2 = \pi_2 - \mu - \phi\pi_1, \hat{\epsilon_3} = \pi_3 - \mu - \phi\pi_2 - \psi\hat{\epsilon_2}, \ldots,$ in which the initial value is $\hat{\epsilon}_0 = 0$. Since the term of ϵ disappears from forecasts after two period, the forecast values can be obtained sequentially by the following relationship:

$$\hat{\pi}_{t+k|t} = \mu + \phi\hat{\pi}_{t+k-1|t}.$$

However, $\hat{\pi}_{t+k}$ is a change rate for one period from $t+k-1$ to $t+k$, so the change rate from t to $t+k$ is the sum of k periods:

$$\hat{\pi}_{t+k,k} = \sum_{i=1}^{k} \hat{\pi}_{t+i}. \tag{4.5}$$

When $k = 2$, it is the 6 months later forecast, and when $k = 4$, it is the 12 months later forecast.

When using ARMA(1,1), the expected squared error of a forecast of one period ahead is always σ^2 regardless of t. Therefore, it is not applicable to the machine model in Section 3.2.2, which assumed the expected error varies depending on the situation. The next section describes RNN that can be used as a machine model for the proposed human-machine ensemble method.

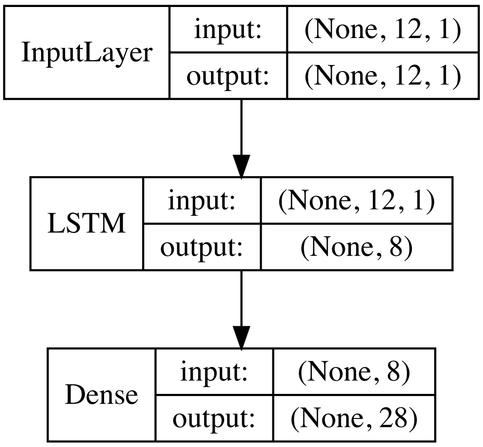


Figure 4.3: The RNN architecture used in the experiment. The inside of parentheses describes the dimentions of data, and the first term represents the batch size 32 instead of "None." The output dimention 28 is for forecasts of annual CPI change rate.

4.2.2 Recurrent neural network model

We used a RNN model that includes *Long short-term memory* (LSTM) units in a hidden layer. Let θ denote the RNN model, which outputs discrete probability distributions $f_\theta(y|\boldsymbol{x})$ when the past 12 months inflation $\boldsymbol{x} = [\pi_{t-11}, \pi_{t-10}, \ldots, \pi_t]$ are given. The forecast value y is an inflation rate for 6 months $\hat{\pi}_{t+6,6}$ or an inflation rate for 12 months $\hat{\pi}_{t+12,12}$. Figure 4.3 shows the RNN architecture that realizes these input and output. By using this architecture, we created five models that forecast annual inflation rates of the four indices, and CPI change rate for 6 months. The rest of this section describes the details of the architecture and how to create the prediction models.

LSTM Ordinary RNN has a problem that, given a long input sequence, the gradient rapidly increases or disappears when backpropagation is applied. It is called *vanishing gradient problem.* The LSTM deals with this problem, and makes long-term memory possible [4, 16]. Replace units in each hidden layer of a basic RNN with LSTM memory blocks shown in Figure 4.4. It looks like there

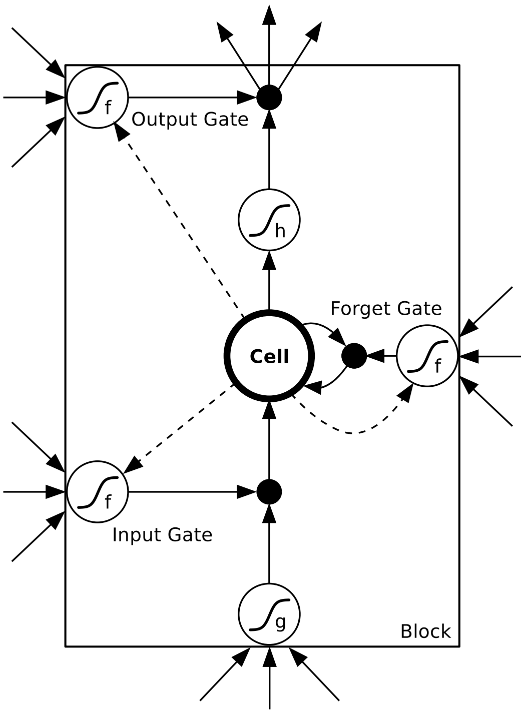


Figure 4.4: LSTM memory block [4]

are four inputs, but they are all the same, and identical to the input of a unit in an ordinary RNN's hidden layer. In the center of the block is a cell. Small black circles around the cell are gates, which have activation function f. As logistic sigmoid is usually used for f, the gate activations are between 0 (gate close) and 1 (gate open). The cell stores a state, and realizes the memory by inputting the state to itself after one time interval. However, a forget gate is on the way, and if the activation is close to 0, the state will be forgotten. The outputs of activation functions g and h at the cell input and output are also transmitted if the activation of the gate is close to 1, and are blocked if it is close to 0. With the above mechanism, the LSTM overcomes the limitation of RNN. The experiment used eight LSTM units in the hidden layer, with tanh for the activation functions g and h.

Output layer As mentioned in Section 3.2.2, the activation function of the output layer uses the *softmax function* so that the prediction model can output probability distributions. Let u_k denote input of unit k, and let o_k denote

Table 4.1: The minimum value, maximum value, number of output labels and difference between labels in each index.

	CPI(12M)	CPI(6M)	CoreCPI	PCE	CorePCE
min	0.1	-0.4	0.7	0.5	0.9
max	13.6	7.4	12.7	11.0	9.9
size	28.0	27.0	25.0	22.0	19.0
step	0.5	0.3	0.5	0.5	0.5

output. Then, the softmax function is

$$o_k = \frac{\exp(u_k)}{\sum_{j=1}^{K} \exp(u_j)}, \tag{4.6}$$

where K is the number of units in the output layer. The sum of outputs $v_1, \ldots, v_K$ is always 1. And the output o_k can be interpreted as the probability of belonging to the corresponding class. Making the labels $y_1, \ldots, y_K$ correspond to each unit of the output layer realizes the output of discrete probability distributions. For the labels, separate every 0.5 from the minimum value to the maximum value in the target outputs of training set. For example, when forecasting the annual CPI change rate, since the minimum value and the maximum value in the training set are 0.13 and 13.62 respectively, the values of labels are $[0.1, 0.6, 1.1, \ldots, 13.6]$, and the number of units is $K = 28$. Table 4.1 shows the minimum and maximum values, the number of labels, and the difference between labels in each series. Only when forecasting the 6 months CPI change rate, we set the step as 0.3 to make the dimensions of the outputs similar to others.

Training We divided the data set into the training set and test set. Creation of models uses only training set. The test set is for evaluation. The pairs of input and output are prepared as follows: First, the input is a sequence of monthly inflation for the past 12 months $\boldsymbol{x}_t = [\pi_{t-11}, \ldots, \pi_t]$ from a time t. And these are normalized with the maximum value in the training set as 1. On the other hand, the output is a vector that has K elements $\boldsymbol{d}_t = [0, 0, \ldots, 1, \ldots, 0]$, where the kth value is 1 and the others are 0 because y_k is closest to the correct value. Furthermore, we used the *cross entropy* as the error function:

$$E(\boldsymbol{w}) = -\sum_{t=1}^{T}\sum_{k=1}^{K} d_{tk} \log v_k(\boldsymbol{x}_t; \boldsymbol{w}). \tag{4.7}$$

The cross entropy is obtained by inverting the log-likelihood of parameters $\boldsymbol{w}$ for the training set $\{(\boldsymbol{x}_t, \boldsymbol{d}_t)\}(t = 1, \ldots, T)$, when assumed the output is posterior distribution. We used the *RMSprop* for the learning algorithm, and set the batch size 32. We stopped learning after 400 epochs, which number is obtained by cross validation.

4.2.3 Survey forecasts

In the Livingston Survey and the SPF, the number of respondents is different for every survey. The minimum number of respondents through the both surveys is 9. Armstrong [2] has proposed the principles for combining forecasts, and one of them is "use at least five forecasts when possible." Hence, we sampled five forecasters randomly from each survey to make the number of forecasters equal. The average values are regarded as the forecasts of the survey.

In addition, it is necessary to convert the surveyed values to appropriate inflation rate. The rest of this section describes the methods to calculate the forecast values based on the documentation of each survey.

Livingston Survey The participants in the Livingston Survey respond with not inflation rate but CPI level. The Livingston Survey presents the index level 2 months before to the participants. Based on that level, the participants forecast the value of the current month, $\hat{P}_t$, the one of 6 months after, $\hat{P}_{t+6}$, and the one of 12 months after, $\hat{P}_{t+12}$. However, it was in 1992 that the survey began to include the forecasts of the current month. According to the presence or absence of the forecast value of the current month, the forecast value of m months after, $\hat{\pi}_{t+m,m}$, is calculated by

$$\hat{\pi}_{t+m,m} = \begin{cases} \log\left(\frac{\hat{P}_{t+m}}{\hat{P}_t}\right) & \text{(if } \hat{P}_t \text{ exists)} \\ \log\left(\frac{\hat{P}_{t+m}}{P_{t-2}}\right)^{\frac{m}{m+2}} & \text{(otherwise).9} \end{cases} \tag{4.8}$$

SPF The participants in the SPF respond with inflation rate, but this is a simple change rate $P_{t+12}/P_t - 1$, not based on the definition of equation (4.1). It is impossible to convert this value to the inflation rate of (4.1), so the experiment regards it as the approximation. Note that since the forecast values are biased somewhat higher by not taking logarithm, the error becomes slightly larger.

4.2.4 Human-machine ensemble

The Human-machine ensemble method combines the RNN models and individual forecasts from each survey. There are the five RNN models, which forecast annual inflation rate of CPI after 6 months and CPI, CoreCPI, PCE, and CorePCE after 12 months. The Livingston Survey forecasts the index level of CPI after 6 months and 12 months, and the SPF forecasts annual inflation of the four indices.

Execution of the human-machine ensemble method requires the three parameters, $\mathrm{var}(\varepsilon_h)$, $\mathrm{cov}(\varepsilon_h)$ and $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$. We estimated these parameters from the forecasts for the training set. Let N denote the number of participants in the training set, and let ε_{h_i} and $\mathrm{var}(\varepsilon_{h_i})$ denote the error and the variance of a participant i respectively. Then, $\mathrm{var}(\varepsilon_h)$ is estimated by

$$\mathrm{var}(\varepsilon_h) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{var}(\varepsilon_{h_i}). \tag{4.9}$$

Also, $\mathrm{cov}(\varepsilon_h)$ and $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$ is estimated by equations (3.1) and (3.4) using the unbiased covariance for $\mathrm{cov}(\varepsilon_{h_i}, \varepsilon_{h_j})$.

Section 3.3.2 describes the optimal number of humans does not always exist. However, in reality there is an upper limit $N_{\max}$ depending on the number of respondents. Therefore, for the human-machine ensemble method, we used the n that minimizes the expected error in the range of $0 \leq n \leq N_{\max}$. In this case, unless $\mathrm{MSE}(Y_{\theta,h}(0)) = \mathrm{MSE}(Y_{\theta,h}(N_{\max}))$ holds, n is specified uniquely. When this equation holds, we set $n^* = N_{\max}$. The experiment fixed $N_{\max} = 5$, which is the same as the members sampled in the survey forecasts. Furthermore, when the expected error of humans only is less than that of combination of humans and a machine, the ensemble does not include the machine forecast.

4.2.5 Assessing forecasting methods

The test set starts from 2008. We also prepared the test set that starts from 1998 only for CPI. The reason is that the SPF has the short sample period since it started from 1981 and the Livingston Survey forecasts only CPI.

Table 4.2: The varieties of the ensembles.

	CPI-LIV	CPI-SPF	CoreCPI	PCE	CorePCE	CPI-6M	CPI-1998
target index	CPI	CPI	CoreCPI	PCE	CorePCE	CPI	CPI
survey source	LIV	SPF	SPF	SPF	SPF	LIV	LIV
forecast period	12M	12M	12M	12M	12M	6M	12M
dataset boundary	2008	2008	2008	2008	2008	2008	1998

The creation of RNN models and estimation of the ensemble parameters uses only the training set. The ARMA(1,1) model is recreated at each forecasting since the creation of the model is easier than RNN. That is, when forecasting annual inflation at 2008Q1, the creation of the model uses the samples up to 2007Q1, and when forecasting it at 2008Q2, the model is recreated using the samples up to 2007Q2. Thus, the creation of ARMA(1,1) models uses the data up to the time of forecasting.

The frequency of forecasts in the test set is quarterly as the same to the benchmark, ARMA(1,1). The number of samples in the test set that starts 2008 is 35 from 2008Q1 to 2016Q3. The Livingston Survey and the ensemble using it can only make forecasts on June and December, so the number of samples in the test set that starts 2008 is 17, and the one that starts 1998 is 37. Although the RNN models can make forecasts monthly, they forecast only on March, June, September and December.

Table 4.2 shows the all ensembles. There are seven different ensembles depending on the target index, the survey source, the period to forecast, and the boundary of the training and test set.

We assess forecast accuracy with the *Root Mean Squared Error (RMSE).* When an actual inflation and forecast value at time t is y_i and $\hat{y}_i$ respectively, and forecasting M times, the RMSE is

$$\mathrm{RMSE}(\hat{\boldsymbol{y}}) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)}. \tag{4.10}$$

In addition, we also report the ratio of RMSEs relative to the benchmark, ARMA(1,1).

Chapter 5 Empirical Results

This chapter describes the results of the experiment. Section 5.1 verifies whether the model assumed in Section 3.2 is applicable even to actual inflation forecasts. Section 5.2 describes the forecast accuracy of each forecasting methods, and Section 5.3 contains the estimated parameter values in the ensembles. We confirm the behavior of the proposed human-machine ensemble method in the real problem in Section 5.4. Finally, Section 5.5 summarizes the results and interpret them.

5.1 Model Verification

We assumed that the forecast errors by humans are unbiased, that is, the average forecast error is 0 in Section 3.2.1. Figure 5.1 shows the error distributions when the individuals forecast the annual CPI change rate. The upper side is the Livingston Survey, and the lower side is the SPF. The figure indicate that the errors of individuals are distributed around zero equally. The means of these distributions are -0.40 in the Livingston Survey and 0.37 in the SPF.

We also assumed that the machine outputs can be regarded as posterior distributions of the target value. Figure 5.2 shows examples of the discrete probability distributions actually outputted by the RNN model that forecast the annual CPI change rate. Since the model learned the training set, this distribution is applicable to the training set well, but to what extent it applies to the test set depends on its generalization ability. However, when applying the human-machine ensemble method, we consider the distributions are reliable, and regard the variance as the expected squared error. Therefore, we examined the relationship between the variance of the distribution outputted by the model and the actual squared error. Figure 5.3 plots those relationships in the test set. Although there are variations, it is considered to be correlated.

5.2 Forecast Accuracy

Table 5.1 shows the relative RMSEs for the test set of each forecasting method. Each entry reports the ratio of its RMSE to the RMSE of the benchmark,

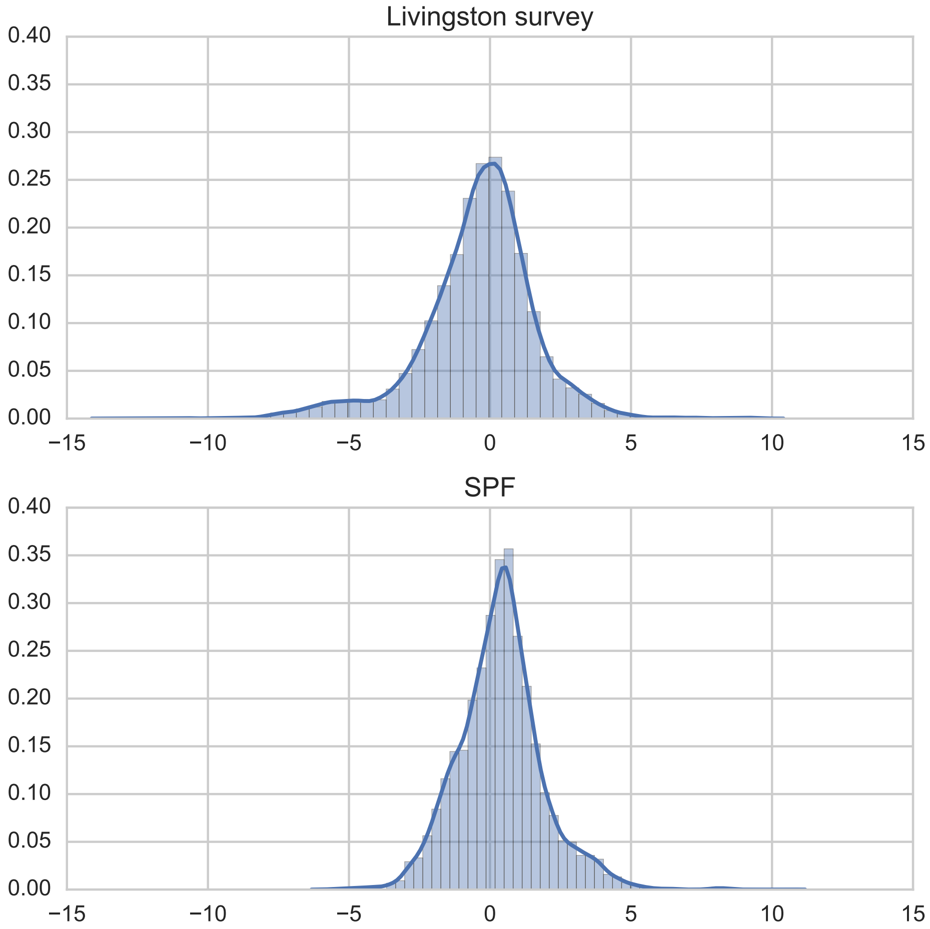


Figure 5.1: Distributions of individual errors in the Livingston Survey and the SPF when forecasting annual CPI change rate. The numbers of samples are 5258 and 4706, and the means are -0.40 and 0.37, respectively.

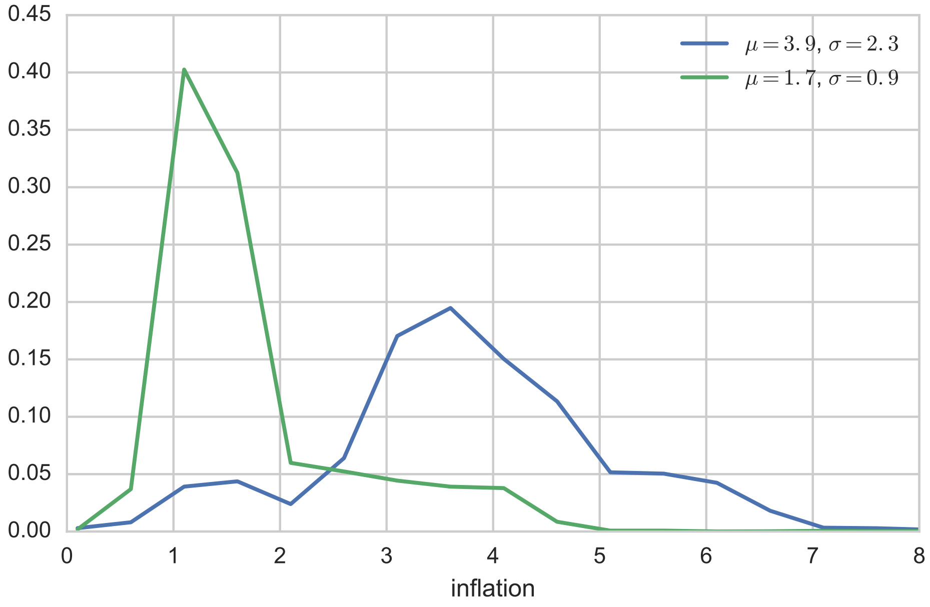


Figure 5.2: Examples of distributions outputted by the RNN models.

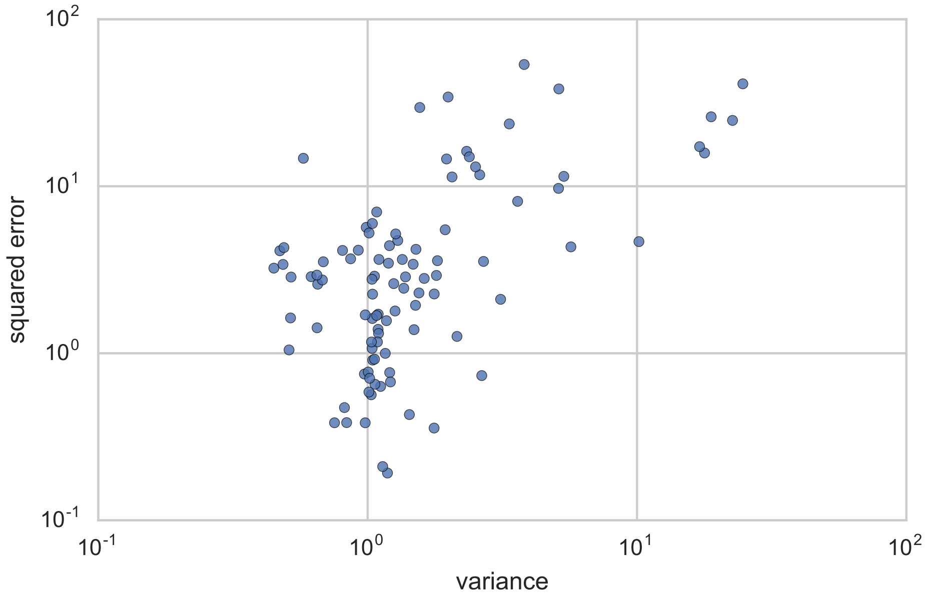


Figure 5.3: Relationship between the variances of the ditributions outputted by the RNN model and the actual squared errors in the test set.

Table 5.1: Relative RMSEs of each forecasting method in the test set. Bold entries are the smallest RMSEs in each column.

	CPI-LIV	CPI-SPF	CoreCPI	PCE	CorePCE	CPI-6M	CPI-1998
ARMA	1.000	1.000	1.000	1.000	1.000	1.000	1.000
RNN	0.889	0.889	0.788	0.933	0.842	1.010	0.941
survey	0.704	**0.736**	**0.767**	**0.711**	0.848	0.939	0.940
ensemble	**0.689**	**0.736**	**0.767**	0.724	**0.827**	**0.931**	**0.755**

ARMA(1,1). That is, if the value is smaller than 1, it is more accurate than the benchmark, and otherwise, it is less accurate than the benchmark.

Most of the values in the row of RNN were less than 1, that is, the RNN models made more accurate forecasts than ARMA(1,1) overall. However, the RNN model that forecasted 6 months CPI change was less accurate than ARMA(1,1). This suggests that ARMA(1,1) is better for shorter term inflation forecasts than RNN, as mentioned in the past study [10].

The average forecasts of five people sampled from the surveys outperformed the benchmark in all cases. And the survey was the most accurate in forecasting PCE. In comparison with RNN, the surveys made more accurate forecasts except for CorePCE. While the RNN models make forecasts based only on the past 12 months sequences, humans can include various information such as economic conditions in the forecasts. This is why the surveys were more accurate than other models.

The ensembles made the most accurate forecasts in 4 out of 7 cases. In other two cases, CPI-SPF and CoreCPI, the RMSEs were the same as the surveys, and it was the best score. This is because the variances outputted by the RNN models were larger than $\mathrm{cov}(\varepsilon_h)$ of the SPF, and the ensembles always selected to use humans only.

5.3 Parameter Estimation

Table 5.2 shows the parameters $\mathrm{var}(\varepsilon_h)$, $\mathrm{cov}(\varepsilon_h)$ and $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$ of each ensemble. They were estimated from the training set. $\mathrm{var}(\varepsilon_h)$ is the average variance of the individual errors in a survey; $\mathrm{cov}(\varepsilon_h)$ is the average covariance in errors

Table 5.2: Parameter values of each ensemble estimated from the training set.

	CPI-LIV	CPI-SPF	CoreCPI	PCE	CorePCE	CPI-6M	CPI-1998
var(ε_h)	2.869	1.887	1.379	1.453	1.134	0.986	2.952
cov(ε_h)	1.846	0.873	0.455	0.450	0.168	0.652	1.924
cov($\varepsilon_\theta, \varepsilon_h$)	1.772	1.096	0.668	0.510	0.056	0.497	1.821

of an arbitrary human pair; cov($\varepsilon_\theta, \varepsilon_h$) is the average covariance in the errors of a RNN model and an arbitrary human. The smaller var(ε_h), the higher the ability of individuals in the group; the smaller cov(ε_h), the more diverse the group; the smaller cov($\varepsilon_\theta, \varepsilon_h$), the more different the forecasts of humans and machine.

The row of var(ε_h) shows that the values of CPI-LIV and CPI-1998 were large. CPI-LIV and CPI-1998 forecasted the annual CPI change rate usigng the Livingston Survey as the forecasts of humans. CPI-6M had the small var(ε_h) as it forecasted shorter period, 6 months. The var(ε_h) of the ensembles that used the SPF were smaller than CPI-LIV. In particular, the values of CoreCPI and CorePCE were small. This suggests that it is easier to forecast indexes that excludes food and energy.

Next, let us look at cov(ε_h) and cov($\varepsilon_\theta, \varepsilon_h$). In CPI-LIV, CorePCE, CPI-6M and CPI-1998, where the ensembles performed well, cov($\varepsilon_\theta, \varepsilon_h$) was smaller than cov($\varepsilon_h$). That is, the forecasts of humans and machines were more diverse than that of only humans. On the other hand, in CPI-SPF, CoreCPI and PCE, where the ensemble did not perform well, cov($\varepsilon_\theta, \varepsilon_h$) was larger than cov($\varepsilon_h$). This means that adding a machine to a group of humans does not increase the diversity of forecasts.

5.4 Behavior of Human-Machine Ensemble

As described in Section 3.1, the basic idea of the human-machine ensemble method is that when the machine receives a pattern not in the past and outputs large variance, the ensemble combines forecasts of humans. This section confirms whether the human-machine ensemble method behaves as expected

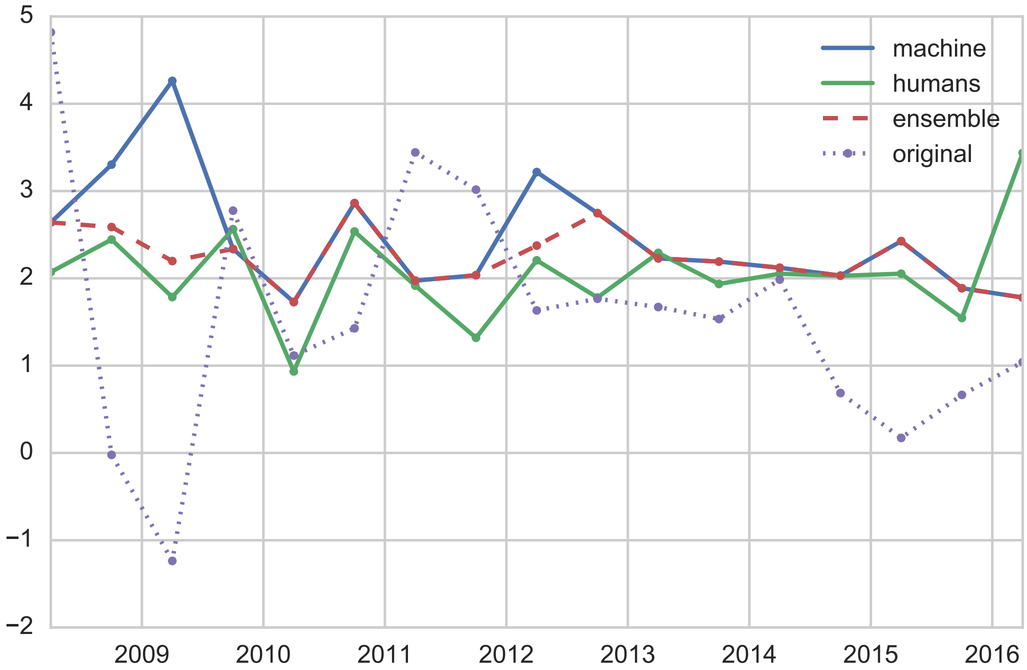


Figure 5.4: Behavior of the ensemble CPI-LIV in the test set. The purple dotted line shows the actual inflation; the blue line shows the forecasts of the RNN model; the green line shows the average forecasts of five people randomly sampled from the Livingston Survey; and the red dotted line shows the forecasts of the ensemble.

through CPI-LIV.

Figure 5.4 is a line graph that shows the annual CPI change rate and the forecasts of the RNN model, the Livingston survey and the ensemble in the test set. The dotted purple line represents the actual inflation, so the closer to this means the forecast is accurate. The blue line shows the forecasts of the RNN model, and the green is forecasts using the Livingston Survey. The forecasts of the ensemble is the red line. When the ensemble used the machine only or humans only, the line overlaps each line.

In the test set, the period from 2008 to 2009 is applicable to the assumed scenario. From the end of 2007 to 2009 was global financial crisis triggered by the subprime mortgage. Particularly, the Bankruptcy of Lehman Brothers in September 2008 decreased the index levels of CPI and PCE. In the training set from 1957 to 2007, the annual CPI change rate never took a negative value, so it is the pattern not in the past. In fact, the forecasts of the RNN model in December 2008 and June 2009 were largely out. However, at this time the forecast of the ensemble depended on forecasts of humans rather than the forecasts of machine. In other words, the ensemble emphasized forecasts of humans because the machine received patterns that were not in the past and outputted distributions with large variances.

Table 5.3 shows the squared error and variance of each forecast of the RNN model, the number of humans and the squared error of each forecast of the ensemble CPI-LIV in the test set. This table also suggests that there is a correlation between variances and squared errors of machine forecasts since when a variance is large, the squared error is also large. In addition, when the variance was large, the ensemble adopted many human forecasts and reduced the error.

Basically, the ensemble behaved as expected, but the number of humans n can only take 0 or $N_{\max} = 5$. This is because the ensemble parameters shown in Table 5.2 did not satisfy the condition (3.9), which is for having a local minimum. If this condition is not satisfied, the optimal n is 0 or $N_{\max}$, as in the two graphs in the lower side of Figure 3.3.

The reason why equation (3.9) is not satisfied is that humans and a machine do not make so different forecasts. The index level at the time of forecasting has a great influence on forecasts both of humans and machines. For example, if the inflation rate at the time of forecasting is 1%, both of humans and machines will forecast around 1% as annual inflation, and when these forecasts goes out, both humans and machines take a similar error. Thus, $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$ can not take a sufficiently small value for $\mathrm{var}(\varepsilon_h)$ and $\mathrm{cov}(\varepsilon_h)$, and can not satisfy the equation (3.9).

5.5 Summary and Interpretations

As a result of model verification, the models proposed in Section 3.2 are applicable to actual inflation forecasts. The mean error of individuals in the Livingston Survey and the SPF was about 0, and there was no bias in the distributions.

Table 5.3: Behavior of the ensemble CPI-LIV in the test set. From the left, the columns show the squared errors of the RNN model, the variances outputted by the RNN model, the numbers of humans the ensemble employed, and the squared errors of the ensemble.

	machine error	variance	n	ensemble error
Jun 2008	4.74	1.53	0	4.74
Dec 2008	11.06	2.15	5	6.81
Jun 2009	30.23	2.81	5	11.80
Dec 2009	0.19	1.63	0	0.19
Jun 2010	0.38	0.90	0	0.38
Dec 2010	2.06	1.72	0	2.06
Jun 2011	2.16	1.15	0	2.16
Dec 2011	0.96	1.22	0	0.96
Jun 2012	2.51	2.12	5	0.55
Dec 2012	0.96	1.65	0	0.96
Jun 2013	0.31	1.35	0	0.31
Dec 2013	0.43	1.33	0	0.43
Jun 2014	0.02	1.28	0	0.02
Dec 2014	1.81	1.21	0	1.81
Jun 2015	5.08	1.45	0	5.08
Dec 2015	1.49	1.07	0	1.49
Jun 2016	0.55	0.96	0	0.55

Thus, the proposed human model, where forecasts of humans are unbiased, is valid. Additionally, the variances outputted by the machine and the squared errors in the test set were distributed approximately along a line with a slope of 1 passing through the origin. Hence, it is possible to create a prediction model that performs as the proposed machine model, which outputs probability distributions of which variance is regarded as expected squared error.

We compared forecast accuracy of the proposed human-machine ensemble method with other forecasting methods. The ensembles made the most accurate forecasts in 4 out of 7 cases. In other two cases, the ensembles used forecasts of only humans, so the scores were the same as the survey forecasts. In the last case, the forecasts of the ensemble were less accurate than the survey forecasts, but were more accurate than ARMA(1,1) and RNN.

As a reason why the proposed ensemble method did not work well in 3 out of 7, it is suggested that $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$ was greater than $\mathrm{cov}(\varepsilon_h)$. In other words, adding a machine to a group of humans did not increase the diversity, and did not reduce the error.

The three cases CPI-SPF, CoreCPI and PCE, where the ensemble method did not perform well, used the SPF as forecasts of humans. Since the SPF began from 1981, its sample period is shorter than the Livingston Survey. This implies that estimation of the parameters $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$ and $\mathrm{cov}(\varepsilon_h)$ did not work well.

Finally, we assumed the human-machine ensemble method takes the advantage of humans forecasts when the expected error of the machine is large as shown in Figure 1.1. Therefore, we verified the behavior of an ensemble after the Bankruptcy of Lehman Brothers. As a result, the ensemble adopted human forecasts as expected, and reduced the error.

Chapter 6 Discussion

In this chapter, we discuss the scope and application of the proposed human-machine ensemble method.

6.1 Scope of Human-Machine Ensemble

The experiment applied the proposed human-machine ensemble method to actual economic forecasts, especially inflation forecasts. As a result, in the majority cases, the ensemble method forecasted more accurately than a time-series model, a machine alone and a survey alone. However, it did not always make accurate forecasts, sometimes made worse forecasts than a survey. Based on this result, this section discusses the conditions for applying the human-machine ensemble method.

First, the forecast target is a numerical value. In the human-machine ensemble method, since the ensemble combines humans and a machine so as to minimize the expected error, it is necessary to forecast numerical values that can be defined errors. However, if the error can be defined, it is possible to extent the proposed method to other problems than regression such as classification. Even in that case, the ensemble should use the machine if the expected error of the machine is small, and use humans if it is large.

Second, the prediction model used as a machine model in the ensemble method must be able to output the expected error according to input. Since an ARMA model assumes the process is stationary, the variance of forecasts is constant regardless of the input. Therefore it can not be used as a machine in the ensemble method.

Third, it is necessary to estimate parameters $\mathrm{var}(\varepsilon_h)$ and $\mathrm{cov}(\varepsilon_h)$ of human population in advance. In the experiment, we estimated these parameter values from long-term survey data, the Livingston Survey and the SPF. The parameter estimation is also possible when a firm makes demand forecasts by known people. Even when using humans whose parameters are unknown such as crowdsourcing, for example, an expectation-maximization (EM) algorithm performs parameter estimation. However, it is difficult to adopt this method to forecasts where the true value is obtained only quarterly or monthly, such as economic forecasts.

Finally, the most important factor is the relationship between humans and a machine. It is necessary for the ensemble method that a machine is more accurate in some cases while a group of humans is more accurate in other cases. If the machine always makes more accurate forecasts than the group of humans, you should use the machine only; if the group of humans always makes more accurate forecasts than the machine, you should not use the machine. Also, if errors of the machine and humans are always similar, the proposed method does not perform well. This is because if the error of humans is also large when the expected error of the machine is large, it is meaningless to combine them. However, if there are cases where either the machine or group of humans is accurate, the proposed method performs well.

Additionally, from the theoretical and empirical results, $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$ should be sufficiently less than $\mathrm{cov}(\varepsilon_h)$. It is necessary to satisfy equation (3.9) for the ensemble to adopt the number of humans between zero and maximum. That is, if the diversity of the humans and machine is sufficiently large relative to diversity of only humans, the ensemble can adopt various number of humans. However, the all seven ensembles used in the experiment did not satisfy equation (3.9), and the three of them, in which $\mathrm{cov}(\varepsilon_h)$ was larger than $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$, did not adopt any machine forecasts at all, or made worse forecasts than those of only humans. Thus, it is not necessary to satisfy equation (3.9), but it is necessary that $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$ is less than $\mathrm{cov}(\varepsilon_h)$ to use the human-machine ensemble method efficiently.

6.2 Application of Human-Machine Ensemble

This section describes the human-machine ensemble method with crowdsourcing and the merits other than forecast accuracy of the human-machine ensemble method.

Crowdsourcing is available as humans in the proposed method instead of the Livingston Survey and the SPF. Crowdsourcing is derived from *crowd* and *outsourcing*, and means outsourcing tasks to an undefined large group of people [5]. Platforms for crowdsourcing such as Amazon Mechanical Turk are on the Web, and you can easily use the labor force on them. The advantage of using crowdsourcing for the human-machine ensemble method is that you can know the required number of workers in advance. On the other hand, whereas the surveys such as the Livingston Survey and the SPF employ specific experts, the crowdsourcing employs unspecific ordinary people. However, Michigan survey, which is a survey for economic forecasts, employs ordinary people, but makes as accurate forecasts as the Livingston Survey and the SPF. This implies that the similar accuracy can be obtained even if crowdsourcing is used for the ensemble method instead of surveys by experts. It is, of course, necessary to estimate the parameters $\mathrm{var}(\varepsilon_h)$, $\mathrm{cov}(\varepsilon_h)$ and $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$. Therefore, it is not possible to apply the ensemble method immediately. You must gather samples for parameter estimation.

The human-machine ensemble method also suggests a form of collaboration for machines and humans. We assumed that the machine makes larger expected error when given a sequence not in the past. Conversely, observing the number of humans that combined in the ensemble method reveals the input that makes the machine error larger. In this way, you can find problems at which the machine is not good but humans are good. It also suggests the direction of improving the machine forecasts, or a new human-machine ensemble method.

Chapter 7 Conclusion

This thesis proposed a human-machine ensemble method for economic forecasts. Short-term economic forecasts are difficult, so various methods have been proposed so far. Machine learning methods such as RNN create a prediction model based mainly on past time series. In surveys for economic forecasts, individual respondents make forecasts considering various information such as policy and economy. Thus, whereas machines are good at forecasts that depends on the past patterns, humans can make forecasts flexibly even when changes not in the past such as financial crisis. However, traditional ensemble methods can not utilize these difference. Therefore, we modeled a machine that outputs expected squared error depending on input by introducing posterior distributions to the output of the machine. Moreover, based on this model, we proposed an ensemble method to dynamically change the combination of the machine and humans depending on input. The ensemble minimizes the expected error.

In addition, we conducted an experiment to apply the proposed method to actual inflation forecasts. The objectives were model verification, evaluation of forecast accuracy and confirmation of the behavior of the proposed method. We created RNN prediction models that output discrete probability distributions for machine forecasts, and used two surveys for human forecasts. From these combinations, we constructed seven ensembles. Evaluation was performed using test set prepared separately from the training set. We obtained following three results. First, the proposed models were applicable to actual inflation forecasts. Second, in 4 out of 7 cases, the ensemble made more accurate forecasts than the machine only and the humans only. Finally, we showed that the ensemble behaved as expected. When machine received a sequence not in the past, the ensemble drew human forecasts.

The proposed ensemble method improved forecast accuracy in the majority cases, but not always reduce the error. The results suggest that the difference of a machine and humans is important for the proposed method to perform well. It is necessary that the group of machines and humans is more diverse than the group of humans only. In the experiment, machines and humans did not make so different forecasts since they are influenced by the index at the time of forecasting. However, there is a possibility of a more efficient human-machine ensemble method by extending the models or improving machine forecasts.

Finally, the main contributions of this thesis are as follows:

Modeling forecasts by humans and a machine and inventing an ensemble method using them We modeled the case where the expected error changes with each forecast by assuming the posterior distribution to the outputs of machines. In addition, we proposed a method which determines the optimal number of humans dynamically depending on the expected error of a machine forecast.

Implementation and evaluation of economic forecasts using the proposed ensemble method We applied the human-machine ensemble method to inflation forecasts. As a result, we confirmed the proposed model can be applied to real inflation forecasts and forecast accuracy improves with 4 out of 7 data.

Acknowledgments

First of all, I would like to express deepest gratitude to my supervisor, Associate Professor Shigeo Matsubara at Kyoto University, for his continuous instruction, invaluable advice, and earnest argument. My special appreciation goes to my advisers, Associate Professor Qiang Ma at Kyoto University and Professor Jun Yoshida at Kyoto University, whose opinions and information have helped me very much throughout the production of this study. I am also indebted to Professor Toru Ishida, who provided helpful comments and suggestions. Finally, I would like to thank all members of Ishida and Matsubara laboratory, especially to Shinsuke Goto, for their various support and valuable discussions.

References

[1] Ang, A., Bekaert, G. and Wei, M.: Do Macro Variables, Asset Markets or Surveys Forecast Inflation Better?, *Journal of Monetary Economics*, Vol. 54, No. 4, pp. 1163–1212 (2007).

[2] Armstrong, J. S.: Combining Forecasts, *Principles of Forecasting*, Springer, pp. 417–439 (2001).

[3] Choudhary, M. A. and Haider, A.: Neural network models for inflation forecasting: an appraisal, *Applied Economics*, Vol. 44, No. 20, pp. 2631–2635 (2012).

[4] Graves, A.: Long Short-Term Memory, *Supervised Sequence Labelling with Recurrent Neural*, Springer, chapter 4, pp. 31–38 (2012).

[5] Howe, J.: *Crowdsourcing*, Crown Publishing Group (2008).

[6] Krogh, A. and Vedelsby, J.: Neural Network Ensembles, Cross Validation, and Active Learning, *Advances in Neural Information Processing Systems*, Vol. 7, pp. 231–238 (1995).

[7] Lamberson, P. J. and Page, S. E.: Optimal Forecasting Groups, *Management Science*, Vol. 58, No. 4, pp. 805–810 (2012).

[8] Michell, T. M.: Artificial Neural Networks, *Machine Learning*, McGraw-Hill, chapter 4, pp. 191–193 (1997).

[9] Moshiri, S. and Cameron, N.: Neural network versus econometric models in forecasting inflation, *Journal of Forecasting*, Vol. 19, No. 3, pp. 201–217 (2000).

[10] Nakamura, E.: Inflation forecasting using a neural network, *Economics Letters*, Vol. 86, No. 3, pp. 373–378 (2005).

[11] Page, S. E.: *The Defference*, Princeton University Press (2008).

[12] Rothe, R., Timofte, R. and Van Gool, L.: Deep expectation of real and apparent age from a single image without facial landmarks, *International Journal of Computer Vision (IJCV)* (2016).

[13] Surowiecki, J.: *The Wisdom of Crowds*, Doubleday (2004).

[14] Ueda, N. and Nakano, R.: Generalization error of ensemble estimators, *Proceedings of International Conference on Neural Networks (ICNN)*,

Vol. 1, pp. 90–95 (1996).

[15] Zhou, Z.-H.: Combination Method, *Ensemble Methods*, CRC Press, chapter 4, pp. 67–98 (2012).

[16] 岡谷 貴之: 深層学習, 講談社 (2015).

[17] 沖本 竜義: 経済・ファイナンスデータの計量時系列分析, 朝倉書店 (2010).