AI-Driven Exploration and Prediction of Company Registration Trends with Registrar of Companies (RoC)

TEAM LEADERS

510521104304: EZHUMALAI P

PHASE-2:DOCUMENT SUBMISSION

OBJECTIVIE:

The problem is to perform an AI-driven exploration and predictive analysis on the master details of companies registered with the Registrar of Companies (RoC). The objective is to uncover hidden patterns, gain insights into the company landscape, and forecast future registration trends.

PHASE 2: Innovation

Consider exploring advanced AI algorithms like time series forecasting or ensemble methods for improved predictive accuracy.

Dataset Link: https://tn.data.gov.in/resource/company-master-data-tamil-nadu-upto-28th-february-2019

ABSTRACT

This paper describes a new type of ensembles that aims at improving the predictive per-formance of these approaches in time series forecasting. Ensembles are recognised as oneof the most successful approaches to prediction tasks. Previous theoretical studies of en-sembles have shown that one of the key reasons for this performance is diversity among ensemble members. Several methods exist to generate diversity. The key idea of the workwe are presenting here is to propose a new form of diversity generation that explores some-specific properties of time series prediction tasks. Our hypothesis is that the resulting en-semble members will be better at addressing different dynamic regimes of time series data.Our large set of experiments confirms that the methods we have explored for generatingdiversity are able to improve the performance of the equivalent ensembles with standard diversity generation procedures.
Keywords: ensemble methods, time series forecasting, bagging, maximum embed

1. Introduction

Ensembles are known to be among the most competitive forms of solving predictive tasks.-Several studies (e.g, Dietterich (2000); Brown and Kuncheva (2010)) have been carriedout to understand and explain the reasons for their competitiveness in a wide range of application domains. Diversity among the individual components of ensembles is known tobe a key element to generate a successful model. Bagging (Breiman, 1996) is a well knownand simple type of ensembles that consists of a large set of standard tree-based modelsthat are grown with the goal of generating a diverse set of models. Diversity in bagging is created through the use of different random bootstrap samples of the original training setto grow each tree. The main idea behind this paper is to propose a variant of bagging wherethe forms of generating diversity are biased towards specific characteristics of time series forecasting tasks. Namely, we aim at trying to have different forms of handling the diverse

2. Problem Description

The standard definition of time series forecasting assumes the existence of a set of time-ordered observations of a variable, y1, y2, · · · , yt, where yi is the value of Y measured attime i, and defines the predictive task as trying to forecast the future values of this variable for time stamps s > t. Many variants of this general task exist, including the use ofother measured variables as potential predictors of the future values of the target series Y .Still, the general assumption is that there is an unknown function that ”maps” the past observations into the future values of Y , i.e. Yt+h = f (〈DescriptorsOf T heP ast〉), and the learning goal is to approximate this function using some prediction error criterion and a-historical record of observed values.The predictors used for forecasting the future values of Y are usually the most recentobservations of Y , as the basic assumption of time series forecasting is that of the existence-of some form of correlation between successive observations of the series. This is the ap-proach used on most approaches to time series forecasting, like for instance the well-knownARIMA models (e.g. Chatfield (2013)). This is also the idea of time delay coordinate em-bedding (Takens, 1981) that is a standard procedure for applying out of the box regression tools to time series forecasting tasks. This strategy assumes that future values of the seriesare only dependent on a limited number of previous values. In this context, delay coordinateembedding consists in using the k past values of a time series as descriptors of the stateof the system at an instant t. If k is appropriate, it’s possible to capture the dynamics of the time series from the embed vectors rt = 〈yt, yt−1, ..., yt−k〉 . Under this assumption, wecan then use any regression tool to obtain a model of the form Yt+h = f (rt) that specifiesthe relationship between a set of predictors (described by the embed vector) and the future values of the serie

3. Bagging for Time Series Forecasting

The approaches based on the use of the recent past values of a time series (the embed) as predictors require setting a critical parameter - how many past values to include, i.e. the size of the embed. Setting this parameter is not trivial and it may involve trying different alternative values for the embed size and use some reliable performance estimation process as a means for deciding the ”optimal” value. The main drawback of these approaches is the fact that frequently there may not exist one single correct answer. In effect, non-stationary
series and the occurrence of different regime shifts along time may lead to the best value being clearly time-dependent. This is one of the main motivations for our work. Another being previous work showing that having diversity in ensembles is a key ingredient to boost

able 1: Summary of the main characteristics of the ensemble variants.

Embed size	Extra predictors
E	All models use kmax. None.
E+S	All models use kmax. All models use μY and $\sigma 2 Y$ calculated with the respective embed.
DE	One third of the models use kmax, another third uses kmax/2, and the last third uses kmax/4.	None.
DE+S	One third of the models use kmax, another third uses kmax/2, and the last third uses kmax/4.	All models use μY and $\sigma 2 Y$ calculated with the respective embed.
DE±S		

One third of the models use kmax,
another third uses kmax/2,
and the last third uses kmax/4.
Half of the models using a certain
embed size use μY and σ2
Y
calculated with the respec

4. Experimental Evaluation

The main goal of our experimental evaluation is to check the validity of the hypothesis that the variants of bagging we have described in Section 3 will outperform standard bagging on time series forecasting tasks. In this context, our baseline benchmark is a standard bagging implementation using the approach tagged as E in the list given in Section 3. All other four variants will use the same base data (the values of the past kmax observations) as trainingset, but will use it in a different way, e.g. by using only part of it in some models or byusing it to generate extra features.We also compare the ARIMA model, a more standard time series forecasting approach,to the same baseline approach to shed more light onto the overall competitiveness of ourproposal. Since ARIMA models usually require a significant parameter tuning effort to

obtain good results, we used the auto.arima function available in the R package forecast

(Hyndman et al., 2014) which automatically searches for an optimal model.

Table 2: Data sets used

ID	Time series	Data source	Data characteristics
1	Temperature	Bike Sharing (Fanaee-T and Gama, 2013)	Daily values from Jan. 1, 2011
2	Humidity		
3	Windspeed		to Dec. 31, 2012 (731 values)
4	Count of total bike rentals		
5	Temperature		
6	Humidity		Hourly values from Jan. 1, 2011
7	Windspeed		to Dec. 31, 2012 (7379 values)
8	Count of total bike rentals		
9	Flow of Vatnsdalsa river	Icelandic river (Tong et al., 1985)	Daily values from Jan. 1, 1972 to Dec. 31, 1974 (1095 values)
10	Minimum temperature	Porto weather1	Daily values from Jan. 1, 2010 to Dec. 28, 2013 (1457 values)

11 Maximum temperature

12 Maximum steady wind

13 Maximum wind gust

14 Total precipitation

All five alternative forms of bagging and the ARIMA model were tested on fourteen real world time series obtained from three different data sources as described in **Table 2**.Each one of these series of data was treated separately from the others in their respective data source (e.g. information on the weather was not used to predict the total number of bike rentals). Moreover, please note that each time series of the Bike Sharing data source is available in two formats (daily and hourly), which were treated as different time series forecasting tasks. All fourteen time series were pre-processed to overcome some well-known issues with this type of data. Specifically, we have created all data sets used in our experiments with the series of the differences between successive values, and not from the original absolute values, in order to avoid trend effects. We have not, however, assumedany shape of these effects, if they exist. The target variable for all tasks was set to the nextvalue of the series of differences.

Table 3:

Paired comparisons results in format Nr.Wins (Statistically Significant Wins)/ Nr.Losses (Statistically Significant Losses) M kmax Variant Wins/Losses

1020

20

E+S 13 (11) / 1 (1)

DE 7 (7) / 7 (3)

DE+S 13 (10) / 1 (0)

DE±S 14 (12) / 0 (0)

ARIMA 7 (3) / 7 (4)

30

E+S 11 (9) / 3 (2)

DE 10 (6) / 4 (3)

DE+S 10 (5) / 4 (2)

DE±S 10 (9) / 4 (2)

ARIMA 6 (3) / 8 (4)

1500

20

E+S 13 (10) / 1 (1)

DE 8 (6) / 6 (3)

DE+S 13 (10) / 1 (0)

DE±S 14 (12) / 0 (0)

ARIMA 7 (3) / 7 (4)

30

E+S 11 (9) / 3 (2)

DE 9 (7) / 5 (3)

DE+S 10 (7) / 4 (2)

Finally, Figure 1 presents more detailed values of the percent difference of MSE withrelation to the baseline for each time series (summarized on **Table 5)**. For visualization

purposes, a signed common logarithm was applied to the results. The represented metric

is, therefore,

sgn(M SEX − M SEE) · log

(|

|

|

|

100 · (M SEX − M SEE)

M SEE

|

|

|

| + 1

)

Once more, our proposals, in general, seem to do well in comparison to standard bagging. The more apparent exceptions to this are the results obtained for kmax = 30 on time series5, 6 and 8, with which the ARIMA model seems to struggle as well. The high variance of performance of the ARIMA approach is well illustrated in this figure. Although it achieves a higher decrease in MSE on datasets 9-14, our proposed approaches are also able to perform well on those while almost always beating the results of the ARIMA model on datasets 1-8with a significant margin.

5. Related Work

Using different predictors on each member of an ensemble is not a novel idea. Random forests (Breiman, 2001) for instance, grow each tree in such a way that for each node a random subset of the features is used to select the best split. Random subspaces (Ho, 1998) is another example of an ensemble method that uses diverse sets of features among the models. This approach consists in randomly selecting subsets (usually of the same size) of the feature space to build each base learner. Contrary to our approach, none of these previous works address time series tasks. Moreover, our subsets of predictors are not

6. Conclusions

This paper describes an initial attempt at proposing ensembles for time series forecasting tasks. The main motivation of this ongoing work is the observation that handling time series tasks requires several decisions in terms of how we describe the recent dynamics of the observed values of the series. Settling on a single answer to these decisions may bedangerous in real world time series where one frequently observes changes in the dynamicproperties of the variable being measured. Ensembles are a well-known answer to this type of problems by taking advantage of diversity among models to reduce both the biasand variance components of the prediction error. Motivated by these observations we haveproposed an initial set of forms of injecting diversity into ensembles that takes into account some specific challenges posed by time series data. Namely, we have considered alternative ways of representing the recent observations of the target series among the members of theensemble. These include the use of different sizes of the embed and also the addition ofvariables summarising the recent observed values.