A Comparative Study of Disease Prediction for Different Population Size and Time Constraints

Dingfei Guo*, Wei Li

[1]Department of Computer Engineering, Taiyuan Institute of Technology, Taiyuan, China
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

*guodingfei0119@163.com, liwei@ict.ac.cn

Abstract. In recent years, the rapid advancement of the internet has propelled humanity into an era characterized by an exponential increase in the volume of information. Decision-making in a variety of economic and social fields has been significantly impacted by the development of unstructured data by big data. Data prediction is now the primary use of big data, driven by advancements in deep-learning neural networks. Examples include typical disease forecasts based on electronic health data, influenza forecasts, traffic forecasts, and more. Therefore, based on the reading and analysis of relevant literature in the past three years, this paper categorizes the development of data prediction into three areas. One is classification based on research questions, another is classification based on research methodology, and the third aspect is classification rooted in measurement methods.

Keywords: Data prediction; Deep learning; Disease Prediction; Classification.

1 Introduction

The importance and use of data has grown significantly due to the expansion of the internet and electronic information. It has been utilized in various areas such as weather forecasting, illness patterns, human health, traffic flow, user activity, and others, as well as mining human needs. Big data predictions can employ an additional 80% of unstructured data to make decisions, whereas people's decisions in the past mostly depended on 20% of structured data. The use of big data prediction technologies has improved both the accuracy of prediction findings and the correctness of decisions.

Particularly, big data technology can make medical researchers study the development and spread of diseases more excellently, thereby improving the ability to predict diseases. By analyzing big data, medical researchers can identify risk factors for diseases, thereby preventing and treating them in advance. In addition, big data can also assist medical researchers in gaining a better learning of the effects and side effects of drugs, thereby improving guidance on their usage.

In this paper, we investigated the prediction of diseases in populations of varying sizes and under different time constraints. More specifically, we used the technique introduced in [20] to construct the question set, the method set and the measurement

M. Yu et al. (eds.), *Proceedings of the 2024 5th International Conference on Big Data and Informatization Education (ICBDIE 2024)*, Advances in Intelligent Systems Research 182,
https://doi.org/10.2991/978-94-6463-417-4_49

set for the above topic. That is, all possible research questions, research methods and measurements of this topic were searched and discussed. Then, literatures of the topic were collected and the research question, research method and measurements of each paper were classified. After labeling existing research questions, research methods and measurements, we identified possible creativities or research opportunities for the topic.

According to the above idea, in Section 2, we classify the research question using two distinct criteria: the range of the predicted time and the predicted target. The first criterion can be classified into two types: short-term and long-term prediction. The second criterion includes two types: individual people and groups of people. Based on these two criteria, a research question set with four elements is constructed. The research questions in related literature are categorized and examined.

In Section 3, we classify the research methods based on two different dimensions: the experimental model and the types of data processed. The experimental model mainly includes two major categories: recursive neural networks and graph neural networks, while some unconventional models are placed in other categories for overall discussion. The data types processed include time series and graph structure data with spatial characteristics. Based on these two dimensions, a set of research methods with six elements was constructed. Then, the research methods in the relevant literature were classified and investigated.

In Section 4, we classify the measures from two different dimensions: System Factors and Metric. System Factors mainly includes Model / Algorithms / Methods, Datasets, and includes some uncommon factors into other categories for overall discussion. Metric mainly includes some indicators of assessment error, recall, AUC and time, and some uncommon indicators into other categories for overall discussion. Based on these two dimensions, a set of 15 elements was constructed. Then, the measures in the relevant literature are classified and investigated.

2 Classification of Research Objects

Table 1. Different Research Objects

Predicted Object	Range of Predicted Time	
	Short-term	**Long-term**
Individual	I. [2][11][13]	II.[2][3][4][7][8][10][11]
Group	III.[6][9][12][13]	IV.[1][5][6][9][14][15][17][18][19][20]

2.1 Criteria

In Table 1, we used two separate and different criteria to classify the research objects into different types:

1) Predicted Object. In some prediction tasks, prediction objects are single objects, such as in healthcare tasks, in clinical prediction is a single patient, and in some time

series tasks, the prediction object is a single node. However, in some prediction tasks, the prediction object is groups, rather than a single point.

2) Range of Predicted Time. Short-term prediction refers to the situation where the predicted node is the next time node of the current time point, or the situation of the predicted object after a few days. Long-term prediction refers to the development of predicted objects after weeks or even months.

2.2 The Classification

Type I: Individual & Short-term.

This type is to predict the performance of individual objects in the short term based on historical information. References ([2][11][13]) belong to this type. For example, reference [2] improves the performance of predictive models by addressing the challenges of insufficient data and presentation inconsistencies in healthcare forecasting with deep learning approaches.

Type II: Individual & Long-term.

This type is to predict the performance of individual objects in the long term based on historical information. References ([2][3][4][7][8][10][11]) belong to this type. For example, reference [11] addresses the problem of decreasing time dependence of RNN when sequence length is large in the task of predicting future health information of patients using historical electronic records.

Type III: Group & Short-term.

This type is to predict the performance of group objects in the short term based on historical information. References ([6][9][12][13]) belong to this type. For example, reference [13] improved the accuracy of data prediction by modeling the dynamics of the data sequence to extract the temporal features in the data prediction. At the same time, the reference [13] also belongs to Type I. It can be applied to both individual prediction task and group prediction task.

Type IV: Group & Long-term.

This type is to predict the performance of group objects in the long term based on historical information. References([1][5][6][9][14][15][17][18][19][20]) belong to this type.

3 Classification of Research Methods

Table 2. Different Research Methods

The Type of Data Processed	Model Classification	
	Recursive Neural Network	Graph Neural Networks
Data Forecasting Based on Time Series	I.[3][5][7][8][10][11][13][18][19][20]	II.[5][6][13]
Data Forecasting Based on Spatial		IV. [6][9][16][18][19][20]

3.1 Criteria

In Table 2, we studied the research methods of the study subjects and classified them into two dimensions:

1) **The type of data processed.** A time series is a chronological sequence of data points. Spatial data is used to describe goals from reality and unify the data to indicate the shape, size, location, and distribution characteristics of spatial entities.

2) **Model classification.** A recurrent neural network (RNN) is a model that can be specifically designed to process time series data. A graph neural network is used to process graph data.

3.2 The Classification

Type I: Data forecasting based on time series & Recursive Neural Networks.

This type is to forecast data based on time series by using Recursive Neural Networks. References ([3][5][7][8][10][11][13][18][19][20]) belong to this type.

Type II: Data forecasting based on time series & Graph Neural Networks.

This type utilizes GNN to predict data based on time series. References ([5][6][13]) belong to this type.

Type III: Data forecasting based on time series & Others.

This type is to forecast data based on time series by using other methods. References ([1] [2] [4]) belong to this type.

Type IV: Data forecasting based on spatial & Graph Neural Networks.

This type is to forecast data based on spatial by using GNN. References ([6][9][16][18][19][20]) belong to this type.

4 Review of Experimental Analysis

Table 3. Experiments with Different Metric and Factors

Metric	System Factors	
	Model/Algorithms/Methods	Datasets
RMSE/MSE/MAE/MAP/MAPE/Accuracy	[1][2][5-13][19][20]	[1][5-13][19][20]
Recall	[7][10]	[7]
AUC	[2-4][10][12][18]	[8][12]
Time	[2][3]	[2]

In table 3, we present different measures of the study subjects and classify them.

4.1 Metric of Evaluation

RMSE (Root Mean Square Error) means the square root of the square of the deviation from the predicted value from the true value to the ratio of the number of observations n.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}$$

MSE (Mean Square Error) means the interpolation of the true and predicted values is then averaged.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2$$

MAE (Mean Absolute Error) means The average of the absolute error between the predicted and observed values.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{m}|\widehat{y_i} - y_i|$$

MAP (Mean Average Precision) means the average of AP. AP is the average precision rate for measuring all recall rates. MAPE (Mean Absolute Percentage Error) means relative error metric value, which uses absolute values to avoid positive and negative errors canceling each other.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\widehat{y_i} - y_i}{y_i}\right|$$

Accuracy represents the ratio of the number of correct decisions to all the decisions

$$\mathrm{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall represents the proportion of those correctly predicted from all positive cases.

$$\mathrm{Recall} = \frac{TP}{TP + FN}$$

TP (True Positive): The number of times that the judgment is actually positive and the judgment is also positive, that is, the number of times that the decision is positive and correct. FP (False Positive): the number of times a negative case but a positive case, that is, the number of positive cases but wrong judgment. TN (True Negative): the number of times that the actual negative case is also negative, that is, the number of times that the negative case is judged and judged correct. FN (False Negative): the

number of times that is actually a positive case but that is, the number of negative cases but wrong judgment.

4.2 System Factors

Model/Algorithms/Methods represents different models, algorithms or methods used for the same prediction task, which are compared to evaluate the task using the same metric. The number of epochs represents the count of rounds of sample iterations in datasets. Let the neural network run on the training data set to observe the value of the loss function. If the value of the loss function is small enough to meet the requirement, it indicates that the neural network fits well. Datasets are collections of data used in prediction tasks. Training models on various datasets can often lead to a more robust evaluation of the model's scientific validity.

5 Conclusions

We classify the results of deep learning networks in recent years for data prediction tasks. The results show that deep learning shows good performance and representation ability in both long-and short-term prediction tasks. Recursive neural network correlation models are more able to process time series data and are widely used in healthcare and clinical prediction. For large disease prediction and traffic prediction tasks, the data is not only dependent in time dimension, but also in spatial and spatial temporal dimension.

References

1. Cao, Defu, et al. "Spectral temporal graph neural network for multivariate time-series forecasting." Advances in neural information processing systems 33 (2020): 17766-17778.
2. Choi, Edward, et al. "GRAM: graph-based attention model for healthcare representation learning." Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017.
3. Choi, Edward, et al. "Using recurrent neural network models for early detection of heart failure onset." Journal of the American Medical Informatics Association 24.2 (2017): 361-370.
4. Choi, Edward, et al. "Mime: Multilevel medical embedding of electronic health records for predictive healthcare." Advances in neural information processing systems 31 (2018).
5. Deng, Songgaojun, et al. "Cola-gnn: Cross-location attention based graph neural networks for long-term ili prediction." Proceedings of the 29th ACM international conference on information & knowledge management. 2020.
6. Gao, Junyi, et al. "STAN: spatio-temporal attention network for pandemic prediction using real-world evidence." Journal of the American Medical Informatics Association 28.4 (2021): 733-743.
7. Gao, Jingyue, et al. "Camp: Co-attention memory networks for diagnosis prediction in healthcare." 2019 IEEE international conference on data mining (ICDM). IEEE, 2019.

8. Gao, Junyi, et al. "Stagenet: Stage-aware neural networks for health risk prediction." Proceedings of The Web Conference 2020. 2020.
9. Huang, Rongzhou, et al. "LSGCN: Long short-term traffic prediction with graph convolutional networks." IJCAI. Vol. 7. 2020.
10. Kumar, Srijan, Xikun Zhang, and Jure Leskovec. "Predicting dynamic embedding trajectory in temporal interaction networks." Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019.
11. Ma, Fenglong, et al. "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks." Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017.
12. Ma, Liantao, et al. "Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 01. 2020.
13. Pareja, Aldo, et al. "Evolvegcn: Evolving graph convolutional networks for dynamic graphs." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 04. 2020.
14. Pei, Sen, and Jeffrey Shaman. "Initial simulation of SARS-CoV2 spread and intervention effects in the continental US." MedRxiv (2020): 2020-03.
15. Qian, Zhaozhi, Ahmed M. Alaa, and Mihaela van der Schaar. "CPAS: the UK's national machine learning-based hospital capacity planning system for COVID-19." Machine Learning 110.1 (2021): 15-35.
16. Veličković, Petar, et al. "Graph attention networks." ar**v preprint ar**v: 1710. 10903 (2017).
17. Yang, Zifeng, et al. "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions." Journal of thoracic disease 12.3 (2020): 165.
18. Yao, Huaxiu, et al. "Modeling spatial-temporal dynamics for traffic prediction." arXiv preprint arXiv:1803.01254 1.9 (2018).
19. Wei Li, Creativity Search and Verification: A Beginner's Guide to Scientific Research, ISBN: 9798838096753, independently published, 2022.
20. Gao, Junyi, et al. "Popnet: Real-time population-level disease prediction with data latency." Proceedings of the ACM Web Conference 2022. 2022.