FUZZY NEAREST NEIGHBOUR METHOD FOR TIME-SERIES FORECASTING[1]

SAMEER SINGH
UNIVERSITY OF PLYMOUTH
SCHOOL OF COMPUTING,
PLYMOUTH PL48AA,
UNITED KINGDOM
E-mail: s1singh@plym.ac.uk

ABSTRACT

This paper explores a nearest neighbour pattern recognition method for time-series forecasting. A nearest neighbour method (FNNM) based on fuzzy membership values is developed. The main aim of the forecasting algorithm is to make single point forecasts into the future on the basis of past nearest neighbours. The nearest neighbours are selected using a membership threshold value. The results include the mean absolute percentage error and the direction error for sales data of three real control products. The forecasts are compared to the actual values over a period of twenty months taking four years of monthly data in the estimation period. The results are very encouraging for further work on the development of fuzzy nearest neighbour methods in forecasting.

1. INTRODUCTION

In the past, conventional statistical techniques such as ARIMA models have been extensively used for forecasting (Delurgio, 1988). It has been realised that statistical techniques have limited capabilities when modelling time-series data, and more advanced methods including neural networks have been frequently used (Azoff, 1994; Refenes et al., 1997). Farmer and Sidorowich (1988) state that local approximation and nearest neighbour techniques can be used in forecasting to give results several orders of magnitude better than conventional statistical techniques. The main characteristic of a nearest neighbour method lies in the fact that it is a very useful tool for identifying the relationship between current and past values with reasonable accuracy. A nearest neighbour method is usually based on the identification of several historic neighbours that may be then used for forecasting by either averaging their contribution or using an extrapolation method. The accuracy of the method is directly dependent on the ability to identify good neighbours.

In order to detail the nearest neighbour procedure for forecasting, we first formalise time-series analysis mathematically. Consider the time series as a vector $\mathbf{y} = \{y_1, y_2, \ldots y_n\}$ where n is the total number of points in the series. Often, we also represent such a series as a function of time, e.g. $y_n = y_t$, $y_{n-1} = y_{t-1}$, and so on. In order to define any neighbour mathematically, we choose to encode the time series y as a vector of change in direction. For this purpose, a value y_i is encoded as 0 if $y_{i+1} < y_i$, as a 1 if $y_{i+1} > y_i$ and a 2 if $y_{i+1} = y_i$. The complete time-series is encoded as $(b_1, \ldots b_{n-1})$ where b_i is either a 0 or a 1 depending on the downward or upward series movement respectively. For a total of k segments in a pattern, it is encoded as a string of k *b* values.

The main aim of the nearest neighbour method is to forecast y_{n+1} if the time-series data $\{y_1, y_2, \ldots y_n\}$ is given. For this purpose, we proceed by identifying the nearest neighbours of y_n in the past data. A total of k neighbours selected are represented as $\{x_{t1}, x_{t2}, \ldots x_{tk}\}$ where x_{tk} is the series value at time tk. The selection of nearest neighbours is based on a fuzzy algorithm which will be detailed in the next section. The forecast $\ddot{y}_{n+1}$ of actual y_{n+1}, is based on either the averaging or extrapolation of values $\{x_{t1+1}, x_{t2+1}, \ldots x_{tk+1}\}$. The next section describes the fuzzy nearest neighbour algorithm (FNNM)

[1] *Singh, S. "Fuzzy Nearest Neighbour Method for Time-Series Forecasting", Proc. 6th European Congress on Intelligent Techniques and Soft Computing (EUFIT'98), Aachen, Germany, vol. 3, pp. 1901-1905 (7-10 August, 1998)*

in both its simple and modified form. This description will be followed by details of the time-series data that will be forecast using FNNM. The results section details the accuracy of the forecast measured using standard error measures including the mean absolute percentage error and the direction error. The paper concludes by highlighting the salient points of the study.

2. FUZZY NEAREST NEIGHBOUR METHOD (FNNM)

The Fuzzy Nearest Neighbour Method is based on the identification of nearest neighbours in the past data. The approach is based on identifying local variation in time-series rather than modelling the global nature of the series behaviour. The relationship between a current series value y_n and the past values may be determined using a fuzzy membership function (Pal and Majumder, 1986). The membership function may be better understood as a proximity function. The function value of 1 denotes the nearest of all neighbours whereas function values towards the tail end near 0 denote poor neighbours. The algorithm gives the user the flexibility to choose the optimal membership threshold λ so that only those neighbours whose membership exceeds λ are allowed to contribute to the further analysis. The following algorithm summarises the FNNM.

① Taking the current time-series value as y_n, identify the proximity of y_n and past values. The proximity between y_n and a past value y_i is calculated as follows:

$$\mu(y_i) = [1 + \{d(y_i, y_n)/F_d\}^F e\,]^{-1.0}$$

Here, $\mu(y_i)$ denotes the fuzzy proximity of y_i and y_n and d is the Euclidean distance between y_i and y_n. The constants F_d and F_e are determined through experimentation.

② Scale the membership/proximity values within the [0, 1] scale using:

$\mu'(y_i) = \mu(y_i) - \min(\mu(y_i))/ ((\max(\mu(y_i)) - \min(\mu(y_i)))$ where min and max are taken for $1 \le i \le n-1$.

③ Identify the nearest neighbours that satisfy the criteria: $\mu'(y_i) \ge \lambda$, where λ is a threshold within the [0, 1] range set by the experimenter. As λ increases, a smaller numbers of nearest neighbours are selected. A total of k nearest neighbours selected from the past data are denoted as $\{x_{t1}, x_{t2}, \ldots x_{tk}\}$.

④ Forecast for time step n+1 by averaging the k nearest neighbours:
$\ddot{y}_{n+1} = (x_{t1+1} + x_{t2+1} + \ldots + x_{tk+1})/ k$

⑤ The algorithm is optimised (optimal λ) for minimal error in prediction. The error measures used in this paper include:

Mean Absolute Percentage Error (MAPE) = $1/p \sum |y_{n+1} - \ddot{y}_{n+1}|/ y_{n+1}$

Direction of change error = error when $y_{n+1} - y_n > 0$ and $\ddot{y}_{n+1} - y_n \le 0$
or error when $y_{n+1} - y_n \le 0$ and $\ddot{y}_{n+1} - y_n > 0$

Here y_{n+1} is the actual value, $\ddot{y}_{n+1}$ is the forecast and p is the total number of forecasts made.

The above algorithm represents the basic philosophy behind the nearest neighbour method. It is possible to modify this further to ensure that nearest neighbours are only selected when the series movement (up or down) in their range matches the series movement around y_n. For this purpose, we may replace the earlier description of step 3 with the following description (modified FNNM):

③ Identify the nearest neighbours that satisfy the criteria:
$\mu'(y_i) \ge \lambda$, and $b_{i-1} = b_{n-1}$ and $b_{i-2} = b_{n-2}$ and $\ldots b_{i-j} = b_{n-j}$

where λ is a threshold within the [0, 1] range and j lies between 1 and 5 and is set by the experimenter. As λ increases, a smaller numbers of nearest neighbours are selected. A total of k nearest neighbours selected from the past data are denoted as $\{x_{t1}, x_{t2}, \ldots x_{tk}\}$. In our study, we have selected j =3. If for a given y_n, we can not find a nearest neighbour using the above criteria, we recursively lower j till one is found.

The FNNM and modified FNNM algorithm will be tested on a total of four different time-series data that is described in the next section.

3. SALES DATA

The sales data for forecasting comes from a manufacturing company ABX located in the South-West England. The company manufactures control equipment. The sales data for three different products A, B, C and D has been selected. Sales data for a total of 70 months is available for all these four products (Figures 1-4). The data seems considerably non-linear and is difficult to predict using linear techniques such as regression analysis. The next section shows the results obtained on predicting this sales data using the FNNM and the modified FNNM method. For this purpose, the overall data will be divided into an estimation period (50 points) and a forecast period (20 points). Results will be shown for the forecast period.

Figure 1. Plot for product A data

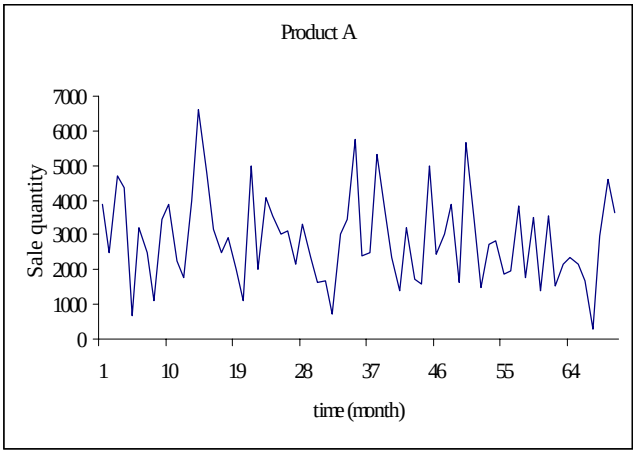


Figure 2. Plot for product B data

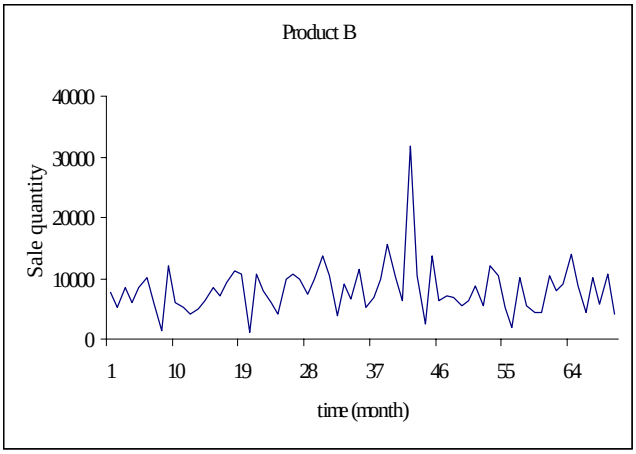


Figure 3. Plot for product C data

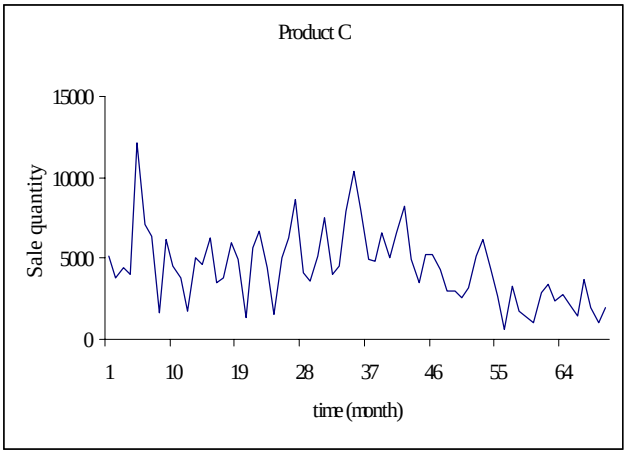
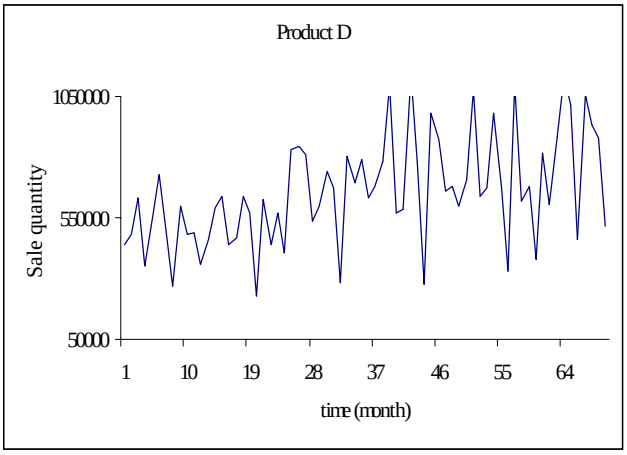


Figure 4. Plot for product D data



4. RESULTS

The results are shown in Table 1. An optimal membership threshold λ was selected which gave the least error by varying λ within the [0, 1] range. The FNNM results for MAPE and % direction error have been compared with the modified algorithm.

Table 1. MAPE and Direction success in % for products A-D

Product	Optimal λ	FNNM MAPE	FNNM Direction Success %	Modified FNNM MAPE1	Modified FNNM Direction Success %
A	.1	90.2	85	111.1	85
B	.2	57.5	80	55.3	85
C	.6	99.1	60	102.0	55
D	.2	37.3	70	31.3	80

The results show that the nearest neighbour method is very good at forecasting the sale quantity of the four products. For products B and D, the modified algorithm has a superior performance compared to the original FNNM method. For products A and C, the original FNNM methods yields better results. We next plot and discuss the predicted product sales with the actual values.

Product A

Figure 5 shows the plot of the predicted and actual values for product A. A total of twenty forecasts were made for the monthly data. Series 1 represents the actual data. Here the FNNM method gives reasonably good forecasts considering the small amount of data used for analysis.

Figure 5. The actual values for product A compared with simple and modified fuzzy nearest neighbour (FNNM) predictions

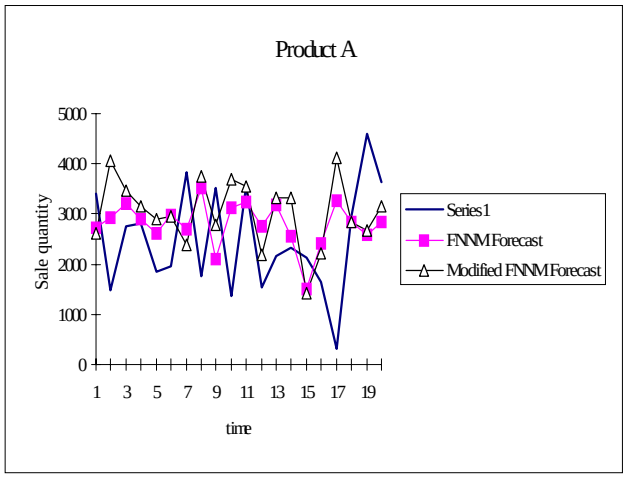
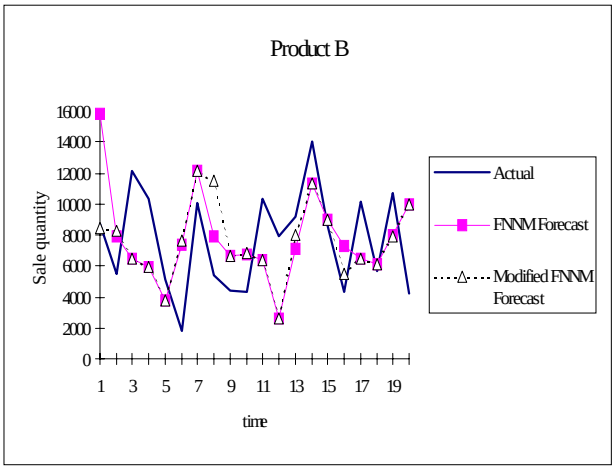


Figure 6. The actual values for product B compared with simple and modified fuzzy nearest neighbour (FNNM) predictions



Product B

Figure 6 shows the plot for product B. Here the modified FNNM is slightly better and the forecasts are more accurate than the previous product. The nearest neighbour forecasts seem to smooth out towards the tail end.

Product C

Figure 7 shows the plot for product C. The forecasts are less smooth and over-estimate for the majority of the months.

Figure 7. The actual values for product C compared with simple and modified fuzzy nearest neighbour (FNNM) predictions

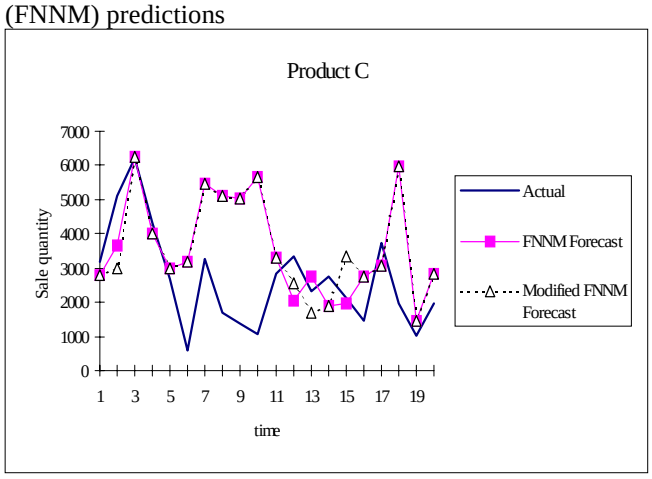
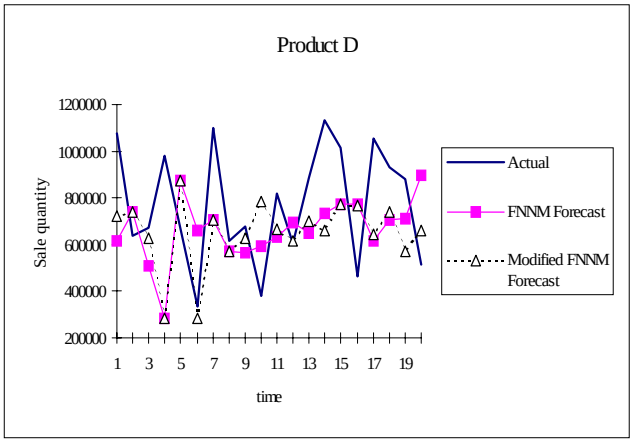


Figure 8. The actual values for product D compared with simple and modified fuzzy nearest neighbour (FNNM) predictions



Product D

The nearest neighbour forecasts are much more smooth here. This smoothing phenomenon leads to conservative forecasts and fails to capture the non-linear series movements. It is expected that this effect can be inhibited by increasing data size. This may also be explored using extrapolation than averaging of nearest neighbours.

5. CONCLUSION

The sales data analysis using the fuzzy nearest neighbour method yields good results. The accuracy of the forecasts depends on the ability of the algorithm to detect good neighbours. Further work is necessary to compare the forecasting accuracy based on nearest neighbour extrapolation. The main conclusions of this study are:

i) The nearest neighbour method for forecasting is a useful tool when forecasting non-linear data.
ii) Results on the real sales data used in this study show that the FNNM is capable of correctly forecasting the direction of future change with more than 80% accuracy on products A, B and D, and with 55% accuracy on product C.
iii) The forecasting accuracy of the FNNM is data dependent and therefore its optimisation parameters including threshold λ should be selected through experimentation.
iv) The nearest neighbour method will, theoretically speaking, work better with large data sets where we have an increased probability of finding good neighbours.

Further work should now be prompted towards advanced nearest neighbour methods for forecasting. It is important to compare the ability of such methods against well established statistical and neural network methods.

REFERENCES

Azoff, M. E. 1994. Neural Network Time Series Forecasting of Financial Markets, John Wiley and Sons.
Delurgio, S. 1998. Forecasting: Principles and Applications, McGraw-Hill.
Farmer, J. D. and Sidorowich, J. J. 1988. Predicting Chaotic Dynamics, in Dynamic Patterns in Complex Systems, J. A. S. Kelso, A. J. Mandell and M. F. Shlesinger (Eds.), pp. 265-292, Singapore: World Scientific.
Pal, S. K. and Majumder, D. D. 1986. Fuzzy mathematical approach to pattern recognition, John Wiley, New York.
Refenes, A. N., Burgess, A. N. and Bentz, Y. 1997. "Neural networks in financial engineering: A study in methodology," IEEE Transactions on Neural Networks, vol. 8, no. 6, pp. 1222-1267.