Forecasting Intermittent Demand by Fuzzy Support Vector Machines

Yukun Bao, Hua Zou, and Zhitao Liu

Department of Management Science and Information System, School of Management, Huazhong University of Science and Technology, Wuhan 430074, China
yukunbao@mail.hust.edu.cn

Abstract. Intermittent demand appears at random, with many time periods having no demand,which is probably the biggest challenge in the repair and overhaul industry. Exponential smoothing is used when dealing with such kind of demand. Based on it, more improved methods have been studied such as Croston method. This paper proposes a novel method to forecast the intermittent parts demand based on fuzzy support vector machines (FSVM) in regression. Details on data clustering, performance criteria design, kernel function selection are presented and an experimental result is given to show the method's validity.

1 Introduction

A fundamental aspect of supply chain management is accurate demand forecasting. We address the problem of forecasting intermittent (or irregular) demand. Intermittent demand appears at random, with many time periods having no demand [1]. Moreover, when a demand occurs, the request is sometimes for more than a single unit. Items with intermittent demand include service (spare) parts and high-priced capital goods, such as heavy machinery. Such items are often described as "slow moving". Demand that it intermittent is often also "lumpy", meaning that there is great variability among the nonzero value. An example of the difference between intermittent demand data and product demand data that is normal, or "smooth", is illustrated in the tables below:

Table 1. Intermittent Demand Data

Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Demand	0	0	**19**	0	0	0	**4**	**18**	**17**	0	0	0	0	0	**3**	0	0

Table 2. Normal, Smooth Demand Data

Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Demand	17	20	18	25	30	68	70	41	32	35	66	26	23	25	25	28	36

M. Ali and R. Dapoigny (Eds.): IEA/AIE 2006, LNAI 4031, pp. 1080–1089, 2006.

Intermittent demand creates significant problems in the manufacturing and supply environment as far as forecasting and inventory control are concerned. It is not only the variability of the demand size, but also the variability of the demand pattern that make intermittent demand so difficult to forecast [2]. The literature, that includes a relatively small number of proposed forecasting solutions to this demand uncertainty problem, can be found in [3-6]. The single exponential smoothing and the Croston methods are the most frequently used methods for forecasting low and intermittent demands [4,6]. Croston [4] proposed a method that builds demand estimates taking into account both demand size and the interval between demand incidences. Despite the theoretical superiority of such an estimation procedure, empirical evidence suggests modest gains in performance when compared with simpler forecasting techniques; some evidence even suggests losses in performance. On the other hand, Bartezzaghi et al. [3] in their experimental simulation found that EWMA appears applicable only with low levels of lumpiness. Willemain et al. [6] concluded that the Croston method is significantly superior to exponential smoothing under intermittent demand conditions.

Recently, support vector machines (SVMs) was developed by Vapnik and his co-workers [7,8]. With the introduction of Vapnik's -insensitive loss function, SVMs have been extended to solve non-linear regression estimation problems and they have been shown to exhibit excellent performance [7,8].

Since SVMs are formulated for two-class problems, some input points may not be exactly assigned to one of these two classes. Some are more important to be fully assigned to one class so that SVM can separate these points more correctly. Some data points corrupted by noises are less meaningful and the machine should better to discard them. SVM lacks this kind of ability. To solve this problem, Fuzzy support vector machines (FSVM) apply a fuzzy membership to each input point of SVM such that different input points can make different contributions to the learning of decision surface and can enhances the SVM in reducing the effect of outliers and noises in data points [9-11].

Our research focuses on the application of FSVM in regression to make a new attempt to novel forecasting method toward the intermittent demand. The results of experiment indicate that FSVM is effective in improving the accuracy of intermittent demand forecasting compared with the Croston method which has been a widely used method in intermittent demand forecasting.

This paper consists of five sections. Section 2 reviews the most widely used approach for forecasting intermittent demand and indicates its limitation and the direction of further improvement. General principles of FSVM and its application in regression are presented in Section 3, together with the general procedures of applying it. Section 4 presents an experiment concerned with the detailed procedures of how to employing FSVM in regression, involving data set selection, data preprocessing and clustering, kernel function selection and so on. Conclusions and discussion for further research hints are included in the last section.

2 Reviews on Forecasting Intermittent Demand Methods

Generally efforts on forecasting the intermittent demand could fall into two categories. One is to find the distribution function and the other is time series forecasting.

2.1 Demand Distribution Estimation

The inventory control method proposed here relies on the estimation of the distribution of demand for low demand parts. It is necessary to use demand distributions rather than expected values of demand because the intermittent patterns characteristic of low demand parts require a probabilistic approach to forecasting that can be incorporated into an inventory management program. Using the demand distributions, it is possible to manage the inventory to maximize readiness given a fixed inventory budget. Other optimization goals are possible as well.

The chief obstacle to reliable demand distribution estimation is the paucity of historical data available for any typical set of low demand parts. Demand typically occurs in small integer numbers of parts. Some parts may have only 3 or 4 non-zero demands among a large number of zero demand periods. This amount of data is not sufficient to construct a robust probability demand distribution. If a probabilistic model of the demand is available, such as a Weibull model or Poisson model, then it is possible to estimate the demand distribution directly from the model. If an empirical estimate of the demand distribution must derived, through bootstrapping or other means, it is necessary to group the data in a way that generates enough non-zero data points to produce robust demand distribution estimates.

2.2 Croston Method

Croston method falls into the time series forecasting category and is the most widely used method, which could be illustrated as follows.

Let Y_t be the demand occurring during the time period t and X_t be the indicator variable for non-zero demand periods; i.e., $X_t = 1$ when demand occurs at time period t and $X_t = 0$ when no demand occurs. Furthermore, let j_t be number of periods with nonzero demand during interval $[0, t]$ such that $j_t = \sum_{i=1}^{t} X_i$, i.e., j_t is the index of the the non-zero demand. For ease of notation, we will usually drop the subscript t on j . Then we let Y_j^* represent the size of the j th non-zero demand and Q_j the inter-arrival time between Y_{j-1}^* and Y_j^*. Using this notation, we can write $Y_j = X_t Y_j^*$.

Croston method separately forecasts the non-zero demand size and the inter-arrival time between successive demands using simple exponential smoothing (SES), with forecasts being updated only after demand occurrences. Let Z_j and P_j be the forecasts of the $(j+1)_{th}$ demand size and inter-arrival time respectively, based on data up to demand j. Then Croston method gives

$$Z_j = (1-\alpha)\, Z_{j-1} + \alpha Y_j^* \tag{1}$$
$$P_j = (1-\alpha)\, P_{j-1} + \alpha P_j \tag{2}$$

The smoothing parameter α takes values between 0 and 1 and is assumed to be the same for both Y_j^* and Q_j . Let $l = j_n$ denote the last period of demand. Then the mean demand rate, which used as the h-step ahead forecast for the demand at time $n+h$ is estimated by the ratio

$$\hat{Y}_{n+h} = Z_l \,/\, P_l \tag{3}$$

The assumptions of Croston method could be derived that (1) the distribution of non-zero demand sizes Y_j^* is iid normal;(2) the distribution of inter-arrival times Q_j is iid Geometric; and (3)demand sizes Y_j^* and inter-arrival times Q_j are mutually independent. These assumptions are clearly incorrect, as the assumption of iid data would result in using the simple mean as the forecast, rather than simple exponential smoothing, for both processes. This is the basic reason for more correction and modification toward Croston method.

3 FSVM for Forecasting

In this section we briefly review the description about the idea and formulations of FSVM in regression problem. In regression problem, the effects of the training points are different. It is often that some training points are more important than others, so we apply a fuzzy membership $0 \le s_i \le 1$ associated with each training point x_i. This fuzzy membership s_i can be regarded as the attitude of the corresponding training point toward the mapping function and the value $(1-s_i)$ can be regarded as the attitude of meaningless. In a result, the traditional SVM was extended as FSVM.

Given a set S of data points with associated fuzzy membership

$$(x_1, y_1, s_1) \quad (x_i, y_i, s_i) \tag{4}$$

where $x_i \in R^n$ is the input vector,$y_i \in R$ is the desired value, and a fuzzy membership $\sigma \le s_i \le 1$ with $i = 1, ..., n$ and sufficient small $\sigma > 0$. The FSVM regression solves and optimization problem:

$$\min_{\omega,p,\xi,\xi^*} \quad C\sum_{i=1}^{n} s_i(\xi_i + \xi_i^*) + \frac{1}{2}\omega^T\omega$$
$$Subject\ to \begin{cases} y_i - \omega^T\phi(x_i) - b_i \le \varepsilon + \xi_i, \xi_i \ge 0 \\ \omega^T\phi(x_i) + b_i - y_i \le \varepsilon + \xi_i^*, \xi_i^* \ge 0 \\ i = 1, ..., n \end{cases} \tag{5}$$

where ϕ is the high dimensional feature space which is non-linearly mapped from the input space x , ξ_i is the upper training error (ξ_i^* is lower), subject to the *varepsilon* -insensitive tube:

$$|y - (\omega^T\phi(x) + b)| \le \varepsilon \tag{6}$$

The parameters that control the regression performance are the cost of error C , the width of the tube ε , the mapping function $\phi(x)$ and the fuzzy membership s_i . Usually, it is more convenient to solve the dual of Eq.(5) by introducing Lagrange multipliers α_i^*, α_i , and leads to a solution of the form

$$\begin{aligned} &f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x, x_i) + b, \\ &0 \le \alpha_i^*, \alpha_i \le s_i C \end{aligned} \tag{7}$$

In Eq. (7), α_i and α_i^* satisfy the equalities $\alpha_i * \alpha_i^* = 0$,$\alpha_i \ge 0$ and $\alpha_i^* \ge 0$ where i = 1,2,,n and are obtained by maximizing the dual function of Eq. (7) which has the following form:

$$\begin{aligned} R(a_i, a_i^*) = &\sum_{i=1}^{n} y_i (a_i - a_i^*) - \varepsilon \sum_{i=1}^{n} (a_i + a_i^*) \\ &-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (a_i - a_i^*) \left(a_j - a_j^*\right) K(x_i, x_j) \end{aligned} \tag{8}$$

with the constraints

$$\begin{aligned} &\sum_{i=1}^{n} (a_i - a_i^*), \\ &0 \le a_i \le C, \quad i = 1, 2, \cdots, n \\ &0 \le a_i^* \le C, \quad i = 1, 2, \cdots, n \end{aligned} \tag{9}$$

Based on the Karush-Kuhn-Tucker (KKT) conditions of quadratic programming, only a certain number of coefficients $(a_i - a_i^*)$ in Eq.(5) will assume nonzero values. The data points associated with them have approximation errors equal to or larger than ε and are referred to as support vectors. Generally, the larger the ε, the fewer the number of support vectors and thus the sparser the representation of the solution.

$K(x_i, x_j)$ is defined as the kernel function. The value of the kernel is equal to the inner product of two vectors X_i and X_j in the feature space $\phi(x_i)$ and $\phi(x_j)$, that is, $K(x_i, x_j) = \phi(x_i) * \phi(x_j)$.Any function satisfying Mercer's condition can be used as the kernel function.

From the implementation point of view, training SVMs is equivalent to solving a linearly constrained quadratic programming (QP) with the number of variables twice as that of the training data points. Generally speaking, application of FSVM for forecasting follows the procedures: (1) Transform data to the format of an FSVM and conduct simple scaling on the data; (2) Generate the fuzzy membership; (3) Choose the kernel functions; (4) Use cross-validation to find the best parameter and ; (5) Use the best parameter and to train the whole training set; (6) Test.

4 Experimental Setting

4.1 Data Sets

Forecasting and inventory management for intermittent demand parts is particularly problematic because of the large number of low demand parts that must be considered. As an experiment setting, of 5,000 unique non-repairable spare parts for the Daya Bay Nuclear station in China, over half of those parts have been ordered 10 time or less in the last ten years. While many of these low demand parts are important for the safe operation of the nuclear reactor, it is simply uneconomical to stock enough spares to guarantee that every low demand part will be available when needed.

4.2 Clustering for Data Preprocessing

Clustering is the process of grouping parts with similar demand patterns. There are several methods to cluster data, among which, agglomerative hierarchical clustering and c-means clustering are two typical methods. We have found that demand patterns can be robustly clustered by using cumulative demand patterns. As the cumulative demand patterns avoids problems invoked by the intermittent pattern of incremental demand. Figure 1 shows one of prototype cumulative demand patterns after clustering 4063 low demand spare parts into 10 clusters. Prototype patterns represent the typical demand pattern for each cluster. The cluster size ranges from 34 parts to 241 parts plotted are the $25th$ and $75th$ percentiles of demand gathered from the cumulative demand of the individual parts in each cluster. Clustering was accomplished using a fuzzy c-means (FCM) clustering routine [12]. The generalized objective function subject to the same fuzzy c-partition constraints[13] is:

$$\text{Min } J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m \|x_k - v_i\|^2 \tag{10}$$

During our experiment, one of the problems associated with clustering is the difficulty in determining the number of clusters, c. Various validity measures have been proposed to determine the optimal number of clusters in order to address this inherent drawback of FCM [13]. In our experiment, the optimal number of terms is defined as the one that has the lowest mean squared error (MSE). The least MSE measure is also used to identify the most appropriate form of membership functions. In summary, the application procedure of the FCM has the following steps:(1) choose $c\,(2 \le c \le n)\,, m\,(1 < m < \infty)$ and initialize the membership matrix. (2) Read in the data set and find the maximum and minimum values. (3) Calculate cluster centers but force the two clusters with the largest and smallest values to take the maximum and minimum domain values. (4) Update the membership matrix (5) Compute the change of each value in the membership matrix and determine whether the maximum change is smaller than the threshold value chosen to stop the iterative process (set at 0.02 throughout this study). If not, return to Step 3. (6) Redistribute erroneous membership

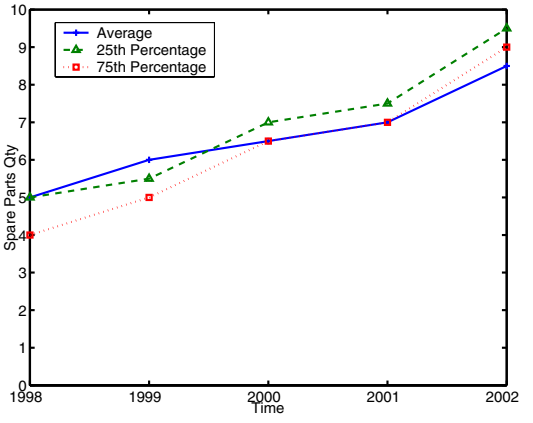

Fig. 1. Cumulative demand pattern of Cluster 1

values to the other two more appropriate terms proportional to their current membership values.

4.3 Defining Fuzzy Membership

It is easy to choose the appropriate fuzzy membership. First, we choose $\sigma > 0$ as the lower bound of fuzzy membership. Second, we make fuzzy membership s_i be a function of time t_i

$$s_i = f(t_i) \tag{11}$$

We suppose the last point x_n be the most important and choose $x_n = f(t_n) = 1$, and the first point x_1 be the most least important and choose $s_1 = f(t_1) = \sigma$. If we want to let fuzzy membership be a linear function of the time, we can select

$$s_i = f(t_i) = \alpha t_i + b = \frac{1-\sigma}{t_n - t_1} t_i + \frac{t_n \sigma - t_1}{t_n - t_1} \tag{12}$$

If we want to make fuzzy membership be a quadric function of the time, we can select

$$s_i = f(t_i) = \alpha(t_i - b)^2 + c = (1-\sigma)\left(\frac{t_i - t_1}{t_n - t_1}\right)^2 + \sigma \tag{13}$$

4.4 Kernel Function Parameters Selection

We use general RBF as the kernel function. The RBF kernel nonlinearly maps samples into a higher dimensional space, so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. Furthermore, the linear kernel is a special case of RBF as (Ref.[14]) shows that the linear kernel with a penalty parameter $\tilde{C}$ has the same performance as the RBF kernel with some parameters (C, γ). In addition, the sigmoid kernel behaves

like RBF for certain parameters[15].The second reason is the number of hyper-parameters which influences the complexity of model selection. The polynomial kernel has more hyper-parameters than the RBF kernel. Finally, the RBF kernel has less numerical difficulties. One key point is $0 < K_{ij} \leq 1$ in contrast to polynomial kernels of which kernel values may go to infinity $\left(\gamma {x_i}^T x_j + r > 1\right)$ or zero $\left(\gamma {x_i}^T x_j + r < 1\right)$ while the degree is large.

There are two parameters while using RBF kernels: C and γ. It is not known beforehand which C and γ are the best for one problem; consequently some kind of model selection (parameter search) must be done. The goal is to identify good (C, γ) so that the classifier can accurately predict unknown data (i.e., testing data). Note that it may not be useful to achieve high training accuracy (i.e., classifiers accurately predict training data whose class labels are indeed known). Therefore, a common way is to separate training data to two parts of which one is considered unknown in training the classifier. Then the prediction accuracy on this set can more precisely reflect the performance on classifying unknown data. An improved version of this procedure is cross-validation.

We use a grid-search on C and γ using cross-validation. Basically pairs of (C, γ) are tried and the one with the best cross-validation accuracy is picked. We found that trying exponentially growing sequences of C and γ is a practical method to identify good parameters (for example, $C = 2^{-5}, 2^{-3}, \cdots, 2^{15}; \gamma = 2^{-15}, 2^{-13}, \cdots, 2^{3}$).

4.5 Performance Criteria

The prediction performance is evaluated using the normalized mean squared error (NMSE). NMSE is the measures of the deviation between the actual and predicted values. The smaller the values of NMSE, the closer are the predicted time series values to the actual values. The NMSE of the test set is calculated as follows:

$$\text{NMSE} = \frac{1}{\delta^2 n} \sum_{i=1}^{n} \left(y_i - \hat{y_i}\right)^2, \tag{14}$$

$$\delta^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2, \tag{15}$$

where n represents the total number of data points in the test set. $\hat{y_i}$ represents the predicted value. $\overline{y}$ denotes the mean of the actual output values. Table 3 shows the NMSE values of different kernel functions compared with Croston method and tells out the best prediction method under our numerical case.

4.6 Experimental Results

Still raise the example of Cluster 1, Figure 2 shows the experimental esults by comparing the forecasting results of actual data, Croston method and FSVM regression. By summing all the clusters' result, SVMs regression method's accuracy is 11.6% higher than Croston method by the computation of standard deviation.

Table 3. NMSE Values of comparative methods

Methods	NMSE
FSVMs_RBF	0.3720
FSVMs_Linear	0.5945
FSVMs_Polynomial	0.6054
Croston	0.5730

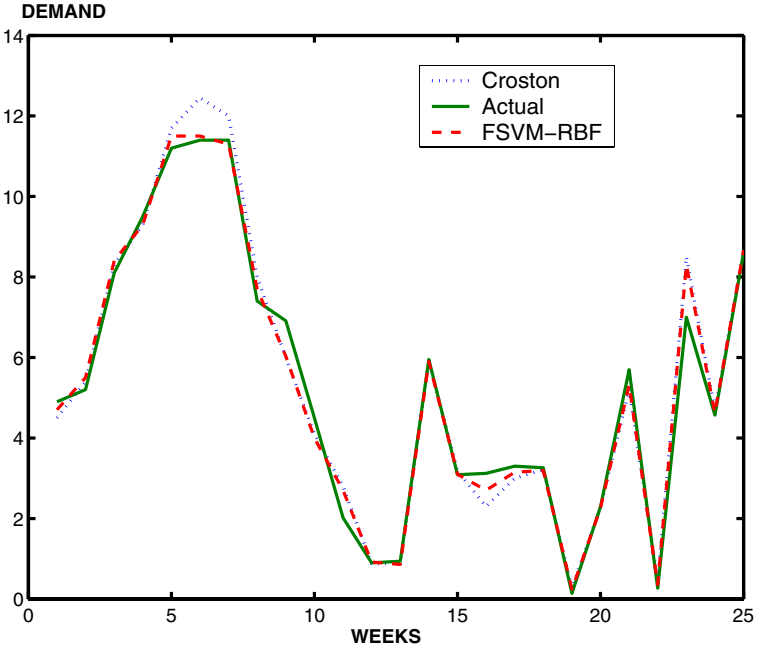

Fig. 2. Forecasting results comparison of Cluster 1

5 Conclusions

The use of FSVM in forecasting intermittent demand is studied in this paper. The study concluded that FSVM provide a promising alternative to forecast the intermittent demand. But further research toward an extremely changing data situation should be done, which means the data fluctuation may affect the performance of this method. In fact, we got confused with the experimental result at the every beginning without the data clustering. Another further research hint is the knowledge priority used in training the sample and determining the function parameters. This is to say, parameters selection is free but affect the performance a lot. A good parameter selection method should be worthy of further research.

Acknowledgements

This research is granted by National Science Foundation of China, No.70401015 and Hubei Provincial Key Social Science Research Center of Information Management.

References

1. Silver, E.A. Operations research in inventory management: A review and critique. Operations Research. **29** (1981) 628-645.
2. Syntetos A.A.,Boylan J.E. On the bias of intermittent demand estimates. International Journal of Production Economics.**71**(2001)457-466.
3. Bartezzaghi E, Verganti R and Zotteri G. A. Simulation framework for forecasting uncertain lumpy demand. International Journal of Production Economics. **59** (1999) 499-510.
4. Croston JD. Forecasting and stock control for intermittent demands. Operational Research Quarterly. **23(3)**(1972)289-303.
5. Rao AV. A comment on: forecasting and stock control for intermittent demands. Operational Research Quarterly. **24(4)**(1973)639-640.
6. Willemain TR, Smart CN, Shockor JH, DeSautels PA. Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method. International Journal of Forecasting.**10(4)**(1994)529-538.
7. Vapnik VN, GolowichSE, Smola AJ. Support vector method for function approximation, regression estimation, and signal processing. Advances in Neural Information Processing Systems.**9** (1996)281-7.
8. Vapnik VN. The nature of statistical learning theory. New York: Springer.(1995).
9. Abe, S., & Inoue, T. Fuzzy support vector machines for multi-class problems. Proceedings of the Tenth European symposium on Artificial Neural Networks (ESANN 2002), 113-118.
10. Chun-Fu Lin, Sheng-De Wang, Fuzzy Support Vector Machines, IEEE Trans on Neural Networks.**13(2)**(2002)464-471.
11. Dug H. H, Changha H.. Support vector fuzzy regression machines.Fuzzy Sets and Systems, 138 (2003) 271-281.
12. Medasani S, Kim J, Krishnapuram R. An overview of membership function generation techniques for pattern recognition. International Journal of Approximation Research. **19(2)**, (1998) 391-417.
13. N.R. Pal, J.C. Bezdek. On cluster validity for the fuzzy c-means model. IEEE Trans. Fuzzy Systems. **3(3)** (1995)370-379.
14. Haykin S. Neural networks: a comprehensive foundation. Englewood CliKs, NJ: Prentice Hall. (1999).
15. Zhang GQ, Michael YH. Neural network forecasting of the British Pound US Dollar exchange rate. Omega.**26(4)** (1998)495-506.