An Effective Method for Data Analysis Using Curve Fitting

Hai Lin[1a]

[1]College of Software, Shenyang Normal University

253 Huanghe North Street, Shenyang, 110034, China

[a]jlu_linhai@163.com

Keywords: curve fitting, data analysis

Abstract. In the last a few decades, there has been growing interest in systems which deal with large amounts of data. In practice, there is a need for analyzing the data so that useful information can be extracted. And people have implemented a lot of systems for this purpose. In this paper, we propose a new approach for data analysis. This new method is based on curve fitting. We conduct some experiments and show that this new approach is good for data analysis and can produce very desirable results.

Introduction

Recently, people deal with systems that process large amounts of data on a daily basis [1,2]. These data is handled via very large database systems. So there is a need to do some analysis on the data so that useful information can be extracted out from the data. For example, people might use large amounts of history data on temperature to predict temperature in near future. There have been a lot of models and methods proposed for this kind of problem [3,4]. Some metrics have been used for comparing across different methods.

In this paper, we argue that in order to solve the proposed problem, the key question is to use some function to fit the existing data and capture the trend of the data. So the method of curve fitting is used in this paper. We conduct some empirical study on sample data, which is from a supermarket. Our experimental results show that this new method is good enough for analyzing data.

This paper is structured as follows. In the next section, we formulate the data analysis problem under consideration and argue that the essence of the problem is to establish some function to fit the existing data and capture the trend of the data. And then we present our proposed method. We then report our experimental results. We give some concluding results in the last section.

Data Analysis Problem

In this section, we will formulate the data analysis problem that we consider. Given a series of random variables $X_0,X_1,\ldots X_{n-1}$, and the goal is to find out the value of the next unknown random variable X_n in such a way that $E(X_n-X^*)^2$ should be minimized, where X* is the computed value for X_n. The goal of this constraint is to make the result as accurate as possible [5,6].

According to the theory of probability, if the goal is to minimize $E(X_n-X^*)^2$, we should use EX_n as an approximation for X^*. However, if we are given a series of random variables, their relationship should be considered somehow. Then we should use conditional probability instead of probability, and use conditional mathematical expectation instead of mathematical expectation.

In this case, conditional mathematical expectation is more accurate than mathematical expectation, since conditional mathematical expectation takes into account the relationship among random variables.

The general problem of computing conditional mathematical expectation is hard since we may not even know enough information about the probability distribution of the individual random variables.

To solve the problem, we use the following equation as the conditional mathematical expectation.

$E(X_n|X_{n-1},X_{n-2},\ldots,X_{n-p})=f(X_{n-1},X_{n-2},\ \ldots,X_{n-p})$

Here is the intuitive interpretation of the above equation. X_n is closely related to "p" random variables $X_{n-1},X_{n-2},\ldots,X_{n-p}$ and the function "f" captures the relationship. The goal of data analysis is thus to compute "f" and "p" from large amounts of data.

Our Proposed Method

We use four types of curves as candidates, which are listed below.

$Y=ab^x$	(b>1)	exponential curve
$Y=k+ab^x$	(a<0,0<b<1)	revised exponential curve
Y=ka^(b^x)	(0<a<1,0<b<1)	Gompertz curve
$1/Y=k+ab^x$	(a>1,0<b<1)	Losistic curve

Here is how we distinguish between these different types of curves. If the differences of the logs of the sequence is approximately a constant, exponential curve is applicable. If the ratio of the differences of the sequence is approximately a constant, one should the revised exponential curve, and this constant is the "b" in the equation. If the ratio of the logs of the sequence is approximately a constant, Gompertz curve should be used. If the ratio of the inverses of the sequence is approximately a constant, one should use Losistic curve.

The control flow of our method is shown in Figure 1.

Experimental Results

We did some experiments to test our method. In this section, we report our experimental results using proposed model. In the first experiment, we analyzed the number of customers for some grocery store. The data from day 1 to day 5 is available. We use it to analyze the most possible value for day 6. The accurate value for Day 6 turns out to be 269.

Table 1. number of customers

Day	1	2	3	4	5	6
number	214	299	230	189	224	278(269)

In the second experiment, we analyzed the sales amount for that same grocery store. Again, the data from day 1 to day 5 is available. We use it to analyze the most possible value for Day 6. The accurate value for Day 6 is 46124.

Table 2. sales amount

Day	1	2	3	4	5	6
amount	41253	51830	46028	39802	42992	47126(46124)

Our experimental results show that our proposed method for data analysis is acceptable.

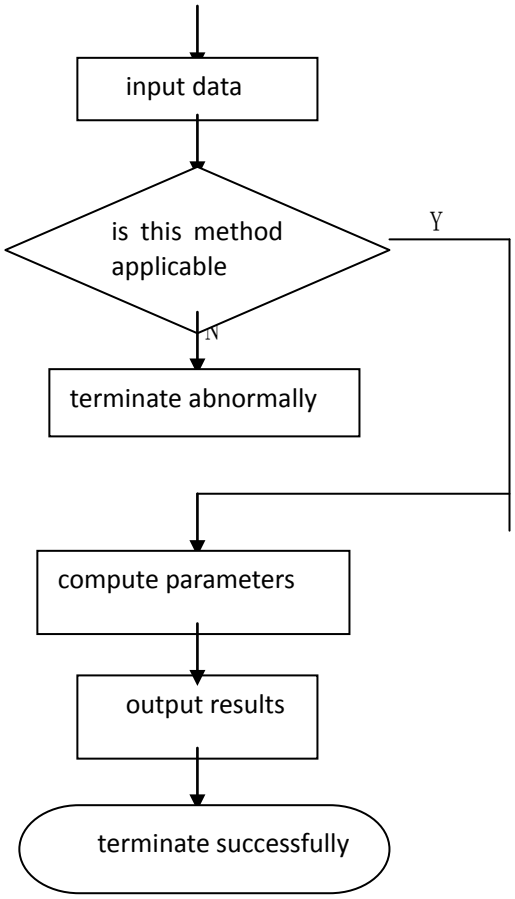

Fig. 1

Summary

The problem of data analysis is receiving more and more attention. People have been using a lot of different methods for solving this problem. In this paper, we propose a new method based on curve fitting. We use different types of curves to approximate the function that relates existing data and future data. We did some experiments to test our method. The experimental results show that our proposed method is quite good.

Acknowledgement

This work is supported by Liaoning Provincial Natural Science Foundation under grant 201202202, Scientific Research Foundation of Liaoning Provincial Education Department under grant L2012388.

References

[1] Kai-Ying Chen, Long-Sheng Chen, Mu-Chen Chen , Chia-Lung Lee.” Using SVM based method for equipment fault detection in a thermal power plant” Computers in Industry 62 (2011) 42–50.

[2] Banerjee A, Merugu S, Dhillon I, Ghosh J, Clustering with Bregman divergences. J Mach Learn Res (2005) 6:1705–1749.

[3] Chen JR, Making clustering in delay-vector space meaningful. Knowl Inf Syst (2007) 11(3):369–385.

[4] Uno T, Asai T, Uchida Y, Arimura H An efficient algorithm for enumerating frequent closed

patterns in transaction databases. In: Proc. of the 7th international conference on discovery science. LNAI vol 3245, Springer, Heidelberg, (2004) pp 16–30.

[5] M.S. Choudhury, S. Shah, N. Thornhill, D.S. Shook, Automatic detection and quantification of stiction in control valves, Control Engineering Practice 14 (12) (2006) 1395–1412.

[6] Bonchi F, Lucchese C On condensed representations of constrained frequent patterns. Knowl Inf Syst (2006) 9(2):180–201.