Outlier Detection in Multiple Linear Regression

Divya D.
Adi Shankara Institute of Engineering & Technology, India

Bhraguram T.M.
Adi Shankara Institute of Engineering & Technology, India

INTRODUCTION

Outlier detection as a branch of data mining has many important applications, and deserves more attention from data mining community. Outliers are normally treated as noise that needs to be removed from a dataset (Ben, 2005). Hawkins (1980) gives the definition for outliers as an outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism. *Outliers* can be caused by different situations. Removing and detecting outliers is very important in data mining, for example error in large databases can be extremely common, so an important property of a data mining algorithm is robustness with respect to outliers in the database.

Removal of outliers is needed for the successful execution of a particular algorithm. Many techniques employed for detecting outliers are fundamentally identical but with different names chosen by the authors. For example, authors describe their various approaches as outlier detection, novelty detection, anomaly detection, noise detection, deviation detection or exception mining (Victoria & Jim, 2004). In the case of clustering algorithms, there may be data points that do not belong to any of the clusters which considered as an outlier. In this case outliers need to be removed for the successful execution of the clustering algorithms. But in some other cases, outlier detection techniques may lead to the discovery of important information in the data. This is because of the fact that "one person's noise is another person's signal" (Varma & Rajesh, 2011). Outliers may be result of variability that is inherent in the data. The manager's salary in a company could naturally stand out as an outlier since it may be extremely higher other employees' salaries. But this outlier should not be removed since it is an important part of the company's payroll. Outlier detection strategies can also be used for data cleaning as a step which is used to clean any data before any traditional mining algorithm is applied to the data.

Many of the researches in outlier detection have focused on datasets that consists of one type of attribute, i.e. only numerical attributes or ordinal attributes that can be directly mapped into numerical values, or only categorical attributes. For example we may have data containing only categorical attributes; it is assumed that the categorical attributes could be easily mapped into numerical values. However, there are cases, where mapping categorical attributes to numerical attributes is difficult (Anna & Michael, 2010).

Today, business is expanding at a rapid pace with changing needs. Business plays a vital role in the capital formation of a country, and people consider it the life blood of a growing economy. Therefore, it is very important to manage business effectively and efficiently. One of the major issues encountered by fund managers today is not just the procurement of funds but also their meaningful deployment to generate maximum returns. Sources of funds are generally the same across all business but then why is it that some businesses are able to do better than the rest? If the logic behind the outstanding performance is a

DOI: 10.4018/978-1-4666-5202-6.ch158

viable business idea, why is it that some companies still fail to achieve success even with ample funds and the right business idea? (KirtiMadan, 2007)The above discussion clearly implies that there is something beyond financial success of business besides great ideas and good geographic presence; this implies the importance of *working capital management (WCM)* in determining the firm's success. Working capital is the proportion of company's total capital which is employed in short term operation.

Many existing research papers noticed that managers spend a considerable time on day to day working of capital decisions since current assets are short-lived investments that are continually being converted to other assets type (Rao, 1989). Excessive working capital leads to un remunerative use of scarce resources and inadequate working capital interrupts the smooth flow of business activity and profitability. The balance allocation of working capital funds between inventories, book debts and other components of working capital is a crucial phase in WCM. The survival and growth of the company depends on the ability to meet two vital aspects of WCM; profitability and liquidity. The company has to maintain an optimum level of liquidity to run the business on a continuous basis without any interruption. If the liquid assets are adequate to pay off the current liabilities, financial soundness is automatically created and its credit reputations are sustained. WCM is concerned with the problems that arise in attempting to manage the current assets, current liabilities and the inter relationship that arise between them.

Some research studies are available on the WCM practices of both large and small firms in India, UK, US and Belgium using either a survey based approach or by empirical analysis (Burns & Walker, 1991; Peel & Wilson 1996) to identify the push factors for firms to adopt good working capital practices or econometric analysis to investigate the association between WCM and Profitability (Shin & Soenen, 1998; Anand, 2001; Deloof, 2003).

There are a lot of business applications for outlier detection including but not limited to finding irregular credit card transaction and credit fraud, or identifying patients with abnormal symptoms by a particular type of disease. Here the paper discusses one application of outlier detection in modeling the profit of a company. Outlier detection has got a lot of applications in sequential data analysis. Mining for Outliers in Sequential Databases includes exception mining on multiple time Series in stock market and nonparametric outlier detection method for financial data. In financial data analysis outlier detection is applied on regression analysis. Regression analysis is trying to find the relationship between a dependent variable and a number of independent variables. In order to increase the accuracy of the regression model we can apply an outlier detection algorithm on the model.

The effect of WCM on profitability is determined by regression analysis is carried out using transport equipment manufacturing industry data. The model is developed by using the different factors affecting the profit of the firm. To improve the efficiency of the model, we are trying to remove the outliers in the data by calculating *residual* of the dependent variable. The residual is the deviation of value of dependent variable calculated from the model and actual value. By removing the outliers in the dataset the accuracy of the regression model is improved which increases the predictability of the model. Main advantage of the model is accurate prediction of profit by developing a regression model by using the dataset which contains no outliers. The disadvantage of the method is that since the dataset considered here is a *high dimensional dataset* complexity of the model is very high which can be reduced by high dimensional *data reduction*.

BACKGROUND

Data objects that show significantly different characteristics from remaining data are known as