Journal of Information Systems and Communication
ISSN: 0976-8742 & E-ISSN: 0976-8750, Volume 3, Issue 1, 2012, pp.-199-202.
Available online at http://www.bioinfo.in/contents.php?id=45

DATA MINING WITH REGRESSION TECHNIQUE

NANHAY SINGH[1], RAM SHRINGAR RAW[1] AND CHAUHAN R.K.[2]

[1]Department of Computer Science & Engineering, Ambedkar Institute of Technology, Delhi, India.
[2]Department of Computer Science & Applications, Kurukshetra University, India.
*Corresponding Author: Email-

Received: January 12, 2012; Accepted: February 15, 2012

Abstract- Extracting patterns and models of interest from large databases is attracting much attention in a variety of disciplines. Knowledge discovery in databases (KDD) and data mining are areas of common interest to researchers in machine learning, pattern recognition, statistics, artificial intelligence, and high performance computing. An effective and robust method, coined regression-class mixture decomposition (RCMD) method, is proposed in this paper for the mining of regression classes in large data sets, especially those contaminated by noise. A new concept, called regression class which is defined as a subset of the data set that is subject to a regression model, is proposed as a basic building block on which the mining process is based. A large data set is treated as a mixture population in which there are many such regression classes and others not accounted for by the regression models. Iterative and genetic-based algorithms for the optimization of the objective function in the RCMD method are also constructed. It is demonstrated that the RCMD method can resist a very large proportion of noisy data, identify each regression class, assign an inliers set of data points supporting each identified regression class, and determine the a priori unknown number of statistically valid models in the data set.
Keywords- Data Mining, Knowledge Discovery in Databases, Regression, Regression-Class Mixture Decomposition, Regression Methodology

Citation: Nanhay Singh, Ram Shringar Raw and Chauhan R.K. (2012) Data Mining with Regression Technique. Journal of Information Systems and Communication, ISSN: 0976-8742 & E-ISSN: 0976-8750, Volume 3, Issue 1, pp.-199-202.

Introduction

Data Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web [1]. Regression is a data mining function that predicts a number. Profit, sales, mortgage rates, house values, square footage, temperature, or distance could all be predicted using regression techniques. For example, a regression model [2] could be used to predict the value of a data warehouse based on web-marketing [3], number of data entries, size, and other factors.

A regression task begins with a data set in which the target values are known. For example, a regression model that predicts data warehouse values could be developed based on observed data for many data warehouses over a period of time. In addition to the value, the data might track the age of the data warehouse, size and number of clusters and so on. Data warehouse value would be the target, the other attributes would be the predictors, and the data for each data warehouse would constitute a case. In the model build (training) process, a regression algorithm estimates the value of the target as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown. Regression models are tested by computing various statistics that measure the difference between the predicted values and the expected values. The historical data for a regression project is typically divided into two data sets: one for building the model, the other for testing the model.

The rest of this paper is organized as follows. In section II, we describe the basics of regression technique. Section III presents the background on data mining. In section IV, useful research methodologies are described. Section V describes some useful regression model used for data mining. Finally in section VI we conclude this paper.

Need Of Regression Technique

There are various reasons for using regression technique in data mining. Some of these are listed below:

- A regression task begins with a data set in which the target values are known. For example, a regression model that predicts children's height could be developed based on observed data for many children over a period of time. The data might track age, height, weight, developmental milestones, family history, and so on. Height would be the target, the other attributes would be the predictors, and the data for each child would constitute a case.
- In the model build (training) process, a regression algorithm estimates the value of the target as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown.
- Regression models are tested by computing various statistics that measure the difference between the predicted values and the expected values.

Background

Traditionally, analysts have performed the task of extracting useful information from recorded data. But, the increasing volume of data in modern business and science calls for computer-based approaches. As data sets have grown in size and complexity, there has been an inevitable shift away from direct hands-on data analysis toward indirect, automatic data analysis using more complex and sophisticated tools. The modern technologies of computers, networks, and sensors have made data collection and organization an almost effortless task. However, the captured data needs to be converted into information and knowledge to become useful. Data mining is the entire process of applying computer-based methodology, including new techniques for knowledge discovery, from data.

Regression Methodology

Regression modeling has many applications in trend analysis, business planning, marketing, financial forecasting, time series prediction, biomedical and drug response modeling, and environmental modeling.

A. Functioning of Regression

Here is not need to understand the mathematics used in regression analysis to develop quality regression models for data mining. However, it is helpful to understand a few basic concepts. The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations that you provide. The following equation expresses these relationships in symbols. It shows that regression is the process of estimating the value of a continuous target (y) as a function (F) of one or more predictors (x_1, x_2 , ..., x_n), a set of parameters (θ_1 , θ_2 , ..., θ_n), and a measure of error (e).

$$y = F(x, \theta) + e \qquad (1)$$

The process of training a regression model involves finding the best parameter values for the function that minimize a measure of the error, for example, the sum of squared errors. There are different families of regression functions and different ways of measuring the error.

1. Linear Regression

The simplest form of regression to visualize is linear regression with a single predictor. A linear regression [4] technique can be used if the relationship between x and y can be approximated with a straight line, as shown in the Fig. 1 below.

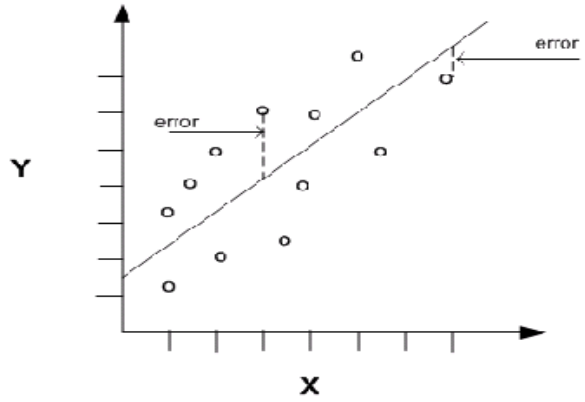

Fig. 1- Linear relationship between x and y

Linear regression with a single predictor can be expressed with the following equation.

$$y = \theta_2 x + \theta_1 + e \qquad (2)$$

The regression parameters in simple linear regression are:

The slope of the line (θ_2) — the angle between a data point and the regression line

The y intercept (θ_1) — the point where x crosses the y axis (x = 0)

2. Multivariate Linear Regression

The term multivariate linear regression refers to linear regression with two or more predictors (x_1, x_2, …, x_n). When multiple predictors are used, the regression line cannot be visualized in two-dimensional space. However, the line can be computed simply by expanding the equation for single-predictor linear regression to include the parameters for each of the predictors.

$$y = \theta_1+\theta_2 x_1+ \theta_3 x_2 +....+ \theta_n x_{n-1}+ e \qquad (3)$$

3. Regression Coefficients

In multivariate linear regression, the regression parameters are often referred to as coefficients. When you build a multivariate linear regression model, the algorithm computes a coefficient for each of the predictors used by the model. The coefficient is a measure of the impact of the predictor x on the target y. Numerous statistics are available for analyzing the regression coefficients to evaluate how well the regression line fits the data.

B. Nonlinear Regression

Often the relationship between x and y cannot be approximated with a straight line. In this case, a nonlinear regression [5] technique may be used. Alternatively, the data could be preprocessed to make the relationship linear. In Fig. 2, x and y have a nonlinear relationship. Oracle Data Mining supports nonlinear regression via the Gaussian kernel of SVM [6].

1. Multivariate Nonlinear Regression

The term multivariate nonlinear regression refers to nonlinear regression with two or more predictors (x_1, x_2, …, x_n). When multiple predictors are used, the nonlinear relationship cannot be visualized in two-dimensional space.

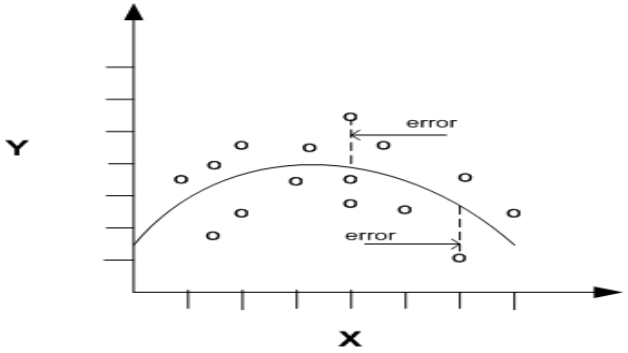

Fig. 2- Nonlinear Relationship between x and y

2. Confidence Bounds

A regression model predicts a numeric target value for each case in the scoring data. In addition to the predictions, some regression algorithms can identify confidence bounds, which are the upper and lower boundaries of an interval in which the predicted value is likely to lie.

When a model is built to make predictions with a given confidence, the confidence interval will be produced along with the predictions. For example, a model might predict the value of a house to be $500,000 with a 95% confidence that the value will be between $475,000 and $525,000.

After undergoing testing, the model can be applied to the data set that you wish to mine. Fig. 4 shows some of the predictions generated when the model is applied to the customer data set provided with the Oracle Data Mining sample programs. Several of the predictors are displayed along with the predicted age for each customer.

case ID | predictors | target

CUST_ID	CUST_GENDER	EDUCATION	OCCUPATION	AFFINITY_CARD	AGE
101501	F	Masters	Prof.	0	41
101502	M	Bach.	Sales	0	27
101503	F	HS-grad	Cleric.	0	20
101504	M	Bach.	Exec.	1	45
101505	M	Masters	Sales	1	34
101506	M	HS-grad	Other	0	38
101507	M	< Bach.	Sales	0	28
101508	M	HS-grad	Sales	0	19
101509	M	Bach.	Other	0	52
101510	M	Bach.	Sales	1	27

Fig. 3- Sample build data for regression

Regression Analysis

Suppose we want to learn more about the purchasing behavior of customers of different ages. We could build a model to predict the ages of customers as a function of various demographic characteristics and shopping patterns. Since the model will predict a number (age), we will use a regression algorithm.

A. Predictions Generation

This example uses the regression model. Fig. 3 shows six columns and ten rows from the case table used to build the model. The affinity card column can contain either a 1, indicating frequent use of a preferred-buyer card, or a 0, which indicates no use or infrequent use.

Oracle Data Miner displays the generalized case ID in the DMR$CASE_ID column of the apply output table. A "1" is appended to the column name of each predictor. Which is choose to include in the output. The predictions (the predicted ages in Fig. 4) are displayed in the PREDICTION column.

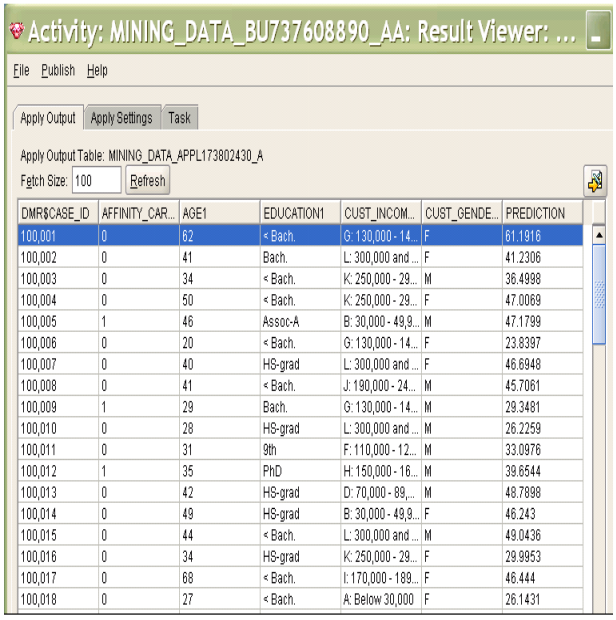

Fig. 4- Regression results in oracle data miner

B. Testing a Regression Model

A regression model is tested by applying it to test data with known target values and comparing the predicted values with the known values.

The test data must be compatible with the data used to build the model and must be prepared in the same way that the build data was prepared. Typically the build data and test data come from the same historical data set. A percentage of the records is used to build the model; the remaining records are used to test the model.

Test metrics are used to assess how accurately the model predicts these known values. If the model performs well and meets the business requirements, it can then be applied to new data to predict the future.

1. Residual Plot

A residual plot [7] is a scatter plot where the x-axis is the predicted value of x, and the y-axis is the residual for x. The residual is the difference between the actual value of x and the predicted value of x.

Fig. 5 shows a residual plot for the regression results shown in Fig. 4. Note that most of the data points are clustered around 0, indicating small residuals. However, the distance between the data points and 0 increases with the value of x, indicating that the model has greater error for people of higher ages.

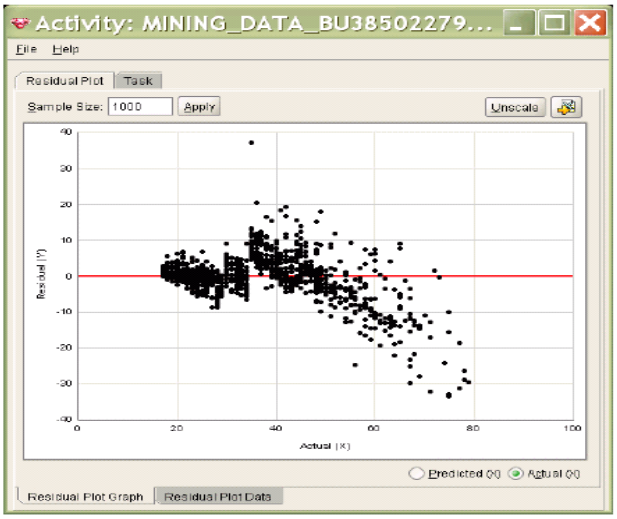

Fig. 5- Residual plots in oracle data miner

2. Regression Statistics

The Root Mean Squared Error and the Mean Absolute Error are commonly used statistics for evaluating the overall quality of a regression model. Different statistics may also be available depending on the regression methods used by the algorithm.

Root Mean Squared Error

The Root Mean Squared Error (RMSE) is the square root of the average squared distance of a data point from the fitted line. This SQL expression calculates the RMSE.

SQRT (AVG ((predicted_value - actual_value) * (predicted_value - actual_value)))

This formula shows the RMSE in mathematical symbols. The large sigma character represents summation; j represents the current predictor, and n represents the number of predictors.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \qquad (4)$$

Mean Absolute Error

The Mean Absolute Error (MAE) is the average of the absolute value of the residuals (error). The MAE is very similar to the RMSE but is less sensitive to large errors. This SQL expression calculates the MAE.

AVG (ABS (predicted_value - actual_value))

This formula shows the MAE in mathematical symbols. The large sigma character represents summation; j represents the current predictor, and n represents the number of predictors.

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \qquad (5)$$

Test Metrics in Oracle Data Miner

Oracle Data Miner [8] calculates the regression test metrics shown in Fig. 6.

Oracle Data Miner calculates the predictive confidence for regression models. Predictive confidence is a measure of the improvement gained by the model over chance. If the model were "naïve [9]" and performed no analysis, it would simply predict the average. Predictive confidence [10] is the percentage increase gained by the model over a naive model. Fig. 7 shows a predictive confidence of 43%, indicating that the model is 43% better than a naive model.

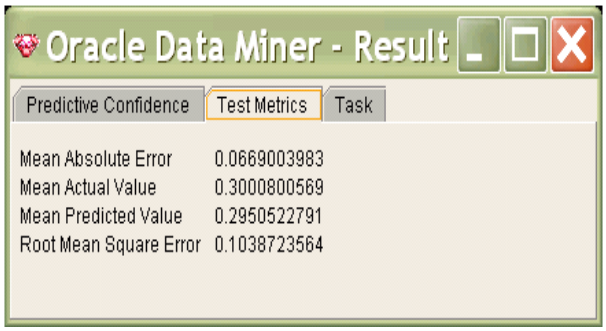

Fig. 6- Test metrics for a regression model

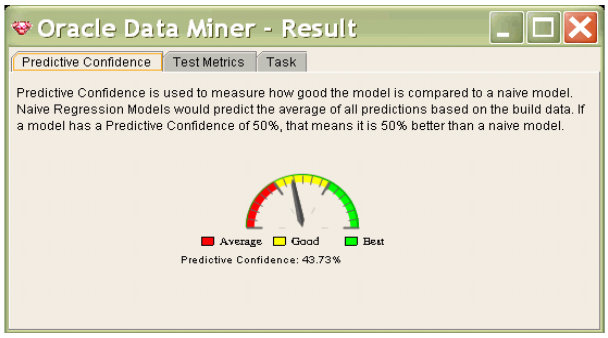

Fig. 7- Predictive confidences for a regression model

A large number of data mining procedures can be considered within a regression framework. A representative sample of the most popular and powerful procedure has been discussed in this paper. But the development of new data mining methods is progressing very quickly, stimulated in part by relatively inexpensive computing power and in part by the data mining needs in a variety of disciplines. Nevertheless, a key distinction between the more effective and the less effective data mining procedures is how over fitting is handled. Finding new and improved ways to fit data is often quite easy.

References

[1] Berners-Lee T.J., Cailliau R., Groff J.F., Pollermann B. (1992) *Electronic Networking: Research, Applications and Policy*, 2 (1).

[2] *http://dss.princeton.edu/training/Regressi on101.pdf.*

[3] Anirban Mahanti, Carey Williamson and Derek Eager, *Traffic Analysis of a Web Proxy Caching Hierarchy, University of Saskatchewan.*

[4] *http://www.bren.ucsb.edu/academics/courses/206/readings/readerch8.pdf.*

[5] Jennrich R.I. (1969) *The Annals of Mathematical Statistics* 40, 633-643.

[6] Blanz B. Scholkopf, ulthoff H.B, Burges C., Vapnik V. and Vetter T. (1996) *ICANN*, 1112, 251-256.

[7] Stromberg A.J. (1993) *J. Am. Stat. Assoc.* 88 (421), 237-244.

[8] *http://www2.tech.purdue.edu/cit/Courses/CIT499d/ODMr% 2011g% 20Tutorial%20for%20OTN.pdf.*

[9] Heckerman D. (1995) *A Tutorial on Learning Bayeisan Networks.*

[10] Bekaert Geert, Robert J. Hodrick and David Marshall (2001) *Journal of Monetary Economics*, 48, 41-270.