Tuning, Diagnostics & Data Preparation for Generalized Linear Models Supervised Algorithm in Data Mining Technologies

Sachin Bhaskar[1], Vijay Bahadur Singh[2] and **A. K. Nayak [3]**

Submitted in July 2013; Accepted in March, 2015

Abstract - Data mining techniques are the result of a long process of research and product development. Large amount of data are searched by the practice of Data Mining to find out the trends and patterns that go beyond simple analysis. For segmentation of data and also to evaluate the possibility of future events, complex mathematical algorithms are used here. Specific algorithm produces each Data Mining model. More than one algorithms are used to solve in best way by some Data Mining problems. Data Mining technologies can be used through Oracle. Generalized Linear Models (GLM) Algorithm is used in Regression and Classification Oracle Data Mining functions. For linear modelling, GLM is one the popular statistical techniques. For regression and binary classification, GLM is implemented by Oracle Data Mining. Row diagnostics as well as model statistics and extensive co-efficient statistics are provided by GLM. It also supports confidence bounds.. This paper outlines and produces analysis of GLM algorithm, which will guide to understand the tuning, diagnostics & data preparation process and the importance of Regression & Classification supervised Oracle Data Mining functions and it is utilized in marketing, time series prediction, financial forecasting, overall business planning, trend analysis, environmental modelling, biomedical and drug response modelling, etc.

Index Terms – GLM, Linear regression, Logistic regression, ODM,Tuning and Diagnostics for GLM

1.0 INTRODUCTION

GLM includes and extends the class of linear models[1]. The set of restrictive assumptions are made by linear models and also the most importantly the conditions are generally distributed on the value of predictors with a constant variance irrespective of predicted response values. An interpretable model form, able to compute specific diagnostic information about the quality and computational simplicity are included by the advantages of linear models and their restrictions. These restrictions are relaxed by Generalized Linear models which are not found is general practice.

[1]*Bihar Institute of Public Administration & Rural Development, Patna, Bihar, India*
[2]*L. N. Mishra Institute of Economic Development & Social Change, Baily Road, Patna, Bihar, India*
[3]*Zakir Hussain National Institute, Patna, Bihar, India.*
E-mail: [1]sachinbhaskar007@yahoomail.com,
[2]vbsinghpat@yahoo.co.in and [3]akn_iibm@yahoo.com

We find that the sum of terms in a linear model typically have large ranges encompassing very negative and positive values. For example in case of binary response, the response of probability is liked in the range [0,1]. We have two mechanism namely variance function and linear function where linear model assumptions are violated by responses which GLM accommodate. The variance function expresses the variance as a function of the predicted response thereby accommodating responses with non-constant variances like binary responses. The linear function transforms the target range to potentially -ve infinity to +ve infinity for maintaining the simple form of linear models. Two widely known members of the GLM family of models with their most popular link and variance functions included in Oracle Data Mining are as follows:

- Linear regression with the identity link and variance function equal to the constant 1 (constant variance over the range of response values). [1]
- Logistic regression with the log it link and binomial variance functions. [2]

GLM is a well established parametric modelling technique. Assumptions about the distribution of the data are made by parametric models. Parametric models become more efficient than the non-parametric models when assumptions are met. Assessing the extent to which the assumptions are met, is involved by the challenge in developing models of this type. That is why for developing quality parametric models, quality diagnostics are the key factors.

2.0 GLM INORACLE DATA MINING

2.1 Interpretability and Transparency

It is easy to interpret the Oracle Data Mining GLM models. Several diagnostics and statistics are generated by each model build. Transparency is also an important characteristic. Model details also describe key characteristics of global details providing high-level statistics and coefficients. **[3][4]**

2.2 Wide Data

To handle wide range of data, GLM is uniquely suited. The algorithm can build and score quality models that uses a virtually limitless number of predictors (attributes). The constraints imposed by the system resources are the only constraints.

2.3 Confidence Bounds

GLM is able to predict confidence bounds. GLM is able to predict confidence bounds. Apart from the predict, the best estimate and a probability, it identifies probability

(classification) and an interval wherein the prediction (regression) will lie. The width of the interval is dependable on the precision of the model and a user-specified confidence level The confidence level is a measure to know the true value that lie within a confidence interval computed by the model. 95% is the popular choice of confidence level. For Example - a model mighty predict that an employee's income is $130K and that we can be sure that around 95% sure that it lies between $85K and $150K. The value so obtained is configurable, although Oracle Data Mining supports 95% confidence by default
It is returned along with the coefficient statistics. PREDICTION_BOUNDS SQL function is to obtain the confidence bounds of a model prediction can also be used

2.4 Ridge Regression
The best regression models are the predictors which correlate highly with the target but there is very little correlation between the predictors themselves. Multi-collinearity is used to describe multivariate regression with correlated predictors. Multi-collinearity is compensated by the technique called Ridge regression. Ridge regression is supported by Oracle Data Mining for both classification and regression mining functions. If the singularity (exact multi-collinearity) in the data is found, ridge is automatically used by algorithm. Information about the singularity is returned in the global mode[6][7].

2.5 Build Settings for Ridge Regression
We can choose to explicitly enable ridge regression by specifying the GLMS_RIDGE_REGRESSION setting. If we exclusively enable ridge, we can use the system-generated ridge parameter or we can supply our own. Explicitly enable ridge regression by specifying the GLMS_RIDGE_REGRESSION setting can be choosen. If ridge explicitly enabled, the system-generated ridge parameter can be used or we can supply our own. The ridge parameter is also calculated automatically in case if the ridge is used automatically.
The build settings for ridge can be summarized as[5]:
• GLMS_RIDGE_REGRESSION — Whether or not to override the automatic choice made by the algorithm regarding ridge regression.
• GLMS_RIDGE_VALUE — The value of the ridge parameter, used only if you specifically enable ridge regression.
• GLMS_VIF_FOR_RIDGE — Whether or not to produce Variance Inflation Factor (VIF) statistics when ridge is being used for linear regression.

2.6 Ridge, Confidence Bounds, Variance Inflation Factor for Linear Regression
Models built with ridge regression do not support confidence bounds[8]. Variance Inflation Factor (VIF) statistics for linear regression models are produced by GLM, unless they were built with ridge. VIF with ridge by specifying the GLMS_VIF_FOR_RIDGE setting can be exclusively requested. VIF with ridge will be produced by the algorithm only in case if enough system resources are available.

2.7 Ridge and Data Preparation
Different data preparations are likely to be produced different results in terms of model coefficients and diagnostics in case the ridge regression is enabled. Automatic Data Preparation for GLM models, especially when ridge regression is used, be enabled; a Oracle Corporation recommends[9].

3.0 TUNING AND DIAGNOSTICS FOR GLM
A number of model builds involved in the process of developing a GLM model. To evaluate and determine the quality of model, each build generates many statistics. To change the model settings or making other modifications can be tried on the basis of these diagnostics.

3.1 Build Settings
Build settings is basically used for the purpose of specification of:

- Coefficient confidence: The default confidence which is used widely is 0.95. The degree of certainty that the true coefficient lies within the confidence bounds computed by the model, is indicated by GLMS_CONF_LEVEL setting.
- Row weights: This checks whether a column is containing a weighting factor for the rows or not and the situation is indicated by - ODMS_ROW_WEIGHT_COLUMN_NAME setting.
- Row diagnostics: This is used to identifya table to contain row-level diagnostics. This is indicated by GLMS_DIAGNOSTICS_TABLE_NAMEsetting.

There are additional build setting which are used for:

- Controlling the utilization of ridge regression[10].
- Handling procedure's specification for missing values which are not present in the training data[9].
- Specification of target values to be used as reference in logistic regression model[7].

3.2 Diagnostics
To evaluate the quality of the model GLM models generate many metrics to help us.

3.3 Coefficient Statistics
Both linear and logistic regression return the same set of statistics but statistics that do not apply to the mining function are returned as NULL [6][7]. The GET_MODEL_DETAILS_GLM function in DBMS_DATA_MININis used for the purpose of returning the coefficient statistics.

3.4 Global Model Statistics
For linear and logistic regression a whole new method of separation is adopted by returning separate high-level statistics which describes the model as a whole[6][7]. Only fewer global

details are returned when the ridge regression is enabled and this makes the adjacent procedures convenient[10].
The GET_MODEL_DETAILS_GLOBAL function in DBMS_DATA_MINING returns global statistics.

3.5 Row Diagnostics

By specifying the name of a diagnostics table in the build settingGLMS_DIAGNOSTICS_TABLE_NAME,
configuration of GLM models could be done to generate per-row statistics[6][7].
Row diagnostics is generated by a case ID which is required by GLM. An exception is raised in the process in casewe provide the name of a diagnostic table but the data does not include a case ID column.

4. 0 DATA PREPARATION FOR GLM

For both linear and logistic regression Automatic Data Preparation (ADP) implements suitable data transformation [11].

4.1 Data Preparation for Linear Regression

The build data are standardized by using a widely used correlation transformation, when ADP is enabled [12]. From the attribute values for each observation the data are first centred by subtracting the attribute means. In an observation by the square root of the sum of squares per attribute across all observations, the data are scaled by dividing each attribute value. For both numeric and categorical attributes, this transformation is used.
Before standardization, When N is the attribute cardinality, categorical attributes are exploded into N-1 columns. During the explosion transformation, the most frequent value (mode) is omitted. The first value in the list is omitted during the explosion and the attribute values are sorted alpha-numerically in ascending order in case of highest frequency ties. Where ADP is enabled or not explosion transformation lies.
The described transformations (explosion followed by standardization) can increase the build data size because the resulting data representation is dense, in case of high cardinality categorical attributes. An alternative approach needs to be used to reduce disk space, memory and processing requirements. Categorical attributes are not standardized for large datasets where the estimated internal dense representation would require approx more than 1Gb of disk space. The VIF statistic should be used with caution under aforesaid circumstances [10][11].

4.2 Data Preparation for Logistic Regression

Categorical attributes are exploded into N-1 columns where N is the attribute cardinality. The Explosion transformation eliminates the most frequent value (mode). The attribute values are sorted alpha-numerically in ascending order in the case of highest frequency ties, and the first value on the list is discarded during the explosion. Explosion transformation takes place irrespective of enabling the ADP and it doesn't depend upon its mode.
Numerical attributes are standardized when ADP is enabled mode. Measure of attribute variability plays a pivotal role in scaling the attribute values and this mechanism helps in the process of standardisation. Computation of the particular measure of variability is done with respect to the standard deviation per attribute with respect to the origin (not the mean)[13].

4.3 Missing Values

In case of applying or building a model, missing values of numerical attributes with the mean and missing values of categorical attributes with the mode are automatically replaced by Oracle Data Mining.
A GLM model can be configured to override the default treatment of missing values. The algorithm can be caused to delete rows in the training data that have missing values instead of replacing them with the mean or the mode, with the ODMS_MISSING_VALUE_TREATMENT setting. However, when we apply the model, Oracle Data Mining performs the usual mean/mode missing value replacement.
As a result, we see that statistics generated from scoring often not match the statistics generated from building the model.
The transformation must be performed explicitly if we want to delete rows with missing values in the scoring the model. The rows with NULLs from the scoring data must be removed before performing the apply operation for making build and apply statistics match.
This can be done by creating a view like

```
CREATE VIEW view_name  AS SELECT * from table_name
   WHERE column_name1 is NOT NULL
   AND   column_name2 is NOT NULL
   AND   column_name3 is NOT NULL.....
```

5.0 CONCLUSION AND FUTURE SCOPE

This paper is about Generalised Linear Models (GLM) Supervised algorithm in data mining technologies specifically in terms of tuning, diagnostics and data preparation process. In today's world GLM is apopular statistical technique, especially for linear modelling since it possesses highly remarkable features with respect to present requirements. Oracle Data Mining implements GLM for binary Classificationand for Regression function. This paper emphasises on explanation of the previous work, which hasbeen reviewed for the understanding of research in the area of Data Mining technologies.

REFERENCES

[1]. Linear Regression for GLM, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/regress.htm#CIHJIFEG
[2]. Logistic Regression for GLM, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#CIACAIFC
[3]. Tuning and Diagnostics for GLM, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#BABBGAHD

[4]. Transparency for GLM, http://download.oracle.com/docs/cd/B28359_01/dAtamine.111/b28129/xform_data.htm#CIAICHGH
[5]. Oracle Database SQL Language Reference, http://download.oracle.com/docs/cd/B28359_01/sErver.111/b28286/functions121.htm#SQLRF20020
[6]. Global Model Statistics for Linear Regression,http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#BABBIADB
[7]. Global Model Statistics for Logistic Regression, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#CHDEBJEB
[8]. Confidence Bounds for GLM, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#BABDBIII
[9]. Data Preparation for GLM, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#CACCHJDC
[10]. Ridge Regression for GLM, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#BABHBBBA
[11]. Automatic and Embedded Data Preparation,http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/xform_data.htm#BABGADFF
[12]. Neter, J., Wasserman, W., and Kutner, M.H., "Applied Statistical Models", Richard D. Irwin, Inc., Burr Ridge, IL, 1990.
[13]. Marquardt, D.W., "A Critique of Some Ridge Regression Methods: Comment", Journal of the American Statistical Association, Vol. 75, No. 369 , 1980, pp. 87-91
[14]. Alex Berson& Stephen J. Smith, "Data Warehousing Data Mining and OLAP", Tata McGraw-Hill Publishing Company Limited, New Delhi.
[15]. Bhaskar Sachin, Dissertation "Managin Data Mining Technologies in Organizations: Techniques and Applications", submitted to Periyar University, Salem, for M.Phil in Computer Science, November, 2007.