Model Performance Improvement with Least Square Method in Highly Imbalanced and Correlated Datasets

Ladan Malazizi
MODEL PERFORMANCE IMPROVEMENT WITH LEAST SQUARE METHOD IN HIGHLY IMBALANCED AND CORRELATED DATASETS

Ladan Malaziz
Computer Engineering Department, Islamic Azad University, Najafabad Branch
Esfahan, Iran
lmalazizi@pco.iaun.ac.ir

Abstract:
The need for accurate, complete and quality data is still a problem in every domain for model development. However, with the volume of stored information growing every day and the necessity of its integration for data analysis and data mining, the need for domain specific applicable methods to overcome the problem is obvious. With our experimental work in this paper the idea is to use Least Square method (LS) to generate artificial data and to show the affect on model improvement. With highly correlated numerical features in imbalanced datasets in any domain the method is applicable.

Keywords: Least Square Method, Highly Correlated, Model Improvement, Artificial Data

1. INTRODUCTION
Mining tools and algorithms have been improved rapidly in the last few years in order to increase learner performance and models accuracy. That is why the knowledge of Computational Statistics, the bridge between computing and statistic has been expanding more and more every day.

With large amount of data collected from distributed sources, the need for use of better data analytical tools and methods is increasing. In this paper the affect of LS method in generating artificial data with domain specific characteristics and improvement of model accuracy have been studied. Number of datasets with highly correlated numerical features is chosen for this experimental work. These datasets are also highly imbalanced in the mean of having different proportion of instances of each class. In this work we use classification algorithms [8] in Weka [4] to perform the modeling and show the affect of each algorithm in the original and new constructed datasets with LS method.

After first introduction about our approach; in section two we introduce number of used methods for data generation. After that we explain about LS method for data generation in our work, section four shows the method implementation followed by the results evaluation in section five finally the conclusions and future work is presented.

2. Methods for Data Generation
There are few methods used in different domains for artificial data generation in which some explained here; they are as follows:

A. Calculation of the Average (Mean): In cases when there is just one attribute value is missing; the average value of all the other attributes in that row is calculated and considered for the missing value. The danger of this is the decrease of the data characteristics variability and balancing out the values. Also when there are outliers exist in our datasets in the case when the related missing values falls in that category then this will affect the mean value calculated parameter. All of these will affect the final result of our mining task especially when the procedure is based on classification strategies [2].

B. Single Imputation: For cases when big proportion of the data is missing this method could be applied. With this strategy missing data in each cell is estimated based on available data in another relevant cell, which satisfies certain matching criteria. For example missing income value estimated by the comparison with existed record from another survey for a person living in the same area and has same educational background and same age. The disadvantage of this method is artificial data which replaces the missing values is exactly same as existing data which as a result reduces data values variation characteristics and increases a specific class or cluster of information that in turn affect the quality of data mining analysis [7].

C. Multiple Imputations: is another method, which instead of filling empty cells with specific values; replaces them with a set of applicable values. The multiply imputed datasets are then analyzed and then the results are combined [9].

D. Expectation Maximization (EM) Algorithm: this method first uses imputation for missing values then re-estimates the missing data values using these estimations and then iterates until convergence. Compare to other methods the method of iteration has been widely applied to missing data problems.

3. Data Generation Using LS
The LS method is a traditional and effective statistical method which has been used for analysis of data in different domains. It has also been used for imputation
of missing data when the existence of complete data is much needed.

In this work a new approach has been taken in order to use the benefits of the two fields, computing and statistics. First the characteristics of the datasets have been studied to prove the correlations between features and then the proportion of missing data have been analyzed. Then with the use of LS method the new data has been generated and added to the original datasets. The machine learning tool Weka has been used in order to model the data and record the prediction accuracy in two situations; first on original datasets with missing data rows removed and then on constructed dataset with generated artificial data included. The results are shown in the following sections.

In our study we highlight the applicability of our approach to domain specific situation when datasets have three important characteristics: first consist of highly correlated features, second being severely imbalanced with missing values for certain classes and last having numeric fields.

One of the most important elements that determine the accuracy level is the proportion of the class members which has direct relationship with the learner performance. Imbalanced or skewed [6] dataset, affect the accuracy of classification algorithms. Even though in these datasets when produced results of the overall accuracy are satisfactory the detailed investigation show that the under presented class samples have been badly classified. In such cases the over presented class dominates the role extended of knowledge presentation for the classifier. Often real world scientific applications encounter this problem [3].

For instance, in some domains like toxicology this problem is severe. When the chemical compounds need to be tested on different species, high toxicity chemicals cannot be sampled as many as low toxicity compounds which leads to creation of rows with missing data. In these datasets the important task of classification has to focus on high toxic chemical compounds since misclassification of high toxic chemicals may lead to disastrous consequences. [5]

To solve the problems mentioned above we propose the use of LS method in generating artificial data in order to achieve two goals; first to balance the data proportion in different classes and then to improve prediction accuracy.

In this experiment we have number of rows with missing values in each file. Considering our datasets with their special characteristics, if there exists strong relationships globally and locally between attributes, with the use of Least Square Method for regression line [1] we can calculate the missing values based on the following formula (considering the straight-line model):

\[ y = \beta_0 + \beta_1x + \epsilon \]  

The least square method involves the determination of \( \beta_0, \beta_1 \) to minimize \( Q \) and they are treated as the variables in the optimization and the predictor variable values, \( x_1, x_2...x_n \) are treated as coefficients.

For this model the least squares estimations of the parameters are computed by:

\[ Q = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \]  

And

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]  

In the case of row having number of missing cells values we need to consider two issues:

1) Measuring distance: (with the use of Euclidean distance squared) the distance between a missing value considered as \( X_i \) and the nearest neighbor \( X_j \) where \( m_{ik} \) and \( m_{jk} \) are missing values for \( x_{ik} \) and \( x_{jk} \) respectively. [9]

2) Neighborhood selection: this can be done with considering the properties of nearest attributes neighboring the target or missing value which corresponds to this entity. We evaluate all the instances in the datasets as the possible candidates. The procedure can be summarized as follows:

Start from first row that contains a missing value named \( X_i \), then find the nearest neighbors and form an \( X_m \) matrix \( (X_m = X_i + k_j) \). Then with the existence of high correlation between neighboring features based on the best fitting straight line found by LS method and with the Regression line analysis we predict the missing entries.

**LS Missing Data Generation Algorithm**

1. Read file (XY: X rows and Y columns)
2. Calculate correlation using (9)
3. Locate missing value \( X_i \)
4. Locate k nearest neighbor form matrix: \( X_m = X_i + k_j \)
5. Using the LS function calculated in step 2, calculate \( Y_i \)
6. Repeat step 3, 4, 5, until row is recovered
7. Repeat until 3 is empty

### 4. Method Implementation

Extensive experimental work has been carried out in order to examine the method for number of toxicity datasets to recover missing values. With the consideration of the datasets possessing the special characteristics requirements, this method has been tested and the results are shown.

**Table1: The Proportion of Missing Values in Each Dataset**

<table>
<thead>
<tr>
<th>After and Before Cleaning</th>
<th>T-t</th>
<th>D-t</th>
<th>B-E</th>
<th>O-Q</th>
<th>D-Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Original</td>
<td>282</td>
<td>264</td>
<td>105</td>
<td>116</td>
<td>123</td>
</tr>
</tbody>
</table>
In this study we used five toxicity datasets. For each dataset, values for six compound features have been considered. For this work the number of chemical instances present in each dataset varies from 105 to 252. Six selected features are those which determine the toxicity of the compounds. As it has been explained earlier these datasets features are highly correlated. In these datasets there are number of rows with missing values. First we cleaned the data (we have deleted rows with missing values) and then we trained number of learner classification algorithms in Weka data mining tool, and recorded learner performance.

In the second step we used the original datasets with missing values and recovered the empty cells based on the correlation of regression line using LS method to reconstruct the data. Finally we tested the same algorithms on these datasets and compared the results.

The results of this experimental work have been shown in table2. Especially the last column in the table should be considered that shows the overall accuracy by percentage. This in the case of O-Q, B-e and T-t datasets shows better results. In general the method has been effective for all datasets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>MLP</th>
<th>IBK</th>
<th>CVR</th>
<th>J48</th>
<th>JRip</th>
<th>LMT</th>
<th>Increase Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>O-Q Original</td>
<td>52.88</td>
<td>34.62</td>
<td>50</td>
<td>53.85</td>
<td>55.77</td>
<td>54.81</td>
<td></td>
</tr>
<tr>
<td>O-Q + Artificial Data</td>
<td>51.72</td>
<td>47.41</td>
<td>56.03</td>
<td>56.9</td>
<td>57.76</td>
<td>56.9</td>
<td>3.6</td>
</tr>
<tr>
<td>D-a</td>
<td>41.39</td>
<td>37.7</td>
<td>37.7</td>
<td>51.23</td>
<td>45.9</td>
<td>47.95</td>
<td></td>
</tr>
<tr>
<td>D-a+ Artificial Data</td>
<td>48.3</td>
<td>39.62</td>
<td>45.66</td>
<td>46.79</td>
<td>46.42</td>
<td>43.77</td>
<td>1.45</td>
</tr>
<tr>
<td>B-e</td>
<td>27.36</td>
<td>38.95</td>
<td>33.48</td>
<td>33.68</td>
<td>37.89</td>
<td>36.84</td>
<td></td>
</tr>
<tr>
<td>B-e+ Artificial Data</td>
<td>34.62</td>
<td>50</td>
<td>35.58</td>
<td>34.62</td>
<td>40.38</td>
<td>38.46</td>
<td>3.39</td>
</tr>
<tr>
<td>D-Q Original</td>
<td>25.84</td>
<td>31.46</td>
<td>34.83</td>
<td>32.58</td>
<td>29.21</td>
<td>32.58</td>
<td></td>
</tr>
<tr>
<td>D-Q + Artificial Data</td>
<td>33.96</td>
<td>32.08</td>
<td>41.51</td>
<td>35.85</td>
<td>35.85</td>
<td>35.85</td>
<td>2.55</td>
</tr>
<tr>
<td>T-t</td>
<td>51.15</td>
<td>52.29</td>
<td>57.25</td>
<td>58.40</td>
<td>54.58</td>
<td>55.73</td>
<td></td>
</tr>
<tr>
<td>T-t + Artificial Data</td>
<td>57.3</td>
<td>62.4</td>
<td>66.7</td>
<td>65.03</td>
<td>57.54</td>
<td>59.6</td>
<td>4.13</td>
</tr>
</tbody>
</table>

For the experiments carried out for our work as it has been explained earlier the data has been cleaned first (table1 shows the proportion of removed data for each dataset). First row of the table shows the number of chemical instances in each dataset. Second row presents number of instances after the data-cleaning task. The third row shows the proportion of lost data after cleaning. As it shown in the table1 for example for T-t dataset the missing data is 7.09% of the whole dataset and in the case of D-Q this is 13%.

5. The Results Evaluation

The results of the experiment are shown in table2. The first row of the table shows the classification algorithms used in Weka which includes; MLP, IBK, CVR, J48, JRip and LMT. The classification task with these algorithms has been performed twice: first on the datasets with rows with missing values removed and second time with added artificial data generated using LS method. In the table first row for each dataset shows the result of classification accuracy on the cleaned datasets and second row shows the accuracy for dataset with added data generated with our method. The last column shows the average increase by percentage. This in the case of O-Q, B-e and T-t datasets shows better results. In general the method has been effective for all datasets.

6. Conclusions and Future Work

This paper presented a new approach with regarding to the use of a classic statistical model for generating artificial data values in order to provide more knowledge for learner algorithm. The method used in this approach has been based on the relationships of compounds features characteristics. It has been used for case studies to fill in missing values and also to show the effect of the data recovery on the prediction accuracy of the models. An algorithm based on the method has been delivered. The method can be effectively used in the cases when the data need to be complete and when better performance is needed for further data analysis.

Based on our previous studies and results, we are trying to find new ways and methods in order to provide quality data in the mean of completeness and accuracy to increase model performance. The use of Artificial Neural Networks would be next step to be considered in the effect of production of such methods.

References:


