A Study of Data Pre-Processing Techniques for Machine Learning Algorithm to Predict Software Effort Estimation

Shaik Shahul¹, Sristi Suneel², M.A. Rahaman³ & Swathi J. N⁴
School of Computing Science and Engineering
VIT University, Vellore

Abstract: Software cost estimation requires high precision, however exact estimations are hard to accomplish. Progressively, data mining is utilized to enhance an association's product process quality, e.g. the exactness of cost estimations. There are countless strategies exists for software cost estimation, selecting the most suitable and pre-processing of data that is used for machine learning based software cost estimation is the subject of this paper. The effort invested in a software project is probably one of the most important and most analysed variables in recent years in the process of project management. Software effort estimates is an important part of software development work and provides essential input to project feasibility analyses, bidding, budgeting and planning. Analogy-based estimates models emerge as a promising approach, with comparable accuracy to arithmetic methods, and it is potentially easier to understand and apply. Studies show all the models are sensitive to the quality and availability data, thus requiring a systematic data treatment. In this paper we investigate various methods of data pre-processing for machine learning techniques based software cost estimation.

Keywords: Software effort estimation, Data pre-processing, binning method, Kilo Line of Code (KLOC)

1. INTRODUCTION

One of the most important purposes of pre-processing spectral data is to linearize them in such a way that linear statistical regression methods give good predictions when the corrected data are used. Another purpose is to try to eliminate the effect of physical phenomena like the light scattering effect of particles of different sizes and shapes. To do this, Martens et al. [1] and Geladi et al. [2] proposed their method of multiplicative scatter correction (MSC). The MSC method is mainly motivated by the empirical fact that for a collection of related samples, the plot of spectral values for each given sample against the corresponding mean spectral values (where the mean is taken over all samples) in many cases is approximately a straight line, such that different samples give different slopes of this line. We will show below that several different scatter correction methods can be related to the same fact, among others the standard normal variate transformation (SNV) proposed by Barnes et al.

Aiming to maximise the predictive accuracy of data mining, research in management science and machine learning is largely devoted to enhancing competing classifiers and the effective tuning of algorithm parameters. Classification algorithms are routinely tested in extensive benchmark experiments, evaluating the impact on predictive accuracy and computational efficiency, using pre-processed datasets. In contrast to this, research in DPP focuses on the development of algorithms for particular DPP tasks. While feature selection, resembling [6] and the discretisation of continuous attributes are analysed in some detail, few publications investigate the impact of data projection for categorical attributes and scaling [19]. More importantly, interactions on predictive accuracy in data mining are not been analysed in detail, especially not within the domain of corporate direct marketing.

2. RELATED WORK

Software administration experience of 50 years of advancement has turned into a far reaching learning of the hypothetical framework. With the advancement of data innovation, business rivalry escalates, the business sector vulnerability and danger expanding the item's life cycle is getting shorter, as a free stable administration framework, software administration in numerous mechanical fields of utilization a noteworthy effect [3]. Following quite a while of advancement, undertaking administration has turned into a more develop field of programming software administration; the result is still a long way from tasteful. Programming software estimation is an
undertaking arrangement in nature[6]. Albeit numerous product organizations are effectively programming software administration into improvement exercises, yet the dominant part of programming advancement associations have not understood the significance of programming estimation. Incorrect cost gauges calamitous understood the significance of programming programming advancement associations have not improvement exercises, yet the dominant part of programming software administration into numerous product organizations are effectively undertaking.

1. 35% appraisal programming advancement cost and time.
2. Four-fifths of the software as far as expense or time and inside of spending plan.
3. Half of the over spending plan can be stayed away from.
4. Just 50% of the task documentation as records under the applicable information in the improvement process.

Before the evaluated expense of the undertaking is just 20 not exactly a rate of the estimation strategy, 60% in totally subjective experience. Software’s that neglect to appraise the mistake is astounding, a large portion of them have more than one hundred percent of these undertakings in the end disappointment brought on issues because of their own evaluation was crossed out. S.D.Conte, E.Dunsmore, V.Y.Shen three educator in 1986, the most widely recognized criteria right now used to assess the appraisal precision: a great estimation strategy ought to have the capacity to ensure 75% of the time the mistake between the evaluations and the genuine worth rate in a quarter of a scope of [9].

Created nations for programming estimation research more than 30 years of history, China’s poor start in such manner, an assortment of estimation strategies is still at an early stage. For the most part as takes after [10]:

1. Not all supervisors perceive the significance of assessments, some have mindfulness however would prefer not to get to the base. With the entrance of programming designing in the business, more groups understood that the product improvement procedure is vital for cost gauges, yet there are likewise numerous cases, senior supervisors unpleasant appraisal can just think about the task.
2. Are usual to depending exclusively on experience to assess, by the subjective impact of the estimator, simple to make the after effects of an absence of objectivity. Despite the fact that we as of now have a percentage of the particular apparatus used to gauge the relative development of estimation techniques, however is not generally accessible in the genuine undertaking.

The usage of ML count requires the region of data in a logically achievable course of action through data pre-preparing. DP strategies contain data decreasing, data projection and missing-data treatment. Data reducing intends to decrease the extent of the datasets by technique for highlight determination (FS) or case decision (CS). Data projection arrangements are required to change the nearness of the data, e.g. scaling, which scales all segments into a predefined same reach. Missing-data meds (MDTs) consolidate deleting missing qualities and/or supplanting them with the evaluations [35]. Likewise, the logarithm change is a great part of the time associated for direct backslides to hold the normality suspicion for the right execution of direct backslides. It is to make certain a crucial step for backslide models to ensure the normality of the remaining [38]. On the other hand, logarithm does not from time to time appear in ML thinks about. In our study, there are only three creations [32] that clearly used logarithm for ML procedures. These study arrangements to look into the ampleness of the standard DP frameworks for ML systems. Consider both scaling and logarithmic change could decrease ranges, we join scaling as a confident DP procedure in our tests.

3. DATA PRE-PROCESSING TECHNIQUES

A. Handling Missing values

In light of the high cost of social event and reporting data from endeavours, change gatherings are less focused on data collection [28]. The divided datasets moreover once in a while appear over the SCE considers (e.g. the ISBSG database and PROMISE datasets) [13]. The missing qualities effects affect ML estimation shows, as reported. There are various MDTs in the written work. They oftentimes include: cancelation systems (list-wise eradication and pair wise deletion), and credit strategies (mean attribution, hot-deck attribution, frigid deck attribution, backslide attribution, etc.) [20]. It is seen that the deletion techniques, especially list-wise cancelation (LD), by and large used as a default approach for overseeing missing qualities, can achieve discarding immense degrees of datasets in cases and introducing biasness [28].

As another game plan of MDT, attribution requires more extensive additionally, jumbled quantifiable and computational examination besides fuses ordinary desire bumble. Mean attribution (MI) credits each missing quality with
the mean of watched qualities additionally, jam the information of data. Regardless, as the slightest complex attribution technique it may achieve to lessen the change of variables [26].

B. Discretization

Discretization helps in Pre-processing of numeric values. It is one of the methods for pre-processing. Not every learning method can handle numeric values. Sometimes the learning methods may not produce the exact output when dealing with numeric values. For this purpose Discretization can help in handling such situations. Discretization convert numbers to a finite number of bins. These bins will have certain range of values for each bin. Simple schemes (like nbins), are surprisingly effective E.g. really helps for NaiveBayes.

C. Outlier data

An outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution. In the former case one wishes to discard them or use statistics that are robust to outliers, while in the latter case they indicate that the distribution has high skewness and that one should be very cautious in using tools or intuitions that assume a normal distribution. A frequent cause of outliers is a mixture of two distributions, which may be two distinct sub-populations, or may indicate 'correct trial' versus 'measurement error'; this is modeled by a mixture model.

4. RESULTS & DISCUSSIONS

We have pre-processed the data set by Kitchenham using Rapid miner. Kitchenham dataset is software cost estimation data set. It has been processed for a variety of techniques such as missing data, noisy data and scaling. The results for missing data in Kitchenham data set are presented as below.

Table 1: Missing values replace using averaging method

<table>
<thead>
<tr>
<th></th>
<th>Replaced Missing value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Duration</td>
<td>204 (Average)</td>
</tr>
<tr>
<td>Actual Effort</td>
<td>3131 (Average)</td>
</tr>
<tr>
<td>Adjusted Function points</td>
<td>528.11 (Average)</td>
</tr>
<tr>
<td>First Estimate</td>
<td>2829 (Average)</td>
</tr>
</tbody>
</table>

Outliers is added to the data set, after adding outliers the classification of different attributes is not possible. Overlapping of different attributes is present.

Figure 1: Presence of outliers

After removing the outlier, the attributes can be classified correctly. It is shown in the below figure.

Figure 2: After removing outliers

Discretization will divide the continuous range of values to partition. This makes the learning methods to easily handle the continuous data. This has two types of discretization. The first type is supervised and the second type is unsupervised. We have considered the unsupervised type of discretization. In this the equal width and equal depth of bins have been considered. This is processed on Weka tool for discretization. We have applied discretization to the Kitchenham dataset for attribute “actual duration”. It divides the values into bins. This can be represented as follows
Figure 3: Discretization applied on actual duration

The number of bins and the values present in the bins can be represented as follows:

Table 2: Discretized values for "actual duration"

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Label</th>
<th>Count</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>-Infinity</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>2.</td>
<td>127.9 – 218.8</td>
<td>59</td>
<td>59</td>
</tr>
<tr>
<td>3.</td>
<td>218.8 – 309.7</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>4.</td>
<td>309.7 – 400.6</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>5.</td>
<td>400.6 – 491.5</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>6.</td>
<td>491.5 – 582.4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>7.</td>
<td>582.4 – 682.3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8.</td>
<td>682.3 – 764.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9.</td>
<td>764.2 – 855.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10.</td>
<td>855.1 – Infinity</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

5. CONCLUSION

Data pre-processing is a crucial phase of ML strategy and has extensive effect on the precision of ML strategies. Be that as it may, there is still absence of exact study in the software cost estimation connection for data pre-processing strategies notwithstanding of their significance. In this study, an organized writing study of data pre-processing strategies is initially led. In this way, an observational study is directed to break down the adequacy of the data pre-processing techniques, i.e. missing data treatment, Discretization and outlier detection and modifying. In addition, the interactions between the pre-processing techniques and ML methods’ predictive accuracies are also studied. In the end, the recommendation of using data pre-processing combination on different datasets and different ML methods is summarized. The effectiveness of an ML method used for software cost estimation can be significantly altered by single data pre-processing steps and their combinations (especially, the missing data treatment, scaling and FS).

6. REFERENCES


[16] Iman Attarzadeh, Amin Mehranazadeh, Ali Barati “Proposing an Enhanced Artificial Neural Network Prediction Model to Improve the Accuracy in Software Effort Estimation” 2012 Fourth International Conference on Computational Intelligence, Communication Systems and Networks


