Statistical Study in Performance Measures for Several Machine Learning Algorithms

With Measures of Classification Complexity

Yan Chen & Vasu Jindal
Extreme Computing Research Center at KAUST

yan.chen.1@kaust.edu.sa — +1 646-866-1082
vasu.jindal@kaust.edu.sa — +1 516-637-1962

Abstract

We summarize the relevant statistical methods and models for evaluating performance measures of different machine learning algorithms. A new model that measures the complexity of classification data is proposed based on the Fisher measures and standard performance level from corresponding algorithms. In the statistical analysis, the final statistics are assessed according to confidence range indicators of classification problems, where the hypothesis test becomes more sensitive for harder problems and less sensitive for easy classifications. In the statistical model, we propose the multi-variable logistic regression by considering different performance metrics as well as complexity. The Wald test is conducted to analyze statistical significance of model’s coefficients. Finally, the statistical experiments demonstrate the rationality of the methodologies.

Introduction

Currently, numerous machine learning algorithms are playing essential roles in solving various classification problems. The problem that arises from popularity of classification algorithms is how to establish an objective and unbiased principle to measure numerical performance of these algorithms and how to conduct a comparative study of evaluating them. Many people contributed to this field and they successfully proposed some crucial measurements such as the accuracy, precision, recall, Fscore and AOC or AUC area under the curve. Effective statistical tests have been successfully implemented to compare significant difference of algorithm groups based on the statistical confidence interval-hypothesis testing. However, they remain under-developed when it comes to comparing the improvement of one method over another one, both statistically and practically. The challenge of this problem is that we generally have different baselines based on complexity of problems. Therefore, it is necessary to consider simultaneously the complexity of problem and baselines of attributes of algorithms or starting levels in our statistical tests or models. In this report, we aim to build a statistical model for standard classification test, and the classification complexity is considered as the crucial factor. Finally, we propose a method of combining Fisher measures and norm distribution to measure this complexity.

Main Objectives

1. Investigate performance metrics of machine learning algorithms.
2. Investigate measures of complexity of classifications.
3. Propose the modifications to general statistical tests to reassess the statistics values.
4. Propose new statistical methods that consider classification complexity to determine improvement difference between two algorithms.

Performance Measures for Machine Learning Algorithms

The performance metrics from confusion matrix are defined as follows:

\[
\begin{align*}
\text{Accuracy} & = \frac{tp + tn}{tp + fp + fn + tn} \\
\text{Precision} & = \frac{tp}{tp + fp} \\
\text{Recall} & = \frac{tp}{tp + fn} \\
\text{Fscore} & = \frac{(\beta^2 + 1)tp}{(\beta^2 + 1)tp + \beta^2 fn + fp} \\
\text{Specifity} & = \frac{tn}{tp + fn} \\
\text{ROC} & = \frac{TP}{TP + FN} = \frac{FP}{FN} \\
\end{align*}
\]

Measures of Problem Complexity

Complexity Range for classification problem is defined as \(CR = (F_1 + F_2)(V)\) where \(V\) is the mean measure in performance formula (1). And:

\[
F_1 = \max_{i=1,\ldots,n} r_j f_i
\]

and

\[
t_j = \left(\frac{1}{\sigma_j^2} - \frac{1}{\sigma'_{j}}\right)^{\frac{1}{2}} \sqrt{\frac{2}{n}}
\]

where \(f_i\) denotes the \(i^{th}\) attribute and \(\sigma\) and \(\sigma'\) are respective mean and variance of \(i^{th}\) attribute.

\[
F_2 = \prod_{i=1}^{n} \left(\frac{\max(max(f_i, c_j), \min(f_i, c_j)) - \min(min(f_i, c_j), \max(f_i, c_j))}{\max(max(f_i, c_j), \min(f_i, c_j)) - \min(min(f_i, c_j), \max(f_i, c_j))}\right)
\]

Normalized CR to CR_{\text{max}}.

Statistical Methodologies

A. Two sample T test \(t_{\text{final}} = \times t_{\text{final}}\)

\[
1 - \frac{1}{\sqrt{2}} \leq \frac{t_{\text{final}}}{\sqrt{2} \text{NanYan}_X} \leq \frac{1 + t_{\text{final}}}{\sqrt{2} \text{NanYan}_X}
\]

And \(S_{X,Y} = \sqrt{\frac{N\text{Yan}_X}{N\text{Yan}_Y}}\) where \(X, Y\) are unbiased estimator of the variance of the two samples.

B. One way ANOVA analysis \(F_{\text{final}} = \frac{\text{max} \left(\frac{\text{SS}_{\text{between}}}{\text{SS}_{\text{within}}}\right)}{\text{df}_{\text{between}}}\).

C. Wilcoxon Signed Rank Test for Matched Pairs \(\text{W}_{\text{final}} = \times t_{\text{final}}\)

\[
W = \sum_{i=1}^{n} \text{sign}(x_{i} - y_{i}) |x_{i} - y_{i}|
\]

D. Friedman Test \(F_{\text{final}} = \times t_{\text{final}}\)

\[
\chi^2 = \frac{(n-1)\sum_{i=1}^{k} (R_i - \bar{R})^2}{\frac{k(k+1)}{4}}
\]

E. Quade Test & post-hoc Test

\[
F = \frac{(k-1)\beta}{(k-1)\gamma^2} \frac{\beta}{\gamma^2} > \chi^2_{\frac{k-1}(k+1)}(\gamma^2)-1
\]

\[
W_j - \frac{\gamma_j}{\gamma_j \left(\frac{n}{\gamma_j} - 1\right)}^2 \leq W_j \leq \frac{\gamma_j}{\gamma_j \left(\frac{n}{\gamma_j} - 1\right)}^2
\]

\[
\beta = \frac{n-j}{\gamma_j \left(\frac{n}{\gamma_j} - 1\right)}^2
\]

\[
\gamma_j = \frac{1}{\left(\frac{n}{\gamma_j} - 1\right)}^2
\]

\[
W_j = \sum_{i=1}^{n} (Q_i + R_{ij})
\]

Results

Table 1: Performance Measures for Algorithms in the Easy Data Set

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>Fscore</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.98</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>Nave Bayes</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>HM</td>
<td>0.96</td>
<td>0.95</td>
<td>0.95</td>
<td>0.96</td>
<td>0.95</td>
</tr>
</tbody>
</table>

We conduct the statistical tests: T test, Anova test, Friedman test and Quade test into 10 machine learning algorithms based on three data sets. For each data set, we score corresponding CRIs \(t_{\text{final}}\) to revise the results of hypothesis tests respectively. The Metrics Performance of Different Algorithms (1) is stated as previously. The algorithms that we investigate are Nave Bayes, Logistic Regression, Multi-layer Perceptron, Random Forest, AdaBoost, Bootstrap aggregating (Bagging), Multi-class Classifier, Support Vector Machines, Decision Tree.

Table 2: Friedman Test Performance

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>Prob</th>
<th>gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>124.49</td>
<td>12</td>
<td>10.37</td>
<td>11.6</td>
<td>0.001</td>
<td>0.5</td>
</tr>
<tr>
<td>Nave Bayes</td>
<td>98.49</td>
<td>12</td>
<td>8.17</td>
<td>9.8</td>
<td>0.007</td>
<td>0.6</td>
</tr>
<tr>
<td>HM</td>
<td>80.06</td>
<td>12</td>
<td>6.67</td>
<td>7.7</td>
<td>0.012</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Consider the adjustment of F statistics, we have \(F_{\text{final}} = 18.015, F_{\text{final}} = 27.08, F_{\text{final}} = 52.68\), difference is clearly significant for all data sets, thus the null hypothesis is sufficiently rejected based on 2% level significance.

Table 3: Quade Test-Multi Comparison Performance for Easy Classification

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>Prob</th>
<th>gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>124.49</td>
<td>12</td>
<td>10.37</td>
<td>11.6</td>
<td>0.001</td>
<td>0.5</td>
</tr>
<tr>
<td>Nave Bayes</td>
<td>98.49</td>
<td>12</td>
<td>8.17</td>
<td>9.8</td>
<td>0.007</td>
<td>0.6</td>
</tr>
<tr>
<td>HM</td>
<td>80.06</td>
<td>12</td>
<td>6.67</td>
<td>7.7</td>
<td>0.012</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>Prob</th>
<th>gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>124.49</td>
<td>12</td>
<td>10.37</td>
<td>11.6</td>
<td>0.001</td>
<td>0.5</td>
</tr>
<tr>
<td>Nave Bayes</td>
<td>98.49</td>
<td>12</td>
<td>8.17</td>
<td>9.8</td>
<td>0.007</td>
<td>0.6</td>
</tr>
<tr>
<td>HM</td>
<td>80.06</td>
<td>12</td>
<td>6.67</td>
<td>7.7</td>
<td>0.012</td>
<td>0.6</td>
</tr>
</tbody>
</table>

The critical value for post-hoc test is 75.8037, which implies there is no statistical difference if the significance value between two algorithms in the table is less than the critical value.

Reference


Acknowledgement

The authors gratefully acknowledge the supervising from Prof. Xin Gao and Prof. David E. Keyes and research work is fully supported by 2016 Supercomputing Workshop at KAUST. We also appreciate the intellectual lectures by experts during the workshop period.