Stratified and Un-stratified Sampling in Data Mining: Bagging

Yousef M. T. El Gimati
Statistics Department, Faculty of Science, Benghazi University-Libya
E-mail: yousef.elgimati@uob.edu.ly

Abstract

Stratified sampling is often used in opinion polls to reduce standard errors, and it is known as variance reduction technique in sampling theory. The most common approach of resampling method is based on bootstrapping the dataset with replacement. A main purpose of this work is to investigate extensions of the resampling methods in classification problems, specifically we use decision trees, from a family of stratification models to improve prediction accuracy by aggregating classifiers built on a perturbed dataset. We use bagging, as a method of estimating a good decision boundary according to a family of stratification models. The overall conclusion is that for decision trees, un-stratified bootstrapping with bagging can yield lower error rates than other sampling strategies for simulated datasets. Based on the results in these experiments, a possible explanation as to why un-stratified sampling is a best is because bagging is itself a method of stratification.

Keywords: bootstrapping, decision boundary, stratification models, resampling, classifier.

1. Introduction

In sampling theory, usually we do not know all the examples of the population of interest. As a result, we often make important decisions about a statistical population based on a relatively small amount of sample data. Typically, we take a sample and compute a quantity (statistic) using the simple random sample, especially when the population is homogenous (similar). In contrast, stratified sampling techniques are generally used when the population is heterogeneous (dissimilar), where certain homogenous or similar subpopulations can be isolated (strata). The objective of using strata information is to reduce standard error of the estimated quantity. For example, in classification problems, we can assume that different classes have their own unique distribution, and that samples within a class share some degree of homogeneity (similarity).

The approach discussed here is a sampling strategy based on resampling the training set. We examine both stratified and un-stratified bootstrapping approaches. For instance, if the original training set had two classes, we then split the data set according to classes into two subsets. Each disjoint subset is independently used to select random examples with replacement, and then the data are combined in one new
training set. Usually, the theory of stratified sampling deals with estimating the population mean from a stratified sample and with the best choice of the sample size \( n_j \), \( j=1,\ldots,J \) for \( j^{th} \) stratum. Here, we considered a sampling strategy in estimating the decision boundary based on several bootstrap resamples. The boundaries are arbitrary, in the sense that no particular structure or class of boundaries (no parametric model) is assumed a priori.

Although much work has been done on evaluating the performance of classification learning, specifically in decision trees, these evaluations have generally used un-stratified cross-validation. However, there was some evidence that stratification of data set with cross-validation may give estimates that are more accurate. This is because cross-validation yields an approximately unbiased estimator (Stone, 1974), while stratification is a variance reduction technique. The idea is this: in stratified cross-validation, the folds are stratified, so that they contain approximately the same proportions of observations according to classes as the original data set. A comparison of stratified and un-stratified cross-validation estimates was given by Breiman(1984) for 10 data sets generated from the same model, but with different random number seeds. They note that in their empirical trials with Classification and Regression Trees CART, (the stratified estimates never do worse than the un-stratified). Kohavi(1995) also notes that stratified cross-validation outperformed un-stratified cross-validation in terms of bias and variance.

In our case study, we follow an alternate line of comparison that is pursued via bootstrap aggregating (or bagging). Most classifiers are designed with the purpose of estimating a good decision boundary. We could employ bagging as a method of estimating the decision boundary in a classification tree according to sampling strategies on the training set, which may yield some insights into the phenomenon of the stratified and un-stratified methods. Our objective is to compare different methods of resampling the data set with bagging to see which give better classifier performance than learning one classifier over the entire data set. Specifically, we use decision tree classifiers.

1.1 Objective of study

A main purpose of this work is to investigate extensions of the resampling methods in classification problems, specifically we use decision trees, from a family of stratification models to improve prediction accuracy by aggregating classifiers built on a perturbed data set. The most common approach of resampling method is based on bootstrapping the data set with replacement, which was introduced by Efron(1993). In the original context, it was applied to obtain better estimates of quantities of interest. The two concepts -bias and variance- are often used to understand the
precision of estimates. For further details, see Hastie et al, (2008). In addition, in this study the aim is to identify the effects of stratified and un-stratified bootstrapping for different training sets by applying weighted bagging that produces better partitions than usual bagging.

1.2 Outline of paper
The rest of this paper is organized as follows. In Section 2, we describe stratified and un-stratified sampling approaches and how they relate to the classification problem. We then introduce a family of stratification models in Section 3; Section 4 contains the simulation study to identify the effects of sampling strategies for different training sets by performing a two-way analysis of variance. Finally, Section 5 concludes the paper.

2. Stratified and un-stratified sampling

The common framework of stratified and un-stratified sampling techniques was used in sample surveys many years ago. We begin with a general description of stratified sampling and how it may be used to estimate an expected value or mean of a population, and then specialize to our context. Suppose that given a set of objects of size \( n \), we divide \( n \) into \( J \) disjoint subsets or strata of sizes \( n_1, \ldots, n_J \). After this division or stratification, suppose that for stratum \( j \) of size \( n_j \), we draw \( m_j \) samples. If \( m \) represents the total number of samples, then \( m=m_1+\ldots+m_J \). Specifically, let \( x_{(i,j)} \) represent the value draw from stratum \( j \). If stratum weight (of stratum \( j \)) is defined as \( W_j = n_j/n \), then the stratified sampling estimate of the sample mean

\[
\bar{x}_{st} = \frac{1}{m} \sum_{j=1}^{J} W_j \bar{x}_j,
\]

where \( \bar{x}_j \) is the sample mean of stratum \( j \).

\[
\bar{x}_j = \frac{1}{m_j} \sum_{i=1}^{m_j} x_{(i,j)}.
\]

The variance of \( \bar{x}_{st} \) is

\[
\text{Var}(\bar{x}_{st}) = \sum_{j=1}^{J} W_j^2 \left( \frac{1-f_j}{n_j} \right) S_j^2,
\]

Where \( S_j^2 \) is the sample variance of stratum \( j \).

\[
S_j^2 = \frac{1}{(m_j - 1)} \sum_{i=1}^{m_j} (x_{(i,j)} - \bar{x}_j)^2.
\]

and \( f_j = \frac{m_j}{n_j} \) is the sampling fraction for \( j \)-th stratum. A good general explanation can be found in Cochran (1977).
An alternative approach is motivated by stratified and un-stratified bootstrapping with replacement in distributed learning, specifically in decision trees. In this context, we examine both stratified and un-stratified random samples of the training set, in which the subsets contain approximately the same proportion of classes as the training set. We describe the two methods in detail below. We consider bagging, bootstrapping the training set and aggregating the bootstrap samples of the original classifier by using: (i) usual bagging (simple majority vote), (ii) weighted bagging (weighted vote).

2.1 Comparison of stratified with un-stratified

Let us consider a problem of two overlapping normal distributions with mean $\mu_i, i=1,2$ and the same variance, as shown in Figure 2.1. If we can further assume that the prior probabilities are equal, then Bayes’ rule is $\mathcal{O}\left(\frac{\mu_1 - \mu_2}{2}\right)$, where $\mathcal{O}(\cdot)$ is a function based on population means which gives discrete values indicting class membership, such that $\mathcal{O}(x) = 1$ if $x < \frac{\mu_1 - \mu_2}{2}$ and 0 otherwise. Let us evaluate the classification efficacy of Bayes’ rule by bootstrap samples which lead to new discriminant rules$\mathcal{O}\left(\frac{\bar{x}_1 - \bar{x}_2}{2}\right)$, where $\bar{x}_1$ and $\bar{x}_2$ represent the two bootstrap sample means.

![Figure 2.1 illustration of two overlapping normal distribution with $\mu_1$ and $\mu_2$ vertical lines (dashed) and same variances. Bayes’ rule (densities intersect) is $\mathcal{O}\left[(\mu_1 + \mu_2)/2\right]$ vertical line (solid).](image)
Our task will now be to illustrate the sample mean and variance for stratified and un-stratified strategies. The result obtained will be put to good use in subsequent sections. Suppose that the sample sizes are equal \( m_1 = m_2 \) from stratified case, then we have \( \mu_1\, E_x \left( \frac{x_1 + x_2}{2} \right) = \frac{\mu_1 + \mu_2}{2} \) and \( \text{Var}_x \left( \frac{x_1 + x_2}{2} \right) = \frac{\sigma^2}{2m} \). For comparison of stratified with un-stratified bootstrapping, let us consider the sample size \( m \), using un-stratified sampling without replacement from a large population, then we have

\[
E_x \left( \frac{x_1 + x_2}{2} \right) = \left( \frac{m_1 + m_2}{2} \right) \text{from population I and} \ (m_1 - m_2) \text{from population II}
\]

where \( m_1 \sim B(m, \frac{1}{2}) \), then \( E_x \left( \frac{x_1 + x_2}{2} \right) = \frac{\mu_1 + \mu_2}{2} \). Therefore, both strategies have the same sample mean, which is unbiased estimator of \( \frac{\mu_1 + \mu_2}{2} \). Now, consider the situation of the variance, we have

\[
\text{Var}_x \left( \frac{x_1 + x_2}{2} \right) = \sum_{m_1=1}^{m-1} \text{Var}_x \left( \frac{x_1 + x_2}{2} \mid t_1 = m_1 \right) p(t_1 = m_1)
\]

\[
= \sum_{m_1=1}^{m-1} \left( \frac{\sigma^2}{4m_1} + \frac{\sigma^2}{4(m - m_1)} \left( \frac{m}{m_1} \right) \left( \frac{1}{2} \right)^m \right)
\]

\[
= \sum_{m_1=1}^{m-1} \frac{m \sigma^2}{4(m - m_1) m_1} \left( \frac{m}{m_1} \right) \left( \frac{1}{2} \right)^m
\]

Then we have

\[
\text{Var}_x \left( \frac{x_1 + x_2}{2} \right) = \frac{1}{4} \sum_{m_1=1}^{m-1} \frac{m}{m_1} \left( \frac{1}{2} \right)^m \times \frac{m}{m}
\]

\[
= \left( \frac{1}{2} \right)^m \sigma^2 \left[ m \times m + \frac{m}{2} \times \frac{m(m - 1)}{2} + \cdots + \frac{m}{m-2} \times \frac{m(m - 1)}{2} + \frac{m}{m-1} \times m \right]
\]

if \( m > 2 \) then \( m^2 > m + 2 \), and we have
This shows that the un-stratified variance is larger than stratified variance. From the result in Equation (2.1), of the two schemes, one can conclude that stratified sampling is a variance reduction technique with unbiased estimate.

3. Family of stratification models

In this section, we illustrate the use of a family of stratification sampling strategies in decision trees with bagging. We now consider a training set consisting of \( n \) observations \( l_i = (x_i, y_i) \), with \( x_i \) being the feature vector, and \( y_i \in [0, 1] \). The training set is split into the two classes with size \( n_1 \) and \( n_2 \), respectively. Bootstrap samples of size \( m_i \) are drawn from \( n_i \) with replacement, in which each observation may appear once, more than once, or never. Since \( m_1 \sim B(n_1, p) \), where \( p = \frac{n_2}{n_1 + n_2} \) for un-stratified samples, and \( m_i \) is fixed for stratified sampling, we can derive a family of possible distributions for \( m_i \) as (truncated) \( N(\mu, \sigma^2) \), where \( \mu = mp \) and \( \sigma = \sqrt{mp(1-p)} \) approximates the binomial. With this procedure, the first bootstrap sample of size \( m_i \) is drawn with replacement from \( n_i \), where \( m_i \) is a positive number, rounded to the nearest integer, generated from truncated random normal distribution with \( \mu = mp \) and standard deviation

\[
\sigma = \begin{cases} 
0, & \text{stratified} \\
\frac{2}{3} \sqrt{mp(1-p)}, & \text{partially stratified} \\
\frac{3}{2} \sqrt{mp(1-p)}, & \text{unstratified} \\
\end{cases}
\]
The second bootstrap sample of size \( m_2 = m - m_1 \) is drawn from \( n_2 \), the results are combined in one new training set with size \( m \) to give a different sample of the required types. Note that the sample in stratified sampling is constant if \( \sigma = 0 \), and partially stratified for small \( \sigma \). As \( \sigma \) increases, stratified sampling tends towards un-stratified, in which the sample size varies from sample to sample. A general introduction to stratified bootstrap sampling without replacement can be found in Davison and Hinkley (1997).

Recall that decision tree classifiers are unstable, i.e. small changes in the training set may lead to large changes in tree topology. Suppose that instead of a single training set of size \( n \), a sequence of \( B \) learning sets \( \{l_b\} \) is given to produce tree classifiers \( c_1(x), \ldots, c_B(x) \). This can be done by growing a tree in each bootstrap sample, and then evaluating in the test set without any pruning (large tree), a slight amount of pruned tree (default pruned tree), and a single split tree with only two terminal nodes (decision stumps), using the R tree algorithm for large tree, and R tree algorithm \texttt{rpart} for the rest of the pruned trees. Interestingly, the gap in performance is affected by the degree of pruning.

4. Simulation study

Experiments are performed based on the various sampling strategies to investigate which methods are the most suitable for reducing the bagging error rate corresponding to the decision boundary of the optimal classifier. These training sets consist of 400 observations of two classes, represented by values \( \{0,1\} \), which is defined by \( \sin(x_1) < x_2 \) as a classifier, approximately equal numbers of observation in each class. An independent test set consisted of 5000 observations generated in the same way as the training set with the same amount of random noise added. A bootstrap sample of different sizes (\( m=100, 200, 300, 400 \) and 500) of observations is selected according to the procedure in Section 2.2. A classifier is learned on each bootstrap sample, and then we use the predictions made by this classifier on the test set to estimate the classifier. The outputs of the sub-classifiers are aggregated by using the weighted bagging of 50 bootstrap iterations and finally averaging those over 50 training sets to obtain test error rate.

4.1 Experimental design

Classical statistical analysis can help to determine whether one method is more accurate than the others by using an appropriate statistical test. Typically in the classification problem, only one measure, that is the error rate, has been considered as a performance measure. For instance, we may want to decide on the basis of different sample sizes whether there is a difference in the effectiveness of different sampling strategies.

For this sort of situation we perform a two-way analysis of variance, in which the total variability of data is partitioned into one component which we ascribe to possible differences due to the sample sizes, a second component which we ascribe to possible
differences due to the sampling strategies, and a third component which we ascribe to possible interaction between the sample sizes and the sampling strategies, while the reminder of the variability is ascribed to chance or error term.

4.2 Two-way analysis of variance

In this section, before we present any formal statistical analysis of the results, we briefly introduce the two-way analysis of variance (ANOVA), which can be used here to test the statistical significance. Let us suppose we have r levels of factor 1 and c levels of factor 2, and that n independent observations can be observed at each rc combinations of levels. Denoting the kth observation at level i of factor 1 and level j of factor 2 by xi,j,k, we specify the underlying model as

\[ x_{i,j,k} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \]

For i=1,…,r, j=1,…,c and k=1,…,n, where \( \sum_{i=1}^{r} \alpha_i = 0 \), \( \sum_{j=1}^{c} \beta_j = 0 \), \( \sum_{i=1}^{r} (\alpha\beta)_{ij} = 0 \) for each j and \( \sum_{j=1}^{c} (\alpha\beta)_{ij} = 0 \) for each i. Here, \( \mu \)

represents an overall level (grand mean), \( \alpha_i \) represents the fixed effect of factor 1,

\( \beta_j \) represents the fixed effect of factor 2, \( (\alpha\beta)_{ij} \) is the interaction between factor 1 and factor 2, and \( \varepsilon_{ijk} \) is the error term which is assumed to be independent \( N(0,\sigma^2) \) for all i,j and k. We are interested in testing the null hypotheses of

\[ H_0: \alpha_i = 0 \text{ for } i=1,\ldots, r \]

\[ H_0: \beta_j = 0 \text{ for } j=1,\ldots, c \]

Note that in our experiment we did not find any evidence for interaction effects. Thus we have no \( (\alpha\beta)_{ij} \) term in our model, and then we can specify the underlying model in Equation (2.2) becomes

\[ x_{i,j,k} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \]

4.3 Experimental results

For our experiments, the numerical results of the two-way layout are presented in Tables 4.1, 4.2 and 4.3 for large, default pruned trees and stumps, respectively. It is observed from these tables that sample size has highly significant effects on the
estimated bagged error rate on the test set on average over 50 training sets for no pruning and default pruned trees, whereas giving a significance probability of 0.0140 for the stumps. Hence the null hypothesis that all sample sizes have the same bagged test error rate is strongly rejected for large and default pruned, and rejected at the 5% level for stumps. The sampling strategy gives a significance probability of 0.0004 and 0.0037 for large and pruned trees, respectively, whilst giving a non-significant difference for the stumps.

Table 4.1 ANOVA for comparison of four different sampling strategies versus different training sets based on large trees of estimated bagged test error rate.

<table>
<thead>
<tr>
<th>Tree size</th>
<th>Source variation</th>
<th>D.F</th>
<th>Mean square error</th>
<th>F-ratio</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>Sample size</td>
<td>4</td>
<td>0.00289</td>
<td>39.056</td>
<td>0.000***</td>
</tr>
<tr>
<td></td>
<td>Sampling strategy</td>
<td>3</td>
<td>0.00047</td>
<td>6.299</td>
<td>0.004***</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>192</td>
<td>0.00007</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Significance codes: *** highly significant, ** significant at 1%, *significant at 5%.

Table 4.2 ANOVA for comparison of four different sampling strategies versus different training sets based on default pruned trees of estimated bagged test error rate.

<table>
<thead>
<tr>
<th>Tree size</th>
<th>Source variation</th>
<th>D.F</th>
<th>Mean square error</th>
<th>F-ratio</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>Sample size</td>
<td>4</td>
<td>0.00915</td>
<td>82.286</td>
<td>0.000***</td>
</tr>
<tr>
<td></td>
<td>Sampling strategy</td>
<td>3</td>
<td>0.00052</td>
<td>4.640</td>
<td>0.0037**</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>192</td>
<td>0.000011</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Significance codes: *** highly significant, ** significant at 1%, *significant at 5%.

Table 4.3 ANOVA for comparison of four different sampling strategies versus different training sets based on stump trees of estimated bagged test error rate.

<table>
<thead>
<tr>
<th>Tree size</th>
<th>Source variation</th>
<th>D.F</th>
<th>Mean square error</th>
<th>F-ratio</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>Sample size</td>
<td>4</td>
<td>0.00232</td>
<td>3.211</td>
<td>0.014*</td>
</tr>
<tr>
<td></td>
<td>Sampling strategy</td>
<td>3</td>
<td>0.01015</td>
<td>0.748</td>
<td>0.525</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>192</td>
<td>0.00291</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Significance codes: *** highly significant, ** significant at 1%, *significant at 5%.

Therefore, the null hypothesis that the sampling strategies are equally accurate on average is rejected, except for the stumps, where there is no difference among the sampling strategies. We build our explanation that there are a large number of splits for the large tree using un-stratified bootstrapping strategy with $\sigma = \sqrt{n} \pi (1 - \pi)$, which leads to a large number of unstable regions where bagging improves over the
original classification tree. On the other hand, the stump has a single split with only two regions, and hence a low variability (stable regions).

Figure 4.1 the top panel is large trees, centre panel is default-pruned trees and bottom panel is stumps. The left plots represent different sample sizes versus estimated error rate for different sampling strategies, whereas the right plots represent different sampling strategies versus estimated error rate for different sample sizes. Note that each value in the figure is average over 50 training sets.
To visualize these results, Figure 4.1 summaries the experimental comparison of four different sampling strategies versus different training sets. The top-left plot shows that un-stratified sampling with \( \sigma = \sqrt{mp(1-p)} \) and large trees would be a useful performance improvement for all sample sizes. The top-right plot in the same figure exhibits the sampling strategy that the use of more information, especially for un-stratified, might also be a useful performance improvement. These results are confirmed by the default pruning showing in the centre plots of Figure 4.1. However, the results from the un-stratified sampling are very encouraging for all training set. The situation for stumps in the same figure at the bottom-left shows that un-stratified sampling performed well, but there is no statistically significant difference among the sampling strategies. Further, in the same figure at the bottom-right displays that sample sizes 200 and 100 performed better than the other sample sizes with all sampling strategies, especially for un-stratified bootstrapping.

5. Conclusions

Our work focused on the effect of sampling strategies to improve the estimated decision boundary of the optimal classifier. Here, we used bagging as a method of estimating a good decision boundary with a decision tree algorithm as the classifier model. The following considerations may provide insight into the pattern of results:

- Accuracy increased with the size of training set for both large and default pruned trees in both sampling strategies.
- Stumps gave reasonable results, especially with training sample of sizes 100 and 200 observations.

In the second experiments, we presented an investigation into alternative sampling strategies using different sample sizes with bagging. The results in these experiments suggest that un-stratified strategy with \( \sigma = \sqrt{mp(1-p)} \) for large and default pruned trees with all sample sizes can perform reasonably well. In addition, un-stratified stumps performed better than stratified for all sample sizes, especially with 100 and 200 observations. In general, it is recognized that increase in the sample sizes above a certain point yields diminishing returns.

In fact, based on the results in these experiments, a possible explanation as to why un-stratified sampling is best is because bagging is also a method of stratification (the stratification may not be disjoint), which means that the stratified strategy and bagging both act to reduce variance. It should come as no surprise that the estimated decision boundary is not significantly accurate by use of bagging in the stratified case. Thus, a significant gain in accuracy can be obtained by applying bagging to un-stratified strategy.
5.1 Potential research directions

The above discussion implies that un-stratified bootstrapping with

$$\sigma = \sqrt{mp(1-p)}$$

of large tree and a slight amount of pruned tree (default-pruned tree) with bagging can yield a better decision boundary than applying other sampling strategies. Results obtained here seem to support the position that bagging results depend simply on the tree levels, bootstrap sample size, and the data distribution, which are strongly affected of the performance of weighted and un-weighted bagging.

In addition to that, we probably need a good collection of data sets for comparing among a family of stratification models based on standard and weighted bagging. We can also recommend use of stratified cross-validation with bagging as an open of research direction.

References


