A Data Mining Study of Weld Quality Models Constructed with MLP Neural Networks from Stratified Sampled Data

T. W. Liao, G. Wang, and E. Triantaphyllou
Department of Industrial and Manufacturing Systems Engineering
Louisiana State University
Baton Rouge, LA 70803, USA

P.-C. Chang
Department of Industrial Engineering
Yuan-Ze University
Nei-Li, Taoyuan, Taiwan 32026

Abstract

A proportional stratified sampling method was implemented to sample radiographic welding data. The sample size was varied at different levels to study its effect on the model quality in terms of training time and its prediction accuracy. The sampled data of each size was then divided into training data and testing data in the ratio of 9 to 1. The training data is used to obtain multi-layer perceptron (MLP) neural network models. The quality of each model was subsequently evaluated with unseen testing data. Moreover, each sampled data set was characterized to show its statistical representation of the population. The correlation between the model quality and the sampled data statistics is also discussed.

Keywords
Data mining, MLP neural networks, Stratified sampling, Radiography, Weld quality.

1. Introduction
Welded structures, especially those for critical applications such as pipelines, offshore drilling platforms, and high-pressure vessels, are often tested non-destructively. Radiographic testing is one of the most commonly used non-destructive evaluation techniques. Today, the interpretation of radiographs still relies very much on human experts. This manual process is subjective, inconsistent, labor intensive, and sometimes biased. It is our belief that radiographs accumulated over time contain rich information that is not fully tapped. For instance, they can be utilized to describe weld quality due to any important welding process factor such as the skills of the welder, materials quality, new materials used, etc. Most importantly, radiographs can be used to mine the human interpretation knowledge, if most (if not all) discriminating features can be extracted from them.

Data mining, an important step in the knowledge discovery process, involves the application of specific algorithms for extracting patterns from data [1]. A wide variety of data mining methods exist and different classification schemes have been used to categorize them [2]. Self-organizing and supervised neural networks are well-established machine-learning techniques that can also be used for data mining. Jovanovic *et al.* [3] included neural networks as one data mining techniques in their integrated system architecture for the operation support in power and process plants. They specifically described the use of MLP neural networks for the prediction of structural component failure. Abidi and Goh [4] presented a Web-based Infectious Disease Cycle Forecaster (IDCF), which comprises a number of distinct neural networks trained on data obtained from long-term clinical observations of 89 types of bacterial infections. One disadvantage of using neural networks for data mining is that knowledge generated is not explicitly represented in the form of interpretable rules. Researchers such as Lu *et al.* [5], who presented an approach to discover symbolic classification rules using neural networks, are addressing this problem. Another disadvantage is that the learning processes are very slow. The efficiency of data mining is a very important issue because of the vastness of the data involved. This issue is especially critical if neural networks are chosen to be the data mining technique.

Several strategies have been proposed to improve the efficiency of data mining, which include data sampling [6], attribute reduction [7], use of high performance computing [8], knowledge-guided discovery [9], and focusing on the discovery of a restricted class of rules or those which appear most interesting [10]. This study follows the first strategy to investigate the efficiency and effectiveness of data sampling. In the next section, a brief description of the data and the data mining technique used for this study is given. Section 3 explains the implemented stratified sampling method for sampling data and testing of the statistical characteristics of the sampled data. The results are discussed in Section 4, followed by the conclusions.

2. Data Description and Mining Tool

This study used some data extracted from radiographic images of industrial welds that are available to us. The data set has 10,500 tuples with each tuple having 25 numeric attributes, which were originally extracted for welding flaw detection [11]. The categorical (or pattern) value of each record is known, which indicates whether a particular tuple is a welding flaw (taking a value of 1) or not (0). Approximately, 14.5% of the tuples represent cases with welding flaws. Refer to [11] for more details about the features and the extraction procedure.

The multi-layer perceptron (MLP) neural network is used as the data mining technique to capture the classification knowledge. Specifically, BrainMaker developed by California Scientific Software is used. BrainMaker uses backpropagation as its engine to modify connection weights. All neural models constructed in this study have three layers: the input layer has 25 nodes corresponding to the 25 numeric attributes, the hidden layer with 25 nodes, and the output layer with one node corresponding to the classification. The sigmoid function is used in every node. The criterion used to stop network training is when 91% training data are good. A training datum is considered "good" when the amount of error is within 25% of the pattern value.

3. Stratified Sampling

Stratified sampling first uses a stratifying strategy to separate the input data called focusing input or F_{in}, into a set of strata $S = \{s_1, \ldots, s_l, \ldots, s_L\}$, and then draws samples from each stratum independently by an application of other sampling techniques such as simple random sampling. The sampled data with size n is called focusing output or *Fout*. The stratifying strategy involves the selection of stratified variables, which must be categorical. For the welding data set studied, the stratified variable is binary: flawed or not flawed. There are four variations of stratified sampling: proportional sampling, equal size, Neyman's allocation and optimal allocation. Proportional stratified sampling ensures that the proportion of tuples in each stratum is the same in the sample as it is in the population. Equal size stratified sampling draws the sample number of tuples from each stratum. Neyman's allocation allocates sampling units to strata proportional to the standard deviation in each stratum. With optimal allocation, both the proportion of tuples and the relative standard deviation of a specified variable within strata are the same as in the population.

Algorithm PSS shows an implementation of proportional stratified sampling that is used in this study. Stratified sampling preserves the strata proportions of the population within the sample. It thus may improve the precision of the fitted models.

Algorithm PSS(F_{in}, n)

begin
    $F_{out} := \varnothing$;
    S := stratify(F_{in});
    $l := 1$;
    while $l \le |S|$ **do**
        $n_l := \lfloor n|s_l|/|F_{in}| \rfloor + 1$;
        $F_{out} := F_{out} \cup \mathrm{RS}\{s_l, n_l\}$;
        $l := l + 1$;
    enddo
    return(F_{out});
end;

Note that in the above algorithm $|\bullet|$ and $\lfloor \bullet \rfloor$ denote the cardinality of $\bullet$ and the largest integer smaller than $\bullet$,

respectively. Algorithm RS implements simple random sampling. Note that in this algorithm the sampling is done with replacement. That is, each tuple has the same chance at each draw regardless whether it has already been sampled or not.

Algorithm RS(F_{in}, n)

begin
  $F_{out} := \varnothing$;
  while $|F_{out}| \leq n$ **do**
    $i := random(1, |F_{in}|)$;
    $F_{out} := F_{out} \cup \{t_i\}$;
  enddo
  return (F_{out});
end;

Several sampling sizes were obtained (750, 1500, 3000, and 4500) in order to study their effect. Both the population and each sampled data set drawn from it are statistically characterized. The sample characteristics are compared with the population characteristics to show the representative-ness of the drawn samples.

Three types of statistical characteristics are distinguished but only the first type is studied here. The first type of characteristics describes the mean and variance of the attribute values. The second type considers the distribution of attribute values for simple attributes. The third type takes into account the joint distribution of attribute values for more than one single attribute. The key procedure used to analyze characteristics about focusing outputs in relation to focusing input is hypothesis testing. The null hypothesis, H_0, is that the sample characteristic is equal to the population characteristic. The alternative hypothesis, H_1, is that the sample characteristic is not equal to the population characteristic.

To test the mean of attribute j in the focusing output with sample size of n (>30), we compute the test statistic $s_{mj} = n^{1/2}(\mu_j(F_{out})-\mu_j(F_{in}))/\sigma_j(F_{out})$. H_0 is rejected at confidence level $1-\alpha$ if $s_{mj} > z_{1-\alpha/2}$ or $s_{mj} < z_{\alpha/2}$. For testing the variance of attribute j in the focusing output with sample size of n (>30), we compute the test statistic $s_{Vj} = (n-1)\sigma_j(F_{out})^2/\sigma_j(F_{in})^2$. H_0 is rejected at confidence level $1-\alpha$ if $s_{Vj} > \chi^2_{1-\alpha/2}(n-1)$.

4. Results and Discussion

4.1 Statistical Characteristics of the Sampled Data

The proportional stratified sampling program was repeated three times for each sample size. The mean and variance of each sampled data set were tested to determine the percentage of features that passed the test. Table 1 summarizes the test results for each group of sampled data sets. For instance, the mean of 12 features was statistically different from the population mean for the data set with 750 samples in the first group (52% = 100x(25-12)/25).

Table 1. Percentage of features passing the hypothesis test for three groups of data sets.

	Mean of Sampled Data				Variance of Sampled Data			
Group	750	1500	3000	4500	750	1500	3000	4500
1	52	72	100	76	92	84	92	88
2	100	72	96	72	84	68	84	80
3	96	84	84	84	84	88	92	80
Avg.	82.7	76.0	93.3	77.3	86.7	80.0	89.3	82.7
S.D.	26.6	6.9	8.3	6.1	4.6	10.6	4.6	4.6

Table 2 summarizes the statistical test results of six selected features. For each sample size, the percentage of its passing the test of its representative-ness of the population (or accepting H_0) is shown for two statistical characteristics: mean and variance. The percentage is derived based on three samples obtained by running the PSS algorithm three times. Subsequently, one-way ANOVA tests were performed to determine whether sample size is statistically significant for each selected feature. For this test, the outcome of two hypothesis tests, instead of three as in Table 2, was used to derive one percentage value. Therefore, three percentage values, instead of one as in

Table 2, were derived. The significance of $\alpha = 0.05$ is consistently used throughout all tests. The test results indicate that sample size is only significant for the means of features 5 and 20 and for the variances of feature 10.

Table 2. Percentage of passing three hypothesis-tests for six selected features.

	Mean of Training Data				Variance of Training Data			
Feature	750	1500	3000	4500	750	1500	3000	4500
1	100	67	100	100	67	100	100	100
5	100	33	100	100	100	67	100	100
10	100	100	67	67	100	0	67	0
15	67	100	100	67	33	67	67	33
20	67	33	67	0	100	67	100	100
25	100	67	67	33	100	100	100	100

4.2 Performance of MLP Neural Models

The performance measures of interest in this study are the time taken to train the MLP neural model (or the training accuracy) and the accuracy of welding-flaw detection. The PC used to learn neural models was a Gateway Enterprise E-4200 equipped with an Intel 450 MHz Pentium III Processor. Table 3 summarizes the results of different MLP neural models with identical number of hidden nodes. In training each neural model, four different random seeds were used to initialize the network weights. Therefore, four replications were obtained.

Table 3. Performances of MLP neural models with identical number of 25 hidden nodes.

		Training Time (seconds)					Testing Accuracy (%)				
Group	Repl.	750	1500	3000	4500	10500	750	1500	3000	4500	10500
1	1	187	110	165[1]	253[1]	1530	89.3	14.7	34.7	31.6	89.1
	2	131	315	160[1]	249[1]	171	90.7	59.3	34.0	27.1	88.4
	3	116	310	161[1]	239[1]	165	92.0	50.0	34.3	32.0	88.7
	4	93	204	167[1]	249[1]	152	86.7	25.3	33.3	30.9	88.2
	Avg.	132	235	163	248	505	89.7	37.3	34.1	30.4	88.6
	S.D.	40	98	3	6	684	2.3	20.8	0.6	2.2	0.4
2	1	135	80[1]	167[1]	245[1]		85.3	38.0	22.0	26.9	
	2	131	450	162[1]	241[1]		85.3	36.7	26.0	26.2	
	3	83	187	158[1]	242[1]		85.3	42.0	25.7	26.9	
	4	138	218	164[1]	242[1]		85.3	58.0	40.0	27.6	
	Avg.	122	234	163	243		85.3	43.7	28.4	26.9	
	S.D.	26	156	4	2		0.0	9.8	7.9	0.6	
3	1	NC[2]	75[1]	162[1]	246[1]		-	40.7	25.7	24.2	
	2	387	80[1,3]	157[1,4]	238[1,5]		85.3	40.7	26.3	28.6	
	3	NC[2]	72[1,3]	164[1,4]	240[1,5]		-	45.3	25.0	30.4	
	4	NC[2]	69[1]	161[1]	243[1]		-	49.3	25.0	31.6	
	Avg.	387	74	161	242		85.3	44.0	25.5	28.7	
	S.D.	0	5	3	4		0	4.1	0.6	3.2	
Avg. of Avg.		214	181	162	244	505	86.8	41.7	29.3	28.7	88.6
S. D. of Avg							2.5	3.8	4.4	1.8	

Notes: (1) The training process was forced to stop at 100 runs despite that almost every run meets the stopping criterion of 91% good (the testing accuracy was very low initially and started to converge beyond 50 runs).
(2) NC denotes not converged. That is, the model does not meet the criterion after 2000 runs.
(3) For replication 2 (3), 44% and 44.7% (63.3% and 48.7%) were obtained if the model was trained for 150 and 200 runs.
(4) Similar to the previous note, 29.7% and 30.7% (24.3% and 28%) were obtained for replication 2 (3).
(5) Similarly, 30% and 30.9% (32.2% and 34.4%) were obtained for replication 2 (3).

From Table 3, the following observations can be made:

1) The training time to reach the termination criterion could be volatile when different random number seeds are used. The training time also greatly varies when different data sets of the same size are used.
2) Data mining based on sampled data could be more efficient and as effective when a proper sample size is used (e.g. 750). However, an in-depth study is needed to determine the optimal sample size.
3) The testing accuracy is quite consistent when different groups of data sets are used (the standard deviation of average testing accuracy is low).
4) The learning pattern of the same sample size is also quite consistent among three different data sets.

The number of hidden nodes is known to affect the performance of MLP neural models. To investigate the magnitude of this effect, the number of hidden nodes is varied for each sample size by fixing approximately the ratio between the number of training data and the number of connection weights in the network. Assuming the network architecture of 25x25x1 is used for the population, the number of hidden nodes was determined to be 2, 4, 7, and 11 for sample sizes 750, 1500, 3000, and 4500, respectively. As a result of reducing the number of connections, the stopping criterion of 91% good training data becomes different to meet for the data set with 750 samples. Therefore, it was decided to introduce a different stopping criterion, i.e., the maximal number of runs. Two different maximal values were used to see the difference.

Table 4 summarizes the training and testing accuracy of MLP neural models constructed with varying number of hidden nodes. These results were obtained based on the first group of sampled data sets in Table 3. In training each neural model, four different random seeds were used to initialize the network weights. Therefore, four replications were obtained. For each sampled data set of a certain size, the coefficient of variance (COV) of performance was also computed. A low value of COV is desirable. The COV index, however, is not a good measure of performance because the average performance is far more important than the variance. The average training time for each sample size was also noted in the last row of Table 4.

Table 4. Performances of MLP neural models with varying number of hidden nodes.

		Training Accuracy (%)					Testing Accuracy (%)				
No. of Runs	Replica-tion	750	1500	3000	4500	10500	750	1500	3000	4500	10500
200	1	86.2	93.9	95.6	97.6	90.7	82.7	31.3	37.7	31.8	88.9
	2	87.0	93.2	95.3	97.6	91.0	22.7	24.7	38.3	36.2	87.7
	3	86.7	93.3	95.8	96.8	90.3	81.3	25.3	39.0	32.9	90.7
	4	86.4	93.0	96.2	97.6	91.1	81.3	28.7	43.7	32.9	88.6
	Avg.	86.6	93.4	95.7	97.4	90.8	67.0	27.5	39.7	33.5	89.0
	S.D.	0.4	0.4	0.4	0.4	0.4	29.5	3.1	2.7	1.9	1.3
	COV	0.004	0.004	0.004	0.004	0.004	0.441	0.112	0.069	0.057	0.014
500	1	83.7	94.0	95.4	97.2		82.7	34.0	39.0	31.1	
	2	88.2	93.0	95.1	97.5		90.7	28.0	42.3	35.1	
	3	83.7	93.3	95.3	97.6		92.0	27.3	38.7	36.4	
	4	86.1	92.8	96.0	97.6		86.7	39.3	45.0	30.0	
	Avg.	85.4	93.3	95.5	97.5		88.0	32.2	41.3	33.2	
	S.D.	2.2	0.5	0.4	0.2		4.2	5.6	3.0	3.1	
	COV	0.025	0.006	0.004	0.002		0.048	0.175	0.072	0.093	
Avg. Training Time											
	- 200 runs	77	157	317	484	1253					
	- 500 runs	195	393	789	1208						

Comparing Table 4 with Table 3, it can be observed that:

1) Using varying number of hidden nodes improves the performance of some sampled data sets (3000 and 4500), but degrades the performance of others (750 and 1500).
2) Using 500 runs as the termination criterion seems to improve the testing accuracy of most sampled data sets, except the 4500 data set.
3) The data set of 750 samples again has comparable testing accuracy as the population.

4.3 Correlation between Statistical Characteristics and Model Performances

The correlation between the testing accuracy and the mean passing percentage is found to be very low with a coefficient of -0.0284. Similarly, the correlation between testing accuracy and variance passing percentage is also found to be low with a coefficient of 0.096. In other words, a sampled data set with representative mean and variance of the population does not always produce better model performance and vice versa. Figures 1 and 2 show the corresponding scatter plots.

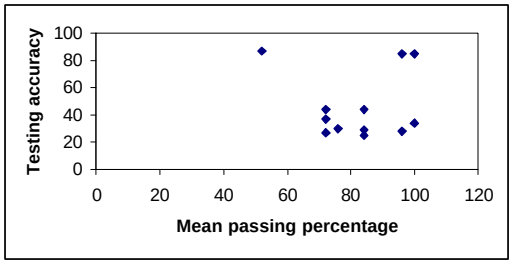

Figure 1. Scatter plot of accuracy vs. mean passing percentage.

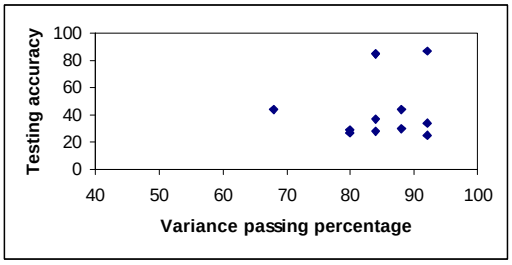

Figure 2. Scatter plot of accuracy vs. variance passing percentage

5. Conclusions

This paper has presented some results obtained in a study of mining weld quality models using MLP neural networks based on sampled data. Major conclusions include the following:

1) Data mining based on sampled data could be more efficient and as effective when a proper sample size is used.
2) There is no correlation between the representative-ness of sampled data, in terms of similar statistical characteristics as those of the population, and the model performance in terms of testing accuracy.

Although there is no guarantee that the MLP neural network models obtained are optimal, it is our belief that the conclusions will not be drastically changed even if each network model is optimized.

References

1. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P., 1996, "From Data Mining to Knowledge Discovery in Databases", *AI Magazine*, Fall, 37-54.
2. Chen, M.-S., Han, J., and Yu, P. S., 1996, "Data Mining: An Overview from a Database Perspective", *IEEE Trans. on Knolwedge and Data Engineering*, 8(6), 866-883.
3. Jovanovic, A., Poloni, M., Psomas, S., and Yoshimura, S., 1996, "Practical Aspects of Extraction of Knowledge from Data in Engineering", *Proc. 3rd World Congress on Expert Systems*, Feb. 5-9, Seoul, Korea, 1367-1376.
4. Adibi, S. S. R. and Goh, A., 1998, "Applying Knowledge Discovery to Predict Infectious Disease Epidemics", *Proc. 5th Pacific Rim Int. Conf. on AI*, November, Singapore, 170-181.
5. Lu, H., Setiono, R., and Liu, H., 1996, "Effective Data Mining Using Neural Networks", *IEEE Trans. on Knolwedge and Data Engineering*, 8(6), 957-961.
6. Reinartz, T., 1999, *Focusing Solutions for Data Mining*, Springer-Verlag, Berlin.
7. Lesh, N., Zaki, M. J., and Ogihara, M., 2000, "Scalable Feature Mining for Sequential Data", *IEEE Intelligent Systems*, March/April, 48-56.
8. Rana, O. F. and Fisk, D., 1999, "A Distributed Framework for Parallel Data Mining Using HP Java", *BT Technology J*, 17(3), 146-154.
9. Chen, Z. and Zhu, Q., 1998, "Query Construction for User-Guided Knowledge Discovery in Databases", *Journal of Information Sciences*, 109, 49-64.
10. Hussain, F., Liu, H., Suzuki, E., and Lu, H., 2000, "Exception Rule Mining with a Relative Interestingness Measure", *Proc. PAKDD 2000*, LNAI 1805, 86-97.
11. Liao, T. W., Li, D.-M., and Li, Y.-M., "Detection of Welding Flaws from Radiographic Images with Fuzzy Clustering Methods", *Fuzzy Sets and Systems*, 108(2), 145-158, 1999.