CHAPTER

An Evaluation of Sampling Methods for Data Mining with Fuzzy C-Means

K. Josien, G. Wang, T. W. Liao*, and E. Triantaphyllou
*ieliao@lsu.edu
Industrial & Manufacturing Systems Engineering Department
Louisiana State University, Baton Rouge, LA 70803

M. C. Liu
Manufacturing R&D, Boeing Company, Wichita, KS

ABSTRACT

Using fuzzy c-means as the data-mining tool, this study evaluates the effectiveness of sampling methods in producing the knowledge of interest. The effectiveness is shown in terms of the representative-ness of sampling data and both the accuracy and errors of sampled data sets when subjected to the fuzzy clustering algorithm. Two population data in the weld inspection domain were used for the evaluation. Based on the results obtained, a number of observations are made.

INTRODUCTION

Data mining is the application of specific algorithms for extracting knowledge from data (Fayyad *et al.*, 1996). Typical kinds of knowledge extracted include association rules, characteristic rules, classification rules, discriminant rules, clustering, etc. Chen *et al.* (1996) surveyed data mining techniques developed in several research communities according to the kinds of knowledge to be mined. This study makes use of a clustering algorithm, specifically the fuzzy c-means (Bezdek, 1987). The possibilistic c-means algorithm (Krishnapuram and Keller, 1993) was tried but eventually not used because of unsatisfactory results.

Clustering or unsupervised classification is the process of grouping physical or abstract objects into classes of similar objects. Consider the partition of a database with N tuples into m clusters. The number of ways in which this can be done, denoted by P(N, m), is as follows (Duran and Odell, 1974):

$$P(N,m) = \frac{1}{m!}\sum_{j=0}^{m}\binom{m}{j}(-1)^{j}(m-j)^{N}. \quad (1)$$

As N increases, P(N, m) grows exponentially. Given this huge search space, much effort has been spent to devise better clustering algorithms. Current clustering algorithms can be broadly classified into two categories: partitional and hierarchical. Partitional clustering algorithms attempt to determine m partitions that optimize a clustering criterion. Algorithms in this category include the popular c-means, CLARANS (Ng and Han, 1994), BIRCH (Zhang *et al.*, 1996), and CLIQUE (Agrawal *et al.*, 1998). A hierarchical clustering algorithm performs a nested sequence of partitions by either an agglomerative or divisive approach. The agglomerative approach starts by placing each object in its own cluster and then merges them into larger and larger cluster until all objects are in one cluster (Guha *et al.*, 1998; Loslever *et al.*, 1996). The divisive approach reverses the process.

Use of a distributed framework for parallel data mining offers another alternative to handle large data sets. Rana and Fisk (1999) described a distributed framework employing task and data parallelism using HPJava. A commercial tool that follows this strategy is Darwin of Oracle. The other alternative, called focusing, is to reduce data before applying data mining algorithms. Data reduction can be achieved by reducing the number of tuples and/or attribtues. Using C4.5 (Quinlan, 1993) and IB in MLC++ (Kohavi *et al.*, 1995) as the algorithms for mining classification rules, Reinartz (1999) analyzed the potentials of focusing tuples in data mining. SAS's Enterprise Miner implements most sampling methods.

Our study applies the same methods used by Reinartz (1999) for focusing tuples, but employs clustering instead of classification algorithms on

different data sets. In the next section, fuzzy c-means is briefly described. Section 3 presents the sampling methods used, followed by a description of the data set and knowledge sought. Section 5 discusses the results obtained in this study. The paper ends with a conclusion section.

FUZZY CLUSTERING

Fuzzy c-means (FCM) is used to serve as the data mining technique in this study. It is an unsupervised classification method, belonging to the partitional clustering category. It was derived from the hard (or crisp) c-means algorithm.

The hard c-means and its variants (Ball and Hall, 1967) are based on the minimization of the sum of squared Euclidean distances between data ($\boldsymbol{x_k}$, k=1, …, n) and cluster centers ($\boldsymbol{v_i}$, i =1, …, c), which indirectly minimizes the variance as follows:

$$Min\ J_1(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^2 \parallel x_k - v_i \parallel^2. \qquad (2)$$

In the above equation, $\mathbf{U} = [u_{ik}]$ denotes the matrix of a hard c-partition and $\mathbf{V}=\{\boldsymbol{v_i}\}$ denotes the vector of all cluster centers. The partition constraints in c-means are: (1) $u_{ik} \in \{0, 1\}\ \forall i, k$, (2) $\sum_{i=1,c} u_{ik}=1, \forall k$, and (3) $0 < \sum_{k=1,n} u_{ik} < n, \forall$i. In other words, each x_k either belongs or does not belong to a cluster and it can only belong to one cluster.

Dunn first extended the hard c-means algorithm to allow for fuzzy partition with the objective function as given in Eq. 3 below (Dunn, 1974):

$$Min\ J_2(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(m_{ik})^2 \parallel x_k - v_i \parallel^2. \qquad (3)$$

Note that $\mathbf{U} = [m_{ik}]$ in this and following equations denotes the matrix of a fuzzy c-partition. The fuzzy c-partition constraints are: (1) $m_{ik} \in [0, 1]\ \forall i, k$, (2) $\sum_{i=1,c} m_{ik}=1, \forall k$, and (3) $0 < \sum_{k=1,n} m_{ik} <$ n, $\forall i$. In other words, each x_k could belong to more than one cluster with each belonging-ness taking a fractional value between 0 and 1. Bezdek (1987) generalized $J_2(\boldsymbol{U}, \boldsymbol{V})$ to an infinite number of objective functions, i.e., $J_m(\boldsymbol{U}, \boldsymbol{V})$, where $1 \le m \le \infty$. The new objective function subject to the same fuzzy c-partition constraints is

$$Min\ J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(m_{ik})^m \parallel x_k - v_i \parallel^2. \qquad (4)$$

Note that both hard c-means and fuzzy c-means algorithms try to minimize the variance of those data within each cluster.

To solve the above model, an iterative procedure is required. Please refer to the original paper for the solution procedure. The FCM solution procedure was implemented in C language for this study.

SAMPLING METHODS

The sampling methods studied include simple random sampling, systematic sampling, and stratified sampling.

Simple random sampling selects n samples tuple-by-tuple from a population of size N by drawing random numbers between 1 and N. Denote the population of N tuples as the focusing input, F_{in}, and the selected samples as the focusing output, F_{out}. Algorithm RS shows an implementation of simple random sampling.

Algorithm RS(F_{in}, n)

---

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

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.

To draw n samples, systematic sampling first determine the step size, next draws the first tuple out of the focusing input at a random position, then iteratively adds each tuple with an index which refers to step positions after the selection position in the previous step. Algorithm SS describes an implementation of systematic sampling. Note that in this algorithm $\lfloor \bullet \rfloor$ denotes the largest integer smaller than $\bullet$.

Stratified sampling first uses a stratifying strategy to separate the focusing input 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 stratifying strategy involves the selection of stratified variables, which must be categorical. For the data set

Algorithm SS(F_{in}, n)

begin
    $F_{out} := \emptyset$;
    $step := \lfloor |F_{in}|/n \rfloor$;
    $i := start$;
    while $i \leq |F_{in}|$ **do**
        $F_{out} := F_{out} \cup \{t_i\}$;
        $i := i + step$;
    enddo
    return (F_{out});
end;

studied, the stratified variable is binary. 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 same 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.

For each sampling method, several sampling sizes were obtained at different levels in order to study their effect. The population as well as each

Algorithm PSS(F_{in}, n)

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

sample data set drawn from it are statistically characterized. The sample characteristics are compared with the population characteristics to show the representative-ness of drawn samples.

Three types of statistical characteristics are distinguished. The first type of characteristics describes the mean and variance of 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 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 $s_{mj} > z_{1-\alpha/2}$. 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 $s_{Vj} < \chi^2_{1-\alpha/2}(n-1)$.

Numeric attributes must be discretized before hypothesis testing for distribution can be performed. Consider attribute *j* with values in domain dom_j and a set of intervals $I = \{I_1, ¼, I_l, ¼, I_L\}$ with $I_l = [b_l, e_l[$, $b_l < e_l$, $1 \le l \le L$-1, and $I_L = [b_L, e_L]$. *I* is discretization of dom_j if $dom_j \subseteq I$, $b_1 = \min dom_j$, $e_L = \max dom_j$, and $b_{l+1} = e_l$. This study employs equal-width discretization, as shown in Algorithm EWD.

Algorithm EWD

```
begin
        I := ∅;
        b1 := min domj;
        eL := max domj;
        width := (eL - b1)/L;
        l := 1;
        while l ≤ L –1 do
                el := bl + width;
                bl+1 := el.;
                I := I ∪ [bl, el[;
                l := l +1;
        enddo
        I := I ∪ [bL, eL];
        return(I);
end;
```

To test the distribution of attribute j in the focusing output with sample size of n after being discretized, we compute the test statistic $s_{Dj} = \sum_{k=1,L}\{[n_{jk}(F_{out})\text{-}n \times n_{jk}(F_{in})/N]^2/\ n \times n_{jk}(F_{in})/N\}$. H_0 is rejected at confidence level $1\text{-}\alpha$ if $s_{Dj} < \chi^2_{1\text{-}\alpha/2}(L\text{-}1)$. This test is valid only if $n \times n_{jk}(F_{in}) \geq 5$ for all k. A similar test can be performed for joint distribution, but the number of combinations could be high as the number of attributes and the number of discretized intervals increase.

DATA AND KNOWLEDGE

Radiographic testing (RT) is one of several commonly used non-destructive techniques to evaluate welded structures such as off shore oil-drilling plate forms and space shuttle external tanks. With the assistance of a view box, a certified inspector interprets radiographs to determine whether a particular weld is sound or not. Although this is the mode of operation in industries today, human interpretation of weld quality is often subjective, inconsistent, labor intensive, and sometimes biased. Attempts have been made to develop a computer-aided system as an assistant to human inspectors. The key in this effort is to come up with a comprehensive set of interpretation knowledge used by human inspectors. It is our belief that this comprehensive set of interpretation knowledge can be extracted from the huge volumes of radiographic images archived. Because the huge amount of raw data involved, a data reduction operation called feature extraction is usually performed. This operation is critical because good knowledge cannot be obtained without discriminate features. Modeling of interpretation knowledge based on these features is yet another critical task, which is the focus of this work.

This study uses some data extracted from radiographic images of industrial welds that are available to us. Two populations of data organized in the form of tables are used. The first population of data has 2,275 tuples with each tuple having 3 numeric attributes, which were originally extracted for weld identification. Refer to Liao *et al.* (2000) for more detailed information about feature extraction. The second population of data has 10,500 tuples with each tuple having 25 numeric attributes, which were originally extracted for welding flaw detection (Liao *et al.*, 1999). For both data sets, the categorical value of each record is known, which indicates whether a particular tuple is a weld (for the first data set) or a welding flaw (for the second data set) or not.

The performance measures of interest here are the accuracy of weld identification or welding flaw detection, the false positive rate, the false negative rate, and the accuracy-falsehood ratio that is defined as the ratio

between the accuracy and the summation of the false positive rate and the false negative rate.

RESULTS AND DISCUSSIONS

For each data set, we first applied each one of the three sampling methods to generate focusing outputs of different sizes. For each sample size, ten focusing outputs were produced. Each focusing output was then statistically characterized and tested in relation to the population characteristic. Subsequently, we applied fuzzy clustering algorithms to each focusing output. The statistical test results are presented first, followed by the clustering results. In each category, the results are organized by data set.

Statistical Test Results

Weld Identification Data Set

Tables 1-3 summarize the statistical test results of the weld identification data by attribute. For each size of focusing output, the percentage of its passing the test of its representative-ness of the population (or accepting H_0) is shown for some statistical characteristics. Note that each entry corresponding to each statistical characteristic has two numbers with the first number derived from the first five focusing outputs and the second number the second five. The significance of $\alpha = 0.05$ is consistently used throughout all tests.
For each statistical characteristic of each feature, analysis of variance was performed to determine the significance of sampling method, sampling size, and their interaction. The results indicate that:

Table 1. Results of statistical test for feature 1 of the weld identification data set.

Sampling Method	*Sample Size*	*Mean*	*Variance*	*Distribution*
Random Sampling	*50*	*100, 100*	*60, 60*	*60, 20*
	100	*60, 100*	*80, 60*	*0, 20*
	200	*100, 100*	*20, 80*	*0, 20*
	300	*100, 80*	*40, 60*	*0, 0*
Systematic Sampling	*50*	*100, 80*	*60, 80*	*40, 20*
	100	*80, 100*	*20, 80*	*0, 0*
	200	*80, 100*	*60, 80*	*0, 0*
	300	*100, 100*	*40, 60*	*0, 0*
Stratified Sampling	*50*	*80, 100*	*80, 60*	*0, 20*
	100	*100, 100*	*80, 60*	*20, 20*
	200	*100, 100*	*40, 40*	*0, 0*
	300	*80, 100*	*80, 40*	*20, 20*

Table 2. Results of statistical test for feature 2 of the weld identification data set.

Sampling Method	*Sample Size*	*Mean*	*Variance*	*Distribution*
Random Sampling	*50*	*60, 100*	*40, 40*	*40, 20*
	100	*80, 80*	*60, 20*	*20, 0*
	200	*60, 100*	*60, 40*	*20, 0*
	300	*100, 100*	*40, 60*	*0, 0*
Systematic Sampling	*50*	*60, 60*	*40, 40*	*20, 40*
	100	*80, 100*	*20, 40*	*0, 0*
	200	*40, 60*	*20, 20*	*0, 0*
	300	*80, 80*	*40, 40*	*0, 0*
Stratified Sampling	*50*	*60, 100*	*40, 40*	*20, 0*
	100	*80, 100*	*0, 20*	*0, 0*
	200	*80, 80*	*20, 0*	*0, 0*
	300	*80, 60*	*0, 40*	*0, 0*

Table 3. Results of statistical test for feature 3 of the weld identification data set.

Sampling Method	*Sample Size*	*Mean*	*Variance*	*Distribution*
Random Sampling	*50*	*100, 100*	*80, 40*	*0, 0*
	100	*100, 100*	*60, 60*	*0, 0*
	200	*100, 100*	*80, 60*	*0, 0*
	300	*100, 80*	*100, 60*	*0, 0*
Systematic Sampling	*50*	*80, 100*	*20, 100*	*20, 20*
	100	*100, 100*	*80, 100*	*0, 0*
	200	*100, 100*	*40, 80*	*0, 0*
	300	*100, 80*	*80, 80*	*0, 0*
Stratified Sampling	*50*	*100, 100*	*40, 80*	*0, 0*
	100	*80, 100*	*60, 80*	*0, 0*
	200	*100, 100*	*60, 100*	*0, 20*
	300	*80, 100*	*40, 60*	*0, 0*

1) The means are statistically indifferent regardless the sampling method and sample size used.
2) For the variance characteristic, the sampling method factor is significant for feature 2 with p-value = 0.018.
3) For the distribution characteristic, sample size is always significant with p-values of 0.02, 0.003, and 0.044 for features 1, 2, and 3, respectively.
4) The interaction between sampling method and sample size is statistically significant with p-value = 0.005 for the distribution characteristic.

Welding Flaw Detection

Because this data set has 25 attributes, it will take up a lot of space to show all of the results. Tables 4-6 summarize the statistical test results of the welding

flaw detection data set for three selected attributes. For each size of focusing output, the percentage of its passing the test of its representative-ness of the population (or accepting H_0) is shown for some statistical characteristics. Note that as in Tables 1-3 the first number is derived from the first five focusing outputs and the second number the second five. The significance of $\alpha = 0.05$ is consistently used throughout all tests.

Table 4. Results of statistical test for feature 5 of the welding flaw detection data set.

Sampling Method	*Sample Size*	*Mean*	*Variance*	*Distribution*
Random Sampling	*100*	*100, 80*	*100, 60*	*40, 20*
	200	*100, 100*	*100, 80*	*20, 20*
	300	*100, 40*	*100, 60*	*20, 0*
	800	*100, 100*	*80, 60*	*0, 0*
	1000	*100, 80*	*80, 60*	*0, 0*
Systematic Sampling	*100*	*100, 100*	*100, 100*	*100, 100*
	200	*100, 100*	*100, 100*	*100, 100*
	300	*100, 100*	*100, 100*	*100, 80*
	800	*100, 100*	*100, 100*	*0, 0*
	1000	*100, 100*	*100, 100*	*0, 0*
Stratified Sampling	*100*	*100, 100*	*80, 100*	*40, 40*
	200	*100, 100*	*80, 100*	*40, 40*
	300	*100, 100*	*100, 100*	*20, 20*
	800	*100, 100*	*80, 100*	*0, 0*
	1000	*100, 100*	*80, 100*	*0, 0*

Table 5. Results of statistical test for feature 15 of the welding flaw detection data set.

Sampling Method	*Sample Size*	*Mean*	*Variance*	*Distribution*
Random Sampling	*100*	*80, 100*	*40, 40*	*20, 20*
	200	*100, 100*	*60, 20*	*0, 0*
	300	*100, 80*	*20, 40*	*0, 0*
	800	*100, 100*	*20, 20*	*0, 0*
	1000	*100, 100*	*20, 20*	*20, 0*
Systematic Sampling	*100*	*60, 80*	*0, 20*	*60, 20*
	200	*60, 100*	*20, 60*	*20, 0*
	300	*100, 100*	*40, 0*	*0, 0*
	800	*100, 100*	*40, 20*	*20, 0*
	1000	*100, 100*	*100, 100*	*0, 0*
Stratified Sampling	*100*	*60, 60*	*0, 0*	*40, 60*
	200	*80, 100*	*40, 20*	*0, 20*
	300	*100, 80*	*0, 40*	*0, 0*
	800	*100, 100*	*40, 40*	*0, 0*
	1000	*100, 100*	*0, 40*	*0, 0*

Table 6. Results of statistical test for feature 25 of the welding flaw detection data set.

Sampling Method	*Sample Size*	*Mean*	*Variance*	*Distribution*
Random Sampling	*100*	*100, 100*	*100, 100*	*0, 0*
	200	*60, 100*	*100, 100*	*0, 0*
	300	*100, 100*	*100, 100*	*20, 0*
	800	*100, 100*	*100, 100*	*20, 0*
	1000	*100, 80*	*100, 100*	*0, 0*
Systematic Sampling	*100*	*100, 100*	*100, 100*	*0, 0*
	200	*100, 100*	*100, 100*	*0, 20*
	300	*100, 100*	*100, 100*	*20, 20*
	800	*100, 100*	*100, 100*	*40, 0*
	1000	*100, 100*	*100, 100*	*0, 0*
Stratified Sampling	*100*	*100, 100*	*100, 100*	*0, 20*
	200	*100, 100*	*100, 100*	*0, 0*
	300	*100, 100*	*100, 100*	*0, 0*
	800	*80, 100*	*100, 100*	*0, 0*
	1000	*80, 100*	*100, 100*	*0, 0*

For each statistical characteristic of each one of the above features, an analysis of variance was performed to determine the significance of sampling method, sampling size, and their interaction. The results indicate that:

1) All three statistical characteristics of feature 25 are indifferent regardless the sampling method and sample size used.
2) The sampling method factor is significant for the variance characteristic of feature 5 with p-values = 0.011. In addition, all factors are significant for the distribution characteristic of the same feature with p-values < 10^{-4}.
3) The sample size factor is significant for the mean and distribution characteristics of feature 15 with p-values = 0.004 and 0.0002, respectively. In addition, the interaction between sampling method and sample size is significant for the variance characteristic with p-value = 0.015.

Clustering Results

Weld Identification Data Set

Table 7 summarizes the clustering results of the weld identification data set obtained by the FCM algorithm. For each size of focusing output clustered, the mean accurate rate (A), mean false negative rate (FN), mean false positive rate (FP), and mean accuracy-falsehood ratio defined as A/(FN+FP) are shown in each table. Each mean value was computed from ten values corresponding to ten focus output data. A weld not identified is a false negative whereas a non-weld identified as a weld is a false positive.

Table 7. Results of FCM clustering of the weld identification data set.

Sampling Method	*Sample Size*	*Mean Accurate Rate (%)*	*Mean False Negative Rate (%)*	*Mean False Positive Rate (%)*	*Mean Accuracy-Falsehood Ratio*
Random Sampling	*50*	*65.8*	*13.2*	*56.3*	*0.95*
	100	*62.4*	*14.2*	*65.3*	*0.79*
	200	*59.5*	*1*	*80.0*	*0.74*
	300	*56.4*	*5.3*	*80.7*	*0.66*
Systematic Sampling	*50*	*64.0*	*20.0*	*55.1*	*0.85*
	100	*58.1*	*10.4*	*73.3*	*0.69*
	200	*60.0*	*0.7*	*79.9*	*0.74*
	300	*57.4*	*0.9*	*81.0*	*0.70*
Stratified Sampling	*50*	*67.6*	*5.9*	*71.5*	*0.87*
	100	*61.4*	*0.2*	*77.0*	*0.80*
	200	*58.8*	*0.5*	*81.9*	*0.71*
	300	*60.0*	*0.5*	*79.8*	*0.75*
Focusing Input	*2275*	*59.2*	*0.6*	*80.5*	*0.73*

For each performance measure, an analysis of variance was performed to determine the significance of sampling method, sampling size, and their interaction. The results indicate that sample size is statistically significant for all performance measures and other factors are all insignificant. Overall, the accuracy, false negative rate, and accuracy-falsehood ratio decrease whereas false positive rate increases as sample size increases. It was surprised to find that for all performance measures except false negative rate, sample sizes of 50 and 100 generally fare better than the population. The performance of sampled data sets with size larger than 200 are more comparable with that of the population for this particular data.

Welding Flaw Detection Data Set

Table 8 summarizes the clustering results of the welding flaw detection data obtained by the FCM algorithm. For each size of focusing output clustered, the mean accurate rate (A), mean false negative rate (FN), mean false positive rate (FP), and mean accuracy-falsehood ratio defined as A/(FN+FP) are shown in each table. Each mean value was computed from ten values corresponding to ten focus output data. A welding flaw not detected is called a false negative. On the other hand, a non-flaw called as a flaw is a false positive.

Table 8. Results of FCM clustering of the welding flaw detection data set.

Sampling Method	*Sample Size*	*Mean Accurate Rate (%)*	*Mean False Negative Rate (%)*	*Mean False Positive Rate (%)*	*Accuracy-Falsehood Ratio*
Random Sampling	*100*	*58.8*	*33.0*	*42.8*	*0.78*
	200	*58.3*	*29.6*	*43.6*	*0.80*
	300	*55.6*	*28.3*	*47.1*	*0.74*
	800	*57.6*	*37.3*	*43.3*	*0.72*
	1000	*58.5*	*28.4*	*43.6*	*0.81*
Systematic Sampling	*100*	*60.5*	*33.8*	*40.3*	*0.82*
	200	*58.2*	*25.1*	*44.3*	*0.84*
	300	*58.8*	*24.9*	*44.0*	*0.85*
	800	*55.0*	*21.6*	*49.0*	*0.80*
	1000	*55.9*	*35.4*	*45.7*	*0.70*
Stratified Sampling	*100*	*60.0*	*24.0*	*40.5*	*0.93*
	200	*61.1*	*31.7*	*40.2*	*0.85*
	300	*61.0*	*25.2*	*41.4*	*0.92*
	800	*60.6*	*26.7*	*41.6*	*0.89*
	1000	*56.2*	*25.7*	*46.9*	*0.77*
Focusing Input	*10,500*	*64.5*	*19.0*	*38.3*	*1.13*

For each performance measure, an analysis of variance was performed to determine the significance of sampling method, sampling size, and their interaction. The results indicate that sampling method is statistically significant for the accuracy and false positive rate. It seems that stratified sampling produces better results than random sampling and systematic sampling in all performance measures. However, no sampling method gives better results than the population for this particular data set.

CONCLUSION

This paper evaluated three sampling methods with respect to the representative-ness and performance of the sampled data. The representative-ness is tested based on three statistical characteristics: mean, variance, and distribution. The performance is measured by using four indices: the accuracy rate, false negative rate, false positive rate, and accuracy-falsehood ratio based on the clustering results of fuzzy c-means. Two population data sets taken from the domain of radiographic testing of welds were used.

It is observed that:

1. Sample means are generally statistically indifferent from the population mean regardless the sampling method and sample size used.
2. The sampling method factor is significant for the variance characteristic for two out of six features tested (feature 2 of weld identification data and

feature 5 of welding flaw detection data). It is also significant for the distribution characteristic for one feature (feature 5 of welding flaw detection data).

3. The sample size factor is significant for the distribution characteristic for five out of six features tested (feature 25 of welding flaw detection data is the only insignificant one).
4. The interaction factor is significant for the variance characteristic of one feature (?) and for the distribution characteristic of two features (features 5 and 15 of welding flaw detection data).
5. For the weld identification data set, sample size is statistically significant for all performance measures and other factors are all insignificant. It was surprised to find that for all performance measures except false negative rate, sample sizes of 50 and 100 generally fare better than the population.
6. For the welding flaw detection data set, sampling method is statistically significant for the accuracy and false positive rate. In addition, stratified sampling seems to produce better results than random sampling and systematic sampling but worse than the population in all performance measures.

Depending upon the data, one factor might be more important than another. More tests on widely different data are needed to reach any definite conclusion. It is also desirable to determine if there is any correlation between the statistical characteristics of drawn samples and the performance measures of interest.

REFERENCES

Agarwal, R., Gehrke, J., Gunopulos, D., and Raghavan, P., "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications," *SIGMOD '98*, Seattle, WA, 94-105, 1998.

Ball, G. H. and Hall, D. J., ISODATA, an iterative method of multivariate analysis and pattern recognition, *Behavior Science*, 153, 1967.

Bezdek, J. C., *Pattern Recognition with Fuzzy Objective Function Algorithms* (Plenum Press, New York and London, 1987).

Chen, M.-S., Han, J., and Yu, P. S., "Data Mining: An Overview from a Database Perspective," *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 866-883, 1996.

Dunn, J. C., A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *J. Cybernet.*, 3, 1974, 32-57.

Duran, B. S. and Odell, P. L., *Cluster Analysis: a Survey*, Volume 100 of *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, 1974.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P., "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, 37-54, Fall 1996,

Guha, S., Rastogi, R., and Shim, K., "CURE: An Efficient Clustering Algorithm for Large Databases," *SIGMOD '98*, Seattle, WA, 73-84, 1998.

Kohavi, R., Sommerfield, D., and Dougherty, J., *Data Mining Using MLC++: A Machining Learning Library in C++*, http://robotics.stanford.edu/~ronnyk.

Krishnapuram, R. and Keller, J. M., "A Possibilistic Approach to Clustering," *IEEE Trans. on Fuzzy Systems*, 1(2), 1993, 98-110.

Liao, T. W., Li, D.-M., and Li, Y.-M., "Extraction of Welds from Radiographic Images Using Fuzzy Classifiers," *Information Sciences*, 126, 21-42, 2000.

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.

Loslever, P., Lepoutre, F. X., Kebab, A., and Sayarh, H., "Descriptive multidimensional statistical methods for analyzing signals in a multifactorial biomedical database," *Med. & Biol. Eng. & Compt.*, 34, 13-20, 1996.

Ng, R. T. and Han, J., "Efficient and Effective Clustering Methods for Spatial Data Mining," in *Proc. of the VLDB Conference*, Santiago, Chile, 144-155, 1994.

Quinlan, J. R., *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann, 1993.

Rana, O. F. and Fisk, D., "A Distributed Framework for Parallel Data Mining Using HPJava," *BT Technology Journal*, 17(3), 146-154, 1999.

Reinartz, T., *Focusing Solutions for Data Mining*, Springer, 1999.

Zhang, T., Ramakrishnan, R., and Livny, M., "BIRCH: An Efficient Data Clustering Method for Very Large Databases, " in *Proc. of the ACM SIGMOD Conference on Management of Data*, Montreal, Canada, June 1996.