Information Sciences 619 (2023) 84–97

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Eigen-Entropy: A metric for multivariate sampling decisions

Jiajing Huang [a], Hyunsoo Yoon [b,*], Teresa Wu [a], Kasim Selcuk Candan [a], Ojas Pradhan [c], Jin Wen [c], Zheng O'Neill [d]

[a] School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85281, USA
[b] Department of Industrial Engineering, Yonsei University, Seoul 03722, Republic of Korea
[c] Department of Civil, Architectural & Environmental Engineering, Drexel University, Philadelphia, PA 19104, USA
[d] J. Mike Walker '66 Department of Mechanical Engineering, Texas A&M University, College Station, TX 77843, USA

ARTICLE INFO

Article history:
Received 9 February 2022
Received in revised form 7 November 2022
Accepted 8 November 2022
Available online 12 November 2022

Keywords:
Information entropy
Correlation coefficient
Eigenvalues
Sampling
Model-free

ABSTRACT

Sampling is a technique to help identify a representative data subset that captures the characteristics of the whole dataset. Most existing sampling algorithms require distribution assumptions of the multivariate data, which may not be available beforehand. This study proposes a new metric called Eigen-Entropy (EE), which is based on information entropy for the multivariate dataset. EE is a model-free metric because it is derived based on eigenvalues extracted from the correlation coefficient matrix without any assumptions on data distributions. We prove that EE measures the composition of the dataset, such as its heterogeneity or homogeneity. As a result, EE can be used to support sampling decisions, such as which samples and how many samples to consider with respect to the application of interest. To demonstrate the utility of the EE metric, two sets of use cases are considered. The first use case focuses on classification problems with an imbalanced dataset, and EE is used to guide the rendering of homogeneous samples from minority classes. Using 10 public datasets, it is demonstrated that two oversampling techniques using the proposed EE method outperform reported methods from the literature in terms of precision, recall, F-measure, and G-mean. In the second experiment, building fault detection is investigated where EE is used to sample heterogeneous data to support fault detection. Historical normal datasets collected from real building systems are used to construct the baselines by EE for 14 test cases, and experimental results indicate that the EE method outperforms benchmark methods in terms of recall. We conclude that EE is a viable metric to support sampling decisions.

1. Introduction

Sampling is a statistical procedure concerning the selection of a subset of individual observations to capture the characteristics of the whole population for different applications [19]. If conducted properly, sampling can save time and cost while supporting statistical inferences [19]. There are, in general, two categories of sampling approaches: *probability* and *non-probability* sampling [19]. (a) In *probability sampling*, each observation from the population is assigned a certain probability of selection and is chosen by incorporating a random mechanism. Some common probability sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling [6,19]. *Simple random sampling* assigns

* Corresponding author.
E-mail address: hs.yoon@yonsei.ac.kr (H. Yoon).

https://doi.org/10.1016/j.ins.2022.11.023
0020-0255/

each observation with an equal probability of being sampled. *Stratified sampling* divides the whole population into several subgroups termed strata; then, each observation within a stratum is selected randomly, and selected observations across the strata become samples. *Cluster sampling* aggregates observations from the population into larger units called clusters. Samples are then randomly selected from the clusters. *Systematic sampling* selects members from a list of population members according to a random starting point and at fixed periodic intervals (b) In *non-probability sampling*, samples are selected subjectively and deliberately. This includes availability sampling, purposive sampling, quota sampling, and respondent-assisted sampling [6]. *Availability sampling* is a procedure in which samples are selected from a target population based on availability, self-selection, and/or discretion of the researchers, while *purposive sampling* selects samples that fit and meet the purpose of the study and specific criteria for inclusion/exclusion. *Quota sampling* realizes sample collection by combining availability sampling and purposive sampling to target specific numbers of observations with characteristics of interest, while *respondent-assisted sampling*, or snowball sampling, selects samples regarding previously selected observations in the population. It is noted that, unlike probability sampling, this set of sampling methods does not involve an explicit stochastic process and mostly relies on subjective judgment.

Sampling methods provide a guideline on which and how many samples must be selected as representatives to guarantee the generalization of study conclusions. Multiple sampling strategy factors include research objective, methodology, definition, and nature of the population, as well as the availability of resources and degree of confidence in generalized conclusions [19]. When comparing probability vs non-probability sampling [6], it is noted that probability sampling is generally preferred in studies requiring confirmatory purposes, quantitative design with a heterogeneous population, representative and unbiased samples capturing essential characteristics of the population, and statistical inferences from the samples. Non-probability sampling is favored when studies are exploratory or descriptive, a qualitative research design is required without the need for statistical inferences or representative samples, or a sampling frame is not available. The focus of this research is on probability sampling, which has been widely used in different fields of study, including machinery safety [23], geology [8], railroads [12], and building engineering [11].

Existing probability sampling research is mostly model-based. That is, they either assume a known distribution a priori or rely on a specific model to extract probability parameters. One example is active learning, which is a machine learning methodology that selects samples to be annotated for training to reduce the labor-intensive efforts of manual annotations [40]. Hajar et al. [23] presented two discrete random sampling strategies, additive random sampling (ARS) and jittered random sampling (JRS), for machine monitoring. Both ARS and JRS sampling show potential for simplified implementation in a remote application having a low-frequency rate while maintaining easy real-time operation management [23].

As the name implies, model-based probability sampling requires probabilistic models or data distribution given a priori to guide data sampling. However, for some real-world applications, there is often no sufficient historical data available to draw a meaningful statistical distribution. In addition, for high-dimensional datasets, it is challenging to determine an appropriate model to extract probability parameters. Model-free probability sampling takes a different approach and divides the dataset into subgroups to guide data sampling decisions. For example, Brus and van den Akker [8] utilized a stratified sampling survey to analyze the seriousness of subsoil compaction problems in the Netherlands. In their study, stratification is accomplished by a map showing five levels of subsoil compaction risks, and stratified sample data are used to estimate areal fraction, an indicator of over-compactness in the subsoil. Chen and Liu [12] proposed a high-dimensional clustering-based stratified sampling (HDCSS) method for roadway asset condition inspection, which yields a relatively small number of samples, potentially leading to inspection cost savings. Chen [11] incorporated the cluster sampling method with a symbolic aggregate approximation-based weather pattern matching (SAX-WPM) model to select samples from historical normal datasets to construct a baseline, serving as ground truth, to detect building anomalies. While promising, it is noted that cluster-based sampling may suffer from increased sampling errors when the base cluster selected already has biases [6]. Stratified sampling may also be challenging when there exist no stratifiable structures in the dataset [6]. In addition, both stratified and cluster sampling require subgroups to be identified first before the sampling.

As reviewed above, both model-based and model-free sampling approaches require pre-processing to either derive the distribution, estimate the probability parameters, or identify the subgroups. In this research, we investigate the use of entropy as a sampling decision metric without extensive pre-processing. Entropy is an information-theoretic measurement to quantify information richness [41] and has been used as a decision criterion for different applications. For example, Wang and Yao [47] proposed the concept of nonlinear correlation information entropy (NCIE) based on Pearson's correlation coefficient to remove redundant objectives in many-objective optimization problems (MaOP). Xia and Liu [49] extended NCIE to a supervised learning algorithm to determine features for synthetic aperture radar (SAR) image recognition. Wang et al. [48] proposed differential correlation information entropy (DCIE) for feature selection in classification problems. In addition, information entropy can also be used as a characteristic measurement, especially in the field of physical optics [45]. For example, Volyar et al. [43] conducted research on Laguerre-Gaussian (LG) beams and found that they produce a fine structure of the Hermite–Gaussian (HG) mode spectrum, and information entropy is used as one of the special integral characteristics. In another study, Volyar et al. [44] used information entropy to monitor the uncertainty in digital sorting perturbed LG beams. However, to the best of our knowledge, research on using entropy for sampling decisions is limited. Although researchers in [40] explore the use of entropy in active learning, as mentioned earlier, active learning requires the distribution of information to be drawn a priori. Some examples include studies by Rossini et al. [37] who used entropy for a locally robust decision on time series smoothing, Li et al. [32] who used entropy-based oversampling (EOS) methods in imbalance learning, Salehi et al. [39] who used relative entropy for semi-supervised section measurement and Xu et al. [50]

who used cross-entropy based noise correction for data and model quality improvement in crowdsourcing; However, these entropy-based approaches under specific distributions may not be suitable for high-dimensional problems [21].

To address the use of entropy for high-dimensional sampling issues, in this research, we propose a new entropy-based sampling metric called Eigen-Entropy (EE), which is based on eigenvalues derived from a correlation coefficient matrix. The proposed approach has three contributions: (1) EE is a model-free decision metric since it relies on data to extract information regarding sample sufficiency without any assumptions on data distributions; (2) Our theoretical analysis demonstrates that EE is able to well characterize the heterogeneity of a dataset; (3) The use of EE can assist sampling decisions specifically on how many samples and which samples should be selected from a massive amount of data. This paper is organized as follows. The design of EE and corresponding mathematical proofs are presented in Section 2. Two types of EE-based samplings, along with case studies and results, are detailed in Sections 3 and 4. Finally, the conclusions are presented in Section 5.

2. Methodology

In this section, we present a detailed description of the proposed Eigen-Entropy (EE) method and how EE can be used to support sampling decision-making. The key idea of EE is to obtain the entropy derived from the eigenvalues of a correlation magnitude matrix from multivariate data. The correlation magnitude matrix is based on the correlation coefficient matrix and takes the absolute values from the correlation coefficients, which measure the strength of the correlations (positive or negative).

2.1. Eigenvalues and positive semi-definite matrices

Let $\mathbf{A} \in R^{m\times m}$ be a matrix with non-negative entries:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{21} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix}, \tag{1}$$

where $a_{jk} \geq 0, \quad j,k = 1,\cdots,m$.

The eigenvalue λ is defined as a scaler such that

$$\mathbf{A}\boldsymbol{v} = \lambda\boldsymbol{v}, \tag{2}$$

where $\boldsymbol{v}$ is the corresponding eigenvector satisfying the equation $\boldsymbol{v}\neq 0$.

There exist m eigenvalues for $\mathbf{A}$ [42]:

$$\lambda_1 + \lambda_2 + \cdots + \lambda_m = tr(\mathbf{A}) = a_{11} + a_{22} + \cdots + a_{mm}, \tag{3}$$

where $tr(\mathbf{A})$ is the trace of $\mathbf{A}$ and λ_i, $i = 1,\cdots,m$ are the corresponding eigenvalues of $\mathbf{A}$. For a symmetric matrix, the eigenvalues are real, and the eigenvectors are orthogonal [42]. A symmetric real matrix $\mathbf{A}$ is positive semi-definite (PSD), denoted by $\mathbf{A}\succcurlyeq 0$, if $\boldsymbol{u}^T\mathbf{A}\boldsymbol{u}\succcurlyeq 0$ for every non-zero vector $\boldsymbol{u} \in \mathbb{R}^m$. For a symmetric matrix $\mathbf{A}$, it is PSD if and only if all its eigenvalues are non-negative [20].

2.2. Information entropy

Entropy is a term from physics that measures the degree of chaotic states in a (heat) system [14]. Shannon [41] extended this concept in information theory to describe the expected volume of information a message contains. Shannon's entropy (H) is defined as

$$H = -\sum_{i=1}^{N} p_i \log p_i, \tag{4}$$

where N is the number of values a random variable can have, and p_i is the probability of the random variable having the value of i ($\sum_{i=1}^{N} p_i = 1$). It is worth noting that entropy reaches a maximum when $p_i = \frac{1}{N}$ for all i's (uniformly distributed) [41].

2.3. Eigen-Entropy

Let $\mathbf{X} \in R^{n\times m}$ denote a dataset with n samples, where each sample has m features. We can represent $\mathbf{X}$ as a matrix:

$$\mathbf{X} = [\mathbf{x_1}, \cdots, \mathbf{x_n}]^{\mathbf{T}} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \tag{5}$$

where $\mathbf{x_i} = [x_{i1}, \cdots, x_{im}], i = 1, \cdots, n$.

Given this, the correlation coefficient matrix on the feature space of $\mathbf{X}$ is defined as

$$\mathbf{C} = \frac{1}{n}\mathbf{X_S^T X_S} = \begin{pmatrix} 1 & c_{12} & \cdots & c_{1m} \\ c_{21} & 1 & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & 1 \end{pmatrix}, \tag{6}$$

where

$$\mathbf{X_S} = \begin{pmatrix} \frac{x_{11}-\mu_1}{\sigma_1} & \frac{x_{12}-\mu_2}{\sigma_2} & \cdots & \frac{x_{1m}-\mu_m}{\sigma_m} \\ \frac{x_{21}-\mu_1}{\sigma_1} & \frac{x_{22}-\mu_2}{\sigma_2} & \cdots & \frac{x_{nm}-\mu_m}{\sigma_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}-\mu_1}{\sigma_1} & \frac{x_{n2}-\mu_2}{\sigma_2} & \cdots & \frac{x_{nm}-\mu_m}{\sigma_m} \end{pmatrix}. \tag{7}$$

In Eq. (7), μ_j denotes the mean and σ_j denotes the standard deviation of feature j. c_{jk} denotes the correlation between features j and k. That is, $\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, $\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \mu_j\right)^2}$, $c_{jk} = \frac{\sum_{i=1}^{n}(x_{ij}-\mu_j)(x_{ik}-\mu_k)}{\sigma_j \sigma_k}$ $(j \neq k, j, k = 1, \cdots, m)$, $c_{jj} = 1$.

Let

$$\mathbf{X_S^*} = \begin{pmatrix} \frac{|x_{11}-\mu_1|}{\sigma_1} & \frac{|x_{12}-\mu_2|}{\sigma_2} & \cdots & \frac{|x_{1m}-\mu_m|}{\sigma_n} \\ \frac{|x_{21}-\mu_1|}{\sigma_1} & \frac{|x_{22}-\mu_2|}{\sigma_2} & \cdots & \frac{|x_{2m}-\mu_m|}{\sigma_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{|x_{n1}-\mu_1|}{\sigma_1} & \frac{|x_{n2}-\mu_2|}{\sigma_2} & \cdots & \frac{|x_{nm}-\mu_m|}{\sigma_m} \end{pmatrix}. \tag{8}$$

We derive the correlation magnitude matrix $\mathbf{C}^*$ as

$$\mathbf{C}^* = \frac{1}{n}\mathbf{X_S^{*T} X_S^*} = \begin{pmatrix} 1 & c_{12}^* & \cdots & c_{1m}^* \\ c_{21}^* & 1 & \cdots & c_{2m}^* \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1}^* & c_{m2}^* & \cdots & 1 \end{pmatrix}, \tag{9}$$

where $c_{jk}^* \geq 0, j, k = 1, \cdots, m$.

We next show that $\mathbf{C}^*$ is positive semi-definite (PSD).

For any $\mathbf{a} \in R^m$ and $\mathbf{a} \neq 0$,

$$E\left(\left(\frac{1}{\sqrt{n}}\mathbf{X_S^*}\right)\mathbf{a}\right)^2 \geq 0. \tag{10}$$

We have that

$$\begin{aligned} E\left(\left(\tfrac{1}{\sqrt{n}}\mathbf{X_S^*}\right)\mathbf{a}\right)^2 &= E\left(\mathbf{a}^{\boldsymbol{T}}\left(\tfrac{1}{\sqrt{n}}\mathbf{X_S^*}\right)^{\mathbf{T}}\left(\tfrac{1}{\sqrt{n}}\mathbf{X_S^*}\right)\mathbf{a}\right) \\ &= E\left(\mathbf{a}^{\boldsymbol{T}}\left(\tfrac{1}{n}\mathbf{X_S^{*T}X_S^*}\right)\mathbf{a}\right) \\ &= E(\mathbf{a^T C^* a}) \\ &= \mathbf{a^T C^* a} \end{aligned} \tag{11}$$

Since $\mathbf{a^T C^* a} \geq 0$, $\mathbf{C}^*$ is positive semi-definite (PSD). According to [20], for symmetric matrix $\mathbf{C}^*$, which is PSD, its eigenvalues are real and nonnegative; that is, $\lambda_i \geq 0, i = 1, \cdots, m$. This nonnegative property of an eigenvalue is important to support the definition of Eigen-Entropy (EE).

Definition 1. Following the form of Shannon's entropy, Eigen-Entropy (EE) is defined as

$$EE = -\sum_{i=1}^{m} \frac{\lambda_i}{m} \log \frac{\lambda_i}{m}, \tag{12}$$

where λ_i is the i^{th} the eigenvalue of the correlation magnitude matrix $\mathbf{C}^*$, $\lambda_i \geq 0, i = 1, \cdots, m$.

We next establish the relationship between the degree of correlation captured by the correlation magnitude matrix and its eigenvalues.

Proposition 1. Without loss of generality, let us consider $\mathbf{C}^*$, where all the non-diagonal entries $c_{ij}^* = c$. λ is one of the corresponding eigenvalues. As the value of c increases, eigenvalue λ increases when $\lambda \in [1, \infty)$ and decreases when $\lambda \in [0, 1)$.

Proof. Let us construct a new correlation magnitude matrix $\mathbf{C}'$ from $\mathbf{C}^*$ by replacing c with αc, $\alpha > 1$:

$$\mathbf{C}' = \begin{pmatrix} 1 & \alpha c & \cdots & \alpha c \\ \alpha c & 1 & \ldots & \vdots \\ \vdots & \vdots & \ddots & \alpha c \\ \alpha c & \alpha c & \cdots & 1 \end{pmatrix}. \tag{13}$$

In [13], it is shown that if $\mathbf{C}^*$ has eigenvalue λ, then $\frac{1}{\alpha}\mathbf{C}^*$ has eigenvalue $\frac{1}{\alpha}\lambda$. Thus, $\lambda\prime$ is an eigenvalue of $\mathbf{C}'$, and $\frac{1}{\alpha}\lambda\prime$ is an eigenvalue of $\frac{1}{\alpha}\boldsymbol{C}\prime$.

Note that we can reconstruct $\mathbf{C}^*$ using $\frac{1}{\alpha}\mathbf{C}'$ and the identity matrix as follows:

$$\frac{1}{\alpha}\mathbf{C}' + \left(1 - \frac{1}{\alpha}\right)\mathbf{I} = \begin{pmatrix} 1 & c & \cdots & c \\ c & 1 & \ldots & \vdots \\ \vdots & \vdots & \ddots & c \\ c & c & \cdots & 1 \end{pmatrix} = \mathbf{C}^*. \tag{14}$$

In [13], it is also shown that if $\mathbf{C}^*$ has eigenvalue λ, then $\mathbf{C}^* + \alpha\mathbf{I}$ has eigenvalue $\lambda + \alpha$. Given this, the left-hand side of the above equation has an eigenvalue $\frac{1}{\alpha}\lambda\prime + 1 - \frac{1}{\alpha}$, while the right-hand side is $\mathbf{C}^*$ with eigenvalue λ. Thus, we have

$$\lambda = \frac{\lambda\prime}{\alpha} + 1 - \frac{1}{\alpha}, \tag{15}$$

or equivalently,

$$\lambda' - \lambda = (\alpha - 1)(\lambda - 1). \tag{16}$$

Given the above, we can conclude that as c increases, the eigenvalue λ' increases when $\lambda \in [1, \infty)$ or decreases when $\lambda \in [0, 1)$.

We next establish the relationship between the correlation magnitude matrix $\mathbf{C}^*$ and eigen-entropy EE.

Proposition 2. Let the correlation magnitude matrix $\mathbf{C}^*$ be such that all the non-diagonal entries $c_{ij}^* = c$. As c increases, Eigen-Entropy (EE) decreases.

Proof. $\mathbf{C}^*$ is PSD and its eigenvalues $\lambda_i \geq 0, i = 1, \cdots, m$,

$$tr(\mathbf{C}^*) = \lambda_1 + \lambda_2 + \cdots \lambda_m = m \tag{17}$$

Thus, we have $\sum_{i=1}^{m} \frac{\lambda_i}{m} = 1$.

Let $p_i = \frac{\lambda_i}{m}$, when we replace the $\frac{\lambda_i}{m}$ term in the EE definition with p_i, we set

$$EE = -\sum_{i=1}^{m} p_i \log p_i \tag{18}$$

As with Shannon's entropy, EE reaches its maximum $p_i = \frac{1}{m}$, or equivalently, when $\lambda_i = 1$.

Now, we connect c with EE. There are two scenarios:

- $\lambda_i \in [1, \infty)$, that is, $p_i = \frac{\lambda_i}{m} \geq \frac{1}{m}$.

  As c increases, λ_i increases. Thus, p_i will move further away from the maximum entropy point ($\frac{1}{m}$), and EE will decrease.

- $\lambda_i \in [0, 1)$, that is, $p_i = \frac{\lambda_i}{m} < \frac{1}{m}$.

  As c increases, λ_i decreases. Thus, p_i will move further away from the maximum entropy point ($\frac{1}{m}$), and EE will decrease.

We conclude that as c increases, EE decreases.

In summary, $\mathbf{C}^*$ records the absolute magnitude of the correlations among the variables that define the feature space, and the relationship between $\mathbf{C}^*$ and EE makes EE a potential metric to guide sampling decisions. Taking a dataset $\mathbf{X} \in R^{n\times m}$ with n samples and m features as an example, we can calculate EE for the first k samples from the dataset $\mathbf{X}$, EE(k). When the $(k+1)$ sample is to be added to the dataset if the new sample increases the variance (σ^2) of the dataset, in other words, the dataset is more diversified (heterogenous), the magnitude of the correlation (c) decreases and EE ($k+1$) increases, and vice versa. We conclude that the proposed EE can be considered as a single metric to determine *which* and *how many* samples to be included.

2.4. Eigen-Entropy based sampling

Given a dataset $\mathbf{X} \in R^{n\times m}$, Algorithm 1 presents the EE-based sampling method to select the subset $\mathbf{S}$ from $\mathbf{X}$. We first normalize the dataset $\mathbf{X}$ on each feature (line 1) and fill $\mathbf{S}$ with some initial samples to calculate the EE. Next, for each of the remaining data samples, we calculate a new EE on the updated $\mathbf{S}$ (with the added sample) to observe the change in EE (increasing or decreasing) and evaluate the rate of change for the EE (line 6). The addition is finalized if the rate of change is above a pre-defined threshold ε, which is a small number. Depending on the applications, in the case where data leading to a more homogeneous dataset is desired (see Section 3), the EE is expected to decrease; in the case where data leading to a more heterogeneous dataset is desired (see Section 4), the EE is expected to increase.

Algorithm 1: EE-based Sampling

Input: $\mathbf{X} \in R^{n\times m}$, n samples, m features
Output: The subset $\mathbf{S}$ from $\mathbf{X}$ determined by EE
Initialization:
1: Normalize $\mathbf{X}$ on each feature to obtain $\mathbf{X}'$
2: Initialize $\mathbf{S}$ with a few samples from $\mathbf{X}'$
3: Calculate EE using $\mathbf{S}$ (see **Definition 1**)
Sampling Decision:
4: For each of the remaining samples
5: Temporarily add the data element into $\mathbf{S}$ and calculate EE for the updated $\mathbf{S}$
6: Calculate the rate of EE change as: $\frac{EE}{\#ofsamplesin\mathbf{S}}$
7: For applications where we seek homogeneous samples
8: If EE decreases, we keep the element in $\mathbf{S}$
9: Otherwise, remove the data element from $\mathbf{S}$
10: If the rate of EE change is greater than a small number ε, the sampling process continues
11: Otherwise, stop
12: For applications where we seek heterogeneous samples
13: If EE increases, keep the element in $\mathbf{S}$
14: Otherwise, remove the data sample from $\mathbf{S}$
15: If the rate of EE change is greater than a small number ε, the sampling process continues
16. Otherwise, stop
17. **Return S**

Remark: Algorithm 1 adopts a greedy search strategy to identify the samples to be included or excluded in the sampling process. It is noted that the greedy search may be trapped at a local optimum if the EE curve is not monotonously increasing or decreasing. Fortunately, the objective here is to continuously update the EE curve with added samples to maintain the monotonous property of the EE curve. For illustration, two example samplings, a homogeneous sampling case study (see Section 3) and a heterogeneous sampling case study (see Section 4), are shown in Fig. 1.

3. Eigen-Entropy for homogeneous sampling

Here, we first focus on the imbalanced learning problem to demonstrate the use of Eigen-Entropy to assist sampling decisions where homogeneous data are to be sampled. Data imbalance is a common problem of machine learning in various domains, e.g., cardiovascular disease studies [17] or credit card fraudulent transactions [9]. As a result, accurate predictions regarding minority classes are of great importance. Extensive research has proposed solutions to address the challenges of an imbalanced dataset by balancing distributions of majority and minority classes [46]. Oversampling techniques are frequently used [25]. Common oversampling techniques include the synthetic minority over-sampling technique (SMOTE) [9], the majority weighted minority over-sampling technique (MWMOTE) [3], SMOTE + Tomek links (SMTL) [5], SMOTE + Edited nearest neighbors (SMENN) [4], EasyEnsemble (EASY) [33], and Balance-Cascade (BC) [33]. The basic idea behind most of

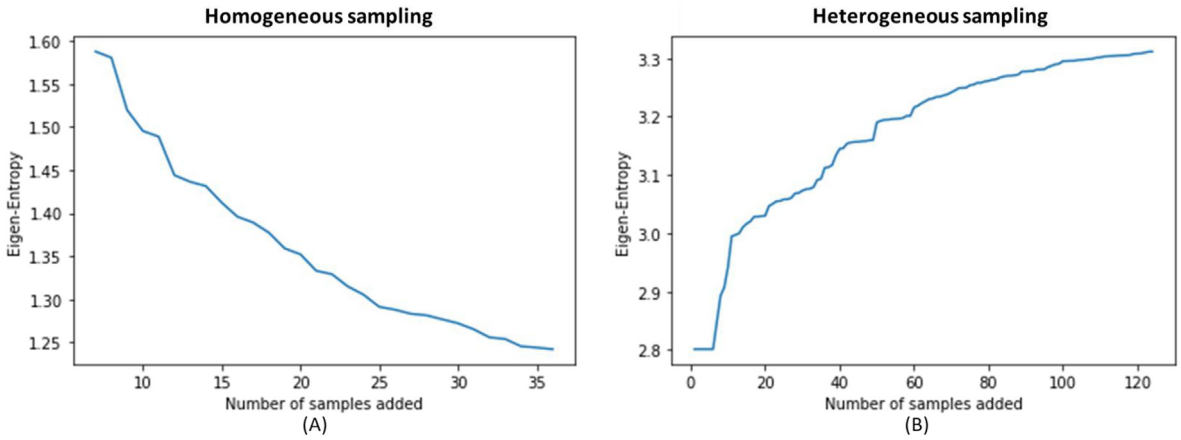

Fig. 1. Monotonous property of the EE curve is maintained while being updated with added samples for (A) a homogeneous sampling case and (B) a heterogeneous sampling case.

these oversampling techniques is to generate synthetic samples based on the distribution of information from the minority class to balance the datasets. However, these approaches may generate noisy or wrong minority samples leading to misclassification [32].

Instead of using distribution information, we argue that if the minority data samples are carefully selected (e.g., homogeneous samples being selected) as the basis for generating synthetic samples, the synthetic dataset may be less noisy, resulting in better classification performance. We propose the use of EE as a metric to guide oversampling decisions. Specifically, given a multi-class imbalanced dataset, for each of the minority classes, we apply Algorithm 1 to identify the homogeneous samples to generate the synthetic samples to obtain a balanced dataset. In this case, the smaller the EE is, the less diverse (more homogeneous) the dataset is. In the next section, where we compare EE against oversampling techniques, we use cosine similarity as a correlation coefficient measurement. This is because cosine similarity overcomes the issues of the nonexistence of correlations (e.g., in Pearson's correlations) [15].

3.1. Datasets

Ten publicly available real-world datasets from KEEL [2] and UCI repositories [18] are used in the comparison experiments (see Table 1). The first 7 datasets, vehicle1, segment0, page_block0, penbased, thyroid, shuttle, and ecoli-0-1-4-7_vs_2-3-5-6 (ecoli) are from KEEL, while the remaining, letter, waveform database generator version 1 (wavefm3), and landsat are from UCI. Among these 10 datasets, 4 (vehicle1, segment0, page_block0, and ecoli) are with binary classes, and the other 6 are multiclass datasets. Features in these datasets are numerical.

3.2. Benchmark algorithms

For comparison purposes, four SMOTE-based oversampling techniques, SMOTE [10], MWMOTE [3], SMTL [5], and SMENN [4], are included because SMOTE has been widely adopted as an oversampling technique [25]. In addition, an entropy-based imbalance degree sampling method, called the entropy-based oversampling approach (EOS), is included owing to its superior performance in imbalance learning [32]. As a result, there are a total of five methods that are compared against our proposed method. It is worth noting that all these techniques are to generate more samples from the non-majority classes to obtain

Table 1
Statistics of 10 experimental datasets (IR: imbalanced ratio).

Datasets	#Instances	#Features	#Classes	IR
vehicle1	846	18	2	2.9
segment0	2308	19	2	6.02
page_block0	5472	10	2	8.79
penbased	1100	16	10	1.95
thyroid	720	21	3	36.94
shuttle	2175	9	7	853
ecoli	336	7	2	10.59
letter	5000	16	26	0.96
wavefm3	5000	21	3	2.04
landsat	2000	36	6	1.98

the same number of samples as that from the majority class. With the same goal, instead of randomly selecting samples as the seed to generate new samples such as in SMOTE, our proposed method is to identify homogeneous samples from the non-majority classes by EE as the seed for new sample generations via SMOTE. Thus, our proposed method is termed EE-SMOTE.

3.3. Evaluation metrics

To investigate the information richness of the sampled data to support imbalanced learning, two commonly used base classifiers, multilayer perceptrons (MLP) [7] and AdaBoost [24] are implemented (Table 2). For each dataset, 5-fold cross-validation is performed. Each classifier is trained 10 times and the output is the average performance over the 10 runs. The One-vs-Rest strategy [36] is applied to multi-class datasets.

Precision [35], recall [35], F-measure [26], and G-mean [22] are used for classifier performance evaluations:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \tag{19}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{20}$$

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}, \tag{21}$$

$$\text{G-mean} = \sqrt{\text{Recall} \times \frac{TP}{TN+FP}}, \tag{22}$$

where *TP* denotes the number of minority samples correctly identified; *FP* denotes the number of non-minority samples incorrectly identified as the minority class; *TN* denotes the number of non-minority samples correctly identified; *FN* denotes the number of minority samples incorrectly identified as the non-minority classes.

3.4. Experimental results

We conduct experiments varying ε from 0.01 to 0.09 with 0.01 increments and observe that ε as 0.08 offers the most satisfactory classification results in terms of performance average and standard deviation (see Fig. 2). Thus, the results reported below are with ε set to 0.08. With the synthetic samples generated from the five benchmark methods and our proposed EE-SMOTE, we implement both the multilayer perceptrons (MLP) and AdaBoost. Fig. 3 illustrates the average performance of the two classifiers in terms of precision, recall, F1-score, and G-mean. The reason we take the average performance is for a fair comparison as in [32]. It is observed that EE-SMOTE outperforms the comparison methods on precision (Fig. 3(A)), recall (Fig. 3(B)), F-measure (Fig. 3(C)), and G-mean (Fig. 3(D)). It is also observed that EE-SMOTE has the smallest standard deviations on all four metrics, indicating the robustness of the algorithm.

In summary, the above results demonstrate that the EE-based sampling method is able to make sampling decisions such as *which* samples and *how many* samples should be included where homogeneous data must be sampled as the basis for synthetic data rendering to support imbalanced classification.

4. Eigen-Entropy for heterogeneous sampling

In this section, we focus on a fault detection problem to demonstrate the use of EE to assist sampling decisions where heterogeneous data must be sampled. Buildings are complex and integrated systems consisting of multiple sensors and subsystems, among which the heating, ventilation, and air conditioning (HVAC) systems are critical for building energy consumption and indoor environment quality. HVAC systems have been reported to be responsible for 20 % of a building's energy consumption [34], and such consumption accounts for 36 % of the global final energy use and 39 % of energy-related carbon dioxide emissions [28]. However, among this primary energy use, approximately 30 % is wasted because of operation faults and malfunctions in the HVAC systems [31]. Studies have shown that automatic fault detection and diagnosis (AFDD) on HVAC systems provide great potential for energy savings [38]. AFDD is a process that includes fault detection, identification, and isolation. Here, faults are typically defined as deviations from normal operating conditions in a

Table 2
Summary of two base classifiers and corresponding parameters.

Base Classifier	Parameters
AdaBoost	100 boosting iterations
MLP	100 epochs, 0.1 learning rate, 10 hidden layer neurons

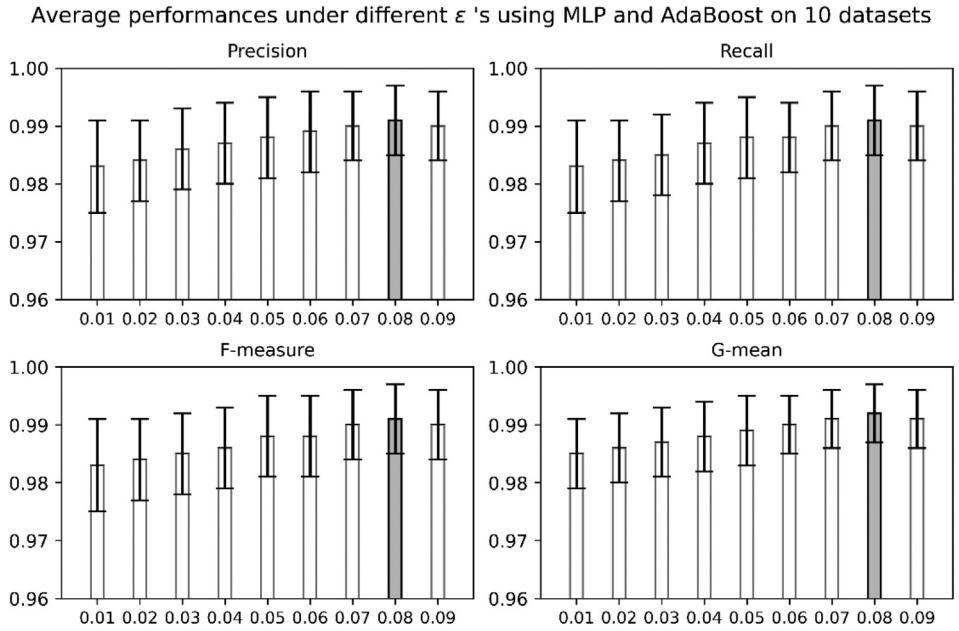

Fig. 2. Average performances under different ε are from 0.01 to 0.09 using MLP and AdaBoost on 10 public datasets in terms of precision, recall, F-measure, and G-mean. The most satisfactory results occur at ε = 0.08 (in grey).

building system [31]. Numerous AFDD methods have been developed over the past decades [31]. Regardless of the AFDD method, a baseline, representing the normal building operation conditions, is needed to enable fault detection. Because building systems are dynamic (e.g., temperature sensor readings vary from the morning to the afternoon) [11], a robust baseline is likely to include completely heterogeneous samples that cover diverse operation conditions [27].

In this paper, we propose the use of EE as a metric to sample heterogeneous data to be included in the building baseline. Specifically, given a building historical dataset (collected under normal operating conditions), we apply Algorithm 1 to identify the heterogeneous samples to be included in the baseline. As previously mentioned, cosine similarity is used for the correlation coefficient measurement on the feature space.

4.1. Datasets

The datasets used for fault detection are real building data collected from Nesbitt Hall at Drexel University [11], among which 14 are fault cases. For each fault case, there is a corresponding baseline candidate set, determined by building domain knowledge, obtained from existing historical normal datasets [11]. Data in each dataset are collected for a one-day period under a 5-min observation rate, and because building systems are complex with different functional components, the number of features in each fault test case varies. Details regarding these fault test cases can be referred to in Table 3.

4.2. Benchmark algorithms

For comparison purposes, our proposed EE-based sampling is compared against the multivariate pointwise mutual information (MPMI) method [16] and symbolic aggregate approximation using weather and schedule-based pattern matching (SAX-WPM) [11]. We choose these two benchmark methods because MPMI is a recently reported entropy-based method (which is of interest in this study) for multivariate Spatio-temporal data sampling, and SAX-WPM is a method developed specifically for building fault detection baseline constructions.

MPMI is a general method to assess the information richness of a multivariate dataset. For a sample $x_i = [f_{i1}, f_{i2}, \cdots, f_{im}]$, MPMI is defined as

$$MPMI(x_i) = \log \frac{p(f_{i1}, f_{i2}, \cdots, f_{im})}{p(f_{i1})p(f_{i2}) \cdots p(f_{im})} \tag{23}$$

where $p(f_{i1}, f_{i2}, \cdots, f_{im})$ and $p(f_{ij})$ is the joint probability of m features and the probability of feature j in x_i, respectively. For each $x_i \in X$, both joint probability and individual feature probability are estimated by histograms under a specific bin number. The MPMI of each sample is then normalized to obtain normalized MPMI (NMPMI). A rejection sampling algorithm [16] is applied for sampling decisions. That is, let s_i be picked from a uniform distribution $\mathcal{U}(0, 1)$: (1) if $NMPMI(x_i) > s_i$, x_i is

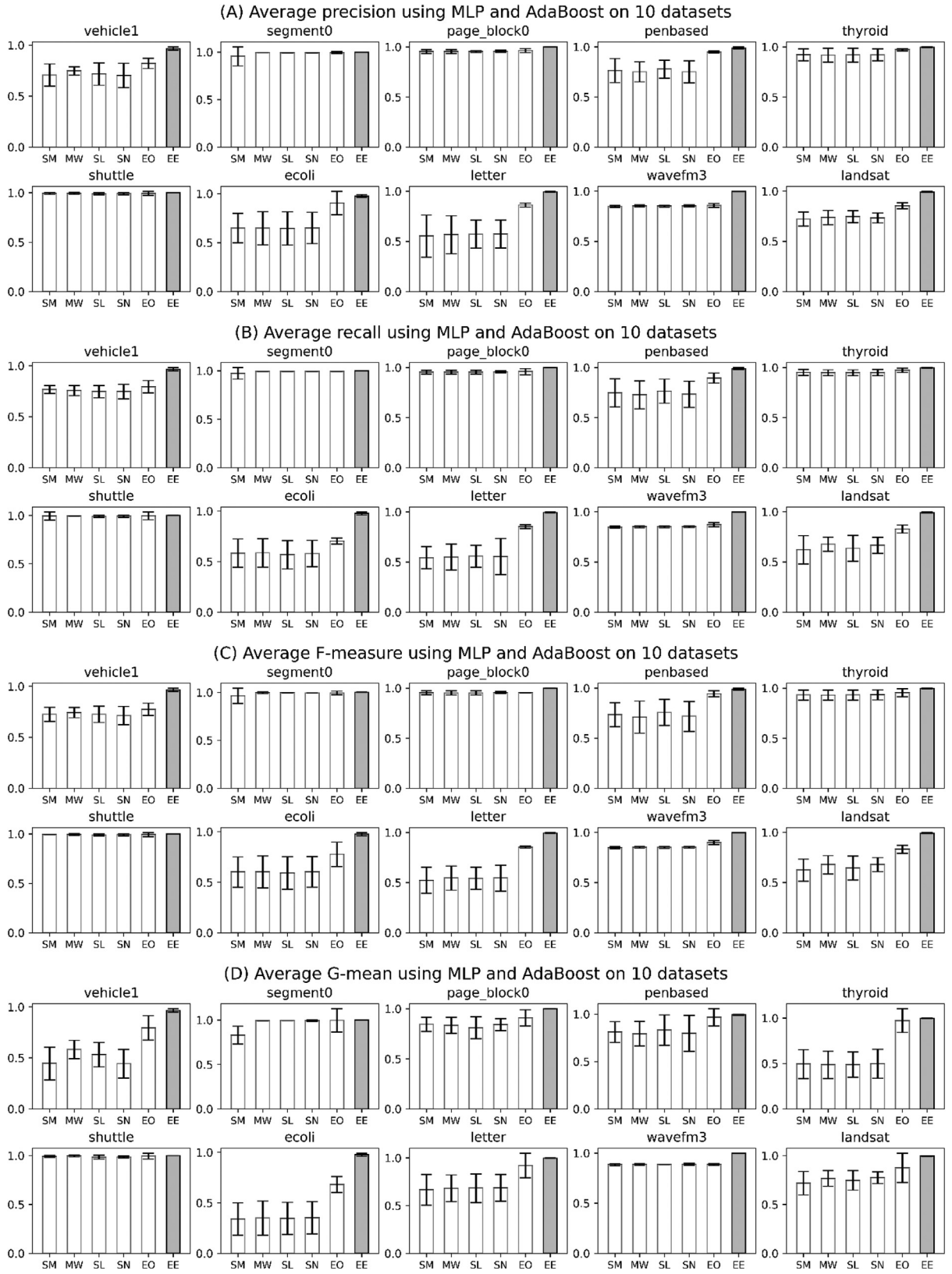

Fig. 3. Average performances using MLP and AdaBoost on 10 public datasets in terms of (A) precision, (B) recall, (C) F-measure, and (D) G-mean. Each subplot represents the results for a specific dataset under six different methods. Solid line segments indicate the standard deviation of performances. In each subplot, the x-axis indicates the methods (SMOTE (SM), MWMOTE (MW), SMTL (SL), SMENN (SN), EOS (EO), and EE-SMOTE (EE) under $\varepsilon = 0.08$, in this order). The results of EE-SMOTE are highlighted in grey.

Table 3
Summary of 14 fault test cases.

Fault test case name	Fault description	# of candidate samples	# of features
20160706	The system stopped from 4:00 PM to 23:30 PM	60,480	100
20160907	AHU-2 supply air temperature sensor negative bias 4°F	60,480	100
20160911	Operator fault, chiller off	60,480	182
20161201	AHU-2 outdoor air damper stuck at 90 % open	60,480	167
20170103	AHU-2 outdoor air damper stuck at 80 % open	60,480	167
20170114	Occupied from 1:30 AM to 7:00 AM	60,480	167
20170811	AHU-2 cooling coil valve position software override at 100 % open	56,448	182
20170915	Chiller chilled water differential pressure sensor positive bias 0.1 psi	56,448	182
20180709	AHU-2 supply air temperature sensor bias fault negative 3.5°F	60,480	182
20180710	AHU-2 OA damper stuck at 30 % open	60,480	182
20180711	AHU-2 cooling coil valve stuck at 80 %	60,480	182
20180718	AHU-2 OA damper stuck at 60 % open	60,480	182
20180722	Change weekend occupied schedule to end at 8:20 PM	60,480	100
20180723	CHWS temperature sensor negative bias 3.0 °F	60,480	182

selected; (2) otherwise, x_i is discarded. There is one parameter in implementing MPMI: the bin number. Here, we set the bin number to 128 as in [16].

In SAX-WPM, the symbolic aggregate approximation method is first employed to find similar patterns within a time series dataset, and these patterns are used to dynamically select qualified samples to generate a baseline. The SAX-WPM method has shown satisfactory performance over conventional data-driven baseline construction methods for high-dimensional building data [11]. There are two parameters in implementing SAX-WPM [11]: (1) the snapshot window size, which, as the name implies, divides a day (24 h) into snapshots (e.g., 1 h) for use in building control and building fault detection; (2) the number of symbolic letters, which is used to categorize weather conditions. As in [11], we set the snapshot window size to 30 min and the number of symbolic letters is set to 10. Following the literature on anomaly detection [1,11,29,30], we adopt Hotelling's T-Square (T^2), incorporating the principal component analysis (PCA) to identify whether a systematic abnormality exists in the building operations with respect to the baseline constructed. Specifically, T_i^2, Hotelling's T-Square for a sample i, is defined as

$$T_i^2 = \mathbf{x}_i^{\mathrm{T}} \mathbf{P} \sum a \mathbf{P}^T \mathbf{x}_i \tag{24}$$

where $\mathbf{x}_i$ is the sample i, $\mathbf{P}$ is a loading matrix obtained from PCA, and $\sum a$ is a set of nonnegative eigenvalues corresponding to a principal components. Because T^2 follows the F distribution, its upper bound can be obtained as

$$T_{threshold}^2 = \frac{a(n-1)}{n-a} F_{a,n-a,\alpha} \tag{25}$$

where n is the number of samples and α is the level of significance. Here, α is set to 0.05, and an abnormal sample i is flagged when $T_i^2 \geq T_{threshold}^2$.

4.3. Evaluation metrics

As in [46], we use recall (see Eq. (20)) to evaluate the performance of the baseline for fault detection, and here, *TP* and *FN* denote the number of abnormal samples correctly identified and the number of abnormal samples incorrectly identified as normal, respectively. A fault test case is said to be detected if its recall ≥ 0.5, following building domain knowledge [11].

4.4. Experimental results

Similar to the experiments on homogeneous sampling, we tested 9 different thresholds ε from 0.01 to 0.09 with 0.01 increments. The results of the fault detection test for 14 fault test cases using EE-constructed baselines are shown in Fig. 4. It can be observed that baselines constructed by our proposed method are able to detect all fault test cases under $\varepsilon = 0.07$ (see Fig. 4(A)). It is worth noting that an average number of baseline samples per case decreases as expected when ε increases (see Fig. 4(B)), mainly because a higher value of ε indicates that additional samples lead to a larger entropy change rate, and thus a smaller number of only qualified samples can be included. Hence, $\varepsilon = 0.07$ is considered as the stopping criterion in our proposed method to generate the final baseline, given that the number of baseline samples is relatively small while all fault test cases are detected under this value.

Table 4 presents the comparison of the fault detection results for the MPMI baselines, SAX-WPM baselines, and those by our proposed EE method under $\varepsilon = 0.07$. As previously mentioned, a fault test case is detected when its recall is greater than or equal to 0.5; therefore, we can observe that among the 14 fault test cases, baselines by MPMI failed to detect 2 cases, cases 20,160,911 and 20,170,114, (see marked rows in Table 4), while those by SAX-WPM and EE are able to detect all cases. For case 20160911, the number of samples in the baseline constructed by EE is 1788, 89.3 % less than that of MPMI, which is

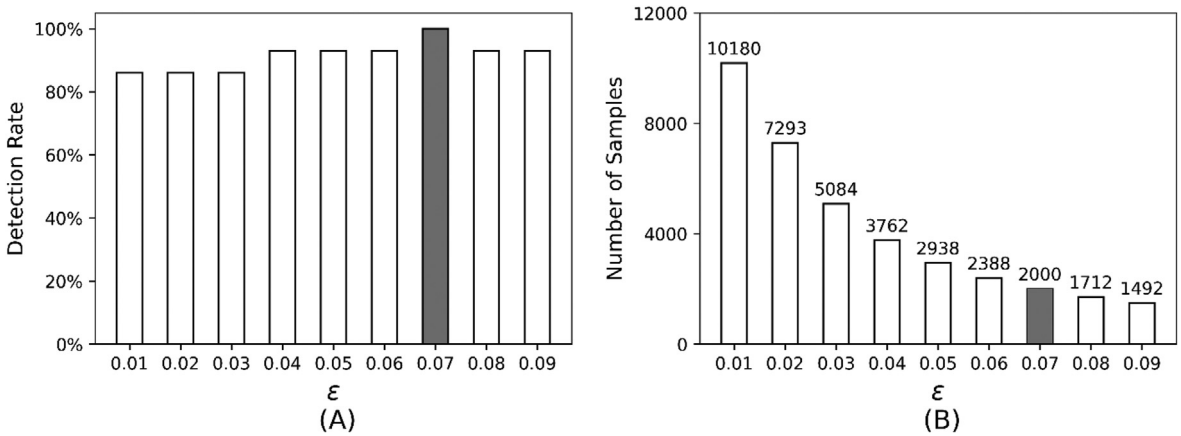

Fig. 4. Detection results for 14 artificial fault injection cases using baselines constructed by EE under different ε's: (A) fraction of detected fault test cases; (B) average number of baseline samples per case. ε = 0.07 is highlighted in grey.

Table 4
Comparison of detection results for 14 fault test cases using MPMI, SAX-WPM, and EE (ε = 0.07) baselines in terms of sensitivity and the average number of constructed baseline samples. The two cases (20160911 and 20170114) were not detected by the MPMI baselines but were detected by the EE baselines (MPMI*: bin = 128; SAX-WPM**: snapshot window size = 30-min, number of symbolic letters = 10).

Fault test case name	Recall			Number of samples		
	MPMI*	SAX-WPM**	EE (ε = 0.07)	MPMI*	SAX-WPM**	EE (ε = 0.07)
20160706	1.00	1.00	1.00	17,232	10,416	2083
20160907	0.90	1.00	1.00	15,216	7344	1905
20160911	0.12	0.92	0.52	22,896	5904	1788
20161201	0.88	1.00	0.88	18,480	8880	2051
20170103	0.99	0.99	0.99	18,672	6432	2142
20170114	0.16	0.78	0.52	16,656	9840	2273
20170811	0.89	0.90	0.94	17,856	7200	1990
20170915	1.00	1.00	1.00	17,952	8016	2024
20180709	1.00	1.00	1.00	22,800	7296	1846
20180710	1.00	1.00	0.98	22,800	10,608	2054
20180711	1.00	1.00	1.00	22,896	10,320	2095
20180718	1.00	1.00	1.00	22,752	8304	2012
20180722	1.00	1.00	1.00	17,376	8352	2053
20180723	0.98	1.00	0.99	22,848	4224	1679

22896, and 69.7 % less than that of SAX-WPM, which is 5904; for case 20170144, the number of samples in the baseline constructed by EE is 2273, 86.4 % less than that of MPMI, which is 16656, and 76.9 % less than that of SAX-WPM, which is 9840. For all 14 cases, the average number of baseline samples per case using MPMI and SAX-WPM are 15,574 and 8099 respec-

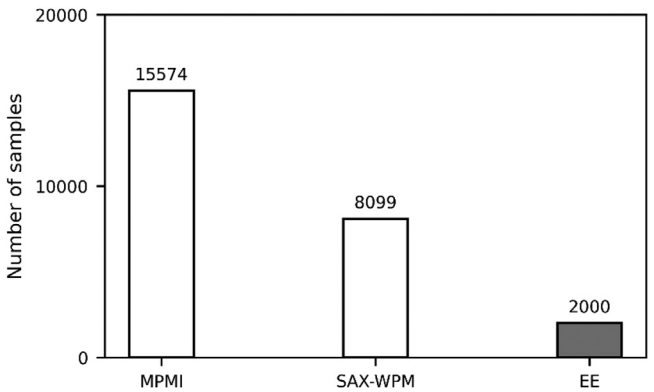

Fig. 5. Comparison of the results using the MPMI, SAX-WPM, and EE (ε = 0.07) methods in terms of the average number of constructed baseline samples across 14 fault test cases.

tively, while that using EE is 2000 (87.2 % and 75.3 % less than those of MPMI and SAX-WPM, respectively, see Fig. 5). Therefore, baselines constructed by the EE method require a significantly smaller number of samples than those by the MPMI method, which also indicates that our proposed method is promising.

In summary, this case study demonstrates that the sampling outputs (baselines) by our proposed EE method (EE-based heterogeneous sampling) outperform the literature-reported method (MPMI) for fault detections. Additionally, baselines constructed by the EE method require significantly fewer samples than those by the MPMI method. We conclude that the EE-based sampling method is able to make sampling decisions such as *which* samples and *how many* samples should be included in the heterogeneous sampling scenario.

5. Conclusion and future work

This study proposes and validates an information entropy-based sampling method. The proposed method uses Eigen-Entropy (EE), defined through eigenvalues from a correlation magnitude matrix using multivariate datasets as a decision-making criterion. Our proposed approach is able to automatically determine *which* samples and *how many* samples should be collected to construct a subset to support specific applications. Theoretical analyses show the relationship between EE and data heterogeneity through two sets of experiments that were conducted using homogeneous and heterogeneous samplings. The imbalance learning case studies show that our proposed homogeneous sampling method outperforms five other methods reported in the literature. The building engineering case studies demonstrate that the sampling results by our proposed heterogeneous sampling method perform better than those of the MPMI and SAX-WPM methods in detecting building operation faults.

Our future work lies in the following directions. In our current work, the stopping criterion ε is determined by empirical experiments. We plan to develop optimization techniques to determine the ε value as our next immediate step. Considering imbalanced learning applications, in the current work, we focused on SMOTE as an oversampling strategy. In the future, we plan to explore other oversampling strategies, such as BC. In addition, because EE as a general metric evaluates the information richness of data, it can be adopted in both supervised, semi-supervised, and unsupervised learning. For future work, we plan to explore the applicability of EE in active learning, identifying the samples to be annotated as a future effort.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by funds from the National Science Foundation Award under grant number IIP #1827757. The U.S. Government is authorized to reproduce and distribute for governmental purposes notwithstanding any copyright annotation of the work by the author(s). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

References

[1] J. Albert, Bayesian Computation with R, 2nd Ed., Springer, New York, NY, 2009.
[2] J. Alcala-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garcia, KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework, J. Multiple-Valued Logic Soft Comput. 17 (2011) 255–287.
[3] S. Barua, M.M. Islam, X. Yao, K. Murase, MWMOTE–Majority weighted minority oversampling technique for imbalanced data set learning, IEEE Trans. Knowl. Data Eng. 26 (2) (2014) 405–425.
[4] G.E. Batista, A. Bazzan, M. Monard, Balancing training data for automated annotation of keywords: a case study, J. Artif. Intell. Res. 3 (2) (2003) 15–20.
[5] G.E. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD Explor. Newslett. 6 (1) (2004) 20–29.
[6] A. Berndt, Sampling methods, J. Hum. Lact. 36 (2) (2020) 224–226.
[7] C. Bishop, Pattern Recognition and Machine Learning, Springer, New York, NY, 2006.
[8] D.J. Brus, J.J.H. van den Akker, How serious a problem is subsoil compaction in the Netherlands? A survey based on probability sampling, Soil 4 (1) (2018) 37–45.
[9] F. Carcillo, Y. Borgne, O. Caelen, Y. Kessaci, F. Oble, G. Bontempi, Combining unsupervised and supervised learning in credit card fraud detection, Inf. Sci. 557 (2021) 317–331.
[10] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[11] Y. Chen, Data driven Whole Building Fault Detection and Diagnostics, Ph.D. dissertation, Dept. Civ. Archit. Env. Eng., Drexel Univ., Philadelphia, PA, USA, 2019.
[12] Z. Chen, X.C. Liu, Roadway asset inspection sampling using high-dimensional clustering and locality-sensitivity hashing, Comput.-Aided Civ. Infrastruct. Eng. 34 (2019) 116–129.

[13] C. Chiang, M. Lin, The eigenvalue shift technique and its eigenstructure analysis of a matrix, J. Comput. Appl. Math. 253 (2013) 235–248.
[14] R. Clausius, The Mechanical Theory of Heat, Macmillan, London, U.K., 1879.
[15] R. Connor, A Tale of Four Metrics. Similarity Search and Applications, Springer, Tokyo, Japan, 2016.
[16] S. Dutta, A. Biswas, J. Ahrens, Multivariate pointwise information-driven data sampling and visualization, Entropy 21 (7) (2019) 669–694.
[17] W. Fan, Y. Si, W. Yang, M. Sun, Class-specific weighted broad learning system for imbalanced heartbeat classification, Inf. Sci. 610 (2022) 525–548.
[18] A. Fernandez, S. Garcia, M.J. del Jesus, F. Herrera, A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets, Fuzzy Sets Syst. 159 (18) (2008) 2378–2398.
[19] W.A. Fuller, Sampling Statistics, 1st Ed., John Wiley & Sons Inc, Hoboken, NJ, 2009.
[20] F.R. Gantmacher, The Theory of Matrices, Vol. 1, Chelsea Publishing Company, New York, NY, 1977.
[21] S. Geyer, I. Papaioannou, D. Straub, Cross entropy-based importance sampling using Gaussian densities revisited, Struct. Saf. 76 (2019) 15–27.
[22] H. Guo, H. Liu, C. Wu, W. Zhi, Y. Xiao, W. She, Logistic discrimination based on G-mean and F-measure for imbalanced problem, J. Intell. Fuzzy Syst. 31 (3) (2016) 1155–1166.
[23] M. Hajar, M.E. Badaoui, A. Raad, F. Bonnardot, Discrete random sampling - Theory and practice in machine monitoring, Mech. Syst. Signal Process. 123 (2019) 386–402.
[24] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, NY, 2009.
[25] H. He, Y. Ma, Imbalanced Learning: Foundations, Algorithms, and Applications, 1st Ed., John Wiley & Sons Inc, Hoboken, NJ, 2013.
[26] G. Hripcsak, A.S. Rothschild, Agreement, the F-measure, and reliability in information retrieval, J. Amer. Med. Informat. Assoc. 12 (3) (2005) 296–298.
[27] J. Huang, H. Yoon, O. Pradhan, T. Wu, J. Wen, Z. O'Neill, K.S. Candan, A cosine-based correlation information entropy approach for building automatic fault detection baseline construction, Sci. Technol. Built Environ. 28 (9) (2022) 1138–1149.
[28] International Energy Agency and the United Nations Environment Programme, Global Status Report: towards a zero-emission, efficient and resilient buildings and construction sector Retrieved from: https://www.worldgbc.org/sites/default/files/2018%20GlobalABC%20Glo-bal%20Status%20Report.pdf, 2018.
[29] I. Jolliffe, Principal Component Analysis, 2nd Ed., Springer, New York, NY, 2002.
[30] M. Kano, S. Hasebe, I. Hashimoto, H. Ohno, A new mulitivariate statistical process monitoring method using principal component analysis, Comput. Chem. Eng. 25 (7) (2001) 1103–1113.
[31] S. Katipamula, M. Brambley, Methods for fault detection, diagnostics, and prognostics for building systems—A review, Part I, HVAC&R Res. 11 (1) (2005) 3–25.
[32] L. Li, H. He, J. Li, Entropy-based sampling approaches for multi-class imbalanced problems, IEEE Trans. Knowl. Data Eng. 32 (11) (2020) 2159–2170.
[33] X. Liu, J. Wu, Z. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Trans. Syst. Man Cybern. Part B (Cybern.) 39 (2) (2009) 539–550.
[34] L. Pérez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption information, Energy Build. 40 (3) (2008) 394–398.
[35] D. Powers, Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation, J. Mach. Learn. Technol. 2 (1) (2011) 37–63.
[36] R. Rifkin, A. Klautau, In defense of one-vs-all classification, J. Mach. Learn. Res. 5 (2004) 101–141.
[37] R. Rossini, S. Poccia, K. S. Candan, M. L. Sapino, CA-Smooth: content adaptive smoothing of time series leveraging locally salient temporal features, in: Proc. 11th Int. Conf. on Management of Digital EcoSystems, Limassol, Cyprus, 2019, pp. 36-43.
[38] K.W. Roth, D. Westphalen, M.Y. Feng, P. Llana, L. Quartararo, Energy impact of commercial building controls and performance diagnostics: market characterization, Energy Impact of Building Faults and Energy Savings Potential. (2005).
[39] F. Salehi, M.R. Keyvanpour, A. Sharifi, SMKFC-ER: Semi-supervised multiple kernel fuzzy clustering based on entropy and relative entropy, Inf. Sci. 547 (2021) 667–688.
[40] B. Settles, Active learning Tech. Rep. 1648, Dept. Comput. Sci., Univ.Wisconsin-Madison, Madison, WI, USA, 2010.
[41] C.E. Shannon, A mathematical theory of communication, Bell Syst. Technol. 27 (1948) 379–423.
[42] G. Strang, Introduction to Linear Algebra, 5th Ed., Wellesley-Cambridge Press, Wellesley, MA, 2016.
[43] A. Volyar, E. Abramochkin, Y. Egorov, M. Bretsko, Y. Akimova, Fine structure of perturbed Laguerre-Gaussian beams: Hermite-Gaussian mode spectra and topological charge, Appl. Opt. 59 (25) (2020) 7680–7687.
[44] A. Volyar, M. Bretsko, Y. Akimova, Y. Egorov, Digital sorting perturbed Laguerre-Gaussian beams by radial numbers, J. Opt. Soc. Am. A 37 (6) (2020) 959–968.
[45] A. Volyar, M. Bretsko, Y. Akimova, Y. Egorov, Orbital angular momentum and informational entropy in perturbed vortex beams, Opt. Lett. 44 (2019) 5687–5690.
[46] Z. Wan, H. He, B. Tang, A generative model for sparse hyperparameter determination, IEEE Trans. Big Data 4 (1) (2018) 2–10.
[47] H. Wang, X. Yao, Objective reduction based on nonlinear correlation information entropy, Soft Comput. 20 (6) (2016) 2393–2407.
[48] W. Wang, Y. Yan, X. Ma, Feature selection method based on differential correlation information entropy, Neural Process. Lett. 52 (2020) 1339–1358.
[49] H. Xia, Z. Liu, Target classification of SAR images using nonlinear correlation information entropy, J. Appl. Remote Sens. 14 (3) (2020).
[50] W. Xu, L. Jiang, C. Li, Improving data and model quality in crowdsourcing using cross-entropy-based noise correction, Inf. Sci. 546 (2021) 803–814.