Stochastic Estimations of the Total Number of Classes for a Clustering having Extremely Large Samples to be Included in the Clustering Engine

Keishu Utimula,* Genki I. Prayogo, Kousuke Nakano, Kenta Hongo, and Ryo Maezono

Numerous reports have elucidated the classification of a large amount of data using various clustering techniques. However, an increase in data size hinders the applicability of these methods. Here, it is investigated how to deal with the exploding number of possibilities to be sorted into irreducible classes by using a clustering technique when its input capacity cannot accommodate the total number of possibilities. This can be exemplified by atomic substitutions in the supercell modeling of alloys. The number of possibilities is sometimes equal to trillions, which is extremely large to be accommodated in a cluster. Thus, it is not practically feasible to identify directly how many irreducible classes exist even though several techniques are available to perform the clustering. In this regard, a stochastic framework is developed to avoid the shortage limitations, providing a method to estimate the total number of irreducible classes (the order of classes), as a statistical estimate. The main conclusion is that the statistical variation of the number of classes, at each sampling trial, can serve as a promising measure to estimate the total number of irreducible classes. Characteristics of this approach is also discussed by comparing with the conventional one based on Polya’s theorem.

1. Introduction

Classifying a large number of data into some groups by individual attributes is a universal task. Beyond the handling based on clear classification rules, the clustering, driven by unsupervised learning techniques, has been developed extensively[1–4] An inherently associated situation with clustering is the large amount of data that needs to be classified. The larger it becomes, the more accurate, in general, the statistical quality of the estimations tends to be, thereby showing the superiority of data science. However, data size could be extremely large, rendering its accommodation into a single clustering process difficult. Let us set up the problem as follows: Suppose we have a clustering tool that can process up to \( l_{\text{max}} \) samples. The tool classifies \( l \) reducible samples (as input) into \( M(l) \) irreducible groups (as output) based on some attributes. Let the maximum possible number of samples be \( L \geq l_{\text{max}} \), and \( G \approx M(L) \) the total number of the irreducible attributes that we want to determine. However, we cannot identify \( G \) when \( l_{\text{max}} \ll L \) owing to the input capacity limitations.

To illustrate the problem more concretely, let us consider a magnetic alloy, \( \text{Nd}_{70.7}\text{Ce}_{0.225}\text{La}_{0.075}\text{Fe}_{14}\text{B} \), which is composed of \( \text{Nd}_{2}\text{Fe}_{14}\text{B} \) by substituting a part of Nd-sites with Ce or La. Supercell methods in ab initio calculations are used to describe such atomic substitutions, which are modeled as a periodic array of a cell structure.[5] The size of the cell (specified by how many Nd have to be accommodated) is selected to ensure that the given concentration \((x, y)\) is captured by the ratio of the number of atoms.[5] To describe \( \text{Nd}_{70.7}\text{Ce}_{0.225}\text{La}_{0.075}\text{Fe}_{14}\text{B} \), 40 Nd-sites are substituted by nine Ce and three La, as shown in Figure 1. The number of possible atomic configurations in this case would amount to \( 40!/((28!9!3!)) \approx 1,229,107,765,600 \approx L \) (total reducible samples to be classified). These “raw” configurations can further be classified into subgroups of irreducible configurations based on the theoretical framework to identify the equivalent structures under the symmetric operations.[6] The number of the subgroup, \( G \), reduces by several orders (typically \( G \approx 100 \)), even for \( L \approx \) trillion of samples. Once we identify \( G \), we can construct concrete \( G \) representative configurations as irreducible structures by using a recursive method. Several packages in materials science provide such a tool to perform the classifications[7,8] but their
input capacity, $l_{\text{max}}$, cannot handle such a large number of $L \approx (\text{trillion})$.

The present study considers the problem of identifying $G$ even for $l_{\text{max}} \ll L$. A stochastic method is proposed to identify $G = M(L)$ using a tool considering $M(l)$ ($l \ll L$) with $l_{\text{max}} \ll L$. The method repeats the numbers of sampling to calculate $M(l)$, thereby obtaining the respective statistics, $\{M(l)\}$ by which $G$ is estimated. The estimation can be achieved in considerably fewer trials than $L/l$, as shown in this study. In the above example that classifies crystal structures, we note that a more powerful deterministic framework exists (Polya’s theorem); however, the present method can properly handle the example. Furthermore, the method described in this paper can handle a broader category of problems. The method only requires a technique to provide $M(l)$, which can be a machine-learning technique to perform the clustering, wherein the classification rule is not necessarily based on any clear principle such as solid-state physics, group theory, as observed in the above-mentioned statistics. Polya’s theorem as a conventional solution for the similar problem relies on algebraic symmetry, and hence its applicability is limited only to such problems with the symmetry. We note again that the present framework has much wider applicability because it does not rely on the symmetry, but only on the existence of any sorting tool that is not necessarily deterministic, for example, machine-learning clustering. When combined with existing clustering techniques, the method can expand the scope of the technique beyond the limitation of input capacities (depending on the memory capacity) for computations.

If $l_{\text{max}}$ is around several times greater than $G$, the convergence of $M(l) \to G$ should be observed to identify the number of irreducible groups $G$, as shown in Figure 2. Nevertheless, this strategy still necessitates considerable computational costs because, in general, it is laborious to identify whether a monotonic behavior has achieved convergence or still requiring a sufficient number of $l$. The proposed method provides a powerful breakthrough, offering another way to identify $G$ as a position of the peak. We derived that $V[M(l)]$ (the variance of $M(l)$) with respect to the random selection of $l$ members of reducible samples becomes maximum when $l \sim G$.

2. Outline of the Results

To formulate the variance of $M(l)$, we shall introduce a probability for a set of $l$ samples to be sorted into $M$ irreducible classes, denoted as $P(l, M, G)$. The expectation value, $\bar{M}$, and its variance $V[M(l)]$ is then given as

$$\bar{M}(l) = \sum_{M=1}^{G} M \cdot P(l, M; G)$$  \hspace{1cm} (1)

$$V[M(l)] = \sum_{M=1}^{G} (M - \bar{M}(l))^2 \cdot P(l, M; G)$$  \hspace{1cm} (2)

For $l = 1$, the sample surely corresponds to an irreducible class, namely $\bar{M}(0) = 1$, regardless of the sample selection; therefore, the following result is achieved: $V[M(1)] = 0$. In the limit when $l$ is large, we know that the convergence $M(l \to L) = G$; hence, $V[M(l \to L)] \to 0$. As such, we expect $V[M(l)]$ to have a maximum peak as a function of $l$ in the range of $1 < l < L$; we define such a $l$ value as $l_0$:

$$l_0(G) = \arg \max_{0 < l \leq L} V[M(l; G)]$$  \hspace{1cm} (3)
We can observe that $V[M(l)]$ has a peak at around $l \sim G$, shown as follows.

With the simplified assumption that the size of each irreducible class (i.e., the number of elements inside each class) is identical, we can derive an asymptotic behavior,

$$P(l, M; G) \sim \left(1 - C \exp \left[- \left(\frac{G - M + 1}{G} \right)(l - M)\right]\right) \times \exp \left[- \left(\frac{G - M}{G} \right)(l - M)\right]$$

$$C = \left(1 - \frac{cC_M \cdot M!}{G^M}\right)^{l}$$

as derived in the Appendix, §A~§B, and shown in Figure B1. By substituting Equation (4) into Equation (2), we can show that the variance $V[M(l)]$ has a peak approximately at $l \sim (G + 1)$ (§C). Even without such an assumption used in the analytical treatments, The numerical verifications have shown that this result remains unchanged, as shown in Figure 3.

We can obtain a rough estimation that $V[M(l)]$ has a peak at approximately $l \sim G$ from the schematic picture shown in Figure 4. For tiny sizes of samplings, $l \approx 1, 2, \ldots$, we expect $M(l) = l$ with little duplications, specifying an increase in the behavior at smaller $l$. The convergence, $M(l) \rightarrow G$, at larger $l$, leads to the schematic behavior shown in Figure 4. Through the vertical red arrows on the dependence $M(l)$, we schematically show the widths of the variances around it. Since the maximum possible value for $M(l)$ is $G$, the tail of the variance cannot undergo

$$M(l) = G.$$ 

Thus, the widest possible tail would occur at the intersection point $P$, shown in the figure and determined as $l = G$.

### 3. Discussion

Although the result is still within the extent of a heuristic finding, we note that $10 \times V[M(l)]$ provides a reasonably reliable estimate for $G$. As shown in Figure 5, the estimate appears rather firm compared to another estimator, $l_{0} \sim G$,” shown in Figure 3.

### 4. Conclusion

We have formulated a stochastic framework to classify vast numbers of samples ($L \approx$ trillion) using any clustering technique to find out the total number of irreducible classes $G$. The situation we considered is when $L$ is extremely large to be accommodated into the capacity of the input size of the technique due to the limitation of the memory- or file-size. Among $L$, the framework randomly picked up a set of samples with the size $l$ ($< L$) to be affordable by the capacity and sorted by a clustering tool into $M$ classes. By repeating the sampling varying the size $l$, we obtained the statistics $\{M(l)\}$. We have derived that the variance of $M(l)$ becomes maximum at $l \sim G$. Although no rigorous mathematical proof was developed, we heuristically obtained that $G$ could be estimated as the quantity approximately ten times of $V[M(l)]$.

In our future study, we will focus on developing a mathematical verification of the proposed method.
Appendix A: Derivation of the Probability

We denote the probability to obtain $M$ irreducible classes after a clustering technique has processed $l$ reducible samples as $P(l, M; G)$, where $l$ samples are randomly chosen from a population with a total of $G$ irreducible structures. The situation can be equivalently described as the “distributing color envelopes into $l$ pigeonholes with one each, where an envelope has $G$-color variations, but only $M$ ($< G$) colors are found over the pigeonholes.” Although the multiplicity for each color (i.e., the number of envelopes with each color) is generally different to each other, we consider the assumption that the multiplicity is the same for each color. We can subsequently evaluate $P(l, M; G)$, in a straightforward way, analytically, even though we shall examine later whether or not the final result is not substantially affected by the considered assumption. Letting the number of cases be $a(l, M; G)$, the probability can be written as

$$P(l, M; G) = a(l, M; G) / \sum_{i=1}^{G} a(l, i; G)$$  \hspace{1cm} (A1)

It can be denoted that

$$a(l, M; G) = c_{CM} \times F$$  \hspace{1cm} (A2)

where $c_{CM} := \frac{G!}{M!(G-M)!}$ and $F \sim M^l$ corresponds to the most rough estimation. The rough estimation still includes the cases where the total colors over the pigeonholes becomes less than $M - 1$. We have to exclude the cases as,

$$F = M^l - [\text{total } (M-1) \text{ colors}] - [\text{total } (M-2) \text{ colors}] - \cdots - [\text{total one color}]$$  \hspace{1cm} (A3)

such that the total number should be $M$. Noticing that $[\text{total } (M-k) \text{ colors}]$ under the factor $C_M$ simple corresponds to $a(l, M-k; M)$ (i.e., even providing $M$ alternatives, $l$ pigeonholes only equal to $(M-k)$ colors in total), we obtain that

$$F = M^l - \sum_{i=1}^{M-1} a(l, i; M)$$  \hspace{1cm} (A4)

leading a recurrence,

$$a(l, M; G) = c_{CM} \cdot \left[ M^l - \sum_{i=1}^{M-1} a(l, i; M) \right]$$  \hspace{1cm} (A5)

We shall further analyze the recurrence relation (A5) by asymptotic analysis as follows. To verify the asymptotic evaluations, we can consider again the relation (A5) recursively generating a series $a(l, M; G)$ by increasing $M = 1, 2, \ldots$ in a straight way.

The denominator of Equation (A1) is evaluated as,

$$\sum_{i=1}^{G} a(l, i; G) = G^l$$  \hspace{1cm} (A6)

because each term under the summation corresponds to the case with “just equals to $i$ colors,” summing them over $i = 1 \sim G$ that amounts to the total cases when assigning $G$ colors into $l$ pigeonholes and allowing any duplications $(\sim G^l)$. Under this assumption, $P(l, M; G)$ is, therefore, evaluated as

$$P(l, M; G) = \frac{c_{CM} \cdot \left[ M^l - \sum_{i=1}^{M-1} a(l, i; M) \right]}{G^l}$$  \hspace{1cm} (A7)

Appendix B: Asymptotic Analysis

We decompose the probability $P(l, M; G)$ (“just $M$ colors”) into

$$P(l, M; G) = A(l, M; G) \cdot D(l, M; G)$$  \hspace{1cm} (B1)

where $A(l, M; G)$ is the probability to obtain “more than $M$ colors ($\geq M$),” while $D(l, M; G)$ denotes the rate to obtain $M$ over the possible range of $(M \sim G)$ for $A$. $D(l, M; G)$ can be easily described as

$$D(l, M; G) = \frac{a(l, M; G)}{\sum_{i=M}^{G} a(l, i; G)}$$  \hspace{1cm} (B2)
Considering the complementary of \( A(l, M; G) \) \((\geq M)\), the \( Q = (1 - A) \) corresponds to the case with \( \{ \leq (M - 1) \} \). If we define

\[
Q(l, M; G) = \frac{\sum_{i=1}^{M} a(l, i; G)}{G^l}
\]

(B3)

the complementary relation is expressed as

\[
A(l, M; G) = 1 - Q(l, M - 1; G)
\]

(B4)

where we intentionally shift the definition of \( M \) for \( Q \) such that the upper limit of the summation in Equation (B3) is \( M \).

Considering \( Q(l + 1, M; G) \) defined in Equation (B3), we can obtain

\[
\frac{\sum_{i=1}^{M} a(l + 1, i; G)}{G^{l+1}} = \frac{G \sum_{i=1}^{M} a(l, i; G) - (G - M) a(l, M; G)}{G^{l+1}}
\]

(B5)

The above equation can be explained as follows: Imagine that the already existing \( l \) pigeonholes include a total of \( M \) colors that corresponds to the first summation appearing in the nominator in the right-hand side. Multiplying \( G \) to the summation implies counting all the possibilities for \( (l + 1) \)-pigeonhole to have any color. That includes the cases with a total of \( (M + 1) \) colors that should be excluded because we are considering the cases with “total \( M \) colors” owing to the argument of \( M \) in \( Q(l + 1, M; G) \). The part to be excluded is the product of \( a(l, M; G) \) (the already existing pigeonholes have \( M \) colors) times \( (G - M) \) \((l + 1)\)-pigeonhole obtains a new color, which is not found in the already existing pigeonholes). Considering the above explanations, we find the nominator of Equation (B5).

From the first term in the nominator of Equation (B5), we rearrange \( Q(l, M; G) \) to obtain,

\[
Q(l + 1, M, G) = Q(l, M; G) - \left( \frac{G - M}{G} \right) a(l, M; G)
\]

(B6)

as a recursion of \( Q(l, M; G) \). Assuming the form,

\[
\left( \frac{G - M}{G} \right) a(l, M; G) = a(l, M; G) \cdot Q(l, M; G)
\]

(B7)

and that \( l \) is treated as a continuous variable, we obtain the differential equation,

\[
\frac{dQ(l, M; G)}{dl} = -a(l, M; G) Q(l, M; G)
\]

\[
\therefore \quad Q(l, M; G) = C \exp \left[ - \int l a(l', M, G) \, dl' \right]
\]

(B8)

From Equation (B7), the integrant \( a \) is asymptotically evaluated as

\[
a(l, M; G) = \left( \frac{G - M}{G} \right) \frac{a(l, M; G)}{\sum_{i=1}^{M} a(l, i; G)} \sim \left( \frac{G - M}{G} \right)
\]

(B9)

because the summation in the denominator is almost dominated by the leading term \( a(l, M; G) \sim M \). This leads to an asymptotic form of \( A(l, M; G) \) as

\[
A(l, M; G) = 1 - C \exp \left[ - \left( \frac{G - (M - 1)}{G} \right) \right]
\]

(B10)

Similarly, we shall derive an asymptotic form of \( D(l, M; G) \) using differential approximations, as given in Equation (B18) by considering \( D(l + 1, M; G) \). We remind that the nominator and the denominator in Equation (B2) are \[Denominator: \{Number of cases with \( (\geq M) \) colors\} and \[Nominator: \{Number of cases with \( (= M) \) color\} \], where we are considering the total number of colors \( M \) or whatever included over pigeonholes. Subsequently, we consider a newly added pigeonhole \((l + 1)\)-th on the already existing ones \((l \) pigeonholes). We have two possibilities for the new pigeonhole, “(a) a new one does not get a new color other than those in the already existing pigeonholes” or “(b) a new one gets a new color.” For the already existing pigeonholes, we sort the cases into three as “(a) Pigeonholes that are getting \( [1 \sim (M - 2)] \) colors,” “(b) those getting \( (M - 1) \) colors,” and “(c) those getting \( M \) colors.” As we are considering the cases with \( \{\geq M\} \) only, the case (a) can be excluded from the present consideration (because this case provides at most \( (M - 1) \) colors when \((a)^{\ast}(y)\) occurs). For (b) and (c), we finally arrive at the conclusion that the contributions to \( D(l + 1, M; G) \) are made as “Denominator; \(([c] + [b]^{\ast}(y)]\)” and “Nominator; \(([b]^{\ast}(y) + [c]^{\ast}(x)]\).” as explained below. For the denominator \( \{\geq M\} \), (c) contributes to this, regardless of \( [c]^{\ast}(x) \) \(= M\) or \( [c]^{\ast}(y) \) \(= (M + 1)\). Thus, the corresponding “number of cases” is

\[
\text{Denominator}[c] = G \cdot \sum_{i=M}^{G} a(l, i; G)
\]

(B11)

where \( G \) is the number of the color-choice for \((l + 1)\)-th pigeonhole. For \((b)\), only \((b)^{\ast}(y)\) contributes to \( M \) colors, as

\[
\text{Denominator}[b]^{\ast}(y) = [G - (M - 1)] \cdot a(l, M - 1; G)
\]

(B12)

where \( [G - (M - 1)] \) is the possible choice for \((l + 1)\)-th one having a new color not appearing in \( l \) pigeonholes. The total contributions for the denominator is equal to

\[
\text{Denominator}[c] + [b]^{\ast}(y) \]

\[
= G \sum_{l=M}^{G} a(l, i; G) + [G - (M - 1)] \cdot a(l, M - 1; G)
\]

(B13)

In a similar way, counting for the nominator (num. of colors \(= M\)) provides the result of the contributions, “Nominator([b]^{\ast}(y) + [c]^{\ast}(x))”:

\[
\text{Nominator}([b]^{\ast}(y) + [c]^{\ast}(x)) = [G - (M - 1)] \cdot a(l, M - 1; G) + M \cdot a(l, M; G)
\]

(B14)
By technically separating the second term in the nominator as \( M = G - (G - M) \), we obtain

\[
\frac{a(l + 1, M; G)}{\sum_{i=M}^{G} a(l + 1, i; G)} = \frac{G a(l, M; G)}{G \sum_{i=M}^{G} a(l, i; G) + \{G - (M - 1)\} a(l, M - 1; G) + \{M - (M - 1)\} a(l, M - 1; G)}
\]

\[
= \frac{a(l, M; G)}{G \sum_{i=M}^{G} a(l, i; G) + \{G - (M - 1)\} a(l, M - 1; G) + \{M - (M - 1)\} a(l, M - 1; G)}
\]

\[
\sum_{i=M}^{G} a(l, i; G) + \{G - (M - 1)\} a(l, M - 1; G) + \{M - (M - 1)\} a(l, M - 1; G)
\]

Because \( a(l, M; G) \sim M \), we can neglect terms with \( a(l, M - 1; G) \) to obtain

\[
\frac{a(l + 1, M; G)}{\sum_{i=M}^{G} a(l + 1, i; G)} \approx \frac{a(l, M; G)}{G \sum_{i=M}^{G} a(l, i; G) + \{G - (M - 1)\} a(l, M - 1; G) + \{M - (M - 1)\} a(l, M - 1; G)}
\]

\[
\sum_{i=M}^{G} a(l, i; G) + \{G - (M - 1)\} a(l, M - 1; G) + \{M - (M - 1)\} a(l, M - 1; G)
\]

leading to

\[
D(l + 1, M; G) = D(l, M; G) - \left(\frac{G - M}{G}\right) D(l, M; G)
\]

\[
\text{Regarding } l \text{ as continuous, we find}
\]

\[
\frac{dD(l, M; G)}{dl} = -\left(\frac{G - M}{G}\right) D(l, M; G)
\]

where the constant of integration is determined such that \( D(l, M = 1; G) = 1 \) can be satisfied.

Substituting Equations (B10) and (B18) into Equation (B1), we obtain

\[
P(l, M; G) \sim \left\{1 - C \exp \left[-\left(\frac{G - (M - 1)}{G}\right)\right]\right\}
\]

\[
\times \exp \left[-\left(\frac{G - M}{G}\right) (l - M)\right]
\]

This corresponds to the counting of possible permutations of \( M \) selected from \( G \) colors. Thus,

\[
C = \left(1 - \frac{cG \cdot M!}{G^M}\right)
\]

As the asymptotic evaluations affect the original normalization for \( P \), we again introduce a normalization factor,

\[
Z(l, G) = \sum_{M=1}^{G} P(l, M; G)
\]

by considering that

\[
P(l, M; G) = \frac{1}{Z[l, G]} \left\{1 - C \exp \left[-\left(\frac{G - M + 1}{G}\right) (l - M)\right]\right\}
\]

\[
\times \exp \left[-\left(\frac{G - M}{G}\right) (l - M)\right]
\]

becomes the probability to be applied to Equation (2) in the asymptotic region. As shown in Figure B1, the asymptotic evaluation performs fairly well to reproduce the behaviors of the original form. Comparisons between the asymptotic behavior (Approx.) and the original one (Equation (A1)) of \( P(l, M; G) \) with \( G = 5 \), as the functions of \( l \) with several different \( M \) is shown in Figure B1.

**Appendix C: Identifying \( l = l_0 \) where the Variance Becomes Maximum**

By using Equation (B24) to evaluate the variance, Equation (2), we can obtain

\[
\frac{d}{dl} V[l] = \frac{1}{Z} \left[Z \sum_{M=1}^{G} M^2 \left(\frac{\partial P}{\partial l}\right) - w \sum_{M=1}^{G} M P\right]
\]

\[
-2M \left[Z \sum_{M=1}^{G} M \left(\frac{\partial P}{\partial l}\right) - w \sum_{M=1}^{G} M P\right]
\]

\[
\text{as given in Equation (4) in the main text, only for } M \leq l, \text{ otherwise}\]

\[
P(l, M; G) = 0.\text{ For } l = M, \text{ the equation is reduced to}
\]

\[
P(l, M; G)_{l=M} = C \frac{G \cdot M!}{G^M}
\]

\[
2000301 \ (6 \ of \ 8) \ 
\text{Adv. Theory Simul.} \ 2021, \ 4, \ 2000301 \ 
\text{© 2021 The Authors. Advanced Theory and Simulations published by Wiley-VCH GmbH}
\]
where we define
\[ \hat{M} = \sum_{M=1}^{G} MP' = \frac{1}{Z} \sum_{M=1}^{G} MP(l, M; G) \quad \text{(C2)} \]

\[ w = \sum_{M=1}^{G} \left( \frac{\partial P}{\partial l} \right) \quad \text{(C3)} \]

and the approximation \( C \approx 1 \) is considered. As shown in \( \S D \), we can observe that \( P \) and its derivative, under the summations in each term of Equation (C1), exhibit the steep peaks as \( \sim \exp(M - M_0)^2 \). Thus, we can evaluate them as

\[ F_2 = \sum_{M=1}^{G} F(M) \cdot \frac{\partial P}{\partial l} \sim F(M_0^{(3)}) \cdot \frac{\partial P}{\partial l} = M_6 \quad \text{(C4)} \]

\[ F_1 = \sum_{M=1}^{G} F(M) \cdot P(l, M; G) \sim F(M_0^{(3)}) \cdot P(l, M; G) \quad \text{(C5)} \]

\[ Z = \sum_{M=1}^{G} P(l, M; G) \sim P(l, M; G_0^{(3)}) \quad \text{(C6)} \]

For \( F_2 \), there are two peaks \( (M_0 = G \text{ and } l) \), as shown in \( \S D \). That is,

\[ \sum_{M=1}^{G} F(M) \left( \frac{\partial P}{\partial l} \right) = \frac{1}{G} \left[ F(G) + (G - l + 1)F(l) \right] \quad \text{(C7)} \]

Using an asymptotic evaluation of \( \partial P/\partial l \) given in Equation (D2), we can obtain

\[ \sum_{M=1}^{G} F(M) \left( \frac{\partial P}{\partial l} \right) = \frac{1}{G} \left[ F(G) + (G - l + 1)F(l) \right] \quad \text{(C8)} \]

For \( F_1 \) and \( Z \), the derivations provided in \( \S D \) lead to

\[ \sum_{M=1}^{G} F(M)P(l, M; G) \sim F(l - 1) \cdot P(l, M; G)_{M=\{l-1\}} \quad \text{(C9)} \]

By defining
\[ P(l, M; G)_{M=\{l-1\}} \approx \left( 1 - e^{-\frac{|l-M|}{c}} \right) e^{-\frac{|l-M|}{c}} =: g(l) \quad \text{(C10)} \]

each term in Equation (C1) is evaluated as

\[ Z \sim g(l), \quad \hat{M} \sim (l - 1), \quad w \sim \frac{G - l + 2}{G} \]

\[ \sum_{M=1}^{G} MP \sim (l - 1)g(l), \quad \sum_{M=1}^{G} M^2 P \sim (l - 1)^2g(l) \]

\[ \sum_{M=1}^{G} M \left( \frac{\partial P}{\partial l} \right) \sim \frac{1}{G} \left[ G + (G - l + 1)l \right] \quad \text{(C11)} \]

\[ \sum_{M=1}^{G} M^2 \left( \frac{\partial P}{\partial l} \right) \sim \frac{1}{G} \left[ G^2 + (G - l + 1)l^2 \right] \]

respectively. Substituting them into Equation (C1) leads to

\[ \frac{\partial V}{\partial l} = \frac{1}{g(l)} \left[ \frac{G^2 + (G - l + 1)l^2}{G} - \frac{(G - l + 2)}{G} (l - 1)^2 \right. \]

\[ -2(l - 1) \left\{ \frac{G + (G - l + 1)l}{G} - \frac{(G - l + 2)}{G} (l - 1) \right\} \]

\[ = \frac{1}{Gg} \left\{ (l - (G + 1))(l - (G + 2)) \right\} \quad \text{(C12)} \]

To find \( l \) values that satisfy \( dV/dl = 0 \), we set \( x = (G - l + 1)/G \). Equation (C12) can be written as

\[ \frac{\partial V}{\partial l} = \frac{x(Gx + 1)}{(1 - e^{-x})e^x} \quad \text{(C13)} \]

Obviously, Equation (C12) is 0 at \( x = -1/G \), namely \( l = G + 2 \).

To consider the behavior of Equation (C12) as \( x \rightarrow 0 \), we approximate \( e^{-x} \) as \( 1 - x \).

\[ \frac{\partial V}{\partial l} \sim \frac{x(Gx + 1)}{x(1 - x)} = \frac{Gx + 1}{1 - x} \quad \text{(C14)} \]

Therefore, Equation (C12) is 1 at \( x = 0 \). We conclude that the variance has the maximum at \( l = (G + 2) \).

**Appendix D: Peaks of \( P \) and Its Derivative**

Considering the derivative of Equation (4), the approximation, \( G - M + 1 \sim (G - M) \), leads to

\[ \frac{\partial l}{\partial l} P(l, M; G) \sim \frac{1}{G} e^{\frac{-(G-M)l-M}{c}} \left[ 2(G - M) + 1 \right] e^{\frac{-(G-M)l-M}{c}} - (G - M) \quad \text{(D1)} \]

The factor \( e^{\frac{-(G-M)l-M}{c}} \sim e^{-M^2} \) achieves its maximum value (1) at the peak; otherwise, it vanishes steeply. The value is realized when \( M = G \) or \( M = l \), along with the value of the derivative,

\[ \frac{\partial l}{\partial l} P(l, M; G) \sim \frac{1}{G} [(G - M) + 1] \quad \text{(D2)} \]

to be picked up from the summation as the product with the \( F(M) \) in Equation (C7).

With \( G - M + 1 \sim (G - M) \) applied to Equation (4), we can obtain

\[ P(l, M; G) \sim \left( 1 - e^{\frac{-(G-M)l-M}{c}} \right) e^{\frac{-(G-M)l-M}{c}} \quad \text{(D3)} \]
Considering $\partial P/\partial M = 0$ to achieve the condition for the peak, we obtain

$$(G - M)(l - M) = G \cdot \ln 2.$$  

$\therefore M^2 - (l + G)M + (G + 1)(l - 1) = 0$  \hspace{1cm} (D4)

To simplify the factorization, we perform further approximations on the term $G(l - \ln 2)$ as $G \sim (G + 1)$ and $\ln 2 \sim 1$, thus, obtaining

$M^2 - (l + G)M + (G + 1)(l - 1) = 0$  \hspace{1cm} (D5)

which can be easily factorized as

$[M - (G + 1)][M - (l - 1)] = 0$  \hspace{1cm} (D6)

This equation provides the conclusion that the peak occurs at $M \approx (l - 1)$, leading to Equation (C9) (another root, $M = G + 1$, does not match with the setting of the problem, $M < G$).

## Acknowledgements

The computations in this work were performed using the facilities of Research Center for Advanced Computing Infrastructure (RCACI) at JAIST. K.U. is grateful for the financial support from the JAIST Research Grant (Fundamental Research) 2019. K.H. is grateful for the financial supports from FLAGSHIP2020 (project numbers hp190169 and hp190167 at K-computer), KAKENHI grant (grant numbers 17K17762 and 19K05029), Grant-in-Aid for Scientific Research on Innovative Areas (16H06439 and 19H05169), PRESTO (JPMJPR16NA), and the “Materials research by Information Integration Initiative” (M2I) project, funded by the Support Program for Starting Up Innovation Hub from Japan Science and Technology Agency (JST). R.M. is grateful for the financial supports from MEXT-KAKENHI (grant numbers 19H04692 and 16KK0097), FLAGSHIP2020 (project numbers hp190169 and hp190167 at K-computer), Toyota Motor Corporation, I-O DATA Foundation, Air Force Office of Scientific Research (AFOSR-AOARD/FA2386-17-1-4049 and FA2386-19-1-4013), and JSPS Bilateral Joint Projects (with India DST).

## Conflict of Interest

The authors declare no conflict of interest.

## Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## Keywords

classification algorithms, clustering techniques, irreducible classes, machine learning, stochastic frameworks

Received: November 24, 2020
Revised: January 24, 2021
Published online: March 26, 2021

1. A. G. Kusne, T. Gao, A. Mehta, L. Ke, M. C. Nguyen, K.-M. Ho, V. Antropov, C.-Z. Wang, M. J. Kramer, C. Long, I. Takeuchi, Sci. Rep. 2014, 4, 6367.
2. S. K. Suram, Y. Xue, J. Bai, R. Le Bras, B. Rappazzo, R. Bernstein, J. Bjorck, L. Zhou, R. B. van Dover, C. P. Gomes, J. M. Gregoire, ACS Comb. Sci. 2017, 19, 37.
3. Y. Iwasaki, A. G. Kusne, I. Takeuchi, npj Comput. Mater. 2017, 3, 4.
4. V. Stanev, V. V. Vesselinov, A. G. Kusne, G. Antoszewski, I. Takeuchi, B. S. Alexandrov, npj Comput. Mater. 2018, 4, 43.
5. K. Utimula, R. Hunkao, M. Yano, H. Kimoto, K. Hongo, S. Kawaguchi, S. Suwanna, R. Maezono, Adv. Theory Simul. 2020, 3, 2000039.
6. P. Paufler, Acta Crystallogr. A 2007, 63, 483.
7. H. T. Stokes, D. M. Hatch, J. Appl. Crystallogr. 2005, 38, 237.
8. Dassault Systèmes BIOVIA, Materials Studio, Version 18.1.0.2017, Dassault Systèmes, San Diego 2018.