Effective search for stable segregation configurations at grain boundaries with data-mining techniques

Shin Kiyohara*, Teruyasu Mizoguchi

Institute of Industrial Science, University of Tokyo, 153-8505 Tokyo, Japan

ARTICLE INFO

ABSTRACT

Grain boundary segregation of dopants plays a crucial role in materials properties. To investigate the dopant segregation behavior at the grain boundary, an enormous number of combinations have to be considered in the segregation of multiple dopants at the complex grain boundary structures. Here, two data mining techniques, the random-forests regression and the genetic algorithm, were applied to determine stable segregation sites at grain boundaries efficiently. Using the random-forests method, a predictive model was constructed from 2% of the segregation configurations and it has been shown that this model could determine the stable segregation configurations. Furthermore, the genetic algorithm also successfully determined the most stable segregation configuration with great efficiency. We demonstrate that these approaches are quite effective to investigate the dopant segregation behaviors at grain boundaries.

1. Introduction

A grain boundary is formed by bonding two crystals with crystallographically different orientations and hence usually has different atomic configurations from those in the bulk. With such atomic configurations, different phenomena from the bulk are often observed [1–12]. For instance, preferential solute solution at the grain boundary, referred to as solute segregation, is often observed and is known to influence the macroscopic properties of materials. Hence, dedicated research for the past several decades has elucidated the effects of solute segregation at grain boundaries on the materials properties [1–5]. For instance, Duscher et al. unraveled the mechanism underlying embrittlement at bismuth-induced copper grain boundaries with theoretical calculations and results from electron microscopy [6]. Guang-Hong Lu et al. examined the grain boundary segregation at aluminum grain boundaries and concluded that a variety of solute atoms largely changes the manner of embrittlement [7]. Aside from mechanical properties, it is also reported that the segregation of multiple elements at grain boundaries influences electrical properties [13–15], ionic conductivity [16,17], and other functional properties [18–20].

To further understand the fundamental mechanisms underpinning grain boundary segregation and its impact on the materials properties, a large number of combinations of solutes and grain boundaries have to be investigated systematically. However, the determination of the segregation configurations at the grain boundary always encounters great difficulties because many atomic sites are present and therefore a large number of combinations of segregation configurations have to be considered. To determine segregation configurations, molecular dynamic and Monte Carlo simulations are often been used [21–23]. Although such simulations do determine stable configurations, a large number of calculation steps, which cannot be executed in parallel, are necessary. Hence, more efficient approaches are required for a systematic investigation of the grain boundary segregation.

In this study, aided by information science, two efficient approaches were applied for this purpose. The first is a machine learning method based on the random-forests algorithm, which is an ensemble learning method for classification and regression [24]. By learning the relationships between the atomistic configuration and segregation energy, the regression model was constructed and then used to predict the stable segregation configurations. The genetic algorithm is similar to a Monte Carlo approach but searches for a stable configuration more efficiently by performing mutation, crossover, and selection operations [25]. Although those methods have been effectively used in materials science, their application to grain boundary segregation has not been exploited to date.

2. Methodology

The two statistical method, random-forests regression and genetic algorithm, were applied to search for the stable segregation configurations. To establish this methodology, an efficient calculation method is suitable. Thus, static lattice calculations using an empirical potential,

* Corresponding author.
E-mail addresses: sin@iis.u-tokyo.ac.jp (S. Kiyohara), teru@iis.u-tokyo.ac.jp (T. Mizoguchi).
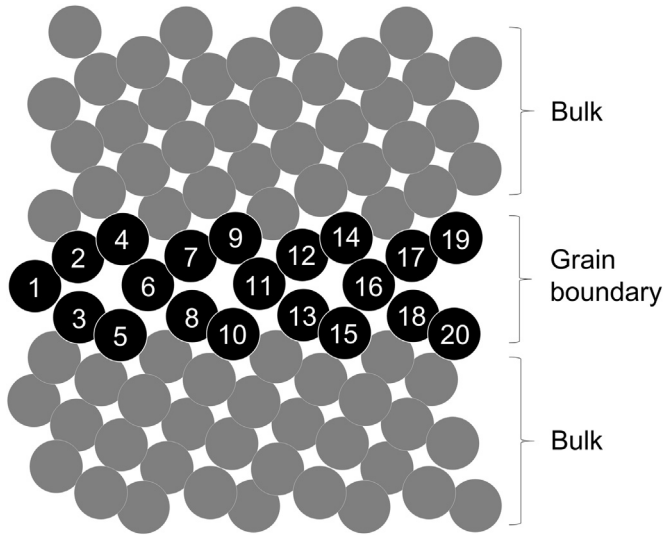
S. Kiyohara, T. Mizoguchi

Fig. 1. Schematic of the supercells including a Σ5[001]/(210) grain boundary of copper. Gray and black filled circles are copper atoms in bulk and grain-boundary regions, respectively. Twenty sites are numbered within the grain-boundary region.

with GULP code [26], was used for structure optimization and energy calculations. Dopant segregation at Cu grain boundaries was investigated because it is well known to influence material properties [6,27]. In our previous work, an empirical potential was established that reproduced the results calculated by the first principles calculation [11].

The segregation behavior of silver atoms at the Σ5[001]/(210) grain boundary of copper was selected as a testing ground for our approach. The supercells, including this grain boundary (Fig. 1), have 20 atomic sites at the boundary. Segregation of up to seven silver atoms were considered because it is known that this grain boundary has four stable atomic site for silver segregation, 1, 6, 11, and 16, from the previous study [11], and hence seven can be considered to be sufficient. In the result, all atomistic configurations are 137,979 as the following equation:

$$N = \sum_{i=1}^{7} C_i^{20}, \tag{1}$$

where N is the number of all combination of the atomic configurations, and i means the segregated silver atoms. In the brute-force method, the energy calculation and structure optimization was performed on these all 137,979 possible atomistic configurations.

The segregation energy, E_{seg}, is defined as

$$E_{seg} = \frac{(E_{n\ seg\ in\ GB} - E_{GB}) - (E_{n\ seg\ in\ Bulk} - E_{Bulk})}{n}, \tag{2}$$

where n is the number of silver atoms; $E_{n\ seg\ in\ GB}$ and $E_{n\ seg\ in\ Bulk}$ represent the total energy of the supercells for the grain boundary and bulk model with n-site segregation (substitution) respectively, and E_{GB} and E_{Bulk} do the total energy of the supercells for the grain boundary and bulk and without segregation (substitution), respectively. Since the segregation energy is estimated on the relative energy of the solution energy between the bulk and GB, the energy of the isolated Ag was canceled out. For the calculation of structure relaxation, the empirical potential was used in the embedded-atom method as an embedding function of type one [11].

Each atomic configuration is described as binary data of twenty-bit length, where 0 and 1 denote copper and silver, respectively. For instance, the binary description of (0,0,1,0,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0,0) means the atomic configuration with silver replaced at sites 3, 6, 12, and 13 in Fig. 1. Details of the random-forests regression and genetic algorithm are given below.
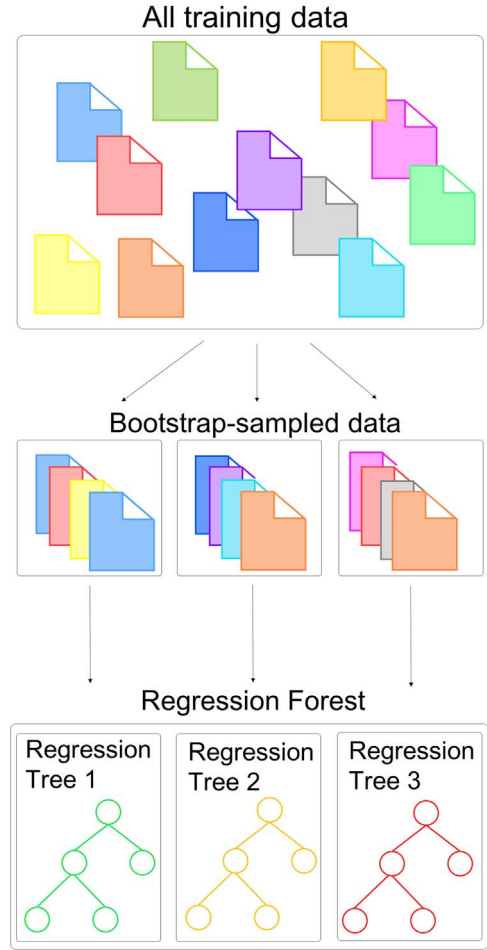

Fig. 2. Schematic diagram of the random forests regression.

2.1. Random-forests regression

The random-forests regression [24] is a kind of ensemble learning based on a regression tree and bagging. Fig. 2 is the schematic diagram of the random-forests regression. A random-forest consists of plural regression trees constructed from the bootstrap-sampled data, which is obtained by the random sampling from all training data allowing duplication. In the random-forests method, predicted values are the averages of the output from each regression tree. Random forests have two parameters in this study, the number of regression trees and the maximum depth of each tree. Larger and deeper forests lead to the prediction accuracy, but overfitting can occur. Furthermore, we changed the size of training data to 0.1%, 0.5%, 1.0%, and 2.0% of all 137,979 data, which equals the number of the atomic configurations. The scikit-learn package [28] for Python was used to perform the random-forests regression analysis.

2.2. Genetic algorithm

A genetic algorithm is a stochastic search technique, which imitates the procedure of evolution. In recent years, the genetic algorithm has been used for some time-consuming problems in materials science [29–31]. The strategy of this algorithm for segregation behavior is shown in Fig. 3. In Fig. 3(a), the first population is generated by randomly constructing atomic configurations with different silver segregations, which are hereafter called "individuals". The segregation energy is calculated for each individual. The populations have 10, 20, 50, 100, and 200 individuals. The operations of selection, crossover, and mutation are then implemented in succession. Selection, illu-