Chemical Physics Letters 728 (2019) 109–114

Contents lists available at ScienceDirect

Chemical Physics Letters

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

Research paper

Data analysis of multi-dimensional thermophysical properties of liquid substances based on clustering approach of machine learning

Gota Kikugawa[a,*], Yuta Nishimura[b], Koji Shimoyama[a], Taku Ohara[a], Tomonaga Okabe[c,d], Fumio S. Ohuchi[d]

[a] Institute of Fluid Science, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan
[b] School of Engineering, Tohoku Univ., 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan
[c] Department of Aerospace Engineering, Tohoku University, 6-6-01, Aramaki Aza Aoba, Aoba-ku, Sendai, Miyagi, 980-8579, Japan
[d] Department of Materials Science and Engineering, University of Washington, BOX 352120, Seattle, WA 98195, USA

HIGHLIGHTS

- Self-organizing map is performed on thermophysical properties of liquid substances.
- Machine learning-assisted screening protocol of liquid substances is proposed.
- Liquid substances are automatically categorized based on various properties.
- Easy visualization technique of relative relationships among liquids is provided.

ARTICLE INFO

Keywords:
Self-organizing map
Clustering analysis
Machine learning
Thermophysical properties
Heat medium

ABSTRACT

In order to develop an efficient framework for global screening in the material exploration, we performed a clustering analysis of machine learning on the multi-dimensional thermophysical properties of the liquid substances. Data mining using a self-organizing map (SOM) based on the unsupervised learning was employed to project high-dimensional thermophysical data onto a low-dimensional space. Here we adopted 98 liquid substances with eight thermo-physical properties for the SOM training in order to group the liquid substances. The present SOM-clustering approach properly categorized liquid substances according to the chemical species characterized by the functional groups.

1. Introduction

Technology to design and synthesize liquid substances having appropriate thermophysical properties for specific applications is an important subject in a wide variety of scientific and technological fields. At the industrial level, efficient development and exploration of heat-transfer fluids or phase-change materials, which are used as a working fluid of heat exchangers in refrigeration and air-conditioning equipment or used for thermal storage, are highly required these days. In these areas, prior knowledge-based survey and/or conventional empirical laws for materials properties are utilized for the design guideline; however, when novel materials with superior properties are developed, a trial-and-error approach is needed since the required specification is so multi-objective, or the empirical laws is not always effective due to the outside of application limit. For example, if required thermophysical properties of the liquids has been realized by mixing multiple liquids, the number of mixing combinations and variation in the composition of each liquid get enormous, which makes the materials exploration almost impossible. Therefore, an alternative approach is highly needed for design and discovery of novel liquid substances in more efficient manner.

One promising approach is a data-driven machine learning (ML) technique, which is known as materials informatics (MI) [1] recently. Materials informatics has drawn considerable attentions in various material design partly because application of ML rapidly prevails to a wide spectrum of the scientific and technological field with the aid of advancement of computational performance for ML. So far, MI is intensively applied for accelerating search of inorganic materials like shape memory alloys [2–4] and piezoelectrics [5,6], and for predicting properties of perovskite crystal and finding one having optimal properties [7,8]. Other example is ML-assisted materials design and property prediction of solid materials such as elpasolite [9,10], metallic alloys

* Corresponding author.
E-mail address: kikugawa@tohoku.ac.jp (G. Kikugawa).

https://doi.org/10.1016/j.cplett.2019.04.075
Received 26 August 2018; Received in revised form 24 April 2019; Accepted 27 April 2019
Available online 29 April 2019
0009-2614/

[11–13], transition metal complex [14], layered double hydroxide [15], high dielectric permittivity materials [16], and thermoelectric materials [17]. As such, these studies have already shown effectiveness of materials informatics, i.e., a use of enormous data from experiments and numerical simulations, in particular, first-principle calculations, gives the accurate prediction of physical properties and development of improved materials.

Besides the inorganic materials, organic materials and organic-inorganic hybrid systems are getting targeted as an application of MI [18–25]. Discovery and design of polymeric materials, metal organic frameworks (MOFs), and organic/inorganic perovskite are good candidates for MI subject and actually require the ML-assisted screening or development. For example, polymer dielectrics were studied by ML with properties provided by DFT (density functional theory) calculations [18,19]. A Bayesian optimization technique was applied for proposing better experimental settings of polymer fiber synthesis [20]. A couple of unsupervised and supervised ML techniques were applied to develop the better prediction of CO_2 absorption properties into amine solutions [21]. For construction of a better quantitative structure-property relationship (QSPR), a large dataset of polymeric materials has been provided online [22,23]. ML-assisted materials development and property prediction for soft matters are also reviewed on Ref. [25].

Applications of MI protocols on the design and exploration of the liquid substances, on the other hand, have been very limited so far. As mentioned earlier, the efficient development and exploration of heat-transfer fluids like a coolant is, however, getting required significantly. When it comes to a coolant, the multi-objective design is needed since the various requirement should be fulfilled as an industrial product, some of which are basic transport properties like thermal conductivity and viscosity, non-flammability, and global warming potential (GWP). Therefore, it is not an easy task to find the improved substances and optimize the various properties at once.

In our study, we aim at building the overall platform for the efficient screening of materials candidates, which is constructed by a multi-stage screening protocol. This platform involves “global screening” which enables to roughly screen a lot of candidates by using data-mining approach based on unsupervised learning, and precise exploration which is realized by commonly adopted MI techniques based on the structure/properties prediction model and/or structure optimization by desired properties. The purpose of this study is to demonstrate that unsupervised approach for grouping materials is effective as the global screening. Our vision regarding the multi-stage screening is mostly motivated from the fact that large amount of dataset is not always available for liquid substances. Therefore, in order to build an accurate prediction model with relatively limited dataset, rough screening prior to precise ML-driven materials finding is effective.

In the present paper, we addressed the global screening stage for exploration of the liquid substances assisted by ML techniques. To this end, we propose to utilize one of the dimensionality reduction techniques called self-organizing map (SOM) to enable visual understanding of the diverse thermophysical properties from various liquid substances. Dimensionality reduction can be used when finding novel materials with meeting a lot of requirements, which is the usual case in the industrial process. SOM is one of the most appropriate techniques for this purpose since we can keep proximity information on the low dimensional mapping unlike the linear dimensionality reduction such as the principal component analysis (PCA). Therefore, SOM has been developed and progressively applied in a variety of engineering fields, such as optimal design of aircraft wings [26] and development of thermosetting polymer resin [27]. Here, we present a combined approach of SOM and clustering applied to a wide variety of the liquid substances, and a possibility of our framework for materials screening to enable us to design and explore better liquid substances more easily.

2. Data analysis

2.1. Self-Organizing Map (SOM)

SOM provides a data-mining technique which helps to understand high dimensional data, such as various thermophysical properties of the liquid species in the present case, by reducing their dimensions of the data to typically a two-dimensional space, so that the relative relationships among input data can be visualized intuitively [28,29]. In this technique, the data are nonlinearly reduced to a low-dimensional space with maintaining the neighborhood relationships among the input data. This data mining technique is now utilized in various fields to clarify relative relationships of complicated data.

In terms of machine learning, SOM is categorized into unsupervised learning, which is based on a neural network (NN). Therefore, it is interpreted as a feed-forward two-layer NN which is composed of an input layer and an output layer. The input layer corresponds to all dimensions of the all input data (thermophysical properties in this study) and all input-layer nodes are connected to all output-layer nodes. Furthermore, each output-layer node has a vector composed of the thermophysical properties. This is called a weighted vector with the same number of dimensions as that of input data, and is updated through a competitive learning process. Each node of the output layer holds the coordinate information at low dimensions (two dimensions in this study) concurrently, and the neighborhood relationships among nodes in the output layer are expressed by these coordinates, i.e., projection onto the low-dimensional map.

Detailed algorithm of the SOM is given in previous literatures, so we only make a brief explanation below. Let total numbers of the input-layer and output-layer nodes be I and J, respectively, and component vectors of the input-layer nodes and the weighted vectors of the output layer at the t-th learning cycle are expressed as $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_I]$ and $\mathbf{m}(t) = [\mathbf{m}_1(t), \mathbf{m}_2(t), ..., \mathbf{m}_J(t)]$, respectively. As the first step of a learning process, a weighted vector from all the output-layer nodes, which has the highest similarity to the input data, is chosen and that node is defined as a winner node, $\mathbf{m}_c(t)$, as follows.

$$\|\mathbf{x}_k - \mathbf{m}_c(t)\| = \min_j \|\mathbf{x}_k - \mathbf{m}_j(t)\|. \quad (1)$$

The weighted vector on the winner node is updated in such a way that the vector gets closer to the input data; at the same time, the output-layer nodes near the winner node also become closer to the input data, depending on their proximity to the winner node. By repeating this learning process, similarities among the weighted vectors on the output-layer nodes near the winner node increase. This update process is repeated until the weighted vectors converge to obtain the final SOM. Liquid species with similar thermophysical properties are finally distributed closer on the SOM. The learning algorithm of SOM can be roughly categorized into two types: online-learning SOM and batch-learning SOM (BLSOM). This study adopted a batch type, in which the learning result is not affected by an input order of the learning data.

For the input data to be used in the present analysis, we have chosen 98 liquid substances which are categorized into alkane, alcohol, aromatic series, carboxylic acid, amine, ketone, ester, and halide. Under standard temperature and pressure conditions, these substances are in a liquid phase. Since the SOM learning does not allow missing data, i.e., SOM cannot be applied if a part of thermophysical properties is not given in databases for a certain liquid substance, only 98 liquids, for which thermophysical properties are fully available in the literatures without any missing data, were examined. In the analysis, the following eight thermophysical properties were chosen; they are mass density, specific heat at constant pressure, melting point, boiling point, saturated vapor pressure, surface tension, viscosity, and thermal

conductivity, each of which was taken from the handbook and database of liquid properties [30–35]. All the datasets are available as Supplementary Material. Note that these thermophysical properties do not have the same averaged value and variance over the tested 98 liquids, i.e., each property has a different data range. Therefore, they were normalized into those with an average of 0 and a variance of 1 before they were used as the input data for learning. All the data analysis codes were implemented using python language (python 3.5) with the SOMPY package [36].

2.2. Clustering approach

The output-layer nodes after the SOM learning have weighted vectors whose components are multi-dimensional thermophysical properties, and the nodes, which are mutually close in distance in a high dimensional space, are placed at mutually near locations in the two dimensional SOM. Furthermore, in order to easily visualize the nodes with similar thermophysical properties, a clustering technique using a k-means method [37] was performed. This unsupervised learning method is one of non-hierarchical clustering techniques, which is used to classify input data having multi-dimensional quantities into preliminarily defined K clusters. In this method, K clusters' gravity center, i.e., centroid, are first given so as to be as far apart as possible each other (the k-means++ method), and then all data points are engaged in a cluster associated with the centroid to which the Euclidian distance is closest. Next, the centroid is calculated from the data points belonging to each cluster again, and the data points belonging to each cluster are updated. This process is repeated until the positions of centroids and the data points belonging to each cluster converge. The overall procedure of SOM-clustering approach described through Sections 2.1 and 2.2 is summarized in Fig. 1.

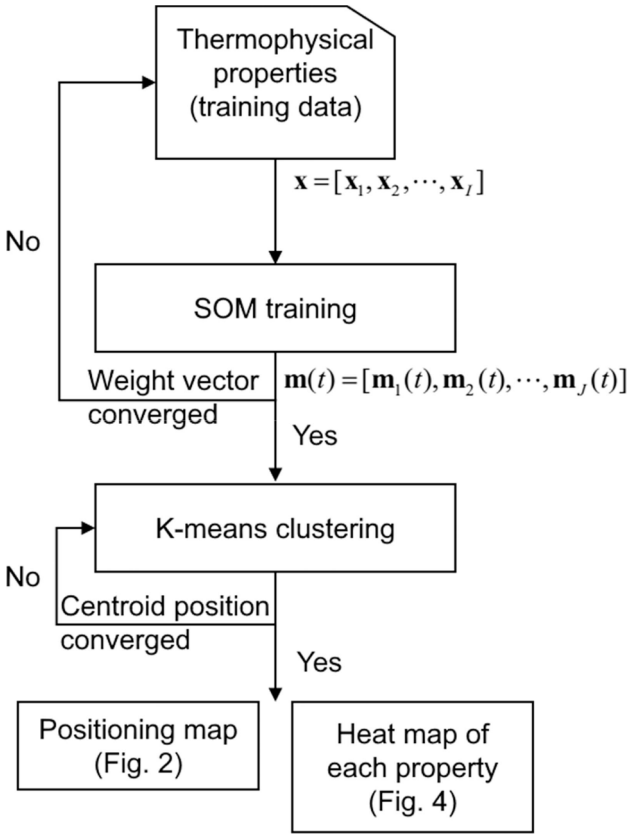

Fig. 1. Flowchart of overall SOM-clustering analysis.

2.3. U-Matrix method

A U-Matrix (unified distance matrix) method [38] is used to visualize the similarity between the adjacent nodes based on the distance information between nearest neighbor output-layer nodes. Here, the U-Matrix is given as

$$U_i = \frac{1}{N_i} \sum_{j}^{N_i} \|\mathbf{m}_i - \mathbf{m}_j\|, \tag{2}$$

where the output-layer node and the nearest neighbor output-layer node are denoted as i and j, respectively, and N_i is the number of nodes adjacent to node i. The U-Matrix gives local distance relationships between the nodes, and a large U-Matrix value would be regarded as a boundary of the clusters where separation between the nodes are large. In this study, the U-Matrix is used for examining "quality" of the clustering result by using the k-means method.

3. Results of SOM-clustering approach

The SOM learning was performed on aforementioned 98 liquid substances with eight thermophysical properties. A positioning map where each substance is assigned to the output-layer node that has the weight vector closest to each substance is presented using a 30 × 30 map in Fig. 2. The clustering result obtained using the k-means method as described in Section 2.2 is represented by the individual color designating each cluster. Originally, an attempt was made to determine the number of clusters through the Elbow method and Silhouette analysis [37]; the number adopted in the present study was 10, selected among those which have relatively good (not best) results of the Silhouette analysis (details not shown here). The liquid substances classified into clusters are outlined as follows.

(1) Water (yellow-green, lower right) and glycerol (red, middle right side) belong to independent clusters.
(2) Liquid species having small molecular weights and likely to form hydrogen bonding, such as ethylene glycol and ethanolamine (yellow, lower right)
(3) Alcohol and alkane liquids with relatively large molecular weights (blue, right)
(4) Alcohol and alkane liquids with relatively small molecular weights (light blue, bottom)
(5) Aromatic series and cycloalkane (purple, top center)
(6) Alkane smaller than 4), unsaturated alkane, and ketone (pink, lower left)
(7) Halogen compounds with relatively small molecular weights (orange, upper left)
(8) Halogen compounds with relatively large molecular weights (green, upper left)
(9) Heavy halogen compounds substituted for two bromine and/or iodine molecules (light purple, top center)

As a whole, the SOM indicates that the liquid substances characterized by each functional group are properly classified into groups. It is found that the liquid species with special thermophysical properties such as water or glycerol are automatically distinguished as an individual cluster and that halogen compounds are clearly separated as other clusters. In addition, cyclohexane and benzene are placed in mutually near positions in SOM, although their electronic structures are different with each other.

Fig. 3 depicts results from the U-Matrix analysis. Red and blue regions of the figure represent higher and lower values of the U-Matrix, respectively. Therefore, liquid substances in a blue valley separated by the red ridge have the larger deviation in thermophysical properties. The result indicates that some liquid substances, which are classified into the same cluster by the k-means method as shown in Fig. 2, have a

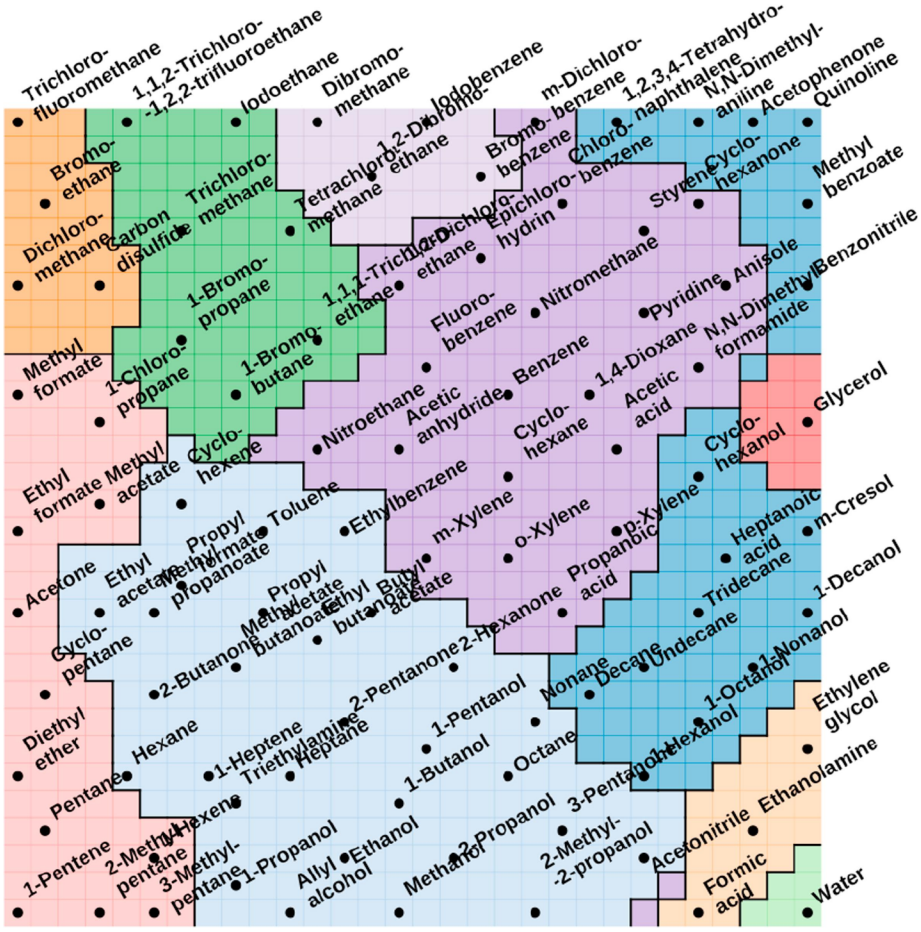

Fig. 2. SOM positioning map obtained from thermophysical properties of various liquids. The output nodes are classified via the k-means method and colored separately by each cluster.

large distance in terms of thermophysical properties. This is because relatively small numbers of the input data are used for learning as compared to the defined number of clusters employed here. It will be necessary to enlarge the number of liquid species for better examination of SOM-clustering approach.

4. Consideration based on each thermophysical property

While the SOM itself can provide information concerning distance relationships of the input data, more detailed information can be extracted by visualizing distributions of each thermophysical properties projected onto the two-dimensional SOM, i.e., each component of the weighted vectors, as discussed later. To briefly visualize the correlations between various thermophysical properties, a correlation matrix for all the thermophysical properties just obtained from the input data is presented in Fig. 4. In this figure, we can find obvious strong correlations, some of which are the positive correlation between boiling point and melting point and negative correlation between boiling point and saturated vapor pressure as expected. Meanwhile, interesting but not straightforward correlations are also found between the specific heat and thermal conductivity, the surface tension and thermal conductivity, and density and specific heat for the tested liquids. In contrast to the strongly correlated properties, several thermophysical properties are less correlated, for example, between vapor pressure and density. This weak correlation suggests that thermophysical properties are not mutually in a trade-off relationship, which implies a possibility in designing the liquid substances with a higher degree of freedom.

Fig. 5 shows various thermophysical properties represented as “heat maps” after the SOM learning. In each figure, the positions of liquid substances and the cluster boundaries are identical to those shown in the SOM positioning map (see Fig. 2). Therefore, one can get a grasp of a whole distribution of various thermophysical properties and correlations of various liquid substances at a glance. For example, these heat maps immediately give a strong negative correlation between density and heat capacity, and specifically this correlation gets stronger near at the cluster of halogen compounds (upper left). The result implies that halogenation of the liquid molecule greatly influences specific thermophysical properties and transport properties. In terms of searching heat-transfer fluids, liquid substances having low viscosity and high thermal conductivity are ideal, i.e., the high heat transfer coefficient can be obtained with low pumping power. Such materials can be readily found in Fig. 5 like methanol and allyl alcohol at the bottom center and formic acid at the lower right as examples. In fact, methanol has thermal conductivity of 0.202 W/(m·K) and viscosity of 0.544 mPa·s, whereas the 2-methyl-2-propanol, which is greatly separated from methanol with large U-matrix ridge (see also Fig. 2 and

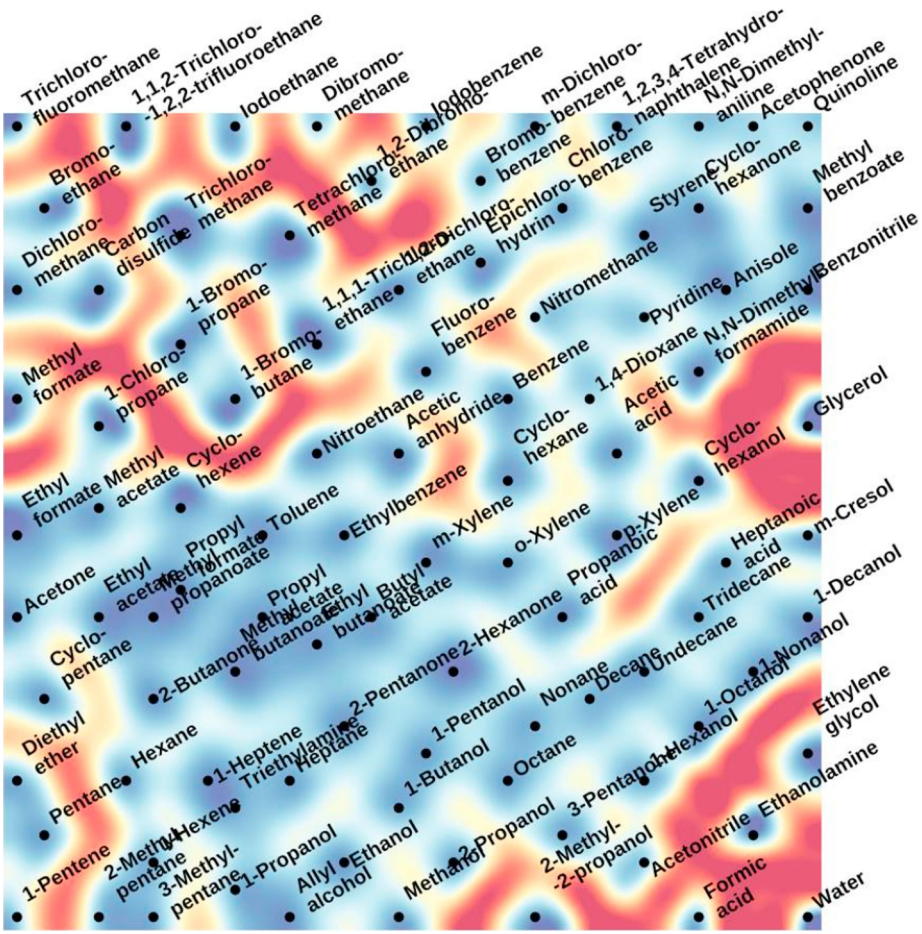

Fig. 3. U-Matrix values on output nodes obtained from the weighted vectors. Red and blue regions of the figure represent higher and lower values of the U-Matrix, respectively. Each liquid species is placed at the same location as in Fig. 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

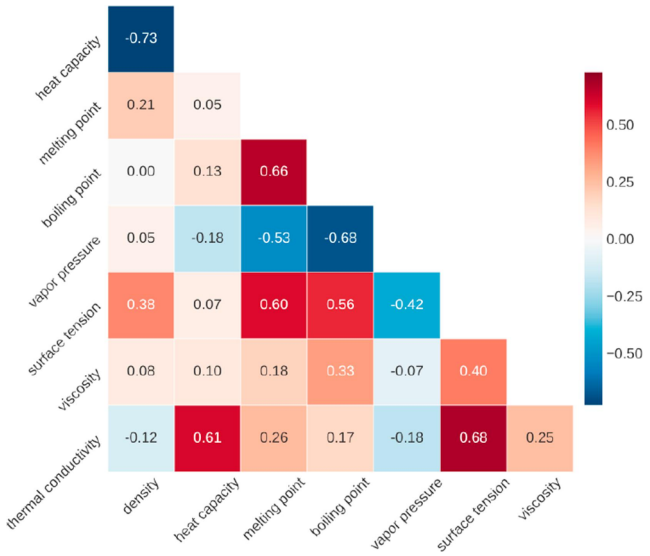

Fig. 4. Correlation matrix for all the thermophysical properties. Each value depicted in the boxes indicates the correlation coefficient between two thermophysical properties.

Fig. 3), has lower thermal conductivity of 0.112 W/(m·K) and higher viscosity of 4.31 mPa·s. Thus, we can easily separate superior substances among inferior ones in a certain practical use. In addition, if heat-transfer fluids are used as a working fluid in a heat pipe (HP), latent heat of vaporization is actively utilized to remove emitted thermal energy effectively. In this case, the boiling temperature should be within an appropriate range, taking into account operating conditions. Assuming a HP is operated under the atmospheric pressure and for cooling of electronic devices, liquid substances having high thermal conductivity and low boiling point (in a temperature range shown in Fig. 5) are desirable like methyl formate or methanol. Although many other restricting conditions, e.g., inflammability, toxicity, and environmental impact, must be considered in actual industrial applications, a highly practicable data mining technique can be achieved by involving such conditions into the framework of SOM.

5. Conclusions

We applied unsupervised machine learning protocols to multi-dimensional thermophysical data of liquid substances, and the liquid substances were mapped in two-dimensional space using the SOM and clustering techniques to visualize the proximity relationship. It was demonstrated that this framework enabled us to easily understand relative relationships of the thermophysical properties among liquid substances. Grouping of the liquid substances was successfully achieved via the clustering approach using the k-means method which was applied to the weighted vectors on output-layer nodes in the SOM. Various thermophysical properties of the tested liquid substances were represented as “heat maps” in the SOM. This visualization gives precise correlations among thermophysical properties of a lot of liquid substances at a glance. Thus, the presented framework is useful for quick design and exploration of the candidates of liquid materials for specific applications.

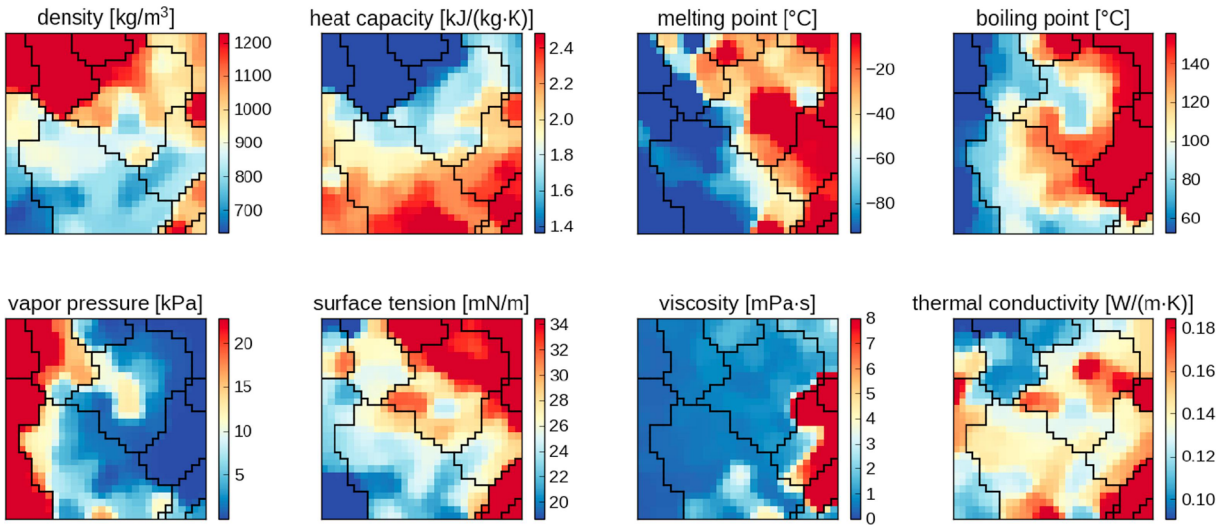

Fig. 5. Heat maps of thermophysical properties projected on the SOM. The cluster boundaries using k-means clustering, which are identical to those in Fig. 2, are also drawn in the figure.

Declaration of interest statement

The authors declared that there is no conflict of interest.

Acknowledgments

We thank Dr. Yutaka Oya for a fruitful discussion. GK and TO thanks the support of the Cross-ministerial Strategic Innovation Promotion Program. We also would like to acknowledge the vitally important encouragement and support made through the University of Washington-Tohoku University: Academic Open Space (UW-TU:AOS).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cplett.2019.04.075.

References

[1] K. Rajan, Mater. Today 9 (2005) 38–45.
[2] D. Xue, P.V. Balachandran, J. Hodgen, J. Theiler, D. Xue, T. Lookman, Nat. Commun. 7 (2016) 11241.
[3] D. Xue, D. Xue, R. Yuan, Y. Zhou, P.V. Balachandran, X. Ding, J. Sun, T. Lookman, Acta Mater. 125 (2017) 532–541.
[4] A.M. Gopakumar, P.V. Balachandran, D. Xue, J.E. Gubernatis, T. Lookman, Sci. Rep. 8 (2018) 3738.
[5] R. Yuan, Z. Liu, P.V. Balachandran, D. Xue, Y. Zhou, X. Ding, J. Sun, D. Xue, T. Lookman, Adv. Mater. 30 (2018) 1702884.
[6] D. Xue, P.V. Balachandran, R. Yuan, T. Hu, X. Qian, E.R. Dougherty, T. Lookman, PNAS 113 (2016) 13301–13306.
[7] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M.A.L. Marques, Chem. Mater. 29 (2017) 5090–5103.
[8] G. Pilania, P.V. Balachandran, C. Kim, T. Lookman, Front. Mater. 3 (2016) 19.
[9] F.A. Faber, A. Lindmaa, O.A. von Lilienfeld, R. Armiento, Phys. Rev. Lett. 117 (2016) 135502.
[10] G. Pilania, J.E. Gubernatis, T. Lookman, Comput. Mater. Sci. 129 (2017) 156–163.
[11] C.C. Fischer, K.J. Tibbetts, D. Morgan, G. Ceder, Nat. Mater. 5 (2006) 641–646.
[12] Y.T. Sun, H.Y. Bai, M.Z. Li, W.H. Wang, J. Phys. Chem. Lett. 8 (2017) 3434–3439.
[13] R. Liu, A. Kumar, Z. Chen, A. Agrawal, V. Sundararaghavan, A. Choudhary, Sci. Rep. 5 (2015) 11552.
[14] J.P. Janet, H.J. Kulik, Chem. Sci. 8 (2017) 5137–5152.
[15] B. Hu, K. Lu, Q. Zhang, X. Ji, W. Lu, Comput. Mater. Sci. 136 (2017) 29–35.
[16] J. Gao, Y. Liu, Y. Wang, X. Hu, W. Yan, X. Ke, L. Zhong, Y. He, X. Ren, J. Phys. Chem. C 121 (2017) 13106–13113.
[17] M.W. Gaultois, A.O. Oliynyk, A. Mar, T.D. Sparks, G.J. Mulholland, B. Meredig, APL Mater. 4 (2016) 053213.
[18] A. Mannodi-Kanakkithodi, G. Pilania, T.D. Huan, T. Lookman, R. Ramprasad, Sci. Rep. 6 (2016) 20952.
[19] A. Mannodi-Kanakkithodi, T.D. Huan, R. Ramprasad, Chem. Mater. 29 (2017) 9001–9010.
[20] C. Li, D.R. de Celis Leal, S. Rana, S. Gupta, A. Sutti, S. Greenhill, T. Slezak, M. Height, S. Venkatesh, Sci. Rep. 7 (2017) 5683.
[21] M.A. Kuenemann, D. Fourches, Mol. Inf. 36 (2017) 1600143.
[22] PolyInfo; URL: http://polymer.nims.go.jp/.
[23] T.D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, R. Ramprasad, URL: Sci. Data 3 (2016).
[24] P. Raccuglia, K.C. Elbert, P.D.F. Adler, C. Falk, M.B. Wenny, A. Mollo, M. Zeller, S.A. Friedler, J. Schrier, A.J. Norquist, Nature 533 (2016) 73–77.
[25] A.L. Ferguson, J. Phys, Condens. Matter 30 (2018) 043002.
[26] K. Shimoyama, K. Kamisori, J. Aircraft 54 (2017) 1317–1327.
[27] Y. Oya, G. Kikugawa, T. Okabe, Macromole. Theory Simul. 27 (2017) 1600072.
[28] T. Kohonen, Self-Organizing Maps, third ed, Springer-Verlag, 2001.
[29] T. Hastie, R. Tibshirani, J. Friedman, The Element of Statistical Learning, Springer-Verlag, 2009.
[30] W.M. Haynes, CRC Handbook of Chemistry and Physics, 93rd ed., CRC Press, 2012.
[31] http://tpds.db.aist.go.jp/tpds-web/, accessed on Jan. 15, 2017.
[32] http://webbook.nist.gov/chemistry/, accessed on Jan. 15, 2017.
[33] https://pubchem.ncbi.nlm.nih.gov/, accessed on Jan. 15, 2017.
[34] http://www.chemspider.com/, accessed on Jan. 15, 2017.
[35] https://scifinder.cas.org/scifinder/, accessed on Jan. 15, 2017.
[36] https://github.com/sevamoo/SOMPY, accessed on Dec. 14, 2016.
[37] S. Raschka, Python Machine Learning, Packt Publishing, 2015.
[38] A. Ultsch, G. Guimaraes, D. Korus, H. Li, Proc. Transputer Anwender Treffen/World Transputer Congress TAT/WTC 93 Aachen, 1993, pp. 194–203.