Songklanakarin J. Sci. Technol.
34 (5), 569-575, Sep. - Oct. 2012

SJST SONGKLANAKARIN JOURNAL OF SCIENCE AND TECHNOLOGY
http://www.sjst.psu.ac.th

Original Article

Novel approach to predict the azeotropy at any pressure using classification by subgroups

Taehyung Kim, Hiromasa Kaneko, Naoya Yamashiro, Kimito Funatsu*

Department of Chemical System Engineering,
The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan

Received 16 December 2011; Accepted 7 May 2012

Abstract

Distillation is one of the dominating separation processes, but there are some problems as inseparable mixtures are formed in some cases. This phenomenon is called as azeotropy. It is essential to understand azeotropy in any distillation processes since azeotropes, i.e. inseparable mixtures, cannot be separated by ordinary distillation. In this study, to construct a model which predicts the azeotropic formation at any pressure, a novel approach using support vector machine (SVM) is presented. The SVM method is used to classify data in the two classes, that is, azeotropes and non-azeotropes. 13 variables, including pressure, were used as explanatory variables in this model. From the result of the SVM models which were constructed with data measured at 1 atm and data measured at all pressures, the 1 atm model showed a higher prediction performance to the data measured at 1 atm than the all pressure model. Thus, for improving the performance of the all pressure model, we focused on intermolecular forces of solvents. The SVM models were constructed with only data of the solvents having same subgroups. The accuracy of the model increased and it is expected that this proposed method will be used to predict azeotropic formation at any pressure with high accuracy.

Keywords: azeotrope, azeotropic prediction, SVM, pressure, subgroup

1. Introduction

Many chemical processes in industrial productions involve purification and separation. Distillation is the dominating separation process and there are many cases to form certain inseparable mixtures where vapor compositions and liquid ones at equilibrium are equal. These specific mixtures are called azeotropes. Information on the occurrence of azeotropes in a mixture is essential in any distillation processes since these azeotropes can make a given separation by using ordinary distillation impossible (Horsley, 1973; Kamath *et al.*, 2005; Modla *et al.*, 2008). Segura *et al.* (1999) used an equation of state as the thermodynamic model to calculate azeotropes for binary mixtures and Dong *et al.* (2010) proposed four methods to predict azeotropes based on the UNIFAC model without any experimental data. UNIFAC (Gmehling *et al.*, 1993) is the thermodynamic equation for the activity coefficient between two liquids and it is one of the most frequently used methods to predict azeotropic formation. But UNIFAC also has drawbacks as its applicability is restricted to low pressures and the total number of the energy parameters is required as many as the squared number of the groups.

Reflecting the importance of whether a given mixture will form an azeotrope in the distillation or not, numerous data of azeotropes have been reported and accumulated by many researchers. In this work, we focused on this large amount of data and we acknowledge that chemoinformatic methods are the best techniques for managing these azeotropic data. Chemistry has produced an enormous amount of data until now and it has been required novel approaches for handling these data. Chemoinformatics is the application of

* Corresponding author.
Email address: funatsu@chemsys.t.u-tokyo.ac.jp

computational methods to solve the chemical problem with the mixing of information resources (Gasteiger *et al.*, 2006). The term was introduced in the late 1990s and there are many areas that can be developed from the application of chemo-informatic methods. In this paper, we constructed a model judging azeotropic formation by using a statistical method, i.e. support vector machine (SVM). Azeotropic data of the Dortmund Data Bank was used for this study (Lohmann *et al.*, 2001).

2. Research Methodology

This chapter gives an introduction to the SVM method to discriminate the occurrence of azeotropy and the method to evaluate the SVM model. Then, we will explain how to classify the solvents for hydrogen bond and apply the proposed method.

2.1 Support Vector Machine

SVM is the method used to train the classifiers which can be applied to classify data in two classes. This method was developed by Vapnik (1999) and has been widely used to solve various pattern recognition and classification problems. Assuming a set of data of two classes, SVM constructs a hyperplane that separates two different classes of vectors with a maximum margin (Vapnik, 1995; Vapnik, 1999). It is separated by finding the vector **w** and the parameter b that minimizes $||\mathbf{w}||^2$. It satisfies the following conditions:

$$\mathbf{w}\mathbf{x}_i + b \geq +1, \text{ for } y_i = +1 \text{ (positive)} \quad (1)$$

$$\mathbf{w}\mathbf{x}_i + b \leq -1, \text{ for } y_i = -1 \text{ (negative)} \quad (2)$$

where **w** is a weight vector to the hyperplane, $\mathbf{x}_i$ and y_i denote training data. In addition, $|b|/||\mathbf{w}||$ shows the perpendicular distance from the origin to the hyperplane, where $||\mathbf{w}||^2$ means the Euclidean norm of **w**. If **w** and b are determined, a vector $\mathbf{x}_i$ can be classified as follows:

$$f(x) = sign[\mathbf{w}\mathbf{x}_i + b] \quad (3)$$

In nonlinear systems, SVM allocates the data into a higher dimensional input space and establishes an optimal separating hyperplane using a kernel function such as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 / 2\sigma^2) \quad (4)$$

where σ is a tuning parameter of the kernel function and represent the width of the Gaussian kernel. The adjustable parameter σ plays a major role in the performance of the kernel. Linear SVM is then applied to this feature space, and then, the decision function is given as follows:

$$f(x) = sign[\sum_{i=1}^{l} \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b] \quad (5)$$

where the coefficients α_i^0 and b are maximized by Lagrangian expression:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (6)$$

$$\text{Subject to : } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^{l} \alpha_i y_i = 0 \quad (7)$$

The positive and negative classes are determined by positive or negative value from Equation 3 or 5.

2.2 Evaluation of SVM models

In this work, to evaluate the performance of a SVM model, accuracy rate, precision, and detection rate were used and defined as follows:

$$\text{Accuracy rate} = \frac{a+d}{a+b+c+d} \quad (8)$$

$$\text{Precision} = \frac{a}{a+c} \quad (9)$$

$$\text{Detection rate} = \frac{a}{a+b} \quad (10)$$

Table 1 shows a confusion matrix for a two-class classifier. A confusion matrix (Kohavi and Provost, 1998) contains information on actual classifications and those which predicted by a classifier. Classification accuracy, precision and detection rate can be defined by using the elements of the confusion matrix. Here, "a" and "d" represent the number of exact prediction about azeotropy and non-azeotropy, respectively, whereas "b" and "c" denote the number of wrong prediction in the same way.

2.3 Classification of solvents

The most important single cause of deviation from ideal behavior in liquid mixtures is hydrogen bonding (Ewell

Table 1. Confusion matrix.

		Prediction data	
		Azeotropy	Non-azeotropy
Experimental data	Azeotropy	a	b
	Non-azeotropy	c	d

Table 2. Classification of organic solvents according to hydrogen bonding.

Class 1	Solvents capable of forming three-dimensional networks of strong hydrogen bonds: water, glycol, glycerol, amino alcohols, amides, etc.
Class 2	Solvents containing both donor atoms and active H atoms in the same molecule: alcohols, acids, phenols, oximes, ammonia, primary and secondary amines, etc.
Class 3	Solvents containing donor atoms but no active hydrogen atoms: ethers, ketones, aldehydes, esters, etc.
Class 4	Solvents containing active hydrogen atoms but no donor atoms such as molecules having two or three chlorine on the same carbon atom and one or more chlorine atoms on adjacent carbon atoms: $CHCl_3$, CH_2Cl_2, CH_2Cl-CH_2Cl, CH_2Cl-$CHCl_2$, etc.
Class 5	All other solvents having no hydrogen bond forming capabilities: hydrocarbons, carbon disulfide, sulfides, non-metallic elements such as iodine, phosphorus, etc.

et al., 1944). Solvents having hydrogen atom bind to the electronegative atom strongly and have abnormal boiling points. This characteristic is caused by hydrogen bonds and hydrogen can coordinate between two molecules such as oxygen, nitrogen, or fluorine (Snyder *et al.*, 1974). Table 2 shows classification of organic solvents according to hydrogen bonding. The lower number of class means higher class in this paper and higher class solvents have stronger hydrogen bonding.

2.4 Classification by subgroups

In this work, we took advantage of the fact that each solvent was composed of a few functional groups. Azeotropes are formed due to differences in intermolecular forces of attraction, like a hydrogen bonding among the mixture components (Ewell *et al.*, 1944). Thus, the solvents are divided into subgroups such as H_2O, ACH and CH_3 to express intermolecular forces before constructing the SVM models.

2.5 Procedure of proposed method

Figure 1 shows the overall strategy of this study. The procedure of an azeotropy discriminant model is largely divided into the part of data grouping to classify solvents that have same subgroups with binary azeotropes and the part of calculation to discriminate whether there is an azeotrope formation or not with the SVM method.

3. Results and Discussion

3.1 Data

The Dortmund Data Bank contains approximately 48,000 datasets for binary azeotropic systems and 3,000 sets for ternary azeotropic systems. It also contains approximately 2,000 kinds of solvents involving physical properties. Duplicated data sets were removed and about 30,000 data sets for binary azeotropic systems were used here. About 20,000 data sets were measured at 1 atm in the binary data sets. Figure 2 shows the number of data sets of major solvents, which are contained in the binary azeotropic data.

In order to build the SVM model, it is necessary to properly quantify the structure and characteristics of each compound regarding azeotropy. Therefore, it is decided to use the structure descriptors as explanatory variables and Table 3 shows the descriptors.

The partial equalization of orbital electronegativity (PEOE) value shown in Table 3 is a descriptor, which has been developed by Gastiger *et al.* (1980), which calculation is based on the electronegativity of the atoms in the molecule. The electronegativity of each atom in the molecule is acquired first, and the charge transfer between the atoms adjacent to each other was calculated by the empirical formula. Next, the electronegativity of each atom is recalculated based on the value of the electric charge, and this cycle is performed until the value of the electric charge is convergent.

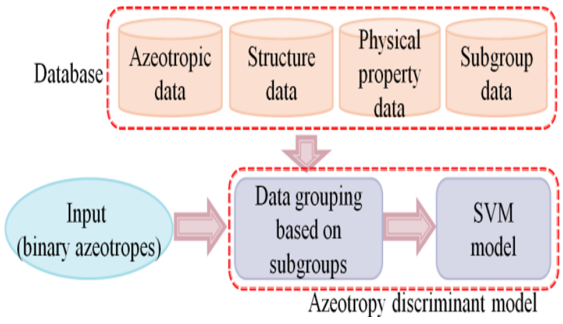

Figure 1. Problem-solving strategy.

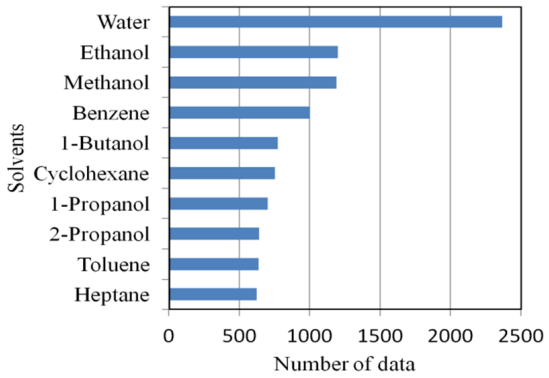

Figure 2. Number of data sets of binary azeotropic solvents.

It makes possible to express an electric charge in the molecule to do such calculation, and PEOE is used for the calculation of pKa, actually. The maximum electric charge of the hydrogen and the maximum negative charge of an atom calculated by PEOE were used as explanatory variables.

Marbin sketch (Miller *et al.*, 1979), a software of Chem Axon company, is used for the calculation of pKa and pKb. This software is able to detect atomic groups having high acidity and alkalinity such as amino group or carboxylic acid. pKa and pKb values are calculated by the empirical formula based on the calculated values of polarizability and PEOE charge in the atomic group.

3.2. Construction of models predicting the presence of azeotropy according to pressure

In order to predict the presence of azeotropy at any pressure, the explanatory variables include pressure. We calculated and compared azeotropic prediction models with data measured at 1 atm and all pressures, respectively, which were constructed with 13 variables described in Section 3.1 using the SVM method. The results are shown in Table 4.

According to Table 4, the estimated results of two data sets were approximately equal but in the case of using all pressure data, estimation performance were slightly higher. Thus, it is expected that this model is able to predict the presence of azeotropy at any pressure levels.

3.3 Analysis of the constructed models

In order to compare the models, which were constructed by using all pressure data ($model_{all}$) and 1 atm data ($model_{1\ atm}$), the presence of azeotropy with data measured at 1 atm was estimated by the $model_{all}$. The results are shown in Table 5. It is obvious that the $model_{1\ atm}$ has a higher prediction performance than the $model_{all}$ for 1 atm data.

From the results of Table 5, the solvents data of two cases were collected. One is the case that the prediction is correct by the $model_{1\ atm}$ but the prediction is incorrect by the $model_{all}$ (Case 1). The other is the case that the prediction is incorrect by the $model_{1\ atm}$ but the prediction is correct by the $model_{all}$ inversely (Case 2). The results are given in Table 6, 7, 8 and 9.

To improve the performance, the prediction about solvents in Table 6 is important. From the above results, it was found that the solvents in the Table 6 almost belong to the low class, which is described in Section 2.3 and the $model_{all}$ could not predict the presence of azeotropy when the solvents belong to the low class. Furthermore, the results of Table 8 show that the $model_{1\ atm}$ could not predict the presence of azeotropy of the low class solvents appropriate-

Table 3. Character of solvents and explanatory variables.

Character	Explanatory variables
Bond (hydrogen bond, polarity)	Calculated values of electric charge of hydrogen by PEOE Number of hydrogen bond acceptor and donor
Size of molecule	Number of atom, Number of ring, Number of double bond, Maximum distance among atoms in the molecule
Acid or base	pKa, pKb Number of amino group, Number of carboxyl group
Pressure	Measured pressure

Table 4. Estimation results according to pressure.

Pressure	Accuracy rate (%)	Precision (%)	Detection rate (%)
1 atm	91.8	93.0	89.3
All pressures	92.6	94.0	92.7

ly. Thus, the method predicting the presence of azeotropy of the low class solvents more precisely was required. On the other hand, although we anticipated that high class solvents would be in Table 9, there were some low class solvents, which are explained in the next chapter.

3.4. Construction of models predicting the presence of azeotropy according to subgroups

The number of acceptors or donors was used as explanatory variables to reflect the influence of hydrogen bonds. But low class solvents show a tendency not to be predicted accurately from the results of Table 6 and 8. Thus, it is important to improve the accuracy of prediction for low class solvents by dividing solvents by subgroups.

First, the solvent names were inputted, then, the binary azeotrope data sets, which have the same subgroups with the input solvents were extracted. On the basis of these data sets, the model to discriminate the presence of azeotropy is calculated with the SVM method. In order to confirm the limit of application of method, the prediction was calculated both for high class and low class solvents. Table 10 shows the results of Sample 1 (low class), Sample 2 (low class), and Sample 3 (high class) as a case study. According to Table 10, the results of each sample improved the predictive accuracy, compared to the results of all pressures in Table 4.

Table 11 shows the results of prediction about the 1 atm data by the model which was constructed with all pressure data after being separated by subgroups, respectively. It was confirmed that the model could predict not only low class solvents but high class solvents from the results above. The prediction performance of each sample achieved remarkably improved compared to the results of the $model_{all}$ in Table 5.

For Sample 3, the number of the ten most solvents data in the case of incorrect predictions is shown in Figure 3. As shown in Figure 3, the presence of azeotropy of the solvents having a ring like benzene and phenol could not be predicted accurately with only subgroup information.

Table 5. Estimation results according to model. Only 1 atm data were estimated.

Model	Accuracy rate (%)	Precision (%)	Detection rate (%)
$Model_{1\ atm}$	91.8	93.0	89.3
$Model_{all}$	72.9	66.1	86.3

Table 6. The ten solvents of the highest proportion in Case 1.

Solvents	Proportion (%)	Class
Bromobenzene	49.12	5
1,2-Dichloroethane	48.36	3
Methoxybenzene	42.72	3
Benzyl chloride	42.31	5
p-Xylene	41.43	5
Acetic acid butyl ester	40.68	3
Tetrachloroethylene	40.17	4
Dimethoxymethane	39.68	3
Carbonic acid diethyl ester	39.66	3
Tetrahydrofuran	39.44	3

Table 7. The ten solvents of the lowest proportion in Case 1.

Solvents	Proportion (%)	Class
1,3-Diisopropylbenzene	6.06	5
Quinoline	5.49	3
N,N-Diethylaniline	5.36	3
Propionamide	4.41	1
Nitromethane	4.26	3
Acetamide	3.13	1
Ethyl carbamate	1.96	1
Glycerol	1.92	1
N-Methylaniline	1.85	2
Monoethanolamine	1.16	2

Table 8. The ten solvents of the highest proportion in Case 2.

Solvents	Proportion (%)	Class
Diethyl sulfide	20.00	5
Nitrotrichloromethane	17.74	4
Glycol monoacetate	16.67	3
Methoxybenzene	16.50	3
4-Bromotoluene	15.69	5
m-Nitrotoluene	14.29	3
Di-n-propyl ether	13.51	3
1,1-Diethoxyethane	13.46	3
Diethyl ether	13.04	3
Isobutyl iodide	12.96	5

Table 9. The ten solvents of the lowest proportion in Case 2.

Solvents	Proportion (%)	Class
Naphthalene	0.96	5
4-Isopropyltoluene	0.93	5
1,2-Ethanediol	0.93	2
1-Pentanol	0.93	2
o-Xylene	0.81	5
Cyclohexanol	0.80	2
Butyric acid	0.79	1
p-Xylene	0.71	5
Indene	0.56	5
Water	0.53	1

Table 10. Estimation results according to subgroups.

	Subgroup	Accuracy rate (%)	Precision (%)	Detection rate (%)
Sample 1	Cl-(C=C)/ C=C/ $CHCl_3$	98.6	98.6	99.9
Sample 2	$CH_2Cl/CHCl_3$	97.6	97.7	97.7
Sample 3	H_2O/ACH	96.6	97.4	95.8

Table 11. Estimation results according to subgroups by using all pressure models.

	Subgroup	Accuracy rate (%)	Precision (%)	Detection rate (%)
Sample 1	Cl-(C=C)/ C=C/ $CHCl_3$	89.0	84.6	98.0
Sample 2	$CH_2Cl/CHCl_3$	88.6	84.3	92.6
Sample 3	H_2O/ACH	81.7	74.4	81.4

Naphthalene, 4-isopropyltoluene, xylene and indene, which have one or more rings, appeared in Table 9. This may be why there are the low class solvents in Table 9. It is conceivable that the solvent having rings need to be predicted accurately.

4. Conclusions

In this paper, we have proposed new approach to predict the occurrence of azeotropy. For predicting an azeotropy at any pressure, 13 variables including pressure were used as descriptors. The constructed model is able to predict the presence of azeotropy with high accuracy. Moreover, the performance of the SVM method improved by classifying solvents according to subgroups. In this work, we investigated only some selected cases, but should confirm all cases. Meanwhile, it was recognized that solvents having rings like benzene have to be handled properly before the prediction.

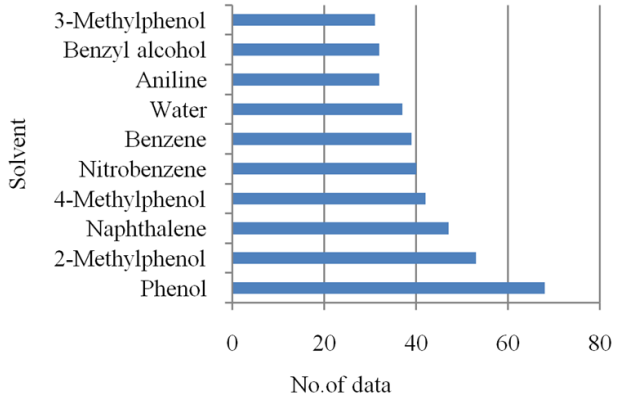

Figure 3. Number of data in case of incorrectness.

References

Dong, X., Gong, M., Zhang, Y., Liu, J. and Wu, J. 2010. Prediction of Homogeneous Azeotropes by the UNIFAC Method for Binary Refrigerant Mixtures. Journal of Chemical & Engineering Data. 55(1), 52–57.

Ewell, R.H., Harrison, M. and Berg, L. 1944. Azeotropic distillation. Industrial & Engineering Chemistry. 36(10), 871-875.

Gasteiger, J. 1980. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. Tetrahedron. 36, 3219.

Gasteiger, J. and Funatsu, K. 2006. Chemoinformatics - An Important Scientific Discipline. Journal of Computer Chemistry, Japan. 5(2), 53-58.

Gmehling, J., Li, J. and Schiller, M. 1993. A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties. Industrial & Engineering Chemistry Research. 32, 178–193.

Horsley L.H. 1973. Azeotropic Data-III. Advances in Chemistry Series. American Chemical Society, Washington DC., U.S.A., 116.

Kamath, G., Georgiev, G. and Potoff, J.J. 2005. Molecular Modeling of Phase Behavior and Microstructure of Acetone-Chloroform-Methanol Binary Mixtures. Journal of Physical Chemistry B. 109, 19463-19473.

Kohavi, R. and Provost, F. 1998. Glossary of Terms. Machine Learning. 30, 271-274.

Lohmann, J. and Gmehling, J. 2001. Modified UNIFAC (Dortmund) : Reliable Model for the Development of Thermal Separation Processes. Journal of Chemical Engineering of Japan. 34, 43-54.

Miller, K. J. and Savchik, J. 1979. A new empirical method to calculate average molecular polarizabilities. Journal of the American Chemical Society. 101, 7206-7213.

Modla, G. and Lang, P. 2008. Feasibility of new pressure swing batch distillation methods. Chemical Engineering Science. 63, 2856-2874.

Segura, H., Wisniak, J., Toledo, P.G. and Mejia, A. 1999, Prediction of azeotropic behavior using equation of state. Fluid Phase Equilibria. 166, 141-162.

Snyder, L.R. 1974. Classification of the Solvent Properties of Common Liquids. Journal of Chromatography. 92, 223-230.

Vapnik, V. 1995. The nature of statistical learning theory. Springer: New York, U.S.A.

Vapnik, V. 1999. An overview of statistical learning theory. Institute of Electrical and Electronics Engineers Transactions on Neural Networks. 10(5), 988–999.