Iranian Journal of Chemical Engineering
Vol. 13, No. 4 (Autumn 2016), IAChE

Liquid-Liquid Equilibrium Data Prediction using Large Margin Nearest Neighbor

M. Pirdashti[1*]***, K. Movagharnejad***[1]***, S. Curteanu***[2]***, F. Leon***[3]***, F. Rahimpour***[4]

[1]*Faculty of Chemical Engineering, Babol University of Technology, PO Box, 484, Babol, Iran*
[2]*Department of Chemical Engineering, Faculty of Chemical Engineering and Environmental Protection,"Gheorghe Asachi" Technical University of Iaşi, Str. Prof. Dr. Doc. DimitrieMangeron, nr. 73, 700050, Iaşi, Romania*
[3]*Department of Computer Science and Engineering, Faculty of Automatic Control and Computer Engineering, "Gheorghe Asachi" Technical University of Iaşi, Str. Prof. Dr. Doc. DimitrieMangeron, nr. 27, 700050, Iaşi, Romania*
[4]*Biotechnology Research Lab, Chemical Engineering Department, Faculty of Engineering, Razi University, Kermanshah 67149-67346, Iran*

ARTICLE INFO

Article history:
Received: 2015-11-08
Accepted: 2016-02-03

Keywords:
Liquid-Liquid Equilibrium
Aqueous Two Phase System
Nearest Neighbor
Large Margin
Regression Analysis

ABSTRACT

Guanidine hydrochloride has been widely used in the initial recovery steps of active protein from the inclusion bodies in aqueous two-phase system (ATPS). Knowledge of the guanidine hydrochloride effect on the liquid-liquid equilibrium (LLE) phase diagram behavior is still inadequate and no comprehensive theory exists for the prediction of the experimental trends. Therefore, the effect of the guanidine hydrochloride on the phase behavior of PEG4000+potassium phosphate+ water system at different guanidine hydrochloride concentrations and pH was investigated in this study. To fill the theoretical gaps, the typical support vector machines were applied was applied to the k-nearest neighbor method in order to develop a regression model to predict the LLE equilibrium of guanidine hydrochloride in the above mentioned system. Its advantage is its simplicity and good performance, with the disadvantage of an increase in the execution time. The results of our method are quite promising; they were clearly better than those obtained by well-established methods such as Support Vector Machines, k-Nearest Neighbor and Random Forest. It is shown that the obtained results are more adequate than those provided by other common machine learning algorithms.

1. Introduction

Aqueous two-phase system (ATPS) is formed by mixing certain amounts of two polymers or a polymer and a salt in water [1]. ATPS, a liquid–liquid extraction (LLE) strategy, is now recognized as a potential technique because of

[*] Corresponding author: pirdashti@yahoo.com

its multiple advantages including: biologic compatibility [2], ease of continuous process [3-8], low interfacial tension [9], short processing time[10], low material cost [11,12], low energy consumption [13,14], good resolution [15], high yield [16], relatively high load capacity [17], scaling up feasibility [18-20], selective extraction [21], separation of metal ions [22-25] and efficient procedure for separation of various biological materials such as recombinant proteins and enzymes [26-32].

Inclusion body refolding processes play a major role in the production of recombinant proteins. One step of the general strategy used to recover active protein from inclusion bodies is the solubilization of aggregated inclusion bodies with denaturant such as guanidine hydrochloride and urea [33]. Several articles have been published about applying guanidine hydrochloride and urea in ATPS for the initial recovery step [26,34], but little has been published about the complex problem of how the guanidine hydrochloride (GuHCl) and urea affect the phase diagram behavior and determine the partition coefficient of guanidine hydrochloride and urea in ATPS.

Novel bio-separation research based on aqueous two-phase systems needs to focus on determining phase diagrams, partition coefficients and other thermodynamic data for the design of industrial-scale processes. Knowledge about the mechanisms involved in the partitioning equilibrium of macromolecules is poor and, consequently, there is no comprehensive theory capable of predicting the experimental trends. The method development is fairly empirical and so, despite these favorable features, ATPSs have not been extensively adopted in either industrial processes or commercial applications [19,20].

Chaotropic agents such as urea and GuHCl are co-solutes that can disrupt the hydrogen bonding network between water molecules and reduce the stability of the native state of proteins by weakening the hydrophobic effect [35, 36].Urea is a powerful protein denaturant as it disrupts the non-covalent bonds in the proteins. GuHCl is also a strong denaturant which can coat the exterior of proteins. Above certain concentrations, GuHCl can fully denature a protein [37]. Guanidine hydrochloride is preferred, because urea solutions may contain and spontaneously produce cyanate[38], which can carbamylate the amino groups of the protein [39].The knowledge of phase systems containing chaotropic compounds is very limited. Rämsch et al. [34, 40] determined phase diagrams of PEG/ sodium sulfate/ urea/ water and PEG/ dextran T-500(DEX)/ phosphate buffer/water at different concentrations of urea and different PEG molecular weights. Rahimpour and Pirdashti [41, 42] used to obtained the partition coefficient of guanidine hydrochloride and the effective parameters such as pH and PEG/Salt weight percent ratio.

In recent times, there has been consistent effort put towards developing generalized correlations to elucidate the physical interactions and develop mathematical models. These models are used to describe the different factors that influence the purification efficiency in ATPS, in addition to the experimental methods which are complex, time, and resource consuming. Unfortunately, these correlations are limited to a narrow range of affected factors such as pH, salt and polymer concentrations. So, many researchers have tried to develop generalized correlations to predict the partition coefficients of chaotropic compounds and their effects on the phase diagram in ATPS systems, especially based on artificial neural networks [43-46]. Also, Pirdashti et al. [47] predicted the partition coefficients of guanidine hydrochloride in PEG–phosphate systems using neural networks developed with differential evolution algorithm. Generally, artificial neural networks have been widely applied, alone or in combination with other artificial intelligence techniques, for modeling and optimization of different complex processes [49-55].

In the present study, the effect of the guanidine hydrochloride on phase behavior of PEG4000/ phosphate/ guanidine hydrochloride/ water at different guanidine hydrochloride concentrations and pH was investigated experimentally and by simulation. A new regression algorithm was proposed, based on the k-nearest neighbor paradigm where the distance metric is optimized using the large margin concept typical of support vector machines. The optimization problem is solved by means of an evolutionary algorithm. The main advantage of the developed algorithm is that it requires virtually no tuning, and its results are better than those obtained by other regression methods. The downside is its increased execution time, caused by the heuristic evolutionary optimization.

The obtained results would be useful to increase the knowledge of aqueous two-phase separation process and improve the yield of protein refolding.

2. Experimental

2.1. Materials

Polyethylene glycol, with a mass average 4000g/mol, di-potassium hydrogen phosphate and sodium di-hydrogen phosphate were of analytical grade (Merck) and were used without further purification. Guanidine hydrochloride was purchased from Sigma-Aldrich. Distilled water was used in all experiments. All other materials were analytical grade.

2.2. Apparatus and procedure

The flow chart of methodology for processing the experimental data is described in Fig. 1. In the first phase, a mixture of PEG 4000, guanidine hydrochloride and phosphate salt solution at a determined pH was prepared. The top and bottom phases were separated in the next phase. In the third section, the components of each phase were analyzed.

2.2.1. Preparation of the aqueous two-phase systems

Biphasic systems were prepared by mixing specific amounts of PEG 4000 and phosphate salt solution at the required pH. The pH of the salt solution was adjusted by mixing appropriate ratio of sodium di-hydrogen phosphate and di-potassium hydrogen phosphate. In this work, for preparation of biphasic system, the experimental data reported by Haghtalab et al. [56] were used as reference. For each of the mentioned systems, four samples including 2.5 %, 5.0 %, 7.5 % and 10.0 % (w/w) of guanidine hydrochloride were prepared. All components were added into a graduated 15 ml test tube as a dry

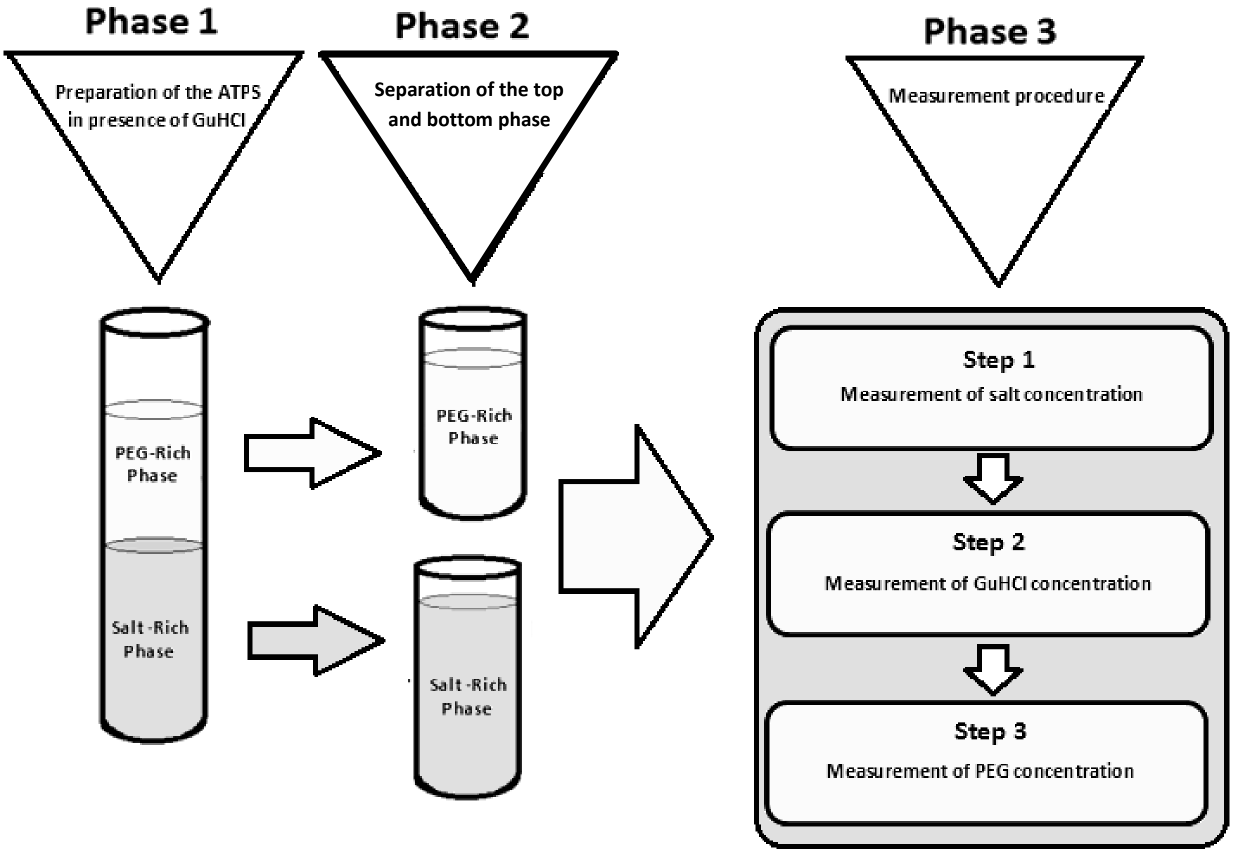
Phase 1
Preparation of the ATPS in presence of GuHCl
Phase 2
Separation of the top and bottom phase
Phase 3
Measurement procedure
PEG-Rich Phase
Salt-Rich Phase
PEG-Rich Phase
Salt-Rich Phase
Step 1
Measurement of salt concentration
Step 2
Measurement of GuHCl concentration
Step 3
Measurement of PEG concentration

Figure1. Methodology for measurement of ATPS components in the presence of guanidine hydrochloride.

powder or as stock solution at constant pH and temperature (298.15 K), resulting in a 10 g system. The pH values of the solutions were measured precisely with a pH meter of JENWAY 3345 model. In order to speed up the phase separation, the resulting solution was mixed by vortex test tube vigorously for 2 minutes and centrifuged at 2400 rpm for 10 min. Then, the tubes were placed in 298.15 K for 24 h; after the solution reached equilibrium, the samples of the top and bottom phases were carefully extracted, in order to leave a layer of solution at least 0.5 cm thick above the interface.

2.2.2. Measurement of salt concentration

The analysis methods for salt concentrations (K_2HPO_4, NaH_2PO_4) were carried out by using atomic absorption spectroscopy (AAS), Shimadzu AA-6300 model.

2.2.3. Measurement of guanidine hydrochloride concentration

The concentration of guanidine hydrochloride was determined by conductometer at 298.15K, using a JENWAY 4510 model. Since the conductivity of phase samples depends on both guanidine hydrochloride and salt concentration, but is independent of PEG concentration, calibration plots of conductivity versus guanidine hydrochloride concentration were prepared for different concentrations of salt. An example of calibration plot for guanidine hydrochloride/K_2HPO_4/ water system is shown in Fig. 2.

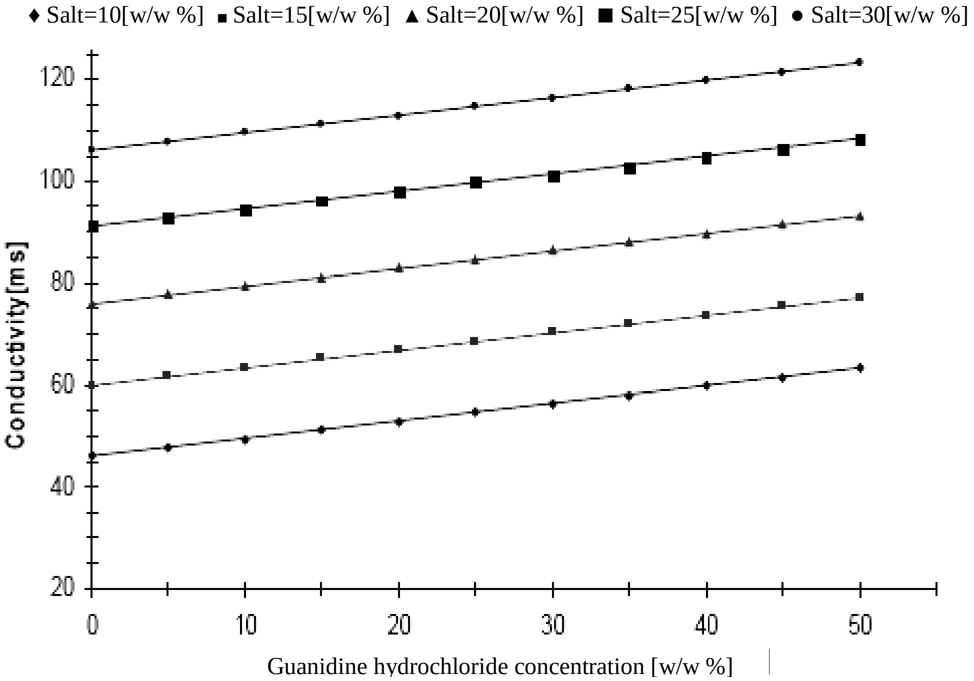

Figure 2. Conductivity calibration curves for PEG (4000) - K_2HPO_4-water at 298.15K.

2.2.4. Measurement of PEG concentration

The concentration of PEG was obtained by refractive index measurements at 298.15K using an ATAGO-DTM1 model. Since the refractive index of phase samples depends on PEG, guanidine hydrochloride, and salt concentration, calibration plots of refractive index versus polymer concentration were prepared for different concentrations of salt and guanidine hydrochloride.

3. Modeling methodology

Regression analysis includes any technique for modeling different kinds of processes with the goal of finding a relationship between a dependent variable and one or more independent variables, given a set of training instances or vectors in the form of ($\mathbf{x}_i$, y_i) pairs, where $\mathbf{x}_i$ are the inputs and y_i is the output of a sample. There are many regression methods; in addition to analytical models, where the task is to find adequate values for the coefficients, usually by means of differential optimization; several machine learning techniques such as: neural networks, support vector machines (ε-SVR, ν-SVR), decision trees (M5P, Random Forest, REPTree) or rules (M5, Decision Table), etc. can be mentioned.

k-Nearest Neighbor (kNN) [49] is a simple, efficient way to estimate the value of the unknown function in a query point, using its values in other known points. A more elaborate version of the method considers a weighted average, where the weight of each neighbor depends on its proximity to the query point and it is usually considered as the inverse of the squared Euclidian distance between the query point and its nearest neighbor. However, the classical approach does not take into account any information about a particular problem. Besides the common practice of normalizing the instance values independently on all dimensions, there is little domain knowledge incorporated into the method.

Because of the importance of the distance metric, researchers seek to find ways to adapt it to the problem at hand in order to yield better performance. This is the issue of distance metric learning [50-52, 57]. The idea of a large margin, one of the fundamental ideas of support vector machines, was transferred to the kNN method for classification tasks [58-60], resulting in the large-margin nearest neighbor method (LMNN). In general, distance metric learning aims at finding a linear transformation $\mathbf{x}' = \mathbf{L}\mathbf{x}$, such that the distance between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ becomes:

$$d_L(\mathbf{x}_i, \mathbf{x}_j) = \left\|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\right\|_2 \tag{1}$$

Since all the operations in k-nearest neighbor classification or regression can be expressed in terms of square distances, an alternative way of stating the transformation is by means of the square matrix: $\mathbf{M} = \mathbf{L}^T\mathbf{L}$, and thus the square distance is:

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j) \tag{2}$$

For a classification problem, the choice of $\mathbf{L}$, or equivalently $\mathbf{M}$, aims at minimizing the distance between a vector $\mathbf{x}_i$ and its k target neighbors $\mathbf{x}_j$, where a target signifies a neighbor that belongs to the same class. At the same time, the distance between a vector and the impostors $\mathbf{x}_l$, i.e. neighbors that belong to a different class, should be maximized. In order to establish a large margin between the vectors that belong to different classes, the following relation is imposed:

$$d_M(\mathbf{x}_i, \mathbf{x}_l) \geq 1 + d_M(\mathbf{x}_i, \mathbf{x}_j) \tag{3}$$

Here, the value of 1 is arbitrary; the idea is to have some minimum value for the margin that separates the classes. However, it is proven that other minimum values for the margin would not change the nature of the optimization problem, but would only result in the scaling of the matrix $\mathbf{M}$.

Overall, the optimization problem is defined as follows:

$$\begin{aligned} \text{Min} \quad & \sum_{ij} \eta_{ij} d_{ij} + \lambda_h \sum_{ijl} \nu_{ijl} \xi_{ijl} \\ \text{such that} \quad & (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j) = d_{ij} \\ & (\mathbf{x}_i - \mathbf{x}_l)^T \mathbf{M}(\mathbf{x}_i - \mathbf{x}_l) - d_{ij} \\ & \qquad \geq 1 - \xi_{ijl} \\ & \mathbf{M} \geq 0,\ \xi_{ijl} \geq 0, \forall i, j, l \end{aligned} \tag{4}$$

where $\eta_{ij} \in \{0, 1\}$ is 1 only when $\mathbf{x}_j$ is a target neighbor of $\mathbf{x}_i$, $\nu_{ijl} = \eta_{ij}$ if and only if $y_i \neq y_l$ and 0 otherwise, d_{ij} is the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, ξ_{ijl} is the hinge loss [61] and $\lambda_h \geq 0$ is a constant.

The two objectives of minimizing the distance to targets and maximizing the distance to impostors are conflicting. Following an analogy to attraction and repulsion forces in physics, Weinberger et al.(2005) [58] introduce two forces, ε_{pull} and ε_{push}, whose balance is reached by setting the value of λ_h. In their work, the authors consider the importance of these forces to be equal, i.e. $\lambda_h = 1$.

In our work, a regression problem is tackled instead of a classification one. Therefore, the optimization problem must be adapted to this case. Thus, the methodology presented [62] is followed to describe the Large Margin Nearest

Neighbor for Regression (LMNNR) method.

L (and thus **M**) is considered to be diagonal matrix. This increases the clarity of the results, because in this case the elements m_{ii} can be interpreted as the weights of the problem input dimensions, and are also a form of regularization.

By using the **M** matrix, the relation between the neighbor weights and the distance is:

$$w_{d_M}(x,x') = \frac{1}{d_M(x,x')} = \frac{1}{\sum_{i=1}^{n} m_{ii} \cdot (x_i - x'_i)^2} \qquad (5)$$

The optimization problem presented in equation 4 is solved by means of an evolutionary algorithm, with the following parameters: 40 chromosomes in the population, tournament selection with 2 candidates, arithmetic crossover with a probability of 95 % and mutation with a probability of 5 % in which a gene value is reinitialized to a random value in its corresponding domain of definition. Elitism was also used such that the best solution in a generation is never lost. As a stopping criterion, a maximum number of generations, 500 in our case, was used. The number of genes in a chromosome depends on the number of prototypes and the dimensionality of the problem, namely: $n_g = n_p \cdot n_i$, where n_g is the number of genes, n_p is the number of prototypes and n_i is the number of inputs.

The domain of the genes, which represents the elements of **L**, is $[10^{-3}, 10]$. All the operations in the software implementation deal with the elements of **M**, i.e. their squares. Thus it can be considered that the corresponding values of the **M** elements lie in the $[10^{-6}, 100]$ domain.

The fitness function F, which is to be minimized, takes into account 2 criteria:

$$F = w_1^F \cdot F_1 + w_2^F \cdot F_2 \qquad (6)$$

where the weights of the criteria are normalized: $w_1^F + w_2^F = 1$.

In order to simplify the expressions of the F_i functions, the following notations are made: where d_M means the weighted square distance function using the weights search is made for: $d_{ij} = d_M(x_i, x_j)$, $d_{ik} = d_M(x_i, x_k)$, $g_{ij} = |f(x_i) - f(x_j)|$ and $g_{ik} = |f(x_i) - f(x_k)|$. Thus, the first criterion is:

$$F_1 = \sum_{i=1}^{n} \sum_{j \in N(i)} d_{ij} \cdot (1 - g_{ij}) \qquad (7)$$

Where $N(i)$ is the set of the nearest k neighbors of instance i, in our case $k = 3$. Basically, this criterion says that the nearest neighbors of i should have similar values to the one of i, and more distant ones should have different values. This criterion tries to minimize the distance between an instance i and its neighbors with similar values. If a neighbor j has a dissimilar value, the second factor, $1 - g_{ij}$, becomes small and is no longer necessary to minimize the distance.

The second criterion is expressed as follows:

$$F_2 = \sum_{i=1}^{n} \sum_{j \in N(i)} \sum_{l \in N(i)} \max(1 + d_{ij} \cdot (1 - g_{ij}) - d_{ik} \cdot (1 - g_{il}), 0) \qquad (8)$$

It takes into account a pair of neighbors, j and l, by analogy to a target and an impostor. However, for the regression problem, these notations are not necessary because there is no class which could be the same or different.

Only the real values of the instance outputs have to be taken into account. The reasoning is the same as for the first criterion, but now the distance to the neighbors has to be minimized with close values (the positive term), while simultaneously trying to maximize the distance to the neighbors with distant values (the negative term). By analogy to equation 4, a margin of at least 1 should be present between an instance with a close value and another with a distant value. The *max* function is used by analogy to the expression of the hinge loss. If it is considered that *j* has a close value and *l* has a distant one, the corresponding hinge loss will be 0 only when $d_{il} \geq 1 + d_{ij}$. The condition $j \neq l$ is implicit because, when the terms are equal, they cancel each other out.

Overall, the following values were used for the weights of the criteria: $w_1^F = w_2^F = 0.5$.

The modeling methodology was implemented in in-house software using the C# programming language.

4. Results and discussion

4.1. Processing the experimental data

In these systems, opposing components were found, considering the lyotropic series $H_2PO_4^-$ and K^+ as so-called structure-making salts, while guanidine hydrochloride is described as structure breaking agent. The combination of these two competing components on phase separation is very interesting and cannot be predicted. For measuring the concentration of components in bottom and top phases of aqueous two-phase systems, standard calibration curves should be built. The PEG concentration has negligible effect on the calibration curve of guanidine hydrochloride concentration in PEG/ guanidine hydrochloride/ phosphate/ water system. In the results, the calibration curves were related to the concentrations of salt and guanidine hydrochloride. A change of PEG concentration in the range of 10- 50 % (w/w) results in 0.05 ms (mili-Siemens) in conductivity of the solution which is negligible relative to 20 ms changes related to changes in GuHCl concentration. A typical calibration curve for guanidine hydrochloride/ K_2HPO_4/ water system is shown in Fig. 2. The guanidine concentrations varied between 0 and 50 (w/w %) and the concentration of phosphate varied between 10 and 30 (w/w %). To increase the knowledge of aqueous two-phase systems containing denaturants, the phase diagrams of a broad range of systems based on PEG4000/ phosphate/ water at 298.15K and different pH (7.2, 9.1, and 10.8) in the presence of guanidine hydrochloride were determined. The experimental data of binodal for these systems at pH 7.2, 9.1 and 10.8 are reported in Tables 1-2 and plotted in Figs. 3-5. Figures 3-5 show the influence of the guanidine hydrochloride concentrations on the binodal curve of PEG4000/ phosphate aqueous two-phase system at constant pH.

The amount of phase components necessary to affect separation of the phases increases only slightly with increasing the guanidine concentrations. From Figs. 3-5, it can be observed that the binodal is displaced toward higher concentrations with increasing guanidine hydrochloride concentrations and the compositions of the polymer-rich phase (and hence the tie-line length) behave in a

Table 1
Phase Compositions for the PEG4000 + Potassium Phosphate + Water System at 25°C, pH 9.1, and different mass % Urea.

Overall composition			Top phase		Bottom phase	
w_s	w_p	w_u	w_s	w_p	w_s	w_p
10.77	12.49	0.0	3.12	29.55	15.74	0.77
10.74	11.17	0.0	3.25	28.78	14.82	1.33
9.74	12.66	0.0	3.62	27.24	14.31	1.21
9.72	11.34	0.0	3.92	25.71	13.98	1.48
9.03	11.36	0.0	5.02	22.83	13.22	1.54
8.70	12.82	0.0	4.29	23.84	13.05	2.23
8.70	11.44	0.0	4.92	21.94	12.8	4.03
8.68	14.24	0.0	3.87	25.53	13.94	1.5.0
9.34	12.52	2.5	4.42	21.72	15.00	1.93
9.66	10.88	2.5	4.59	21.56	13.46	2.85
9.44	11.95	2.5	4.63	19.89	12.96	2.67
9.65	10.76	2.5	4.97	16.46	12.63	2.95
8.60	10.29	2.5	5.17	16.74	12.44	3.08
8.77	12.11	2.5	4.69	16.37	11.87	3.45
8.66	10.69	2.5	5.22	17.33	11.54	5.13
8.21	12.30	2.5	4.65	19.63	12.71	3.06
8.15	12.70	5.0	4.71	21.15	14.49	2.51
9.43	10.92	5.0	5.35	19.78	12.89	3.69
8.11	12.46	5.0	4.94	19.21	12.29	3.55
8.49	11.13	5.0	5.19	15.82	12.11	3.78
6.97	8.91	5.0	5.44	16.23	11.66	3.81
7.93	10.41	5.0	4.98	15.98	11.34	3.95
8.68	10.42	5.0	5.32	17.01	10.93	6.00
8.85	10.76	5.0	4.82	19.22	12.23	3.65
8.65	12.52	7.5	5.23	20.12	13.61	3.48
9.67	10.48	7.5	5.65	18.89	12.22	4.46
8.66	11.87	7.5	5.29	18.86	11.44	4.36
8.14	11.18	7.5	5.39	15.5	11.46	4.75
8.54	9.88	7.5	5.76	15.72	10.93	4.88
8.09	10.18	7.5	5.43	15.63	10.57	5.11
8.40	10.48	7.5	5.56	16.46	10.12	6.88
8.51	10.88	7.5	5.13	18.89	11.11	4.71
9.28	11.85	10.0	5.88	19.28	12.36	4.59
9.41	10.40	10.0	6.38	17.59	11.18	5.78
8.74	11.35	10.0	5.77	18.13	10.39	5.61
7.77	11.15	10.0	5.75	14.72	10.41	5.99
8.13	10.14	10.0	6.12	14.78	9.88	6.11
7.42	11.15	10..0	5.83	14.81	9.51	6.35
7.82	11.14	10.0	6.01	15.29	9.23	7.92
8.41	10.59	10.0	5.93	17.56	10.02	6.05

Table 2
Phase Compositions for the PEG4000 + Potassium Phosphate + Water System at 25°C, pH 10.8, and different mass % Urea.

Overall composition			Top phase		Bottom phase	
w_s	w_p	w_u	w_s	w_p	w_s	w_p
12.00	12.00	0.0	2.12	31.10	17.03	2.62
11.00	13.50	0.0	2.46	31.86	16.73	1.51
10.00	13.53	0.0	2.59	34.16	15.73	1.50
9.98	11.97	0.0	4.90	32.82	12.58	2.10
8.99	12.00	0.0	5.16	26.84	11.32	1.09
8.98	13.50	0.0	4.54	25.93	14.74	0.99
9.07	12.17	2.5	4.16	25.8	15.24	3.97
8.55	13.58	2.5	4.36	24.44	14.94	2.87
9.42	13.24	2.5	4.46	21.31	13.93	2.86
9.19	9.46	2.5	5.89	17.81	10.78	3.45
8.05	10.8	2.5	4.86	23.56	9.89	3.45
8.89	12.96	2.5	4.53	18.38	12.95	2.34
9.79	12.07	5.0	4.61	24.02	14.86	4.58
8.88	13.53	5.0	4.95	22.50	14.46	3.41
9.49	12.86	5.0	5.32	18.05	13.34	3.39
9.54	8.97	5.0	6.15	14.85	10.37	3.99
8.18	10.24	5.0	5.24	22.69	9.45	3.84
7.37	13.5	5.0	4.66	18.02	12.56	2.85
9.04	12.19	7.5	5.54	21.58	14.21	5.24
7.92	13.82	7.5	5.59	20.42	13.78	4.13
9.01	12.93	7.5	5.70	16.59	12.71	4.09
9.17	9.21	7.5	6.27	14.20	9.81	4.81
8.17	9.50	7.5	5.71	20.73	8.87	4.45
7.64	13.25	7.5	4.83	17.88	11.94	3.12
10.15	11.72	10.0	6.34	19.81	13.35	6.31
8.86	13.35	10.0	6.35	17.35	12.81	5.22
7.52	13.70	10.0	6.11	14.56	11.86	5.11
8.47	8.12	10.0	7.15	14.00	9.00	5.79
7.18	11.14	10.0	6.22	18.82	8.08	5.55
7.68	13.01	10..0	5.06	17.44	11.07	4.11

slightly shorter manner. This effect is also seen for urea [63]. It can be related to the structure breaking effect of guanidine hydrochloride on the water or the preferential interaction with aqueous interface [64]. But, it seems that further investigations are necessary to study the effect of guanidine hydrochloride on water structure changes. PEG is a hydrophilic polymer and, due to the effect of guanidine on water structure, the depletion

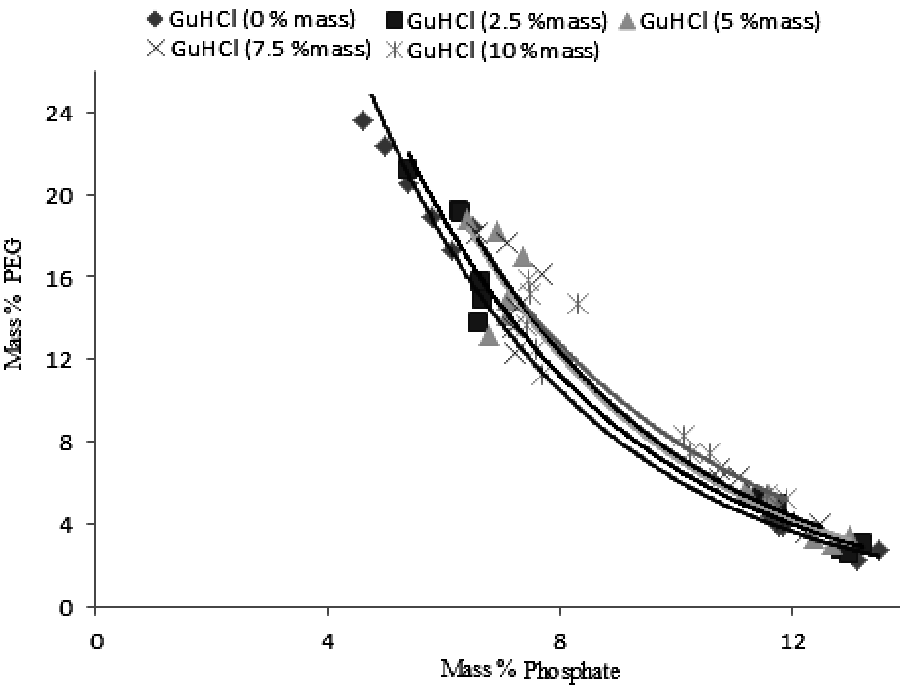

Figure 3. Binodal of PEG 4000/ phosphate systems at 298.15 K and pH 7.2 at different guanidine hydrochloride concentrations.

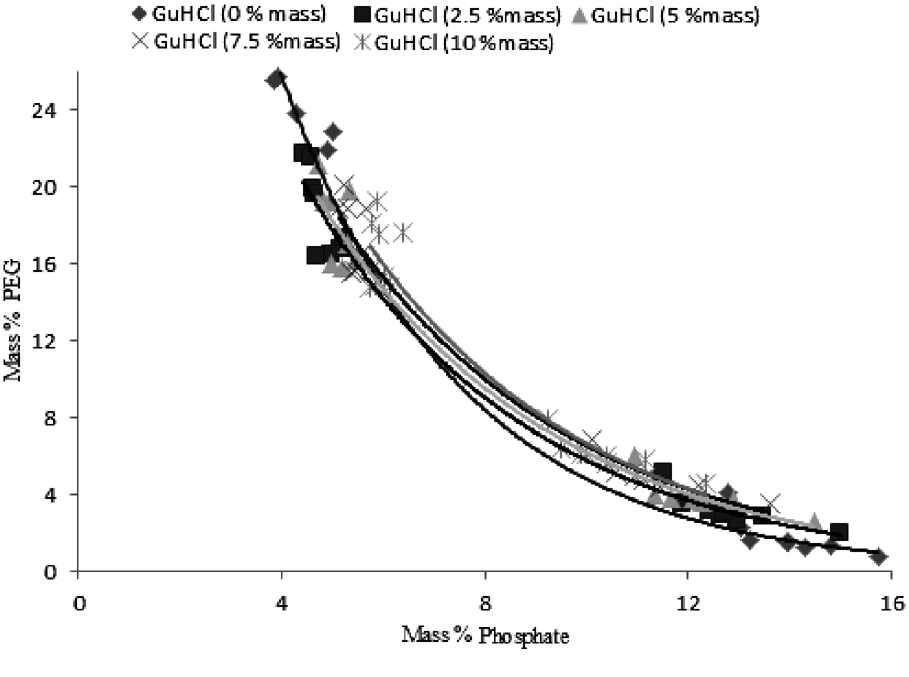

Figure 4. Binodal of PEG 4000/ phosphate systems at 298.15 K and pH 9.1 at different guanidine hydrochloride concentrations.

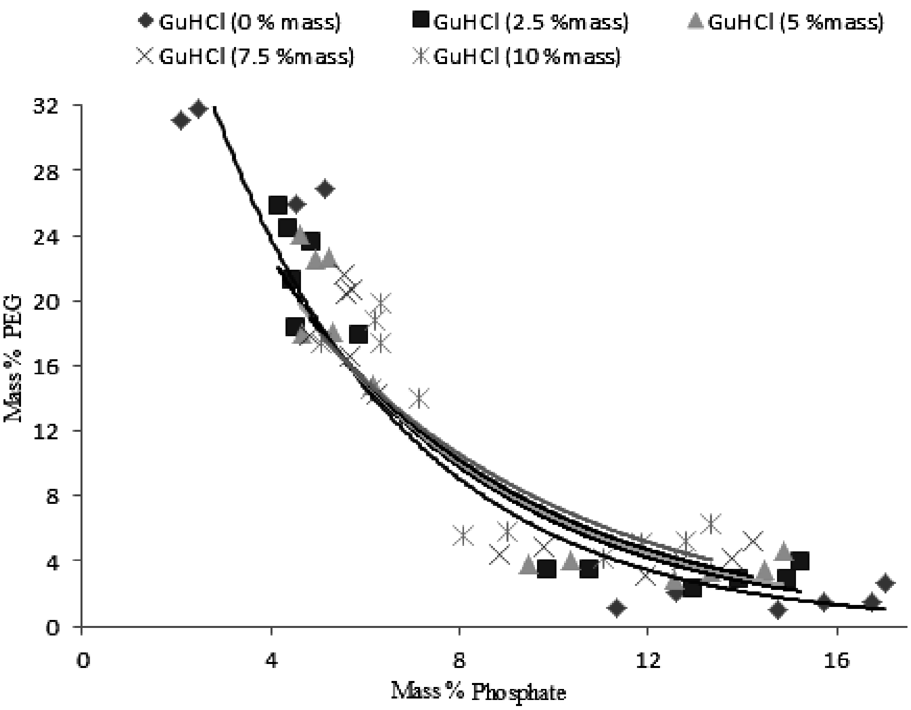

Figure 5. Binodal of PEG 4000/ phosphate systems at 298.15 K and pH 10.8 at different guanidine hydrochloride concentrations.

force between PEG-GuHCl and water increases. As a result, the ratio of PEG in the two phases is changed and the two phase region shrinks.Increasing the pH results in a slightly larger two-phase area. With increasing the pH, the charge of salt ions changes and, for a constant PEG concentration, lower concentration of salt is needed.

4.2. Modeling the experimental data

In order to compare different algorithms and models, the cross-validation method was applied with 10 folds. According to this method, the complete dataset is divided into 10 equal groups. One of those groups is considered as the testing set, and the other nine compose the training set. A model is thus built on 90 % of the data, and tested on the remaining 10 %. This procedure is repeated 10 times, each time with a different group as the testing set. In the end, the quality metric of the whole model is computed as the average of the quality metrics of the 10 models.The coefficient of determination (r^2) and the squared coefficient of correlation is used as a metric to compare the performance of different models.

When using our software implementation for the regression problem, 10 neighbors were considered for computing the output. Fewer neighbors do not provide enough information to compute a precise value, while if more neighbors are used, the influence of the more distant ones becomes negligible, because the neighbor weights are proportional to the inverse of the square distance. In the computation of the fitness function, three reference neighbors were used. This number should be at least two, in order to provide some contrast between an instance with a close value and another with a distant value (by analogy with the target and impostor instances for classification). With two reference neighbors, the results are far worse than with three, while with more reference neighbors, the computation time greatly increases and the results are no better than in case of three reference neighbors.

Table 3 presents the results obtained with various algorithms implemented in the popular collection of machine learning algorithms using Weka [65]. The results are given for each of the four outputs of the model, i.e. O1-O4.

The comparison is made with the nearest neighbor algorithm, taking into account one or *k* nearest neighbors, where the optimal value of *k* is obtained by cross-validation. Also, the ε-Support Vector Regression algorithm with an RBF kernel is employed. These algorithms were mainly chosen for comparison because of their good results in general. The proposed method, LMNNR, namely the Large Margin Nearest Neighbor for Regression, is also based on concepts from both nearest neighbor paradigm and large margin from support vector machines.

Table 3
Cross-validation performance of different regression algorithms.

Algorithm	O1	O2	O3	O4
LMNNR average	0.8918	0.8858	0.9289	0.9094
LMNNR best	0.9361	0.9384	0.9603	0.9291
NN	0.8626	0.7380	0.7552	0.7974
k-NN (k determined by cross-validation)	0.8928 (k = 3)	0.8024 (k = 3)	0.8658 (k = 3)	0.8848 (k = 5)
ε-SVR (RBF kernel)	0.9039	0.8194	0.8570	0.8640
Random Forest	0.8780	0.8111	0.7573	0.8178
M5 rules	0.9062	0.8132	0.7233	0.8840

Also, the Random Forest algorithm [66] was used because it is also considered to have very good performance in general. It relies on building a certain number of random trees (e.g. 10) using bagging, i.e. each tree is built for a slightly different set of training instances. Finally, after the trees are built, each of them chooses the value of the new instance to be processed.

A representative of the decision rules family was included as well, i.e. the M5 rules algorithm. It generates a decision list for regression problems using the separate-and-conquer method. In each iteration, it builds a model tree using M5 and transforms the best leaf into a rule [67].

In Table 3, regarding the two rows for LMNNR, since an evolutionary algorithm is used for optimization, and this is a heuristic method, finding the best solution is not guaranteed every time. Therefore, several trials are performed for each problem. The row “LMNNR average” contains the average of the results obtained over all these trials. The row “LMNNR best” contains the best result found by our algorithm for each output, i.e. problem. One can see that these values, marked in bold letters, are the best and exceed even the results given by very good algorithms such as the ε-Support Vector Regression or the Random Forest. Also, the results are better than those of the traditional kNN, even when the optimal number of neighbors *k* is determined itself by cross-validation.

Regarding the parameters of the LMNNR

model, as a nearest-neighbor model, all the training data is stored. In this paradigm, a good generalization is not obtained by trying to find a model as simple as possible, e.g. with a lower number of parameters, like with polynomial regression or even decision trees relying on homogeneity criteria such as entropy. An increased generalization capability is ensured by maximizing the margin between instances with different output values by learning an appropriate space metric.

After the model is created, it can be applied for new data using the original training set and the parameters learned by solving the optimization problem: m_{ii} in equation (5) and, if the problem requires multiple prototypes, the position vector of the prototypes in the input space and the corresponding m_{ii} for each one.

5. Conclusions

The phase behaviour of PEG4000/phosphate/ guanidine hydrochloride/water system at different guanidine hydrochloride concentrations and pH was investigated. The increase of the guanidine hydrochloride concentration displaced the binodal of the system toward the higher concentrations of the components. The increase of the pH resulted in a slightly larger two-phase area.

A new algorithm, the large margin nearest neighbor regression (LMNNR) is presented and applied for the modelling of the liquid-liquid equilibrium (LLE) of guanidine hydrochloride in the PEG4000/phosphate/ guanidine hydrochloride/water system.

It belongs to the class of instance-based methods, but the distance metric, which is crucial for the performance of this type of algorithm is not fixed, but depends on the problem at hand. The distance metric is obtained by solving an optimization problem which tries to decrease the distance between the instances with similar output values and increase the distance between the instances with different output values. This process is actually equivalent to maximizing the margin between the instances with different output values. An evolutionary algorithm is used to solve the optimization problem. Its advantage is its simplicity and good performance, with the disadvantage of an increase in the execution time. The results of our method are quite promising: they were clearly better than those obtained by well-established methods such as Support Vector Machines, k-Nearest Neighbour and Random Forest.

Acknowledgment

This work was supported by the “Partnership in priority areas – PN-II” program, financed by ANCS, CNDI - UEFISCDI, project PN-II-PT-PCCA-2011-3.2-0732, No. 23/2012.

References

[1] Albertsson, P. Å., Partition of cell particles and macromolecules, 3rd ed., New York, Wiley, USA, (1986).

[2] Raghavarao, K., Ravganathan, T., Srinivas, N. and Barhate, R., “Aqueous two phase extraction: An environmentally benign technique”, Clean Technol. Environ. Policy.,**5**(2), 136 (2003).

[3] Biazus, J. P., Santana, J. C., Souza, R. R., Jordao, E. and Tambourgi, E. B., “Continous extraction of alpha- and beta-amylases from Zea mays malt in a

PEG4000/$CaCl_2$ ATPS", J. Chromatogr. B., **85**(1-2),227 (2007).

[4] Cavalcanti, M.T.H., Carneiro-da-Cunha, M. G., Brandi, I.V., Porto, T.S., Converti, A., Lima Filho, J.L., Porto, A.L.F. and Pessoa, A., "Continuous extraction of á- toxin from a fermented broth of Clostridium perfringens Type A in perforated rotating disc contactor using aqueous two-phase PEG–phosphate system", Chem. Eng. Prog., **47**,1771 (2008).

[5] Vázquez-Villegas, P., Aguilar, O. and Rito-Palomares, M., "Study of biomolecules partition coefficients on a novel continuous separator using polymer-salt aqueous two-phase systems", Sep. Purif. Technol., **78**(1),69(2011).

[6] Rosa, P., Azevedo, A., Sommerfeld, S., Bäcker, W. and Aires-Barros, M., "Continuous aqueous two-phase extraction of human antibodies using a packed column", J. Chromatogr. B., **880**,148(2012).

[7] Rosa, P.A.J., Azevedo, A.M., Mutter, M., Bäcker, W. and Aires-Barros, M.R., "Continuous purification of antibodies form cell culture supernatant with aqueous two-phase systems: From concept to process", Biotechnol. J., **8** (3),352 (2013).

[8] Espitia-Saloma, E., Vázquez-Villegas, P., Aguilar, O. and Rito-Palomares, M., "Continuous aqueous two-phase systems devices for the recovery of biological products", Food Bioprod. Process,**92** (2),101 (2014).

[9] Luechaua, F., Ling, T. and ChandLyddiatt, A., "A descriptive model and methods for up-scaled process routes for interfacial partition of bioparticles in aqueous two-phase systems", Biochem. Eng. J., **50**(3),122 (2010).

[10] Hu, R., Feng, X., Chen, P., Fu, M., Chen, H., Guo, L. and Liu, B-F., "Rapid, highly efficient extraction and purification of membrane proteins using a microfluidic continuous-flow based aqueous two-phase system", J.Chromatogr. A., **1218** (1),171 (2011).

[11] Rodrigues, G.D., Teixeira, L.D., Ferreira, G.M.D., da Silva, M.D.H., da Silva, L.H.M. and de Carvalho, R.M.M., "Phase diagrams of aqueous two-phase systems with organic salts and F68 triblock copolymer at different temperatures", J. Chem. Eng. Data., **55**(3),1158(2010).

[12] Rosa, P.A.J., Azevedo, A.M., Sommerfeld, S., Bäcker, W. and Aires-Barros, M.R., "Aqueous two-phase extraction as a platform in the biomanufacturing industry: Economical and environmental sustainability", Biotech. Adv., **29**(6),559 (2011).

[13] Naganagouda, K. and Mulimani, V.H., "Aqueous two-phase extraction (ATPE): An attractive and economically viable technology for downstream processing of Aspergillusoryzae á-galactosidase", Process Biochem., **43** (11),1293(2008).

[14] Hatti-Kaul, R., Aqueous two-phase systems: Methods and protocols, methods in biotechnology, 11, Humana Press, (2000).

[15] Bradoo, S., Saxena, R.K. and Gupta, R., "Partitioning and resolution of mixture of two lipases from Bacillus stearothermophilus SB-1 in aqueous two-phase system", Process Biochem., **35** (1-2),57(1999).

[16] Rito-Palomares, M., “Practical application of aqueous two-phase partition to process development for the recovery of biological products”, J. Chromatogr. B., **807** (1),3(2004).

[17] Asenjo, J.A. and Andrews, B.A., “Aqueous two-phase systems for protein separation: Phase separation and applications”, J. Chromatogr. A., **1238**,1(2012).

[18] Selber, K., Tjerneld, F., Collén, A., Hyytiä, T., Nakari-Setälä, T., Bailey, M., Fagerström, R., Kan, J., van der Laan, J., Penttilä, M. and Kula, M. R., “Large-scale separation and production of engineered proteins, designed for facilitated recovery in detergent-based aqueous two-phase extraction systems”, Process Biochem., **39** (7),889(2004).

[19] Yan-Min, L., Yan-Zhao, Y., Xi-Dan, Z. and Chuan-Bo, X., “Bovine serum albumin partitioning in polyethylene glycol (PEG)/potassium citrate aqueous two-phase systems”, Food Bioprod. Process., **88** (1),40 (2010).

[20] Rocha, M.V. and Nerli, B.B., “Molecular features determining different partitioning patterns of papain and bromelain in aqueous two-phase systems”, Int. J. Bio. Macromol., **61**,204 (2013).

[21] Rodrigues, G.D., de Lemos, L.R., da Silva, L.H.M. and da Silva, M.C.H., “Application of hydrophobic extractant in aqueous two-phase systems for selective extraction of cobalt, nickel and cadmium”, J. Chromatogr. A., **1279**,13(2013).

[22] Wang, Z.H., Song, M. and Ma, Q., “Two-phase aqueous extraction of chromium and its application to speciation analysis of chromium in plasma”, Mikrochim. Acta., **134**(1),95(2000).

[23] Gao, Y.T. and Wang, W.W., “Distribution behavior and extraction mechanism of gold(III) in polyethylene glycol ammonium sulphate aqueous biphasic system”, Chin. J. Appl. Chem., **19**(6),578 (2002).

[24] Patrício, P.R., Mesquita, M.C., da Silva, L.H.M. and da Silva, M.C.H. “Application of aqueous two-phase systems for the development of a new method of cobalt(II), iron(III) and nickel(II) extraction: A green chemistry approach”, J. Hazard Mater., **193**,311(2011).

[25] Bulgariu, L. and Bulgariu, D., “Selective extraction of Hg(II), Cd(II) and Zn(II) ions from aqueous media by a green chemistry procedure using aqueous two-phase systems”, Sep. Purif. Technol., **118**,209(2013).

[26] Rahimpour, F., Feyzi, F., Maghsodi S. and Kaul, R.H., “Purification of plasmid DNA with polymer-salt aqueous two-phase system: Optimization using response surface methodology”, Biotech. Bioeng., **95** (4),627 (2006).

[27] Rahimpour, F., Mamo, G., Feyzi, F., Maghsoudi, S. and Kaul, R.H., “Optimization refolding and recovery of active recombinant bacillus haloduransxylanase in polymer-salt aqueous two-phase system using surface response analysis”, J. Chromatogr. A., **1141**(1),32 (2007).

[28] Zaveckas, M., Zvirblieñe, A., Zvirblis, A. Chmieliauskaite, V., Bumelis, V. and Pesliakas, H., “Effect of surface histidine mutations and their number on the partitioning and refolding of recombinant

human granulocyte-colony stimulating factor (Cys17Ser) in aqueous two-phase systems containing chelated metal ions", J. Chromatogr B., **852** (1-2),409 (2007).

[29] Shahbaz Mohamadia, H. and Omidinia, E., "Purification of recombinant phenylalanine dehydrogenase by partitioning in aqueous two-phase systems", J. Chromatogr. B., **854**(1-2),273 (2007).

[30] Shahbaz Mohamadia, H. and Omidinia, E., "Process integration for the recovery and purification of recombinant Pseudomonas fluorescensproline dehydrogenase using aqueous two-phase systems", J. Chromatogr. B., **929**,11 (2013).

[31] Lan, J. Ch-W., Yeh, C-Y., Wang, C-C., Yang, Y-H. and Wu, H-S., "Partition separation and characterization of the poly hydroxyalkanoates synthase produced from recombinant Escherichia coli using an aqueous two-phase system", J. Biosci. Bioeng., **116** (4),499 (2013).

[32] Ibarra-Herrera, C., Aguilar, O. and Rito-Palomares, M., "Application of an aqueous two-phase system strategy for the potential recovery of a recombinant protein from alfalfa (Medicago sativa)", Sep. Purif. Technol., **77** (1),94(2011).

[33] Clark, ED., "Protein refolding for industrial processes", Curr. Opin. Biotechnol., **12** (2),202 (2001).

[34] Rämsch, C., Kleinelanghorst, L. B., Knieps, E., Thommes, M. R. and Kula, A. J. "Aqueous two-phase systems containing urea: Influence of protein structure on Protein Partitioning", Biotechnol. Bioeng., **69**,83 (2001).

[35] Salvi, G., De Los Rios, P. and Vendruscolo, M., "Effective interactions between chaotropic agents and proteins", *protein*., **61** (3),492(2005).

[36] Vemić, A., Stojanović, B. J., Stamenković, I. and Malenović, A., "Chaotropic agents in liquid chromatographic method development for the simultaneous analysis of levodopa, carbidopa, entacapone and their impurities", J. Pharm. Biomed. Anal., **77**,9 (2013).

[37] Parnica, J. and Antalika, M. "Urea and guanidine salts as novel components for deep eutectic solvents", J. Mol. Liq., **197**,23(2014).

[38] Hagel, P., Gerding, J.J.T., Fieggen, Wand Bloemendal, H., "Cyanate formation in solutions of urea I. Calculation of cyanate concentrations at different temperature and pH", Biochim. Biophys. Acta., **243** (3),366(1971).

[39] Cejka, J., Vodražkaand, Z. and Salgk, J., "Carbamylation of globin in electrophoresis and chromatography in the presence of urea", Biochim. Biophys. Acta., **154**(3),589 (1968).

[40] Rämsch, Ch., Kleinelanghorst, L.B., Knieps., E.A., Homes, J. and Kula, M.R., "Aqueous two-phase system containing urea; Influence on phase separation and stabilization of protein conformation by phase components", Bitechnol. Prog., **15**(3),493(1999).

[41] Rahimpour, F. and Pirdashti, M., "The effect of guanidine hydrochloride on phase diagram of PEG-phosphate aqueous two-phase system", World Academy of Science, Engineering and Technology, **1** (5), 29(2007).

[42] Rahimpour, F. and Pirdashti, M. "Effective parameters on the partition

coefficient of guanidine hydrochloride in the poly ethylene glycol + phosphate + water system at 298.15 K". Iranian J. Chem. Eng., **7**(1),67 (2010).

[43] Gautam, G. and Simon, L. "Prediction of equilibrium phase compositions and â-glucosidase partition coefficient in aqueous two-phase systems", Chem. Eng. Commun., **194** (1),117 (2007).

[44] Pazuki, G.R., Taghikhani, V. and Vossoughi, M., "Prediction the partition coefficients of biomolecules in polymer-polymer aqueous two-phase systems using the artificial neural network", Particulate Sci. Technol., **28** (1),67 (2010).

[45] Pazuki, Gh. and Seyfi Kakhki, S., "A hybrid GMDH neural network to investigate partition coefficients of Penicillin G Acylase in polymer–salt aqueous two-phase systems", J. Mol. Liq., **188**,131 (2013).

[46] Abdolrahimi, Sh., Nasernejad, B., Pazuki, Gh., "Prediction of partition coefficients of alkaloids in ionic liquids based aqueous biphasic systems using hybrid group method of data handling (GMDH) neural network", J. Mol. Liq., **191**,79(2014).

[47] Pirdashti, M., Movagharnejad, K., Curteanu, S., Dragoi, E.N., Rahimpour, F., "Prediction of partition coefficients of guanidine hydrochloride in PEG–phosphate systems using neural networks developed with differential evolution algorithm", J. Ind. Eng. Chem., **27**,268(2015).

[48] Thomas Cover, M. and Hart, P.E., "Nearest neighbor pattern classification". IEEE T Inform Theory, **13**(1),21(1967).

[49] Shental, N., Hertz, T., Weinshall, D. and Pavel, M., "Adjustment learning and relevant component analysis", Proceedings of the 7th European Conference on Computer Vision, ECCV-02,4,776-792, London, UK, Springer-Verlag,776(2002).

[50] Shalev-Shwartz, Sh., Singer, Y. and Ng, A.Y., "Online and batch learning of pseudo-metrics", Proceedings of the 21st International Conference on Machine Learning, ICML-04, Banff, Canada, pp. 94–101 (2004).

[51] Chopra, S., Hadsell, R. and LeCun, Y., "Learning a similarity metric discriminatively, with application to face verification", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR-05, San Diego, CA, USA, pp. 349-356 (2005).

[52] ElSehiemy, R., Abou El-Ela, A. and Shaheen, A., "Multi-objective fuzzy-based procedure for enhancing reactive power management", IET Gener. Transm. Dis., **7**(12),1453 (2013).

[53] Precup, R.E., Rdac, M.B., Tomescu, M.L., Petriu, E.M. and Preitl, S., "Stable and convergent iterative feedback tuning of fuzzy controllers for discrete-time SISO systems", Expert Syst. Appl., **40** (1), 188 (2013).

[54] Khmelev, A. and Kochetov, Yu., "A hybrid local search for the split delivery vehicle routing problem", IJ-AI., **13**(1), 147(2015).

[55] Kazakov, A. L. and Lempert, A. A., " On mathematical models for optimization problem of logistics infrastructure, IJ-AI., **13**(1), 200(2015).

[56] Haghtalab, A. and Mokhtarani, B., "The new experimental data and a new

thermodynamic model based on group contribution for correlation liquid-liquid equilibria in aqueous two-phase systems of PEG and (K_2HPO_4 or Na_2SO_4)", Fluid Phase Equilib., **215**(2),151 (2004).

[57] Goldberger, J., Roweis, S., Hinton, G. and Salakhutdinov, R., Neighbourhood components analysis, Advances in Neural Information Processing Systems, 17, Cambridge, MA, USA, MIT Press, pp.513-520(2005).

[58] Weinberger, K.Q., Blitzer, J. and Saul, L.K., Distance metric learning for large margin nearest neighbor classification, Advances in Neural Information Processing Systems, 18, MIT Press, Cambridge, MA, USA, pp. 1473-1480 (2006).

[59] Weinberger, K.Q. and Saul, L.K., "Fast solvers and efficient implementations for distance metric learning", Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, pp. 1160-1167(2008).

[60] Weinberger, K.Q. and Saul, L.K., "Distance metric learning for large margin nearest neighbor classification", J. Mach. Learn Res.,**10**,207 (2009).

[61] Moore, R.C. and DeNero, J. "L1 and L2 regularization for multiclass hinge loss models", Proceedings of the Symposium on Machine Learning in Speech and Language Processing, pp. 1-5 (2011).

[62] Leon, F. and Curteanu, S., "Evolutionary algorithm for large margin nearest neighbour regression", 7th International Conference on Computational Collective Intelligence Technologies and Applications, ICCCI, Spain. Madrid, pp. 21-23 (2015).

[63] Estapé, D., Rinas, U., "Optimized procedures for purification and solubilization of basic fibroblast growth factor inclusion bodies", Biotechnol. Tech., **10** (7),481 (1996).

[64] Annuziata, O., Asherie, N., Lomakin, A., Pande, J., Ogun, O., Benedek, G.B., "Effect of polyethylene glycol on the liquid-liquid phase transition in aqueous protein solutions", Proc. Natl. Acad. Sci., **99** (22),14165(2002).

[65] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H., "The WEKA data mining software: An update". ACM SIGKDD Explor., **11**(1),10(2009).

[66] Breiman, L., "Random Forests", Mach. Lear., **45** (1),5 (2001).

[67] Holmes, G., Hall, M. and Frank, E., "Generating rule sets from model trees", Twelfth Australian Joint Conference on Artificial Intelligence, p. 1-12 (1999).