UDC 514.742.4:556.3(497.5 Odra)
Original Scientific Paper

Classification of Hydrochemical Data in Reduced Dimensional Space

Jasminka Dobša *jasminka.dobsa@foi.hr*
University of Zagreb, Croatia
Faculty of Organization and Informatics

Petr Praus *petr.praus@vsb.cz*
VSB-Technical University, Ostava, Czech Republic
Department of Analytical Chemistry and Material Testing

Ch. Aswani Kumar *cherukuri@acm.org*
VIT University, Vellore, India
School of Information Technology and Engineering

Pavel Praks *pavel.praks@gmail.com*
VSB-Technical University, Ostrava, Czech Republic
Department of Applied Mathematics

Abstract

The main objective of this paper is to systematically analyze the performance of water sample classifications for different data representations. We compare the classification of full data representation to the classification of data items in a lower dimensional space obtained by the projection of the original data on the space formed by first principal components, and further, on the space of centroids of classes. We use linear support vector machines for classification of ground water samples collected from five different localities of the Odra River basin. The obtained results are evaluated by standard measures including recall, precision and F_1measure.
Keywords: concept decomposition, dimensionality reduction, principal components analysis, support vector machines

1. Introduction

Water is incredibly important for every living organism to survive. It covers two thirds of the Earth's surface, which makes our planet unique among other known celestial bodies. Water is primarily a chemical substance. Analyzing and interpreting the chemistry of water provides valuable insights into this phenomenon. According to ISO standards, the quality of water is determined by testing for a specific chemical composition of a particular water sample. Water is generally described in terms of its nature, usage, or origin. In that respect, distinction is made between water as a natural resource, drinking water, water for industrial use, sewage water, to name but a few examples. The category of the water being analyzed determines parameters by which it will be tested. The analysis of water samples is characterized by parameters forming an *n*-dimensional data space. Dimensionality Reduction (DR) techniques are often applied on such data to reduce dimensionality, thereby achieving noise reduction and allowing for recognition of latent similarities between data items. This preprocessing step in data mining improves the accuracy of the data analysis and the efficiency of the mining algorithm [11].

In their previous work, Praus and Praks [23] proposed applying the Latent Semantic Indexing (LSI) method for information retrieval of hydrochemical data. LSI includes two steps. In the first step, the dimensionality of the data is reduced by the projection on first principal components (which is equivalent to the application of SVD decomposition). The second step calculates the similarity between representations of water samples [3]. The representation of data in a lower dimensional space by concept decomposition is an intuitive approach that was originally developed for the purpose of text mining, where data are projected on centroids of clusters or classes [12]. Using concept decomposition instead of SVD is computationally less complex. Moreover, the literature suggests that the results of information retrieval of text documents are comparable to that of representation by LSI [13]. In our research we will apply dimensionality reduction by PCA and also use concept decomposition in a supervised setting to test this approach on the task of hydrochemical data classification. Motivated by the analyses by Praus and Praks [23], Dobsa and Dalbelo-Basic [13], and Aswani Kumar and Srinivas [5], we classify ground water samples collected from the Odra River basin in Ostrava, Czech Republic, using linear support vector machines (SVM). The paper is organized as follows. Section 2 provides problem description and reviews related work. Principal components analysis, concept decomposition and the algorithm of support vector machines are explained in Section 3. The methodology we adopted to solve the problem is introduced in Section 4. We demonstrate the proposed methodology and analyze the results obtained in Section 5.

2. Problem Description and Related Work

Hydrochemical datasets are represented as data matrices where each column of a given data matrix represents a sample composition and can be expressed as vector $X=(x_1,x_2,\ldots, x_n)$, where x_i is the i^{th} chemical parameter and n is the total number of chemical parameters being considered for analysis. Real hydrochemical data samples are noisy so the retrieval of similarities among such data items can lead to incorrect findings. Such multidimensional data requires DR techniques to remove the noise by reducing dimensionality and enable the interpretation of latent information. In the literature there are reviews of various DR techniques [2], [11]. Praus and Praks [23] applied a SVD based approach for removing noise and retrieval of similar water samples from a ground water dataset. Several applications of SVD can be found in the literature including rule mining [6].

Applications of data mining and artificial intelligence techniques for hydro data analysis are also well discussed in the literature. Ouyang applied PCA and Principal Factor Analysis (PFA) to evaluate the effectiveness of a surface water quality monitoring network in a river where monitoring stations were used as variables [20]. Brodnjak-Voncina et al. studied the quality of the Mura River water by applying standard multivariate statistical methods and PCA [9]. They concluded that the PCA method is discriminant enough even given the weak correlation among the variables. The case study by Kunwar P. Singh et al. presented the usefulness of multivariate statistical techniques for evaluation and interpretation of large complex data sets to obtain better information about water quality [18]. Together with these statistical techniques, several hybrid methods are also discussed in the literature. Sarbu and Pop proposed a fuzzy PCA method for measuring the water quality of the Danube River [25]. Their analysis demonstrated that the fuzzy PCA achieved better compressible results than the classical PCA. Razmkhah et al. used PCA and cluster analysis methods to investigate the water quality of the Jajrood River [24]. To study the spatial variation and source apportionment of water pollution in the Qiantang River, Huang et al. used Fuzzy Comprehensive Assessment (FCA) and Factor Analysis [15]. Liu et al. combined the information entropy theory and the fuzzy mathematics method for water quality assessment in the Three Gorges region in China [19]. Their analysis showed that the improved fuzzy comprehensive evaluation method is superior to the traditional model. In light of the studies cited above, the present research aims to compare the performance of classification of water samples represented in the original and a lower dimensional space obtained by PCA and

concept decomposition. The classification of groundwater samples with similar composition will be elaborated through a practical case study of the Odra River basin.

3. PCA, Concept Decomposition and SVM

In this section we briefly review the methodologies we adopted in our analysis for dimensionality reduction and classification.

3.1. Principal Components Analysis

PCA is one of the most popular multivariate statistical methods for information extraction and data compression by retaining only the important information followed by analyzing the variables structure. PCA transforms the original data variables into new, uncorrelated variables (axes), called principal components. These components in the new coordinate system are the linear combinations of the original variables and describe different sources of variation. The first principal component (PC) contains the largest variance of the original variables and passes through the center of the data. The second PC is orthogonal to the first PC and contains the second largest variance in the original data variables set. A detailed explanation and illustration on PCA can be found in [1].

3.2. Concept Decomposition

Dhillon and Modha proposed to use spherical *k*-means clustering as a means of identifying latent concepts in document collections [12]. They also proposed concept decomposition (CD), where documents are represented as projections on centroids of clusters which are called concept vectors. Dobsa and Dalbelo-Basic improved CD by using a fuzzy k-means (FCM) algorithm and tested its appropriateness for the task of information retrieval [13]. FCM algorithm is the most widely used algorithm among family of algorithms which are based on iterative optimization of fuzzy objective function [8]. Karypis and Hong [17] and Dobsa et al. [14] used a supervised version of CD which calculates concept vectors as centroids of the existing classes for the task of classification. They showed that such an approach results in better classification performance than using representation of full data matrix and representation by LSI especially when the size of the lower dimensional space is small. A weighted FCM algorithm for CD has recently been proposed in [5] as well. In [26] is proposed algorithm for detecting the principal allotment among fuzzy clusters and its application as a technique for dimensionality reduction.

Let A be $n \times m$ matrix of the hydrochemical data, where n is the number of parameters and m is the number of samples. The *concept matrix* is an $n \times k$ matrix whose j-th column is concept vector c_j, that is, $C_k = [c_1, c_2, ..., c_k]$. If we assume linear independence of the concept vectors, then it follows that the concept matrix has rank k. Now we define concept decomposition D_k of data matrix A as the least-squares approximation of A on the column space of the concept matrix C_k. Concept decomposition is an $n \times m$ matrix

$$D_k = C_k Z^*, \tag{1}$$

where Z^* is the solution of the least-squares problem given as

$$Z^* = \left(C_k^T C_k\right)^{-1} C_k^T A. \tag{2}$$

The columns of data matrix A are represented as a linear combination of the concept vectors, thereby reducing data space dimensionality.

3.3. Support Vector Machines

Support vector machine (SVM) finds a hyperplane which separates positive and negative training examples with the maximum possible margin [10]. This means that the distance between the hyperplane and the corresponding closest positive and negative examples is

maximized. A classifier of the form $sign(w \cdot x + b)$ is learned, where w is the weight vector or normal vector to the hyperplane and b is the bias. The goal of margin maximization is equivalent to the goal of minimization of the norm of the weight vector when the margin is fixed to be of the unit value . Let the training set be the set of pairs $(X_j, y_j),\ j = 1,2,\ldots,m$, where X_j are vectors representing samples, y_j are labels which take value 1 if the sample is in the observed class and –1 if the sample is not in the class, and m is the total number of samples. The problem of finding the separating hyperplane is reduced to the optimization problem of the type $\min_{w,b} \langle w, w \rangle$ subject to

$$y_j \left(\langle w, X_j \rangle + b \right) \geq 1,\ j = 1,2,\ldots,m. \tag{3}$$

4. Methodology

The classification of water samples will be performed for different representations of samples: full representation of data, representations by supervised concept decomposition and representation by projection on principal components. The methodology used for the representation of water samples in the concept space is as follows:

1. Compute centroids of classes to obtain the concept matrix C_k.
2. Compute the reduced concept decomposition matrix D_k of the data matrix as $D_k = C_k Z^*$, where Z^* is the solution of the least squares problem given as $Z^* = (C_k^{\mathrm{T}} C_k)^{-1} C_k^{\mathrm{T}} A$.

The SVM algorithm performs binary classification, classifying test samples into two classes: positive and negative. Since we are dealing with the problem of classification in five classes (i.e. five different locations) we shall transform that problem into five binary classifications and perform the classification by SVM for each class separately.

5. Experimental results and Discussion

5.1. Data Description

In this section we report on the experiments conducted on ground water samples collected from five different localities in the Ostrava region spreading across the Odra River basin. The area of this basin is approximately 6252 km^2 and the total watercourse is approximately 1360 km in length. All ground water samples were analyzed based on the parameters that are in accordance with the ISO standards, including: pH, Ammonium, Nitrate, Chloride, Sulfate, Hardness, Electric Conductivity (EC), Alkalinity, Acidity, Chemical Oxygen Demand by Permanganate (COD-Mn), Iron, Manganese, Dissolved oxygen and Aggressive carbon dioxide. Descriptive statistics of these parameters in the samples considered for the analysis are given in Table 1. In the data matrix we created the rows were constructed from the 14 aforementioned parameters, while the 95 water samples represented the columns. Projection on principal components columns of the data matrix was standardized (to have meant equal to 0 and the standard deviation equal to 1). The classification was performed on three-fold cross validation. We used the SvmLight v.5.0 (http://svmlight.joachims.org/) software with default parameters [16].

	Mean	Standard deviation	Minimum	Maximum	Standard skewness	Standard kurtosis
Ammonia (mg/l)	0.74	0.98	0.014	3.62	5.17874	1.41348
Chloride (mg/l)	34.9	15.5	12.3	90	4.73363	3.2008
COD-Mn (mg/l)	0.84	0.52	0.21	2.36	2.69893	-0.73901
CO2 aggressive (mg/l)	44.1	24.3	0.21	91.3	-0.73533	-1.99214
Nitrate (mg/l)	17.1	18.2	0.50	81.7	4.21018	1.23388
Iron (mg/l)	5.76	7.57	0.06	27.8	4.19133	-0.18692
Alkalinity (mmol/l)	1.5	0.93	0.25	4.1	1.6625	-1.69405
Manganese (mg/l)	0.463	0.490	0.06	1.76	3.09299	-1.39442
pH	6.33	0.35	5.63	7.01	-0.17751	-2.34029
Dissolved oxygen (mg/l)	4.04	3.23	0.49	11.9	3.69017	-0.78504
Sulfate (mg/l)	147	74.4	37.7	367	3.05774	-0.09845
Hardness (mmol/l)	2.20	0.83	0.83	4.4	1.49623	-1.30413
Conductivity (mS/m)	50.3	17.4	24.5	95.8	2.26373	-0.53798
Acidity (mmol/)	1.20	0.45	0.25	2.25	2.13862	-0.86485

Table 1. Descriptive statistics of the ground water samples

5.2. Results

We evaluated the classification accuracy using recall, precision and the F_1 measure. Precision *p* is a proportion of data items predicted positive that are actually positive. Recall *r* is defined as a proportion of positive data items that are predicted positive. The F_1 measure is defined as $F_1 = 2pr/(p+r)$. Table 2a and Table 2b present the results of evaluation of classification: precision, recall and F_1 measures for each class and each representation, i.e.; representation by full data matrix (full), projection on 2 centroids (CD2), projection on 5 centroids (CD5) and projection on *n* principal components P*n*, *n*=2,3,4,5,6. In the case of C2 representation, the first centroid is a joint centroid of classes 1 and 2, while the second one is a joint centroid of classes 3, 4 and 5. The best results (for precision, recall and F_1 measure for all representations) are highlighted for every class. The classification using the representation of full data yielded high results on the F_1 measure (above 80% for all classes but class 4). Classes are joined on the basis of similarity between them which is determined graphically by scatter plots, as shown in Figures 1 and 2. Figure 1 illustrates the projection on two centroids, while Figure 2 displays the projection on first two principal components.

Class 4 is the hardest one to recognize: recall for that class was significantly lower than for other classes. This can be explained by the fact that samples from that class were dispersed and are not linearly separable, which can be seen in scatter plots in Figures 1 and 2. Generally, the highest results on the F_1 measure were achieved for representation by full data for classes 1 and 2 and for representation by projections on principal components for classes 3, 4 and 5. This can be explained by the fact that measures from locations 3, 4 and 5 were interwoven and therefore the semantic representation by principal components produced better classification results in this case. This can also be substantiated by the fact that recall for classes 4 and 5 improved for representation P6 and P4, respectively, in comparison to representation by full data.

Generally, representation by projections on centroids of classes is less effective than representation by projections on principal components. Nevertheless, representation by projections on two centroids gives best results of recall for classes 1 and 3 (i.e., the largest classes forming centroids). Performance of the classification by this representation for class 2 was much lower, while classes 4 and 5 are not amenable to this representation since they are not linearly separable in a two dimensional space. For representation by projections on five centroids all classes were recognized. However, the results for measures (recall, precision and F_1 measure) were lower than representation by full data, except for precision for classes 1 and 5, which were the best for all representations. For representation achieved by projections on the first two principal components, classes 4 and 5 are not amenable, while for representation by projections on the first three components only class 4 is not amenable.

Class	Measure	Full	CD2	CD5
1	Precision	88.89 ±7.86	80.16 ± 6.25	**91.11 ± 6.85**
	Recall	**100.00 ± 0.00**	**100.00 ± 0.00**	85.56 ± 13.96
	F_1	**93.94 ± 4.29**	88.85 ± 3.94	86.97 ± 4.94
2	Precision	**100.00± 0.00**	38.33 ± 30.64	47,22± 33.56
	Recall	**87.78± 8.75**	30.00 ± 21.60	30.33± 21.60
	F_1	**93.27± 4.83**	33.33 ± 24.94	30.33± 21.60
3	Precision	81.94± 5.20	80.47± 14.97	82.22± 13.70
	Recall	81.55± 7.19	**91.07± 6.36**	32.14± 18.21
	F_1	81.47± 4.44	84.15± 6.06	42.68± 17.58
4	Precision	**100.00 ± 0.00**	0.00 ± 0.00	50.00 ± 40.82
	Recall	52.22± 22.17	0.00 ± 0.00	12.22 ± 8.75
	F_1	66.02± 17.84	0.00 ± 0.00	19.05 ± 13.47
5	Precision	91.67± 11.79	0.00± 0.00	**100.00± 0.00**
	Recall	91.67± 11.79	0.00± 0.00	55.56± 21.87
	F_1	91.67± 11.79	0.00± 0.00	68.57± 20.34

Table 2a. Results of classification for full representation and representation by concept decomposition

All the classes were recognized in the process of classification by representation P4, by which 86.38% of the data variation was explained. By projection of the original data onto the first five principal components 90.38% of the data variation was explained. Despite the fact that data variation was explained to a great extent by representation P5, classification results were significantly improved for classes 3 and 4 by representation P6 (where 93.20% of data variation was explained).

Class	Measure	P2	P3	P4	P5	P6
1	Precision	86.75 ± 9.73	81.99± 3.71	86.75± 9.73	86,25± 9.93	86.25 ± 9.93
	Recall	**100.00 ± 0.00**	**100.00± 0.00**	**100.00 ± 0.00**	96.67± 4.71	96.67 ± 4.71
	F_1	92.62± 5.46	90.06± 2.27	92.62± 5.46	90.89± 6.46	90.89 ± 6.46
2	Precision	100.00 ± 0.00	91.67± 11.79	93.33± 9.43	**100.0± 0.00**	**100.00± 0.00**
	Recall	75.55 ± 17.50	81.11± 16,41	**87.78± 8.75**	81.11± 16.41	81.11± 16.41
	F_1	85.00 ± 10.80	85.86± 14.07	90.30± 8.18	88.64± 10.33	88.64± 10.33
3	Precision	**83.81± 11.97**	**83.81± 11.97**	73.99± 8.07	81.35± 4.59	82.87± 3.98
	Recall	58.93± 2.53	58.93± 2.53	82.14± 12.71	77.38± 6.07	86.31± 11.69
	F_1	68.69± 2.86	68.69± 2.86	76.57± 1.16	79.21± 4.66	**84.20± 6.83**
4	Precision	0.00 ± 0.00	0.00 ± 0.00	**100.00± 0.00**	**100.00± 0.00**	**100.00± 0.00**
	Recall	0.00 ± 0.00	0.00 ± 0.00	46.67± 14.40	46.67± 14.40	**63.33± 17.85**
	F_1	0.00 ± 0.00	0.00 ± 0.00	62.38± 12.80	62.38± 12.80	**76.02± 14.07**
5	Precision	0.00 ± 0.00	91.67± 11.79	93.33± 9.43	93.33± 9.43	93.33± 9.43
	Recall	0.00 ± 0.00	83.33± 11.79	**100.0± 0.00**	**100.0± 0.00**	**100.00± 0.00**
	F_1	0.00 ± 0.00	86.90± 10.24	**96.30± 5.24**	**96.30± 5.24**	**96.30± 5.24**

Table 2b. Results of classification for representation by projection on principal components

Figure 3 shows macro-averages through classes of F_1 measure of SVM classification for all representations of data including full data matrix (full), projection on 2 centroids (CD2), projection on 5 centroids (CD5) and projection on *n* principal components (P*n*, *n*=2,3,4,5,6). The performance of representation by projection on four, five and six principal components was comparable to the performance of representation by full data matrix. Only P6 representation result exceeded (by 2%) the macro-average of F_1 measure of full data matrix representation.

Generally speaking, results were better for representation by projection on principal components, but separate results of recall and precision show that representation by projection on centroids of classes can also be useful in combination with other representations. Concept decomposition was not as effective in this case as it was in the case of text mining because the dimensions of data representation of hydrochemical data are generally lower. Furthermore, representations of textual documents contain more redundancy than representations of hydrochemical data. On the other hand, the advantage of CD is that it is computationally more efficient than SVD based projection on principal components [12]. The complexity of projection on principal components and concept vectors is the same, but computation of centroids of classes is much more effective than computation of principal components. Future work should investigate classification by SVM using different kernel functions and classification using decision trees.

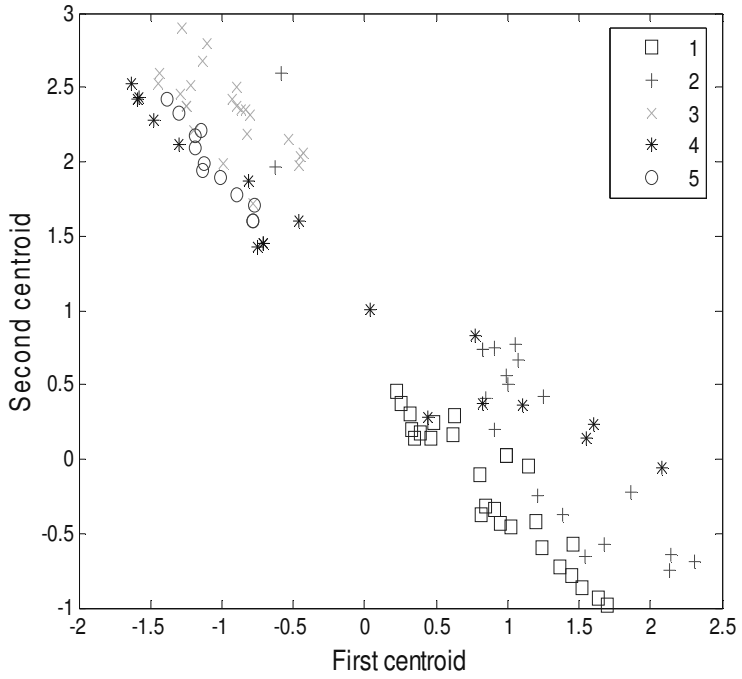

Figure 1. Projection on two centroids

6. Conclusions

This paper confirms that the representation of data in a lower dimensional space can improve classification performance by capturing latent relations between variables. Two methods for lowering the dimension of the original data representation were used: concept decomposition and projection on principal components. Representation in a lower dimensional space can improve precision (which is improved for classes 1 and 5 by concept decomposition and class 3 by projection on principal components) and recall (which is improved by concept decomposition for class 3 and for classes 4 and 5 by projection on principal components).

Acknowledgements

One of the authors, Ch. Aswani Kumar gratefully acknowledges the financial support from the National Board of Higher Mathematics, Department of Atomic Energy, Government of India, under the grant number 2/48(11)2010-R&D 11/10806. This work was also supported by the Regional Materials Science and Technology Centre (CZ.1.05/2.1.00/01.0040) and also by the framework of the IT4Innovations Centre of Excellence project, reg. no. CZ.1.05/1.1.00/02.0070 supported by the Operational Programme 'Research and Development for Innovations' funded by the European Union's Structural Funds and the state budget of the Czech Republic.

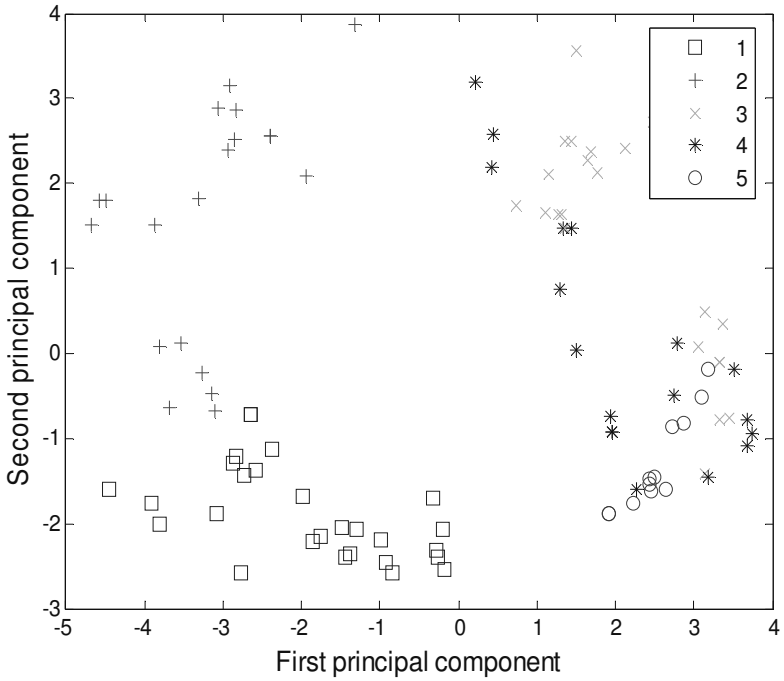

Figure 2. Projection on principal components

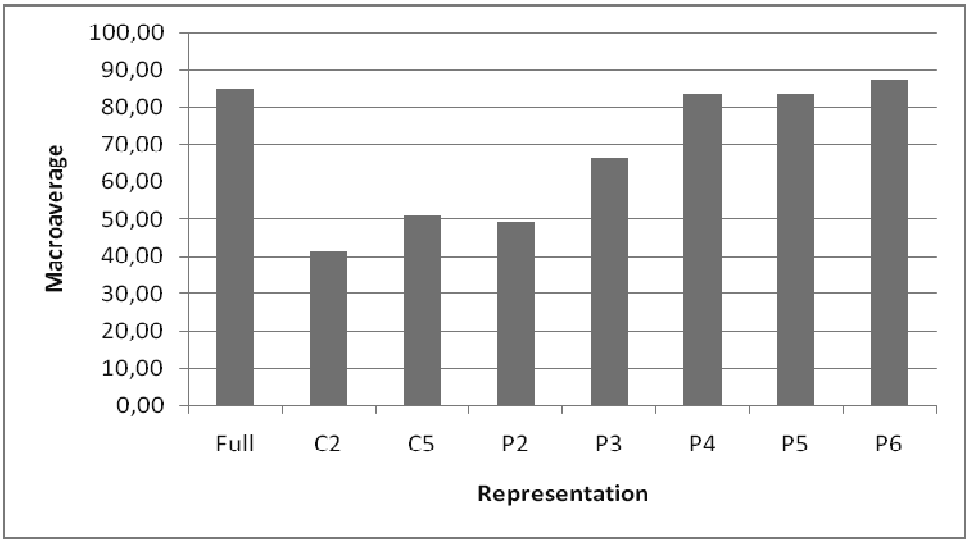

Figure 3. Macro-average of F_1 measure of classification by SVM

References

[1] Abdi, H; Williams, L.J. Principal component analysis. *WIREs Computational Statistics*, 2, 433-459, 2010.

[2] Aswani Kumar, Ch. Analysis of unsupervised dimensionality reduction techniques. *Computer Science and Information Systems*, 6(2), 217-227, 2009.

[3] Aswani Kumar, Ch; Srinivas, S. Latent semantic indexing using eigenvalue analysis for efficient information retrieval. *International Journal of Applied Mathematics and Computer Science*, 16(4), 551-558, 2006.

[4] Aswani Kumar, Ch; Srinivas, S. Concept lattice reduction using fuzzy k means clustering. *Expert Systems with Applications*, 37, 2696-2704, 2010.

[5] Aswani Kumar, Ch; Srinivas, S. A note on weighted fuzzy k means clustering for concept decomposition. *Cybernetics and Systems*, 41, 455-467, 2010.

[6] Aswani Kumar, Ch; Srinivas, S. Mining association rules in health care data using formal concept analysis and singular value decomposition. *Journal of Biological Systems*, 18(4), 787-807, 2010.

[7] Babaoglu, I., Findik, O ; Bayrak, M. Effects of principle component analysis on assessment of coronary artery disease using support vector machine. *Expert Systems with Applications*, 37(3), 2182-2185, 2010.

[8] Bezdek, J.C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[9] Brodnjak-Voncina, D; Dobcnik, D; Novic, M; Zupan, J. Chemometrics characterization of the quality of river water. *Analytica Chimica Acta,* 462, 87-100, 2002.

[10] Cristianini, N; Shawe-Taylor, J. *Support vector machines and other kernel based learning methods*. Cambridge University Press, 2000.

[11] Cunningham, P. Dimension Reduction. Technical Report: UCD–CSI _2007 -7, University College of Dublin, Ireland, 2007 .

[12] Dhillon, I.S; Modha, D.S. Concept decomposition for large sparse text data using clustering, *Machine Learning*. 42, 143-175, 2001.

[13] Dobsa, J; Dalbelo-Basic, B. Concept decomposition by fuzzy k-means algorithm. In *Proceedings of IEEE/WIC International Conference on Web Intelligence*, pages 684-688, Halifax, Canada, 2003.

[14] Dobsa, J; Mladenic, D; Grobelnik, M; Brank, J. Experimental evaluation of documents classification represented using concept decomposition. In *Proceedings of the 27th International Conference on Information Technology Interfaces*, pages 187-192, Cavtat, Croatia, 2005.

[15] Huang, F; Wang, X ; Lou, L ; Zhou, Z ; Wu, J. Spatial variation and source apportionment of water pollution in Qiantang river (China) using statistical techniques. *Water Research*, 44, 1562-1572, 2010.

[16] Joachims, T. *Learning to Classify Text Using Support Vector Machines*. Dissertation, Universität Dortmund, Kluwer Academic Publishers, 2001.

[17] Karypis, G; Hong, E. Fast dimensionality reduction algorithm with applications to document retrieval & categorization. In: *Proceedings of 9th International Conference on Information and Knowledge Management*, McLean, VA, 12-19, 2000.

[18] Kunwar P; Singh, Malik, A; Mohan, D; Sinha, S. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti river (India) – a case study, *Water Research*. 38, 3980-3992, 2004.

[19] Liu, Li., Zhou, J; An, X., Zhang, Y; Yang, Li. Using fuzzy theory and information entropy for water quality assessment in three Gorges region. *Expert Systems with Applications*, 37(3), 2517-2521, 2010.

[20] Ouyang, Y. Evaluation of river water quality monitoring stations by principal component analysis. *Water Research,* 39, 2621-2635, 2005.

[21] Praks, P; Dvorsky, J; Snasel, V; Cernohorsky, J. On SVD free latent semantic indexing for image retrieval for application in hard real industrial environment. In: *IEEE International Conference on Industrial Technology*. Slovenia, 466-471, 2003.

[22] Praks, P; Machala, L; Snasel, V. On SVD free latent semantic indexing for iris recognition of large databases. In: V. A. Petrushin and L. Khan (Eds.) *Multimedia Data Mining and Knowledge Discovery,* Springer, 472-486, 2006.

[23] Praus, P; Praks, P. Information retrieval in hydrochemical data using the latent semantic indexing approach. *Journal of Hydroinformatics*, 9(2), 135-143, 2007.

[24] Razmkhah, H; Abrishamchi, A ; Torian, A. Evaluation of spatial and temporal variation in water quality by pattern recognition techniques : A case study on Jajrood river (Tehran, Iran). *Journal of Environnemental Management*, 91, 852-860, 2010.

[25] Sarbu, C; Pop, H.F. Principal component analysis versus fuzzy principal component analysis A case study : the quality of Danube water (1985-1996), *Talanta*, 65, 1215-1220, 2005.

[26] Viattchenin, D.A. An Algorithm for Detecting the Principal Allotment among Fuzzy Clusters and its Application as a Technique of Reduction of Analyzed Features Space Dimensionality. *Journal of Information and Organizational Sciences*, 33(1), 205-217, 2009.