Pol. J. Environ. Stud. Vol. 25, No. 4 (2016), 1685-1690
DOI: *10.15244/pjoes/61671*

Original Research

Classification of Regions with Endemic Diseases Based on Trace Element Concentrations in Groundwater

Aida Sahmurova[1]*, Bunyad Shahmurzada[2]

[1]Department of Environmental Health Okan University and Department
of Environmental Engineering Azerbaijan Khazar University
[2]Kocaeli University, Faculty of Medicine, Kocaeli, Turkey

Received: 29 December 2015
Accepted: 2 February 2016

Abstract

This study measured trace elements and assessed fluoride levels in groundwater in Azerbaijan. We investigated endemic diseases in regions of Azerbaijan using the aforementioned data. Geographic regions were classified as an endemic region or not by using a support vector machine (SVM). Classification accuracy for the SVM classifier was determined to be 76.92%.

Keywords: fluoride concentration, trace elements, endemic disease, geographic classification, support vector machine

Introductıon

Natural concentrations of trace elements in water and soil change according to the geology, geomorphology, climate, and other features of a region. Since the middle of the last century, industrialization and exponential-like population growth have increased the adverse effects created by anthropogenic-related compounds and residuals that may be found in natural waters with some as trace concentrations. Trace elements are found in very low levels in an organism's structure. However, these same elements may create toxic conditions in tissue based on relatively high concentrations present in the environment. Also, regarding some organisms (including ourselves), low levels of these elements are relatively higher than required nutrient-based levels that may cause serious health problems [1].

Trace elements have both positive and negative impacts on human health. Although some elements are considered to have only toxic effects, many trace elements provide beneficial effects up to certain exposures before exhibiting detrimental effects, and in some cases the concentration window of beneficial effects is extremely narrow [2].

Human intake of drinking water can serve as a significant source of toxic chemicals. Seasonal variations, age relationships, and diet and socioeconomic determinants also can be factors that determine toxic chemical exposure [3, 4].

Groundwater can become contaminated from natural sources or numerous types of human activities. Natural sources of contamination are dissolved rocks or soils. Iron, manganese, arsenic, chlorides, fluorines, sulfates, or radionuclides are the main contamination compounds of groundwater. If the groundwater contains unacceptable concentrations of these substances, it should not

*e-mail: aida.sahmurova@okan.edu.tr

be used for drinking water or other domestic water uses unless it is treated to remove these contaminants [5, 6]. We determined trace elements in drinking water (Offa Metropolis) samples collected from tap, well, borehole, stream, and sachet water that was analyzed using atomic absorption spectrophotometry. These concentrations are within the permissible limits of the WHO drinking water quality guidelines (except some samples above the WHO threshold limit and some samples within the limit) [7].

Fluoride concentrations in drinking water are important for both bone and tooth health. Optimum concentration of fluoride is considered to be 1 mg/l with a maximum of 1.5 mg/l permitted in regulations [8-9]. Fluoride when present as an electronegative ion may create areas of high fluoride concentration in the body by connecting to calcium in both bones and teeth. Health problems that occur at various concentrations are as follows [10]:

- < 0.5 mg/l tooth sensitivity
- 0.5-1.5mg/l beneficial for tooth health
- 1.5-4mg/l dental fluorosis diseases in dental and skeleton system
- 10mg/l fluorosis diseases and breaking bones

Fluorosis has been on the rise since the 1950s and is related to high concentrations of fluoride present in drinking water or produced during coal-burning, causing damage to the human body characterized by a vast array of symptoms and pathological changes in addition to the well-known effects on the skeleton and teeth. Excessive intake of fluoride leading to fluorosis is a slow degenerative disorder affecting the structure and function of skeletal muscle, the brain, and the spinal cord [11, 12].

High fluoride concentrations are found in subsurface waters of India, China, Sri Lanka, Holland, Mexico, and in the countries of both North and South America. According to a study conducted in India, maximum fluorine concentration was measured at approximately 5.2 mg/l in 62 million people, including 6 million children who have suffered from fluorosis due to consumption of water having relatively high fluoride concentrations [13, 14].

Some studies have characterized the relationship and spatial variability of physio-chemical parameters in drinking water using multivariate statistical techniques to identify the natural anthropogenic factors controlling the distribution of these parameters and to predict levels in water [15].

Classification of geological areas to endemic-based diseases is one of the machine-learning applications. The machine-learning task has become very attractive in the last decade [16-19]. Here we have applied SVM as a tool to provide classification of endemic disease potential areas. Recently support vector machines, developed by Vapnik [20], have been used for a range of problems including pattern recognition [21, 22], bioinformatics [23, 24], and text categorization [25, 26]. The use of classification in this facet and in medical diagnosis has been gradually increasing [27-30]. SVM provides a novel approach for a two-variable classification problem [31].

SVMs are supervised learning methods used for classification and regression. An SVM will construct a separating hyperplane in a space and can be observed via display of input data as two sets of vectors in n-dimensional space in which one maximizes a margin between two data sets. Two parallel hyperplanes are created to calculate a margin, and one is on each side of the separating hyperplane placed against the two data sets. Intuitively, good separation is achieved by a hyperplane with the largest distance to neighboring datapoints for both classes. Therefore, in general the larger a margin the lower the generalization error for the classifier.

SVMs have established themselves as a technique useful for variants in a least-squares SVM and have gained increased attention in recent times due to computational benefits. Although considered high-performance models, it is consensual that the applicability of vector machines heavily depends on coping with non-trivial machine learning problems. SVM depends on proper selection of control parameters as is the case with many different models.

Materials and Methods

Groundwater samples were obtained from 10 different regions of Azerbaijan, including 65 towns according to the possibility of endemic diseases based on climate, geography, soil structure, and water features. Measurements were made in 10 different areas of the Kuba-Hachmaz and Sheki-Zakatala regions, where endemic diseases are recorded (e.g., endemic goiter) as well as in the Apsheron Peninsula, which has no recorded instance of any endemic disease.

A comparison of regions in which diseases did or did not occur were made. Samples were kept in plastic vessels. Colorimetric measurements were carried out according to Standard Methods [32] and conducted in the Water Hygiene and Sanitation Laboratories of the National Medical Prophylactic Research Institute.

Fluoride was analyzed following sodium fluoride transformation. In order to control interference during analyses, chlorine, sulfate, nitrate, and phosphate parameters were determined. In cases of high interference by these parameters, liquid samples were analyzed by colorimetric methods after a distillation process. For all other parameters, solutions were prepared according to Standard Methods.

Support Vector Machine

SVM [33] is a tool that utilizes a statistical approach. SVM has been used successfully for pattern recognition and regression tasks [34]. SVM is formulized under the concept of the structural risk minimization rule (SRM) [35], which minimizes the possible upper margin error rate over all samples. It was originally designed for binary classification in order to construct an optimal hyperplane so that the margin of separation between a negative and positive data set will be maximized. Generally, SVM is used for a two-variable pattern classification problem.

Assume there is a hyperplane that separates positive and negative values from each other. That is, there exists a linear function of the form:

$$f(x) = w^T x + b \tag{1}$$

… such that for each training example x_i, $f(x) \geq 0$ for $y_i = +1$ and $f(x) < 0$ for $y_i = -1$. Thus

$$f(x) = w^T x + b = 0 \tag{2}$$

SVM finds a hyperplane that maximizes the separating margin between two classes. (Fig. 1). Mathematically, this hyperplane can be found by minimizing the following cost function:

$$P(w) = \frac{1}{2}\|w\|^2 \tag{3}$$

Subject to separability constraints:

$$w^T x_i + b \geq +1 - \varepsilon_i \quad \text{for} \quad y_i = +1$$

or

$$w^T x_i + b \leq -1 + \varepsilon_i \quad \text{for} \quad y_i = -1 \tag{4}$$

…where ε_i represents slack variables and $\varepsilon_i \geq 0\ \forall i$.

If $\varepsilon_i > 1$, then an error occurs. Therefore, a C regularization parameter has been added to increase the cost function. If the value of the C parameter is high, then more penalties will be given to errors.

Accordingly, the cost function in (3) can be modified as follows:

$$P_c = \frac{1}{2}\|w\|^2 + C\sum_i \varepsilon_i \tag{5}$$

subject to

$$y_i = (w^T x_i + b) \geq 1 - \varepsilon_i \text{ and } \varepsilon_i > 0 \ \forall_i \tag{6}$$

In order to find a solution for the problem above, w and b values which minimize the Lagrange function below have to be calculated:

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_i \varepsilon_i - \sum_i \alpha_i \left\{ y_i (w^T x + b) - 1 + \varepsilon_i \right\} - \sum_i \mu_i \varepsilon_i \tag{7}$$

…where μ_i has been added to make ε_i positive and α_i is the positive Lagrange multiplier. Lagrange multipliers α_i are solved from the dual form of (7), which is expressed as

$$L_D = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j \tag{8}$$

and subject to

$$0 \leq \alpha \leq C \text{ and } \sum_i \alpha_i y_i = 0 \tag{9}$$

One will notice that cost function L_D is convex and quadratic in terms of the unknown parameters α_i. In practice this problem is solved numerically through quadratic programming.

In the nonlinear case, SVM will not achieve sufficient or useful classification. Kernel approaches were developed to overcome this limitation with SVM. In nonlinear SVM, the input data set is converted into a high-dimensional linear feature space via equation (10), and the exponential radial basis function (ERBF) kernel is displayed in equation (11):

$$K(x, x_i) = (x.x_i + 1)^p \tag{10}$$

$$\mathrm{K}(\mathrm{x}, \mathrm{x}_i) = \exp[\gamma|\mathrm{x}\text{-}\mathrm{x}_i|] \tag{11}$$

In our experiment, p = 2 (i.e., equivalent to a quadratic classifier), C = 200, and γ = 0.5. These values were selected via a trial-and-error approach. As a result, SVM can be expressed as follows:

$$f(x) = sign\left(\sum_i y_i \alpha_i k(x, x_i) + b\right) \tag{12}$$

Parameters are found by maximizing the function below:

$$L_D = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j k(x_i x_j) \tag{13}$$

…subject to (9).

Results

SVM Training and Model Selection

We have applied SVM in a fashion useful for predicting the presence or absence of endemic diseases for regions

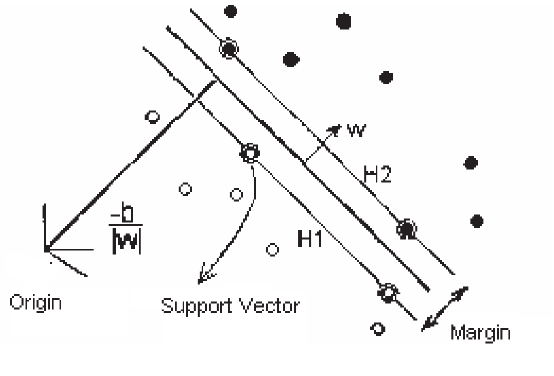

Fig. 1. Linear separation of hyperplanes for the separable case.

Table 1. Performance comparison using different k-fold cross-validation values.

Number of k-fold cross-validation	Classification accuracy (%)
k = 5	70.76
k = 8	72.45
k = 10	76.92
k = 15	73.84

of Azerbaijan. A total of 65 subjects were used for training and testing. We aimed to classify regions based on seven inputs (i.e., trace element concentrations). Outputs are represented by 0 (i.e., not an endemic disease region) or 1 (i.e., an endemic disease region). A 10-fold cross-validation procedure was used for training and testing the SVM classifier under various model and parametric settings. Different k fold cross-validation results are provided in Table 1, where the best result for classification accuracy is obtained for *k = 10*. Therefore, a 10-fold cross validation procedure was chosen for this study.

We used the polynomial function kernel and the RBF kernel for training the SVM classifier. For these kernels we found that for the best selection of parameter values for an acceptable error level for this case use p = 2, C = 200, and $\gamma = 0.5$.

Performance Evaluation

We have used four methods for performance evaluation of classification of regions. These are analysis of sensitivity, specificity, total classification accuracy, and confusion matrix. Total classification accuracy, sensitivity, and specificity are defined as [36]:

- Total classification accuracy is the number of correct decisions / total number of cases.
- Sensitivity is the number of true positive decisions / number of actual positive cases.
- Specificity is the number of true negative decisions / number of actual negative cases.

Classification results from this study are also displayed using a confusion matrix. In a confusion matrix, each cell contains the raw number of exemplars classified for a corresponding combination of desired and actual network outputs. The confusion matrix is provided in Table 2.

Table 2. Confusion matrix results using SVM.

Actual	Predicted		
	Results (normal)	Results (having endemic)	Correctly Predicted (%)
Results (normal)	38	8	82.6
Results (having endemic)	7	12	63.15

According to the confusion matrix results, seven out of 19 subjects having endemic disease potential were classified incorrectly by the network as normal subjects. However, eight out of 56 of the normal subjects were classified incorrectly by the network as subjects having endemic disease potential. Finally, those subjects without and with endemic disease potential were classified correctly as 82.6 and 63.15%, respectively. Total classification accuracy for the SVM for this case was determined to be 76.92%.

Discussion

Literature reviews show that:

- Trace elements play an important role in the functioning of normal physiological processes in human and animal organisms.
- High or low regional trace element concentrations may provide appropriate conditions for the development of various endemic diseases.
- The relationship among trace elements and also between each living organism and trace element is specific in nature.
- Negative relationships between trace elements and organisms (failure of adaptation) result in the development of non-infectious somatic diseases (endemic goiter, caries, dental fluorosis, etc.).

By taking into consideration the relationship between humans and environmental factors, researchers have determined the daily trace element requirement met by the water and nutrients consumed by living things in regions free from endemic diseases such as biogeochemical states; i.e., regions with high trace element concentrations.

The natural concentration of trace elements in water and soil changes according to the geological, geomorphological, and climatic characteristics of the region.

In the present study, trace elements were measured in groundwater samples obtained across 10 regions of Azerbaijan and the level of fluorine was assessed and regions were investigated with the considered data obtained, and these regions were classified as either “endemic” or “non-endemic” using SVM.

Conclusion

Trace element concentrations in a region are related to geologic and environmental characteristics. However, in recent years due to industrialization and exponential human population growth, natural resources have become increasingly polluted in terms of quality and are decreasing in terms of quantity [37]. The results indicate that – depending on the natural characteristics and soil structure of a region – none of the fluoride concentrations in artesian waters exceeded the limits specified by the regulations. Regarding spring waters: mean fluoride levels in the Karabagh region exceeded the limit of “2.4 mg/l” specified in the regulations, while other element concentrations did not exceed the specified limits. Results for well waters showed that, in various provinces of the Apsheron Peninsula, fluoride concentrations exceeded the "2.4 mg/l" maximum permissible value specified in the regulations. This is thought to result from the characteristics of the mineral water springs and the geological structure of the region as the Apsheron Peninsula – one of the regions with high fluoride concentration that is surrounded on three sides by the Caspian Sea.

One result of this is that we are seeing an increase of trace elements present in natural waters, and tooth fluorosis diseases are observed in these regions. Therefore, we can conclude that SVM is a useful method for classifying regions with endemic disease potential based on concentrations of trace elements found present in groundwater for cases like this one.

Acknowledgements

Authors wish to express gratitude to Dr. Niyazı Kilic for useful help in constraction of mathematical model.

References

1. LU S.Y., ZHANG H.M., SOJINU S.O., LIU G.H., ZHANG J.Q., NI H.G. Trace elements contamination and human health risk assessment in drinking water from Shenzhen, China. Environ Monit Assess. **187** (1), 4220. doi: 10.1007/s10661-014-4220-9. Epub **2014** Dec 17.
2. FRANCESCONI K.A., KUEHNELT D., KOKARNIG S,. RABER G. Elemental speciation analysis in human health assessment Journal of Trace Elements in Medicine and Biology, **27** (1), 5, **2013**.
3. JANE L. Drinking Water Intake Evaluations-a Review for Developed and Developing Countries Journal of Trace Elements in Medicine and Biology, **27** (1), 29, **2013**.
4. ŁAGOCKA R., SIKORSKA-BOCHIŃSKA J., NOCEŃ I., JAKUBOWSKA K., GÓRA M., BUCZKOWSKA-RADLIŃSKA J. Influence of the mineral composition of drinking water taken from surface water intake in enhancing regeneration processes in mineralized human teeth tissue Pol. J. Environ. Stud. **20** (2), 412, **2011**.
5. BLANES P.S, BUCHHAMER E.E, GIMÉNEZ M.C. Natural contamination with arsenic and other trace elements in groundwater of the Central-West region of Chaco, Argentina.J Environ Sci Health A Tox Hazard Subst Environ Eng. **46** (11), 1197, **2011**.
6. TOKATLI B.C., KÖSE E., ÇIÇEK A., Groundwater quality of Türkmen mountain,Turkey Pol.J.Environ.Stud. **22** (4), 1197, **2013**.
7. JIMOH W.L.O. AND SHOLADOYE Q.O. Trace elements as ındıcators of qualıty of drınkıng water ın offa metropolıs, kwara state, nıgerıa Bayero Journal of Pure and Applied Sciences, **4** (2), 103, **2011**.
8. AINCHIL K. Fluoride variations in groundwater of an area in Buenos Aires Province, Argentina. Environmental Geology **44**, 86, **2003**.
9. FRENCKEN J.E. (editor). Endemic Fluorosis in developing countries, causes, effects and possible solutions. NIPG-TNO, Leiden, The Netherlands, Publication number 91.082, **1992**.
10. DISSANAYEKE C.B. The fluoride problem in the groundwater of Sri Lanka-environmental management and health. Intl. J. Environ. Studies. **19**, 195, **1991**.
11. ADEBAYO O.L., SHALLIE P.D., SALAU B.A. AJANI E.O, A. ADENUGA G. Comparative study on the influence of fluoride on lipid peroxidation and antioxidants levels in the different brain regions of well-fed and protein undernourished rats" Journal of Trace Elements in Medicine and Biology **27**, 370, **2013**.
12. LIU L., ZHANG Y., GU H., ZHANG K., MA L. Fluorosis induces endoplasmic reticulum stress and apoptosis in osteoblasts in vivo, Biol Trace Elem Res. **164**, 64, **2015**.
13. LATHA S., AMBIKA S.R., PRASAD S.J. Fluoride contamination. status of ground water in Karnataka. Curr. Sci. **6**, 730, **1999**.
14. BA Y., ZHANG H., WANG G., WEN S., YANG Y., ZHU J., REN L., YANG R., ZHU C., LI C., CHENG X., CUI L. Association of Dental Fluorosis with Polymorphisms of Estrogen Receptor Gene in Chinese Children, Biol Trace Elem Res. **143**, 87, **2013**.
15. MUSTAPHA A., ARIS A.Z. Multivariate statistical analysis and environmental modeling of heavy metals pollution by industries. Pol. J. Environ. Stud. **21** (5), **2012**.
16. GUMUS E., KILIC N., SERTBAS A., UCAN O.N. Evaluation of face recognition techniques using PCA, wavelets and SVM, Expert Systems with Applications **37** (9), 6404, **2010**.
17. TARTAR A., KILIC N., AKAN A. Classification of Pulmonary Nodules by Using Hybrid Features, Computational and mathematical methods in medicine, vol. Article ID 148363, **2013**.
18. MERT A., KILIÇ N., AKAN A., Evaluation of bagging ensemble method with time-domain feature extraction for diagnosing of arrhythmia beats, Neural Computing and Applications **24** (2), 317, **2014**.
19. MERT A., KILIÇ N., AKAN A. An improved hybrid feature reduction for increased breast cancer diagnostic performance, Biomedical Engineering Letters **4** (3), 285, **2014**.
20. VAPNIK V. The nature of statistical lerning theory. (New York, Springer-Verlag). **1995**.
21. YEH Y.R., HUANG C.H., WANG Y.C.F., Heterogeneous Domain Adaptation and Classification by Exploiting the Correlation Subspace, IEEE Transactions on Image Processing, **23** (5), **2014**.
22. MARSICO M., NAPPI M., RICCIO D., WECHSLER H., Robust Face Recognition for Uncontrolled Pose and Illumination Changes, IEEE Transactions on Systems, Man, and Cybernetics: Systems, **43** (1), **2013**.
23. DAI H.L. Imbalanced Protin Data Classification Using Ensemble FTM-SVM, IEEE Transactions on NanoBioscience, **14** (4), **2015**.
24. ANGADI U.B., VENKATESULU M., Structural SCOP Superfamily Level Classification Using Unsupervised Machine Learning, IEEE/ACM Transactions on Computational Biology and Bioinformatics, **9** (2), **2012**.
25. CAI D., HE X., ADAPTIVE M. Experimental Design for Text Categorization, IEEE Transactions on Knowledge and Data Engineering, **20** (4), **2012**.
26. KUMAR M.A., GOPAL M. An Investigation on Linear SVM and its Variants for Text Categorization, International Conference onMachine Learning and Computing, **2010**.
27. NIAF E., FLAMARY R., ROUVIERE O., LARTIZIEN C., CANU S. Kernel-Based Learning From Both Qualitative and Quantitative Labels: Application to Prostate Cancer Diagnosis Based on Multiparametric MR Imaging, IEEE Transactions on Image Processing, **23** (3), **2014**.

28. ALVAREZ I., GORRIZ J.M., RAMIREZ J., SALAS-GONZALEZ D., LOPEZ M., PUNTONET C.G., SEGOVIA F. Alzheimer›s diagnosis using eigenbrains and support vector machines, Electronics Letters, **45** (7), **2009**.
29. WANG Y., CROOKES D., ELDIN O.S., WANG S., HAMILTON P., DIAMOND J. Assisted Diagnosis of Cervical Intraepithelial Neoplasia (CIN), IEEE Journal of Selected Topics in Signal Processing, **3** (1), **2009**.
30. BARAKAT N., BRADLEY A.P., BARAKAT M.N.H. Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus, IEEE Transactions on Information Technology in Biomedicine, **14** (4), **2010**.
31. POLAT K., GUNEŞ S. Breast cancer diagnosis using least square support vector machine, Digital Signal Processing, **17** (4), 694, **2007**.
32. GREENBERG A.E., CLESCERI L.S, EATON A.D. Standarts methods for the examination water and wastewater. (Washington, American Public Health Association) **1992**.
33. BURGES C. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery. **2**, 1, **1998**.
34. VAPNIK, V. An overview of statistical learning theory. IEEE Transaction on Neural Network. **10**, 989, **1999**.
35. CHANG C.C., LIN C.J. LIBSVM: a library for support vector machines. software available at http://www.csie.ntu.edu.tw/~cjilin/libsvm, **2001**.
36. UBEYLI E.D. Combining eigenvector methods and support vector machines for detectin variability of Doppler ultrasound signals. Computer methods and programs in biomedicine, **86**, 181, **2007**.
37. SAHMUROVA A., ONGEN A., ELMASLAR E., BALKAYA N. Determination of Micro-elements in Surface Water Resources. Journal of Residuals Science & Technology, **4** (3), July **2007**.