35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Half-Against-Half Structure in Classification of Benthic Macroinvertebrate Images

Henry Joutsijoki[1]

Abstract— **Benthic macroinvertebrates play a key role when water quality assessments are made. Benthic macroinvertebrates are difficult to identify and their identification need special expertise. Furthermore, manual identification is slow and expensive process. This paper concerns benthic macroinvertebrate classification when Half-Against-Half (HAH) structure was applied to Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Minimum Mahalanobis Distance Classifier (MMDC) classifiers. Especially, LDA, QDA and MMDC classifiers were for first time applied with HAH structure to benthic macroinvertebrate classification. We performed thorough experiments altogether with ten methods. In the case of HAH-SVM we managed to improve classification results from the earlier research by using a different approach to class division problem. We obtained 96.1% classification accuracy with Radial Basis Function (RBF) kernel. Moreover, new variants of LDA, QDA and MMDC classification methods achieved 89.5% and 91.6% classification accuracies which can be considered as a good result in such a difficult classification task.**

I. Introduction

Pure water is not a matter of course in our Globe although we use it in our daily needs. Major disasters and different kinds of environmental problems are a constant threat and they remind us about the importance of pure water since it is an essential part for all living organisms. Benthic macroinvertebrates are small organisms without backbones that inhabit the bottom substrates of their habitats, for at least part of their life cycle [22]. Common habitats for benthic macroinvertebrates are rivers, lakes and streams. Benthic macroinvertebrates are used in water quality assessments due to their ability to react changes in freshwater ecosystems. They are excellent indicators of water quality changes for long-term studies while commonly used chemical samples give only a short-term point of view of the situation of a freshwater ecosystem.

The use of benthic macroinvertebrates in water quality assessments requires their identification. Currently, identification process is made manually by biologists or taxonomists. Automated benthic macroinvertebrate identification [6]–[10], [12]–[14], [16], [17], [19], [20] has gained a scant attention among computer scientists, but it can save resources and enable wider and more efficient biomonitoring.

SVM has attracted researchers and practitioners for several years and it was chosen as a primary classification method for this paper due to its excellent results in earlier researches (see, for example, [6]–[10], [14], [19]). In this paper we used HAH-SVM [10], [15], HAH-LDA, HAH-QDA and HAH-MMDC to classify benthic macroinvertebrate images. In [10] HAH-SVM was used successfully for the classification and the aim of this paper was to improve the results of [10] and to apply, for the first time, LDA, QDA and MMDC classifiers with HAH structure in benthic macroinvertebrate classification. Applying HAH structure with LDA, QDA and MMDC classifiers give a new and an interesting point of view to these traditional classification methods. Moreover, HAH-SVM and the application of HAH structure with other classification methods are still quite less researched area and the construction of HAH structure itself consists of a fascinating and challenging theoretical problem.

In Section II we explain briefly the theory of binary Support Vector Machines and Half-Against-Half structure. In Section III dataset and test arrangements are explained and the results are presented and analyzed. Section IV is left for discussion and conclusions.

II. Methods

A. Support Vector Machine

Let us have a given training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ where $\mathbf{x}_i \in \mathbb{R}^n$ are the training examples and $y_i \in \{-1, 1\}$ are the corresponding class labels of $\mathbf{x}_i$, $i = 1, 2, \ldots, l$. SVM finds a hyperplane separating classes with maximum margin. Support vectors, which are the closest points of hyperplane, determine the value of margin $\frac{2}{\|\mathbf{w}\|}$ (see for details [1], [3], [21]). An optimal hyperplane can be found by solving an optimization problem in the dual form:

$$\max L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{l} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \quad (1)$$

subject to $\sum_{i=1}^{l} \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$ where C is a user-defined parameter (also called box constraint). A new example $\mathbf{x}$ can be classified according to the sign of the decision function

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i \langle \mathbf{x}, \mathbf{x}_i \rangle + b. \quad (2)$$

For linearly non-separable data kernels can be used. More specifically, kernels are $K(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$ where ϕ is a nonlinear transformation. Commonly used (also in this paper) kernels are: linear kernel $\langle \mathbf{x}, \mathbf{z} \rangle$, polynomial kernels $(1 + \langle \mathbf{x}, \mathbf{z} \rangle)^{deg}$ where $deg \in \mathbb{N}$ is the order of the kernel. Furthermore, there are Radial Basis Function (RBF) $\exp(-\|\mathbf{x} - \mathbf{z}\|^2 / 2\sigma^2)$ with $\sigma > 0$ and Sigmoid kernel $\tanh(\kappa \langle \mathbf{x}, \mathbf{z} \rangle + \delta)$ with $\kappa > 0$ and $\delta < 0$. All valid kernels need to satisfy

[1]H. Joutsijoki is with School of Information Sciences, University of Tampere, FI-33014 Tampere, Finland Henry.Joutsijoki@uta.fi

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

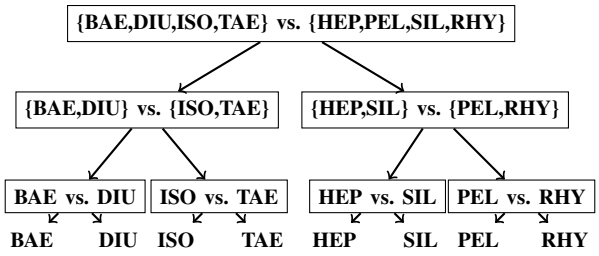

Fig. 1. Half-Against-Half structure used for classification.

Fig. 2. Example images on benthic macroinvertebrates. The order of taxonomic groups of benthic macroinvertebrates from top left to down right is BAE, DIU, HEP, PEL, SIL, ISO, RHY and TAE.

the conditions of Mercer's theorem [1], [3], [21]. When a kernel is used, the inner products in (2) are replaced with $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ and a new example is classified according to the sign of decision function. More details concerning theory of binary SVMs can be found, for instance, from [1], [3], [21].

B. Half-Against-Half Architecture

HAH-SVM is a multi-class extension of SVM introduced by Lei and Govindaraju [15]. Generally speaking, HAH structure uses a binary tree form where in each node there is a binary classifier. In this paper we used HAH structure with SVM [1], [3], [21], LDA [2], QDA [2] and MMDC [2] classifiers. From used classifiers LDA, QDA and MMDC are for first time applied with HAH structure in this paper and especially in benthic macroinvertebrate classification.

Classification begins from the root node and continues via the left or right edge until a leaf is reached where the final class label for a test example can be found. Theoretically, the most challenging problem in HAH structure is to find the right class subsets in nodes. In [15] hierarchical clustering was used for solving this problem and in [6] Scatter method [11] and random division were applied to the problem. In this paper, we used a different approach to this problem compared to methods in [10] and [15]. The idea was to collect those taxonomic groups together which had similar exterior features in images and, furthermore, at the same time class subsets within a node would be as balanced as possible. By this means, we obtained an HAH structure which can be seen in Figure 1.

III. Experimental results

A. Data Description and Test Arrangements

Our dataset contains 1350 images from eight taxonomic groups of benthic macroinvertebrates. Table I shows more specifically the taxonomic groups of the dataset and their sizes and proportions in the dataset. Seven of all taxonomic groups were identified to a species level and one, *Isoperla* sp., was identified only to a genus level being also the largest taxonomic group in the dataset. Benthic macroinvertebrate samples were collected first by biologists and the specimens were identified by biologists. After that the specimens were scanned by a flatbed scanner and each one of them was saved as an individual image. More details concerning preprocessing of the images can be found from [20]. An example image of each taxonomic group included to the dataset is in Figure 2.

The dataset contains altogether 25 features and 15 of them was selected to be used in this paper. The selected 15 features were the best ones according to the results in [10] for HAH-SVM. Moreover, we use a 15D abbreviation on this feature set. The 15D feature set included seven statistical features (they can also be called intensity based features) and eight geometrical features. These were {Mean, Standard deviation, Mode, Median, Integrated density, Kurtosis, Skewness} and {Area, Perimeter, Width, Height, Feret's Diameter, Major, Minor, Circularity}. All the features were extracted from the images by using ImageJ program. Detailed information concerning the features and the ImageJ program can be found from [5].

The dataset was divided 100 times into training, validation and test sets so that 10% was left to validation and testing and 80% for training. Training, validation and test sets were the same as used in [10]. Before classification the columns of the whole dataset were standardized to have a mean of zero and unit variance. Other transformations were not made. Optimal parameter values for HAH-SVM were determined according to the mean accuracy (accuracy is here determined as a trace of a confusion matrix divided by the sum of all elements in confusion matrix) of validation sets. When the optimal parameters were found, SVMs were trained again with the training data including validation set. In addition, since the HAH-SVM includes several binary SVMs, for each classifier we consider that parameter values are the same. This approach was proposed in [4].

Polynomial kernels including the linear kernel were tested with 100 parameter values and RBF and Sigmoid kernels were tested with 10000 parameter value combinations. For Sigmoid we made an agreement of $\kappa = -\delta$ due to the computational reasons. Parameter value space for box constraint (C), σ and κ was $\{0.1, 0.2, \ldots, 10\}$. For δ the corresponding parameter value space was $\{-10.0, -9.9, \ldots, -0.1\}$.

As a final result a mean confusion matrix was evaluated. Results are presented in percentages and classification rates (also known as true positive rate or sensitivity) and accuracy were the main measures. These two measures are presented in Table III. Also, we present standard deviations of accuracies and classification rates. In Table III we boldfaced the best classification rate for each class and the best accuracy for making reading easy for a reader. All the tests were made by using Matlab 2010b with Bioinformatics Toolbox and Statistics Toolbox. Furthermore, in the case of HAH-SVM we applied the binary SVM implementation of Matlab in Bioinformatics Toolbox and LDA, QDA and MMDC

implementations in Statistics Toolbox of Matlab in our tests. Moreover, Least Squares method [18] was used in finding optimal hyperplane for SVM. All the tests were performed with an Asus G53SX laptop having 16GB of memory and Core i7 2.0GHz processor.

TABLE I
FREQUENCIES AND PERCENTAGES OF BENTHIC MACROINVERTEBRATE CLASSES IN DATASET.

Class		Size	%
Baetis rhodani	BAE	116	8.6
Diura nanseni	DIU	129	9.6
Heptagenia sulphurea	HEP	172	12.7
Hydropsyche pellucidulla	PEL	102	7.6
Hydropsyche siltalai	SIL	271	20.0
Isoperla sp.	ISO	311	23.0
Rhyacophila nubila	RHY	83	6.1
Taeniopteryx nebulosa	TAE	166	12.3

TABLE II
OPTIMAL PARAMETER VALUES FOR HAH-SVM WITH DIFFERENT KERNEL FUNCTIONS.

Kernel	C	σ	κ	δ
Linear	7.4	–	–	–
Polynomial $deg = 2$	2.5	–	–	–
Polynomial $deg = 3$	0.2	–	–	–
Polynomial $deg = 4$	0.1	–	–	–
Polynomial $deg = 5$	0.1	–	–	–
RBF	9.3	2.1	–	–
Sigmoid	0.6	–	0.1	−0.1

B. Results

Table III shows the results of HAH-LDA, HAH-QDA, HAH-MMDC and HAH-SVM. Let us consider first the three discriminant based classification methods. With HAH-LDA all classes were recognized well except class HEP which had classification rate of 77.7±7.7%. Otherwise, classification rates were at the interval of 85.6%-93.9% where the highest classfication rate was for class ISO, the largest class in the dataset. Mean accuracy of 89.5±2.2% was gained by HAH-LDA and it was the fifth highest among all mean accuracies. HAH-MMDC is an interesting case because it achieved the same mean accuracy than HAH-LDA, but the classification rates had in most cases significant changes. From class BAE to class ISO a trend was seen compared to HAH-LDA results that HAH-MMDC classified those classes better what HAH-LDA did not and vice versa. Classes RHY and TAE achieved classification rates within 1% compared to HAH-LDA. HAH-QDA succeeded better than HAH-LDA or HAH-MMDC in the classification. Mean accuracy was around 2% higher than those of the previous methods. Classes BAE and HEP were classified at the level between HAH-LDA and HAH-MMDC. An interesting detail was that the smallest class in the dataset, class RHY, was now classified with 96.3±5.9% classification rate which is significantly higher than with HAH-LDA or HAH-MMDC. Mean accuracy of 91.6% was the fourth highest among all methods.

HAH-SVM with the Sigmoid kernel was again the poorest alternative for classification. The same situation also occurred in [6]–[10] where other multi-class extensions of SVM were examined. Also, HAH-SVM with the 5th degree polynomial kernel achieved clearly a lower mean accuracy compared to other methods tested. With the linear kernel classification rates were very close to HAH-LDA results in many cases. Only a bit larger difference to HAH-LDA results was in class BAE where the difference was around 2%. Mean accuracy was 89.4% with the linear kernel. A mean accuracy of 89.3% was achieved by the 4th degree polynomial kernel but in classes DIU, PEL and RHY differences to the linear kernel results were significant.

Results with the quadratic, cubic and RBF were very good. Class BAE was classified with the highest classification rate (92.4%) among all methods tested with quadratic kernel. All other classes were also identified above 90% classification rates except class DIU which obtained 88.8% result. Mean accuracy was 94.3% and it was third highest of all methods. Cubic kernel achieved 94.2% mean accuracy in [10] so almost the same accuracy was achieved by the new class division. Quadratic and RBF kernels were the only ones which gained above 95% mean accuracy. In [10] quadratic kernel obtained 93.3% mean accuracy now 2.0% improvement was achieved. Furthermore, classes HEP, PEL and TAE were classified best with the quadratic kernel. RBF kernel was the best alternative for classification of benthic macroinvertebrates. It obtained the highest classification rates in classes DIU, SIL, ISO and RHY. Every class gained above 92.0% classification rate. Moreover, mean accuracy was 96.1±1.4% and it was 0.2% higher than in [10].

IV. CONCLUSIONS

In this paper we applied Half-Against-Half structure to SVM classifiers and to LDA, MMDC and QDA methods for the first time in benthic macroinvertebrate image classification. Our aim was to classify benthic macroinvertebrate images and to improve earlier results. Benthic macroinvertebrates play an important role in water quality monitoring and, thus, they are important also for humans. We performed wide experimental tests with 15D feature set. This feature set was chosen to this paper, because it obtained the best results in [10] where HAH-SVM was examined with four different feature sets and with two class division methods which were Scatter method [11] and random division for point of comparison. Compared to the results in [10], we obtained better results with all kernel functions. The new class division method used in this paper was based on visual information gained from the images and taking care that class subsets within a node have as equal number of instances as possible.

The new variant for LDA, QDA and MMDC succeeded relatively well in the classification having 89.5% or higher accuracy. However, the best accuracies were achieved by HAH-SVM with quadratic and RBF kernels which both gained above 95.0% accuracies. HAH structure is a very promising general technique for classification problems and

TABLE III

RESULTS (%) OF HAH-SVM WITH DIFFERENT KERNELS AND HAH-LDA, HAH-MMDC AND HAH-QDA.

Method/Class	BAE	DIU	HEP	PEL	SIL	ISO	RHY	TAE	Mean accuracy
HAH-LDA	91.9	85.6	77.7	87.0	92.1	93.9	89.4	92.6	89.5
StdDev	7.8	9.5	8.7	6.8	3.8	4.3	9.7	5.6	2.2
HAH-MMDC	81.8	94.1	91.8	96.2	85.3	89.0	90.2	92.3	89.5
StdDev	10.7	6.1	6.8	4.7	6.4	6.0	8.8	6.0	2.3
HAH-QDA	88.8	94.5	83.9	95.4	94.4	89.6	96.3	93.7	91.6
StdDev	8.6	5.8	7.8	4.9	3.9	5.6	5.9	4.8	2.0
HAH-SVM Linear	89.7	84.8	79.1	86.7	92.2	93.6	89.4	93.3	89.4
StdDev	9.0	10.3	8.7	6.9	3.7	4.4	9.8	5.0	2.2
HAH-SVM Pol. $deg = 2$	88.1	91.7	**95.6**	**98.5**	97.4	95.1	96.8	**97.4**	95.3
StdDev	8.8	7.1	5.6	3.8	3.0	3.8	5.6	3.6	1.6
HAH-SVM Pol. $deg = 3$	**92.4**	88.8	90.9	94.3	96.4	95.4	96.9	97.0	94.3
StdDev	7.3	8.1	5.6	7.0	3.8	3.4	5.3	3.9	1.8
HAH-SVM Pol. $deg = 4$	89.8	76.1	81.1	81.0	93.4	94.7	92.6	94.4	89.3
StdDev	9.0	11.8	8.8	11.1	4.7	3.9	7.9	5.9	2.5
HAH-SVM Pol. $deg = 5$	79.8	52.2	63.3	53.7	88.1	85.8	87.9	85.4	77.3
StdDev	11.4	13.6	12.8	12.7	6.0	5.5	8.8	9.2	3.1
HAH-SVM RBF	92.1	**95.5**	93.7	98.0	**98.5**	**96.3**	**97.8**	95.8	**96.1**
StdDev	7.9	4.7	5.0	4.4	2.2	3.3	4.8	4.7	1.4
HAH-SVM Sigmoid	49.6	54.6	35.1	40.5	44.1	45.9	40.6	51.3	45.3
StdDev	17.8	16.2	13.9	15.4	13.9	10.6	20.0	17.0	7.9

it can be applied to numerous applications. In future concentration will be given to examination of larger benthic macroinvertebrate dataset.

ACKNOWLEDGMENT

The author is thankful for Maj and Tor Nessling Foundation and Oskar Öflund Foundation for the support. The author wants to thank Finnish Environment Institute, Jyväskylä, Finland for the data.

REFERENCES

[1] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition,"' *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 121–167, 1998.

[2] K.J. Cios, W. Pedrycz, R.W. Swiniarski and L.A. Kurgan, *Data Mining: A Knowledge Discovery Approach*, Springer-Verlag, 2007.

[3] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, Vol. 20, No.3, pp. 273–297, 1995.

[4] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines,"' *IEEE Transactions on Neural Networks*, Vol. 13, No. 2, pp. 415–425, 2002.

[5] ImageJ: public domain Java-based image processing program. Available: http://rsbweb.nih.gov/ij/

[6] H. Joutsijoki and M. Juhola, "Automated benthic macroinvertebrate identification with decision acyclic graph support vector machines,"' *Proceedings of the 2nd International Conference on Computational Bioscience*, pp. 323–328, 2011.

[7] H. Joutsijoki and M. Juhola, "Comparing the one-vs-one and one-vs-all methods in benthic macroinvertebrate image classification,"' *Lecture Notes in Artificial Intelligence*, 6871, Springer-Verlag, pp. 399–413, 2011.

[8] H. Joutsijoki and M. Juhola, "DAGSVM vs. DAGKNN: An experimental case study with benthic macroinvertebrate dataset,"' *Lecture Notes in Artificial Intelligence*, 7376, pp. 439–453, 2012.

[9] H. Joutsijoki and M. Juhola, "Kernel selection in multi-class support vector machines and its consequence to the number of ties in majority voting method,"' *Artificial Intelligence Review*, DOI:10.1007/s10462-011-9281-3, In press.

[10] H. Joutsijoki, "Half-Against-Half multi-class support vector machines in classification of benthic macroinvertebrate images,"' *Proceedings of 2012 International Conference on Computer and Information Science (ICCIS 2012)*, IEEE, Vol. 1, pp. 414–419, 2012.

[11] M. Juhola and M. Siermala,"A scatter method for data and variable importance evaluation," *Integrated Computer-Aided Engineering*, Vol. 19, No. 2, pp. 137–149, 2012.

[12] S. Kiranyaz, M. Gabbouj, J. Pulkkinen, T. Ince and K. Meissner, "Network of evolutionary binary classifiers for classification and retrieval in macroinvertebrate databases," *Proceedings of 2010 IEEE 17th International Conference in Image Processing*, pp. 2257–2260, 2010.

[13] S. Kiranyaz, M. Gabbouj, J. Pulkkinen, T. Ince and K. Meissner, "Classification and Retrieval on Macroinvertebrate Image Databases using Evolutionary RBF Neural Networks," *Proceedings of the International Woshop on Advanced Image Technology*, 2010.

[14] S. Kiranyaz, T. Ince, J. Pulkkinen, M. Gabbouj, J. Ärje, S. Kärkkäinen, V. Tirronen, M. Juhola, T. Turpeinen, K. Meissner, "Classification and retrieval on macroinvertebrate image databases," *Computers in Biology and Medicine*, Vol. 41, No.7, pp. 463–472, 2011.

[15] H. Lei and V. Govindaraju, "Half-against-half multi-class support vector machines," *Lecture Notes in Computer Science*, 3541, Springer-Verlag, pp. 156–164, 2005.

[16] D.A. Lytle, G. Martinez-Muñoz, W. Zhang, N. Larios, L. Shapiro, R. Paasch, A. Moldenke, E.N. Mortensen, S. Todorovic and T.G. Dietterich, "Automated processing and identification of benthic invertebrate samples," *Journal of North American Benthological Society*, Vol. 29, No.3, pp. 867–874, 2010.

[17] M.J. Sarpola, R.K. Paasch, E.N. Mortensen, T.G. Dietterich, D.A. Lytle, A.R. Moldenke and L.G. Shapiro, "An aquatic insect imaging system to automate insect classification," *Transactions of the ASABE*, Vol. 51, No.6, pp. 2217-2225, 2008.

[18] J.A.K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, Vol. 9, pp. 293–300, 1999.

[19] V. Tirronen, A. Caponio, T. Haanpää and K. Meissner, "Multiple order gradient feature for macro-invertebrate identification using support vector machines," *Lecture Notes in Computer Science*, 5495, pp. 489–497, 2009.

[20] J. Ärje, S. Kärkkäinen, K. Meissner and T. Turpeinen, "Statistical classification and proportion estimation - an application to a macroinvertebrate image database," *Proceedings of 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010)*, pp. 373–378, 2010.

[21] V.N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd edition, Springer-Verlag, 2000.

[22] Watershedss: A decision support system for nonpoint source pollution control. http://www.water.ncsu.edu/watershedss/ Accessed 11.1.2013.