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

Machine Learning Approach to an Otoneurological Classification Problem

Henry Joutsijoki[1], Kirsi Varpa[2], Kati Iltanen[3] and Martti Juhola[4]

Abstract— **In this paper we applied altogether 13 classification methods to otoneurological disease classification. The main point was to use Half-Against-Half (HAH) architecture in classification. HAH structure was used with Support Vector Machines (SVMs), *k*-Nearest Neighbour (*k*-NN) method and Naïve Bayes (NB) methods. Furthermore, Multinomial Logistic Regression (MNLR) was tested for the dataset. HAH-SVM with the linear kernel achieved clearly the best accuracy being 76.9% which was a good result with the dataset tested. From the other classification methods HAH-*k*-NN with cityblock metric, HAH-NB and MNLR methods achieved above 60% accuracy. Around 77% accuracy is a good result compared to previous researches with the same dataset.**

I. Introduction

Vertigo can be a symptom of many different diseases having overlapping symptoms which makes diagnosis of a vertiginous patient challenging [9]. Machine learning methods can be a valuable tool for diagnostic purposes. By means of machine learning and data mining we can find patterns from datasets which are collected from the earlier cases. A physician can use the information obtained from a dataset when making a final diagnosis. However, the difficulty of otoneurological diseases set up a challenge for the methods used.

In [16] One-vs-All (OVA) and One-vs-One (OVO) methods were used with SVM [2] and *k*-NN [12] classification methods. Compared to OVA and OVO methods HAH structure [7], [10] has advantages, especially in the low number of classifiers and the computational efficiency. HAH has been applied for instance to benthic macroinvertebrate classification in [7] and the promising results were a motivation for this paper. The use of HAH structure is a novel approach to otoneurological disease classification for this dataset.

In this work we applied Half-Against-Half structure with SVMs, *k*-NN and NB [11] classifiers to the classification. Furthermore, we tested also MNLR [1] for the classification problem. The paper has following structure. In Section II Half-Against-Half method is explained and SVM is introduced. Section III is left to results and data description. Section IV is for conclusions and discussion.

II. Methods

A. Half-Against-Half Structure

Half-Against-Half (HAH) architecture was originally developed for Support Vector Machines and it was introduced in the article by Lei and Govindaraju [10]. HAH is a general classification architecture which also can be used with other classification methods than SVM. It uses a binary tree where each one of the nodes includes a binary classifier. In this paper we applied SVM, *k*-NN and NB classifiers to HAH architecture.

Classification of a test example begins from the root and continues via edges until a leaf is reached. In the leaf there is a predicted class label for a test example. The main issue with HAH structure is to find the correct way to divide classes into two subsets in nodes. Some examples for solving this problem are, for instance, hierarchical clustering [10] and Scatter method [8]. When the number of classes in a dataset is small even a random division can be used.

At first, to test how the HAH structure works with the above mentioned classifiers on the otoneurological data, we created one example HAH architecture (Figure 1) that we tested with all the methods. The class divisions into subsets were made based on the disease descriptions and their similarities described in [5], [9] and confusion matrices of previous researches to search for the most similar disease class. The diseases reported similar were collected into the same group. At the same time two groups within a node were tried to keep as balanced by the number of cases as possible. The class division into similar groups is challenging with otoneurological data because all of the diseases have more or less similar symptoms and, in addition, data contains also cases having confounding symptoms, for example, benign positional vertigo cases can have age-related hearing loss.

B. Support Vector Machine

Suppose that we have a labeled training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$. Optimal hyperplane minimizes $\frac{1}{2}\|\mathbf{w}\|^2$ with respect to constraints $y_i[\langle \mathbf{w}, \mathbf{x} \rangle + b] \geq 1$. A hyperplane can be found with the easiest way by solving so called Wolf dual form [2], [14]:

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

with respect to $\sum_{i=1}^{l} \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$ where C is a user-defined parameter and α_i's are the same Lagrange coefficients as in primal form (see details [2]). A new

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

[2] K. Varpa is with School of Information Sciences, University of Tampere, FI-33014 Tampere, Finland Kirsi.Varpa@uta.fi

[3] K. Iltanen is with School of Information Sciences, University of Tampere, FI-33014 Tampere, Finland Kati.Iltanen@uta.fi

[4] M. Juhola is with School of Information Sciences, University of Tampere, FI-33014 Tampere, Finland Martti.Juhola@uta.fi

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

example $\mathbf{x}$ can be classified according to the sign of the decision function

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

Kernel functions, $K(\mathbf{x},\mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$ where ϕ is a nonlinear mapping to a higher dimensional space, are used for classification when the data is not linearly separable. Typically in literature used kernels which are used in this paper as well 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 degree of the kernel, 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 $\delta < 0$. Valid kernels satisfy conditions presented in Mercer's theorem [2], [3]. The use of kernels modifies a decision function so that the inner products in (2) are replaced with $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. Otherwise, the classification of a new example goes similarly as before.

III. Experimental results

A. Data Description and Test Arrangements

An otoneurological dataset used in this research contains 1030 vertigo cases from nine different vertigo diseases (Table I). There are 94 attributes from which 77 attributes are qualitative, mostly binary, and 17 quantitative. Attributes describe a patient's health status: occurring symptoms, medical history and clinical findings (otoneurologic, audiologic and imaging tests). Clinical tests are not done to every patient and, therefore, there were missing values in several test results. In total, the data had about 11% missing values, which allowed using imputation. Imputation was needed due to the calculation of the SVM method. Missing values of qualitative attributes were substituted with the class modes and other attributes with the class medians.

Dataset was first split to training and test sets by using 10 times 10-fold cross-validation. For the search of optimal parameters for SVMs, every training set was split to training and validation sets by using 3-fold cross-validation. 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 (shown in Table II), SVMs were trained again with the full training data. Because HAH-SVM includes several binary SVMs, we applied the procedure given in [6] where all SVMs are trained with the same parameter value.

Polynomial kernels including the linear kernel were tested with 100 parameter values and RBF and Sigmoid kernels were tested 10000 parameter value combinations. For Sigmoid we made an agreement of $\kappa = -\delta$. The parameter value space for C, σ and κ was $\{0.1, 0.2, \ldots, 10\}$. For parameter δ it was $\{-10.0, -9.9, \ldots, -0.1\}$. In the case of k-NN, odd values from 1 to 9 were tested with each distance measure. The distance measures used with the k-NN classifier were Cityblock, Correlation, Cosine and Euclidean measures [17]. The best k value was determined according to the mean accuracy of validation sets.

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 and k-NN implementations of Matlab in Bioinformatics Toolbox and Naïve Bayes and Multinomial Logistic Regression implementations in Statistics Toolbox. For Naïve Bayes kernel density estimation [4], [13] was applied. Least Squares method [15] was used in finding optimal hyperplane for SVM.

TABLE I

Frequencies and percentages of disease classes in the dataset.

Disease name		Size	%
Acoustic Neurinoma	ANE	131	12.7
Benign Positional Vertigo	BPV	173	16.8
Menière's Disease	MEN	350	34.0
SUDD Deafness	SUD	47	4.6
Traumatic Vertigo	TRA	73	7.1
Vestibular Neuritis	VNE	157	15.2
Benign Recurrent Vertigo	BRV	20	1.9
Vestibulopatia	VES	55	5.3
Central Lesion	CL	24	2.3

TABLE II

Optimal parameter values for HAH-SVM with different kernel functions.

Kernel	C	σ	κ	δ
Linear	0.1	–	–	–
Polynomial $deg = 2$	0.1	–	–	–
Polynomial $deg = 3$	0.1	–	–	–
Polynomial $deg = 4$	0.1	–	–	–
Polynomial $deg = 5$	0.1	–	–	–
RBF	8.5	10.0	–	–
Sigmoid	0.3	–	0.1	-0.1

B. Results

As a final result a mean confusion matrix was evaluated. True positive rates (TPR) and total accuracies (in percentages) were the main evalution measures. These measures are presented in Table III. Moreover, the standard deviation of accuracies and TPRs are shown. In Table III we boldfaced the best TPR and accuracy for each class to ease the analysis of results.

Total accuracies for different HAH-k-NN combinations varied from 47.6% to 61.5%. HAH-NB and MNLR had total accuracies of 65.9% and 68.3%, respectively. For different HAH-SVM combinations, total accuracies varied from 7.8% to 76.9%. HAH-SVM with linear kernel yielded the highest total accuracy (76.9±3.5%) and the highest TPRs for six classes (BPV, SUD, TRA, VNE, BRV and CL). For the classes ANE, MEN and VES, HAH-SVM produced the second, third and fourth highest TPRs, respectively. The highest total accuracy is at the same level with OVA SVM

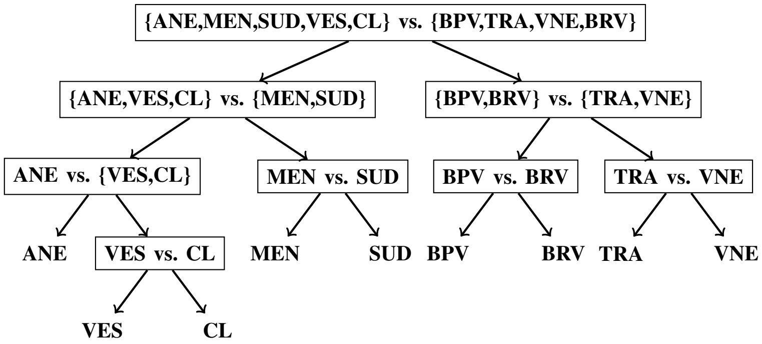

Fig. 1. Half-Against-Half structure used for otoneurological disease classification.

linear result in [16], but other classifiers in [16] obtained higher mean accuracies than 76.9%.

The dataset had a dichotomy class distribution. Four classes (ANE, BPV, MEN and VNE) had above 130 instances in the dataset while the rest of the classes had only below 80 instances in a dataset. With HAH-*k*-NN for all measures, HAH-NB, HAH-SVM with the quadratic, cubic and RBF kernels the largest classes obtained the highest TPRs. Furthermore, in HAH-SVM results with the linear kernel only class TRA obtained higher TPR than the aforementioned classes. The smallest classes, BRV and CL, were in most cases recognized below 20% TPR. The only exception was HAH-SVM with the linear kernel by which 38.0±32.7% result was obtained for class BRV. For class BRV in [16] the highest TPR was 21.0% with the linear kernel and OVA method, so improvement was gained for this class with HAH structure. In [16] 28.5% TPR was achieved on class CL with the RBF kernel and OVO method and now with the linear kernel 17.5±25.7% result was obtained.

In the case of VES higher TPRs were achieved in [16] with the RBF kernel and OVO method (TPR 22.8%) and with the 5-NN OVA classifier (TPR 20.7%) than with any of the methods used in this research. Now, with the quadratic kernel the highest TPR 17.1±16.2% was obtained for VES. For class VNE TPRs in [16] were ranging from 81.4% to 88.1%. Only HAH-SVM with the linear kernel reached a similar level having 87.6±6.6% TPR. Class TRA had interesting results because only two methods, HAH-SVM with the linear kernel and MNLR, were able to beat the limit of 70.0% whereas in [16] all seven methods were able to gain above 70.0% TPR. A surprising results was gained for class SUD where HAH-SVM with the linear kernel and MNLR were the only methods, which got above 57.0% TPR. Especially, the results of *k*-NN with all measures were exceptionally low compared to results in [16] where 94.3% TPR was achieved with 5-NN and OVO method.

Menière's disease (class MEN) was recognized better with HAH-NB (96.4%) and HAH-SVM with the RBF kernel (95.5%) than with OVO or OVA methods in [16]. Further, HAH-*k*-NN with cityblock metric and HAH-SVM with the linear kernel reached above 80.0% TPR. BPV was identified with 70.9% TPR by using HAH-SVM with the linear kernel when in [16] most of the results were better than 70.9%. Class ANE achieved 90.8±9.0% TPR with MNLR and 89.4±8.3% result with HAH-SVM and linear kernel. The results of these two methods were at the same level as OVA and OVO methods in [16].

IV. Discussion and Conclusions

This study showed the preliminary results with one HAH tree combined with SVM, *k*-NN and NB classifiers. HAH structure with *k*-NN classifier did not achieve as good results as were achieved with 5-NN OVA and OVO in [16]. One reason might be the used distance measures. In [16], *k*-NN was used with HVDM [17] measure whereas in this study HAH-KNN was used with four other measures. HVDM takes into account qualitative attributes properly, apart from the measures used in this study. Furthermore, the size of *k* (especially for the classes BRV, VES and CL) and the structure of the HAH tree itself could have effected to HAH-*k*-NN results.

In the case of HAH-SVM the simplest method was the best alternative. The linear kernel achieved 76.9±3.5% accuracy being the best method in this study and the only method which was comparable with the results of OVA and OVO methods in [16]. The large deviation of mean accuracies ranging from 7.8% to 76.9% shows the importance to search widely and thoroughly for the best classification method and to try to develop new perspectives to traditional classification methods.

Overall, the smallest classes were also the most difficult classes to recognise in this study. The two smallest classes, BRV and CL, were in most cases recognized below 20% TPR which is understandable since the 10-fold cross-validation was used and then these classes might have only from 2 or 3 cases in a test set. If one of these cases was misclassified, it decreased TPR greatly. Moreover, BRV, VES and CL are difficult to identify due to nature of the disease. For instance patients with BRV have been reported to develop typical Meniere's disease (MEN) and benign positional vertigo during years.

Because the structure of the HAH tree is crucial for the classification results, in the future we will test several combinations of different tree structures and classification methods to find the most appropriate way to divide the

TABLE III

RESULTS (%) WITH DIFFERENT CLASSIFICATION METHODS.

Method/Class	ANE	BPV	MEN	SUD	TRA	VNE	BRV	VES	CL	Accuracy
HAH-k-NN (Cityblock) $k = 7$	85.5	59.9	80.8	10.1	17.7	72.5	0.0	7.0	0.0	61.5
StdDev	8.6	10.2	6.2	12.1	13.4	10.8	0.0	10.1	0.0	3.8
HAH-k-NN (Correlation) $k = 9$	68.3	44.7	68.8	10.4	2.3	46.6	0.0	2.1	8.7	47.6
StdDev	10.0	9.9	6.5	13.5	5.9	12.1	0.0	5.8	17.3	3.5
HAH-k-NN (Cosine) $k = 9$	68.5	44.6	69.5	11.0	1.6	46.9	0.0	2.0	8.7	47.9
StdDev	10.1	10.0	6.9	13.4	4.5	12.1	0.0	5.6	17.3	3.6
HAH-k-NN (Euclidean) $k = 7$	72.2	41.5	66.9	4.3	11.5	55.9	0.0	6.9	0.0	48.8
StdDev	11.3	10.4	7.1	8.7	10.7	11.8	0.0	10.6	0.0	4.1
HAH-NB	66.6	52.9	**96.4**	1.9	46.9	80.9	0.0	0.0	3.7	65.9
StdDev	12.1	12.1	3.0	6.7	16.5	9.8	0.0	0.0	12.7	3.5
HAH-SVM Linear	89.4	**70.9**	86.1	**66.3**	**90.8**	**87.6**	**38.0**	9.0	**17.5**	**76.9**
StdDev	8.3	9.2	5.1	22.3	11.1	6.6	32.7	12.0	25.7	3.5
HAH-SVM Pol. $deg = 2$	63.6	52.3	54.3	38.0	43.8	57.3	10.5	**17.1**	13.8	50.4
StdDev	13.3	11.8	8.6	21.0	16.9	12.2	20.5	16.2	23.7	4.7
HAH-SVM Pol. $deg = 3$	48.3	47.6	40.5	19.8	39.4	53.6	11.0	7.5	12.2	40.7
StdDev	14.4	11.0	9.0	18.7	17.6	12.2	23.1	10.4	20.9	4.6
HAH-SVM Pol. $deg = 4$	33.9	39.2	31.3	14.6	35.3	45.9	9.5	9.2	10.2	32.6
StdDev	14.0	10.1	7.6	16.0	16.9	12.4	21.0	11.9	19.4	3.9
HAH-SVM Pol. $deg = 5$	25.8	32.7	24.8	12.5	26.4	38.4	10.5	6.8	6.8	26.3
StdDev	11.6	10.8	7.5	14.4	16.7	12.4	21.7	10.6	16.9	3.9
HAH-SVM RBF	17.2	22.5	95.5	0.0	8.8	30.9	0.0	5.6	0.0	44.0
StdDev	9.8	8.3	2.9	0.0	8.9	11.0	0.0	9.6	0.0	2.8
HAH-SVM Sigmoid	4.5	20.0	0.0	18.0	42.0	0.0	7.0	1.0	8.0	7.8
StdDev	18.9	40.2	0.0	38.6	49.6	0.0	25.6	10.0	26.3	4.7
MNLR	**90.8**	65.1	73.6	57.9	70.4	78.2	1.5	16.0	17.2	68.3
StdDev	9.0	24.1	25.2	28.5	27.3	27.1	8.6	15.9	28.1	18.9

classes into subsets. The Scatter method, hierarchical clustering and confusion matrices of different classification methods will be used to form class divisions. A special interest will also be given to a hybrid HAH structure where all nodes in a tree do not consist of the same classification method, but for every node an optimal method is searched for. This may improve classification results.

ACKNOWLEDGMENT

The first author is thankful for Maj and Tor Nessling Foundation and Oskar Öflund Foundation for the support. The second author is thankful for Päijät-Häme Regional Fund of Finnish Cultural Foundation for the support. The authors are thankful to Erna Kentala, M.D., and prof. Ilmari Pyykkö, M.D., for their help in data collection during these years and their valuable aid in domain expertise.

REFERENCES

[1] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, 1990.

[2] 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.

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

[4] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition, Springer, 2008.

[5] M. Havia, "Menière's Disease Prevalence and Clinical Picture,"' Ph.D. dissertation, Department of Otorhinolaryngology, University of Helsinki, Finland, 2004.

[6] 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.

[7] 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.

[8] 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.

[9] E. Kentala, "A Neurotologic Expert System for Vertigo and Characteristics of Six Otologic Diseases Involving Vertigo,"' Ph.D. dissertation, Department of Otorhinolaryngology, University of Helsinki, Finland, 1996.

[10] 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.

[11] D.D. Lewis, "Naive Bayes at forty: The independence assumption in information retrieval,"' *Lecture Notes in Computer Science*, 1398, pp. 4-15, 1998.

[12] T. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[13] Y. Murakami and K. Mizuguchi, "Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites,"' *Bioinformatics*, Vol. 26, No. 15, pp. 1841–1848, 2010.

[14] W. Ruihu, F. Bin, H. Zhangping, C. Liang and W. Weihua, "Cascaded SVMs in Pattern Classification for Time-Sensitive Separating,"' *Proceedings on Third international Symposium on Intelligent information Technology and Security Informatics (IITSI)*, pp. 640–644, 2010.

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

[16] K. Varpa, H. Joutsijoki, K. Iltanen and M. Juhola, "Applying one-vs-one and one-vs-all classifiers with k-nearest neighbour method and support vector machines to an otoneurological multi-class problem,"' *Studies in Health Technology and Informatics*, IOS Press, Vol. 169, pp. 579–583, 2011.

[17] R.D. Wilson and T.R. Martinez, "Improved heterogeneous distance functions,"' *Journal of Artificial Intelligence Research* Vol. 6, pp. 1–34, 1997.