Active Learning Algorithm Using the Discrimination Function of the Base Classifiers

Robert Burduk[(✉)]

Department of Systems and Computer Networks, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
`robert.burduk@pwr.edu.pl`

Abstract. The goal of the Active Learning algorithm is to reduce the number of labeled examples needed for learning. In this paper we propose the new AL algorithm based on the analysis of decision profiles. The decision profiles are obtained from the outputs of the base classifiers that form an ensemble of classifiers. The usefulness of the proposed algorithm is experimentally evaluated on several data sets.

Keywords: Active Learning · Query by Committee · Multiple classifier system

1 Introduction

Supervised learning is one of the types of machine learning [1]. Classification methods are applied in many practical tasks [5,8,10]. Generally, the recognition algorithm maps the feature space to the set of class labels. This process requires a sufficiently large number of training examples in the learning set. Typically these examples are manually labelled by the expert (sometimes called - oracle). On the other hand unlabelled examples are much easier to be acquired.

Active Learning (AL) [6] is a special case of machine learning in which a learning algorithm is able to interactively query the expert to obtain a label for unlabelled examples. These labelled examples are further used to improve a classifier. The key issue is to select the most informative examples. In this paper we use Query by Committee (QBC) approach [11]. This approach to AL is based on using the ensemble of classifiers to select the most informative examples.

The text is organized as follows: after this introduction, in Sect. 2 the idea of an ensemble of classifiers is presented. Section 3 contains the description of the proposed AL scheme based on the QBC approach. The experimental results on several data sets are presented in Sect. 4. Finally, conclusions from the experiments and future research proposals are presented.

2 Ensemble of Classifiers

The classification task can be accomplished by a single classifier or by a team of classifiers. In the literature, the use of the multiple classifiers for a decision

problem is known as the multiple classifier systems (MCS) or the ensemble of classifiers EoC [7,12]. The construction of MSC consists of three phases: generation, selection and integration [3]. In the second phase, which is discussed in this paper, one classifier or a subset of the base classifiers is selected to make the final decision which is to assign an object to the class label.

The output of an individual classifier can be divided into three types [13]:

– The abstract level – the classifier ψ assigns the unique label j to the given input x.
– The rank level – in this case for each input x, each classifier produces an integer rank array. Each element within this array corresponds to one of the defined class labels. The array is usually sorted with the label at the top being the first choice.
– The measurement level – the output of a classifier is represented by a measurement value that addresses the degree of assigning the class label to the given output x. An example of such a representation of the output is a posteriori probability returned by Bayes classifier.

Let us assume that we possess K of different classifiers $\Psi_1, \Psi_2, \ldots, \Psi_K$. Such a set of classifiers, which is constructed on the basis of the same learning sample is called an ensemble of classifiers or a combining classifier. However, any of Ψ_i classifiers is described as a component or a base classifier. As a rule K is assumed to be an odd number and each of Ψ_i classifiers makes an independent decision. As a result, of all the classifiers' actions, their K responses are obtained. Having at the disposal a set of base classifiers one should determine the procedure of making the ultimate decision regarding the allocation of the object to the given class. It implies that the output information from all K component classifiers is applied to make the ultimate decision.

In this work we consider the situation when each base classifier returns the estimation of a posteriori probability. This means that the output of all the base classifiers is at the measurement level. Let us denote a posteriori probability estimation (most often discrimination function – DF) by $p_k(\omega|x)$, $k = 1, 2, \ldots, K$, $\omega = 1, 2, \ldots, \Omega$, where Ω is the number of the class labels. One of the possible methods for such outputs is the linear combination method. This method makes use of the linear function like Sum, Prod or Mean for the combination of the outputs. In the sum method the score of the group of classifiers is based on the application of the following sums:

$$s_i(x) = \sum_{k=1}^{K} p_k(\omega|x), \qquad \omega = 1, 2, \ldots, \Omega. \tag{1}$$

The final decision of the group of classifiers is made following the maximum rule and is presented accordingly, depending on the sum method (1):

$$\Psi_S(x) = \arg\max_i s_i(x). \tag{2}$$

In the presented method (2) DF obtained from the base classifiers take an equal part in building MCSs. This is the simplest situation in which we do not

need additional information on the testing process of the base classifiers except for the models of these classifiers. One of the possible methods in which weights of the base classifier are used is presented in [4].

3   Proposal of Active Learning Algorithm Using the Decision Profile

The general AL scheme is defined as follows [15]:

Input: Learning algorithm - A; Set of labeled training examples - L; Set of unlabeled training examples - U; Number of active learning iterations - n; Number of selected examples - m.

Repeat n times

1. Generate a committee of classifiers, $C^* = EnsembleMethod(A, L)$
2. $\forall x_j \in U$ compute $InformationValue(C^*, x_j)$, based on the current committee
3. Select a subset S of m examples that are the most informative
4. Obtain a label for examples in S from "oracle" or an expert
5. Remove examples in S from U and add to L

Return $EnsembleMethod(A, L)$

3.1   Proposal of Information Value Calculations

For K base classifier their outputs are arranged in the decision profile:

$$DP(x) = \begin{bmatrix} p_1(1|x) & \vdots & p_1(\Omega|x) \\ \vdots & \vdots & \vdots \\ p_K(1|x) & \vdots & p_K(\Omega|x) \end{bmatrix}. \tag{3}$$

During learning of the base classifiers we obtain m decision profiles, where m is the number of objects from the learning set. From the decision profiles we calculate the decision scheme according to the formula:

$$DS = \begin{bmatrix} ds_{11} & \vdots & ds_{1\Omega} \\ \vdots & \vdots & \vdots \\ ds_{K1} & \vdots & ds_{K\Omega} \end{bmatrix}, \tag{4}$$

where

$$ds_{k\omega} = \frac{\sum_{n=1}^{m} I(\Psi_k(x_n) = \omega_n)\, p_k(\omega_n|x_n)}{\sum_{n=1}^{m} I(\Psi_k(x_n) = \omega_n)}, \tag{5}$$

where $I(\cdot)$ is the indicator function. The value of $ds_{k\omega}$ is calculated only from those DFs for which the classifier k did not make an error.

The information value is calculated for the new object on the basis of its decision profile according to the formula:

$$IV(x) = \sum_{k=1}^{K} \sum_{\omega=1}^{\Omega} I(p_k(\omega|x) < ds_{k\omega}). \tag{6}$$

The obtained information value IV for the new object from a set of an unlabelled training example is used to indicate the object x to be labelled. The labelled object is the one that exceeds the established value of the IV. In the experiments the algorithm using the proposed above method is denoted as Ψ_{AL}^{IV}.

4   Experimental Studies

In the experiential research we used two data sets from the UCI repository [9], one data set form Keel repository and the two generated randomly – they are the so called Banana and Higleyman sets. The numbers of attributes and examples are presented in Table 1. In the experiments we have used the standard 10-fold-cross-validation method and the feature selection process [14] was not performed.

Table 1. Description of data sets selected for the experiments

Data set	Example	Attribute
Banana	2000	2
Highleyman	400	2
MAGIC Gamma Telescope	19020	11
Phone	961	6
Pima Indians Diabetes	768	8

In the experiments 9 base classifiers were used. One of them (labelled as Ψ_1) used the decision trees algorithms, with the splitting rule based on entropy, the number of branches equal to 2 and the depth of the precision tree having at most 6 levels. Two of them work according to $k - NN$ rule where k parameter is equal to 3 or 5 and they are labelled as Ψ_2 and Ψ_3 respectively. The classifier labelled as Ψ_4 is the rule induction classifier. This classifier uses a tree-based modelling in which a tree is run and models for which the purity is at or above a specified threshold are removed from the training data set and placed on the side. The fifth classifier Ψ_5 uses Support Vector Machines models with the Decomposed Quadratic Programming estimation method. The sixth classifier Ψ_6 uses the least squares regression model to predict the class label. The last of the base classifiers Ψ_7 is the multilayer perceptron model with 3 hidden units.

Tables 2, 3, 4, 5, 6, 7, 8, 9, 10 and 11 show the results of the classification for all base classifiers and two ensemble methods. The results refer to the first iteration of AL scheme and are for information value equal 12 or 15.

Table 2. Classification accuracy for the base classifiers $(\Psi_1, ..., \Psi_9)$, Majority Voting and Sum method for AL scheme with $IV = 12$ and random labelling of unlabelled training examples - Pima data set

AL scheme	Ψ_1	Ψ_2	Ψ_3	Ψ_4	Ψ_5	Ψ_6	Ψ_7	Ψ_8	Ψ_9	Ψ_{MV}	Ψ_{AL}^{12}
$k = 0$	0.25	0.25	0.22	0.21	0.40	0.28	0.26	0.27	0.28	0.18	0.21
$k = 1$	0.22	0.22	0.40	0.27	0.22	0.27	0.25	0.31	0.30	0.26	0.25
Random	0.25	0.25	0.40	0.22	0.35	0.28	0.26	0.35	0.30	0.25	0.28

Table 3. Classification accuracy for the base classifiers $(\Psi_1, ..., \Psi_9)$, Majority Voting and Sum method for AL scheme with $IV = 15$ and random labelling of unlabelled training examples - Pima data set

AL scheme	Ψ_1	Ψ_2	Ψ_3	Ψ_4	Ψ_5	Ψ_6	Ψ_7	Ψ_8	Ψ_9	Ψ_{MV}	Ψ_{AL}^{15}
$k = 0$	0.25	0.25	0.22	0.21	0.40	0.28	0.26	0.27	0.28	0.18	0.21
$k = 1$	0.25	0.25	0.40	0.21	0.40	0.28	0.26	0.27	0.28	0.18	0.21
Random	0.34	0.34	0.40	0.23	0.35	0.43	0.26	0.25	0.23	0.30	0.26

Table 4. Classification accuracy for the base classifiers $(\Psi_1, ..., \Psi_9)$, Majority Voting and Sum method for AL scheme with $IV = 12$ and random labelling of unlabelled training examples - Banana data set

AL scheme	Ψ_1	Ψ_2	Ψ_3	Ψ_4	Ψ_5	Ψ_6	Ψ_7	Ψ_8	Ψ_9	Ψ_{MV}	Ψ_{AL}^{12}
$k = 0$	0.28	0.28	0.45	0.42	0.16	0.14	0.27	0.14	0.13	0.21	0.12
$k = 1$	0.18	0.18	0.45	0.51	0.12	0.14	0.32	0.12	0.13	0.15	0.13
Random	0.24	0.24	0.45	0.42	0.13	0.14	0.28	0.11	0.11	0.16	0.12

Table 5. Classification accuracy for the base classifiers $(\Psi_1, ..., \Psi_9)$, Majority Voting and Sum method for AL scheme with $IV = 15$ and random labelling of unlabelled training examples - Banana data set

AL scheme	Ψ_1	Ψ_2	Ψ_3	Ψ_4	Ψ_5	Ψ_6	Ψ_7	Ψ_8	Ψ_9	Ψ_{MV}	Ψ_{AL}^{15}
$k = 0$	0.28	0.28	0.45	0.42	0.16	0.14	0.27	0.14	0.13	0.21	0.12
$k = 1$	0.18	0.18	0.45	0.45	0.15	0.14	0.30	0.14	0.12	0.18	0.13
Random	0.31	0.31	0.45	0.41	0.17	0.14	0.25	0.13	0.13	0.22	0.12

Table 6. Classification accuracy for the base classifiers $(\Psi_1, ..., \Psi_9)$, Majority Voting and Sum method for AL scheme with $IV = 12$ and random labelling of unlabelled training examples - Higleman data set

AL scheme	Ψ_1	Ψ_2	Ψ_3	Ψ_4	Ψ_5	Ψ_6	Ψ_7	Ψ_8	Ψ_9	Ψ_{MV}	Ψ_{AL}^{12}
$k = 0$	0.05	0.05	0.2	0.17	0.12	0.05	0.05	0.12	0.05	0.05	0.05
$k = 1$	0.05	0.05	0.5	0.15	0.05	0.05	0.07	0.02	0.05	0.05	0.05
Random	0.07	0.07	0.5	0.2	0.12	0.07	0.1	0.05	0.05	0.05	0.05

Table 7. Classification accuracy for the base classifiers $(\Psi_1, ..., \Psi_9)$, Majority Voting and Sum method for AL scheme with $IV = 15$ and random labelling of unlabelled training examples - Higleman data set

AL scheme	Ψ_1	Ψ_2	Ψ_3	Ψ_4	Ψ_5	Ψ_6	Ψ_7	Ψ_8	Ψ_9	Ψ_{MV}	Ψ_{AL}^{15}
k = 0	0.05	0.05	0.2	0.17	0.12	0.05	0.05	0.12	0.05	0.05	0.05
k = 1	0	0	0.5	0.17	0.12	0.1	0.12	0.07	0.1	0.1	0.1
Random	0.17	0.17	0.67	0.17	0.1	0.05	0.07	0.1	0.1	0.1	0.07

Table 8. Classification accuracy for the base classifiers $(\Psi_1, ..., \Psi_9)$, Majority Voting and Sum method for AL scheme with $IV = 12$ and random labelling of unlabelled training examples - Magic data set

AL scheme	Ψ_1	Ψ_2	Ψ_3	Ψ_4	Ψ_5	Ψ_6	Ψ_7	Ψ_8	Ψ_9	Ψ_{MV}	Ψ_{AL}^{12}
k = 0	0.16	0.16	0.20	0.20	0.20	0.18	0.17	0.20	0.20	0.16	0.16
k = 1	0.15	0.15	0.36	0.20	0.21	0.17	0.18	0.19	0.19	0.16	0.15
Random	0.15	0.15	0.36	0.21	0.21	0.17	0.16	0.18	0.18	0.16	0.15

Table 9. Classification accuracy for the base classifiers $(\Psi_1, ..., \Psi_9)$, Majority Voting and Sum method for AL scheme with $IV = 15$ and random labelling of unlabelled training examples - Magic data set

AL scheme	Ψ_1	Ψ_2	Ψ_3	Ψ_4	Ψ_5	Ψ_6	Ψ_7	Ψ_8	Ψ_9	Ψ_{MV}	Ψ_{AL}^{15}
k = 0	0.16	0.16	0.20	0.20	0.20	0.18	0.17	0.20	0.20	0.16	0.16
k = 1	0.15	0.15	0.36	0.25	0.19	0.18	0.16	0.20	0.19	0.15	0.14
Random	0.15	0.15	0.36	0.21	0.20	0.19	0.17	0.19	0.19	0.16	0.16

Table 10. Classification accuracy for the base classifiers $(\Psi_1, ..., \Psi_9)$, Majority Voting and Sum method for AL scheme with $IV = 12$ and random labelling of unlabelled training examples - Phone data set

AL scheme	Ψ_1	Ψ_2	Ψ_3	Ψ_4	Ψ_5	Ψ_6	Ψ_7	Ψ_8	Ψ_9	Ψ_{MV}	Ψ_{AL}^{12}
k = 0	0.25	0.25	0.26	0.27	0.30	0.26	0.21	0.26	0.26	0.23	0.22
k = 1	0.22	0.22	0.32	0.27	0.17	0.19	0.20	0.13	0.14	0.17	0.15
Random	0.20	0.20	0.32	0.28	0.19	0.22	0.21	0.15	0.15	0.18	0.17

Table 11. Classification accuracy for the base classifiers $(\Psi_1, ..., \Psi_9)$, Majority Voting and Sum method for AL scheme with $IV = 15$ and random labelling of unlabelled training examples - Phone data set

AL scheme	Ψ_1	Ψ_2	Ψ_3	Ψ_4	Ψ_5	Ψ_6	Ψ_7	Ψ_8	Ψ_9	Ψ_{MV}	Ψ_{AL}^{15}
k = 0	0.25	0.25	0.26	0.27	0.30	0.26	0.21	0.26	0.26	0.23	0.22
k = 1	0.23	0.23	0.32	0.27	0.23	0.20	0.22	0.2	0.21	0.21	0.19
Random	0.22	0.22	0.32	0.27	0.22	0.18	0.20	0.18	0.19	0.18	0.17

The obtained results are promising. For a given data sets we received improvement of classification quality. This improvement relates to the proposed AL algorithm compared to random labelling of the unlabelled training examples. For example, algorithm Ψ_{AL}^{12} in our AL approach is about two percent higher than the same algorithm with random labelling in the case of MAGIC Gamma Telescope and Phone data sets. Made experiments also show that the selection of IV parameter values significantly affect the quality of classification. For example algorithm Ψ_{AL}^{IV} obtained the best results for $IV = 15$ and $IV = 12$ for MAGIC Gamma Telescope and Phone data sets respectively.

5   Conclusion

The paper presents the new AL algorithm based on the QBC approach. In the information value calculating process the decision profiles are used. In the paper experiments on several data sets were carried out. The obtained results are promising. This is due to the improvement of classification quality when we use the proposed AL algorithm compared to random labelling of unlabelled training examples.

In the future studies we plan to discuss the impact of using another value of the information value. In addition, further experiments should apply imbalanced data set [2] and not only to the binary data set.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Information Science and Statistics. Springer, New York (2006)
2. Borowska, K., Topczewska, M.: New data level approach for imbalanced data classification improvement. In: Burduk, R., Jackowski, K., Kurzyński, M., Woźniak, M., Żołnierek, A. (eds.) Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015. Advances in Intelligent Systems and Computing, vol. 403, pp. 283–294. Springer, Switzerland (2016)
3. Britto, A.S., Sabourin, R., Oliveira, L.E.: Dynamic selection of classifiersa comprehensive review. Pattern Recogn. **47**(11), 3665–3680 (2014)
4. Burduk, R.: Classifier fusion with interval-valued weights. Pattern Recogn. Lett. **34**(14), 1623–1629 (2013)
5. Choraś, M., Kozik, R.: Machine learning techniques applied to detect cyber attacks on web applications. Logic J. IGPL **23**(1), 45–56 (2015)
6. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Mach. Learn. **15**(2), 201–221 (1994)
7. Cyganek, B.: One-class support vector ensembles for image segmentation and classification. J. Math. Imaging Vis. **42**(2–3), 103–117 (2012)
8. Forczmański, P., Łabędź, P.: Recognition of occluded faces based on multi-subspace classification. In: Saeed, K., Chaki, R., Cortesi, A., Wierzchoń, S. (eds.) CISIM 2013. LNCS, vol. 8104, pp. 148–157. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40925-7_15

9. Frank, A., Asuncion, A.: UCI machine learning repository (2010)
10. Frejlichowski, D.: An algorithm for the automatic analysis of characters located on car license plates. In: Kamel, M., Campilho, A. (eds.) ICIAR 2013. LNCS, vol. 7950, pp. 774–781. Springer, Heidelberg (2013). doi:10.1007/978-3-642-39094-4_89
11. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Mach. Learn. **28**(2–3), 133–168 (1997)
12. Giacinto, G., Roli, F.: An approach to the automatic design of multiple classifier systems. Pattern Recogn. Lett. **22**, 25–33 (2001)
13. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley, New Jersey (2004)
14. Rejer, I.: Genetic algorithm with aggressive mutation for feature selection in BCI feature space. Pattern Anal. Appl. **18**(3), 485–492 (2015)
15. Stefanowski, J., Pachocki, M.: Comparing performance of committee based approaches to active learning. In: Recent Advances in Intelligent Information Systems, pp. 457–470 (2009)