On Reliability of Majority Voting

Agus Budi Raharjo, Mohamed Quafafou, Faicel Chamroukhi

On Reliability of Majority Voting

Agus Budi Raharjo*, Mohamed Quafafou*
Faicel Chamroukhi**

*Aix-Marseille University, CNRS, LSIS UMR 7296
13397, Marseille, France
agus-budi.raharjo@etu.univ-amu.fr, mohamed.quafafou@univ-amu.fr
**Caen University, LMNO UMR CNRS 6139
14000, Caen, France
faicel.chamroukhi@unicaen.fr

Abstract. In ensemble learning field, the voting of different experts can produce an optimal solution. However, the quality of voting depends on the participant expertise. In this paper, an expert selection algorithm is proposed by considering reliability measure extracted from the confidence score. Our method has been applied based on the combination of 6 algorithms. Experimental result using 8 datasets shows that the proposed reliable majority voting algorithm provides a better average accuracy than the ordinary majority voting and the base classifiers.

keyword: reliable majority voting, classification, ensemble learning.

1 Introduction

The combination of Multiple classifiers has been investigated to obtain efficient solution by adapting previous experiences. There are two approaches to combine classifiers, i.e. weighting and meta-learning methods (Rokach, 2010). Weighting method could be useful in the case when several classifiers have similar performance.While meta-learning works by looking at different problems, domains, tasks or contexts or simply past experience (Lemke et al., 2015). However, these existing methods generally focus on processing the label prediction and ignoring each prediction class score. Prediction score of each class can be used to understand the confidence degree of classifier. The study of confidence score in supervised learning is previously conducted by Zadrozny and Elkan (2002), where they transform a confidence score of predictor into a useful reliability information. In this paper we present a modification of this approach by using confidence score as reliability degree of each classifier, then we apply this approach in case of ensemble learning.

2 The Combination Method of Classifiers

Base learning algorithm in classification has several approaches to represent a problem as follow : decision table, linear model, decision tree, rule, and instance-based representation based on knowledge representation used (Witten et al., 2011). All these categories are implemented into several algorithms, Bayesian classifiers, trees, rules, functions, lazy classifiers, multi-instance classifiers, and miscellaneous. Since classifiers have different approaches to

predict dataset, it is difficult to decide which classifier can be used for certain dataset. Therefore ensemble learning is proposed to solve this drawback by combining heterogeneous base classifiers in several characteristics of datasets. Majority Voting (MV), Stacked Generalization (Stacking), and Multischeme are some examples of combination method that consider all base classifiers as the input. In another way, some ensemble learning like Random Forest, Bootstrap aggregating (Bagging), and Boosting tend to build several learning models based on one classifier to improve the performance of their prediction. In section 3, we extends a weighted MV in ensemble learning. Various base classifiers are applied as a set of input which is further discussed in section 4.

3 The Reliable MV Algorithm

In this section, we specify our problem and basic notation used in binary classification. We also provide an overview of reliability diagram according to the previous literatures. Then, the workflow of reliable MV algorithm is proposed.

A typical scenario for training process of ensemble learning consists of a set of instances $X = \{x_1, x_2, ..., x_i, ..., x_N\}$, where x_i is predicted by a set of T classifiers. For each classifier prediction, contains its independent label estimation denoted as y and confidence score denoted as s. T classifier prediction with its confidence value for N instances can be noted as follow: $D = \left\{(y_i^1, s_i^1), (y_i^2, s_i^2), ..., (y_i^t, s_i^t), ..., (y_i^T, s_i^T)\right\}_{i=1}^{N}$ such that $y_i^t \in \{0, 1\}$, $s_i^t \in \mathbb{R}$, and $s_i^t \in [0, 1]$. For certain classifier, the confidence score of its prediction is in interval $[-a, b]$, where $a \in \mathbb{R}$ and $b \in \mathbb{R}$. Therefore, the score needs to be re-scaled into the value between 0 and 1 by applying Platt scaling as in (Platt, 1999).

Confidence score can be considered to measure classifier reliability by applying reliability diagram. The term of reliability diagram is firstly used in (Murphy and Winkler, 1977) to represent probability forecast of precipitation and temperature at Chicago. They proposed an x-y diagram that displays forecast probability as the axis of argument and observed relative frequency as the axis of values (Y-axis). In supervised learning, observed relative frequency can be referred to empirical class membership probability as it is proposed in the study of (Zadrozny and Elkan, 2002). This probability can be denoted as $P(c|s)$, where $c \in \{0, 1\}$ and c is a class label. This diagram is implemented in training step, so that the true label for each instance denoted as z_i is known, with $z_i = \{0, 1\}$. In order to reflect the distribution of s, confidence score is converted according to the class c, which is $s(c = 1)_i^t = 1 - s(c = 0)_i^t$ for binary case. Then, a set of this score is split into bins with smaller interval. Let V be a set of interval limit, where $V = \{v_j | v_{j+1} - v_j = v_j - v_{j-1}, v_j \in \mathbb{R}, v_j \in [0, 1 + \Delta v[\}$. Then, s_i^t will be categorized in interval j when $v_j \leq s_i^t < v_{j+1}$. $P(c|s)$ is defined as the number of true label corresponded to c divided by the number of all prediction in interval j.

Our contribution is using this reliability representation as knowledge base to weight classifier in majority voting decision. Our method framework is explained through an example given in table 1. We suppose there are 9 instances predicted by 3 classifiers as shown in table 1a. In the first step, we convert label prediction and confidence score based on class 1 as is represented in table 1b. Then, we separate each bin with $\Delta v = 0.25$ and calculate the number of prediction for each bin in table 1c. Table 1d figures the total number of true label that appropriates with class 1. In the last step, $P(1|s)$ is computed in table 1e by dividing the result in step 3 and step 2. In this example, confidence score space is divided into 4 bins. In another case, there is a condition when reliability representation produces unbalance distribution. Therefore, the

X	A	B	C	Z
x_1	(0,0.7)	(0,0.9)	(1,0.6)	1
x_2	(0,0.8)	(1,0.6)	(1,0.8)	1
x_3	(1,0.6)	(0,0.6)	(0,0.9)	0
x_4	(1,0.7)	(1,0.7)	(1,0.6)	1
x_5	(0,0.7)	(0,0.8)	(0,0.7)	0
x_6	(1,0.9)	(1,0.9)	(1,0.7)	1
x_7	(1,0.6)	(1,0.8)	(1,0.6)	1
x_8	(0,0.6)	(0,0.6)	(1,0.7)	0
x_9	(1,0.7)	(1,0.6)	(1,0.8)	1

(a) X is predicted by A, B, and C.

X	A	B	C	Z
x_1	(1,0.3)	(1,0.1)	(1,0.6)	1
x_2	(1,0.2)	(1,0.6)	(1,0.8)	1
x_3	(1,0.6)	(1,0.4)	(1,0.1)	0
x_4	(1,0.7)	(1,0.7)	(1,0.6)	1
x_5	(1,0.3)	(1,0.2)	(1,0.3)	0
x_6	(1,0.9)	(1,0.9)	(1,0.7)	1
x_7	(1,0.6)	(1,0.8)	(1,0.6)	1
x_8	(1,0.4)	(1,0.4)	(1,0.7)	0
x_9	(1,0.7)	(1,0.6)	(1,0.8)	1

(b) Step 1: the conversion of s^t where $c = 1$.

$Interval$	A	B	C
0 - 0.25	1	2	1
0.26 - 0.5	3	2	1
0.51 - 0.75	4	3	5
0.76 - 1	1	2	2

(c) Step 2: $|y|$ for each bin.

$Interval$	A	B	C
0 - 0.25	1	1	0
0.26 - 0.5	1	0	0
0.51 - 0.75	3	3	4
0.76 - 1	1	2	2

(d) Step 3: $|z = 1|$ for each bin.

$Interval$	A	B	C
0 - 0.25	1	0.5	0
0.26 - 0.5	0.33	0	0
0.51 - 0.75	0.75	1	0.8
0.76 - 1	1	1	1

(e) Step 4: $P(1|s)$ for each bin.

TAB. 1: The framework of proposed algorithm through example given.

size of subspace needs to be adjusted carefully in order to reflect the distribution of the sample. Once we get the information of reliability representation for each classifier, a threshold is defined to filter reliable classifiers which is denoted as RC. RC is defined in formula 1 by considering two conditions.

$$RC = \begin{cases} P(c|s_i^t) \leq \lambda_1 & \text{if } s_i^t < 0.5 \\ P(c|s_i^t) \geq \lambda_2 & \text{otherwise} \end{cases} \quad (1)$$

where $\lambda_1 \in \mathbb{R}$, $\lambda_1 \in [0, 1]$, $\lambda_2 \in \mathbb{R}$, and $\lambda_2 \in [0, 1]$.

According to the previous example, given a new test instance x_i, where x_i is predicted by A, B, C, so that $D_{x_i} = \{(0, 0.6)^A, (0, 0.8)^B, (1, 0.7)^C\}$, and unknown true label with $z = 1$. We set $\lambda 1 = ((v_j + v_j + 1)/2) - 0.1$ and $\lambda 2 = ((v_j + v_j + 1)/2)$. Label 1 is defined to be the default class. Once the thresholds are applied in table 1e, reliable classifier can be concluded as $RC = \{C\}$, since A and B do not pass the threshold. In contrary, we will get the wrong answer if we consider MV. Based on our knowledge, the parameter of threshold has to be optimized iteratively in training step to obtain the best group of reliable classifier.

4 Experimentation and Result

In order to perform our experiments, six algorithms are applied based on different knowledge representations to obtain diversity among the models combination. We used Weka [1] library to build models of J48 (Decision Tree), Naive Bayes (Bayesian), JRip (Rule), Sequential minimal optimization (Function), k-nearest neighbors (Lazy), and Hyperpipes (Miscellaneous). Then, we compare our proposed algorithm with the original MV since it is built from the voting based approach. This experiment have been tested on eight dataset from UCI repository (Lichman, 2013), i.e. Diabetes, EEG Eye State (EES), MAGIC Gamma Telescope (MAGIC), Diabetic Retinopathy Debrecen (DRD), Musk (Version 1), Occupancy Detection, Phishing Websites, and Spambase dataset. Table 2 shows the accuracy comparison between each classifier and combination methods. Both MV and Reliable MV provides the best average results than all base classifiers. Reliable MV proves that score reliability selection can

1. http://www.cs.waikato.ac.nz/ml/weka/

	J48	NB	JRip	SMO	IBk	HP	MV	Reliable MV
Diabetes	0.7451	0.7516	0.7124	0.7451	0.7451	0.6013	0.7516	0.7516
EEGEyeState	0.7804	0.6943	0.7413	0.7353	0.8064	0.5367	0.7814	0.7921
Gamma Telescope	0.8283	0.7618	0.8136	0.8215	0.8068	0.3567	0.8312	0.8323
Diabetic Retinopathy	0.6348	0.6304	0.6435	0.6478	0.6261	0.4957	0.6435	0.6478
Musk	0.8211	0.7684	0.7368	0.8947	0.7895	0.6632	0.8737	0.8947
Occupancy	0.994	0.9884	0.9935	0.9935	0.9944	0.8718	0.9935	0.994
Phising Websites	0.9498	0.9213	0.9426	0.9349	0.9634	0.5807	0.9507	0.9543
Spambase	0.9217	0.8946	0.9174	0.9348	0.9272	0.612	0.9467	0.9478
mean	0.8344	0.8014	0.8126	0.8385	0.8324	0.5898	0.8465	**0.8518**

TAB. 2: Accuracy comparison between base classifiers and ensemble methods.

duplicate the best prediction from different knowledge representation in Diabetes, DRD, and Musk. Reliable MV also presents better results for eight dataset compared to MV and reveals the best average accuracy among the others.

5 Conclusion

The combination of several learning algorithms can be considered to obtain the better accuracy prediction. However, the result quality of combination method depends on the performance of its base classifiers. Therefore, we propose a weighted voting algorithm called Reliable Majority Voting in binary classification problem. This extended version of majority voting has been investigated with the different characteristics of datasets. The result shows that Reliable MV obtains higher accuracy compared to majority voting and its base classifiers. This approach opens a new chance to improve multiple annotators method in several problems. In the future works, we focus on how to divide interval subspace of each classifier score distribution optimally and how to optimize the classifier selection threshold.

References

Lemke, C., M. Budka, and B. Gabrys (2015). Metalearning: a survey of trends and technologies. *Artificial Intelligence Review 44*(1), 117–130.

Lichman, M. (2013). UCI machine learning repository.

Murphy, A. H. and R. L. Winkler (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 26*(1), 41–47.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*. MIT Press.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review 33*(1), 1–39.

Witten, I. H., E. Frank, and M. A. Hall (2011). Chapter 11 - the explorer. In I. H. Witten, E. Frank, and M. A. Hall (Eds.), *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)* (Third Edition ed.)., The Morgan Kaufmann Series in Data Management Systems, pp. 407 – 494. Boston: Morgan Kaufmann.

Zadrozny, B. and C. Elkan (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, New York, NY, USA, pp. 694–699. ACM.