Rank thresholds in classifier ensembles in medical diagnosis

K. ANTCZAK
karol.antczak@wat.edu.pl

Military University of Technology, Faculty of Cybernetics
Institute of Computer and Information Systems
Kaliskiego Str. 2, 00-908 Warsaw, Poland

Classification methods have multiple applications, with medical diagnosis being one of the most common. A powerful way to improve classification quality is to combine single classifiers into an ensemble. One of the approaches for creating such ensembles is to combine class rankings from base classifiers. In this paper, two rank-based ensemble methods are studied: Highest Rank and Borda Count. Furthermore, the effect of applying class rank threshold to these methods is analyzed. We performed tests using real-life medical data. It turns out that specificity of data domain can affect classification quality depending on classifier type.

Keywords: medical diagnostics, classifier ensemble, rank threshold.

1. Introduction

A machine learning is evolving now faster than ever before. Rapid growth of available computational power allows to use in practice many methods which were not much more than just a theory not so long ago. This growth also applies to one of the most important ML branches: a classification theory.

The classification is a very broad topic which can be roughly defined as the task of assigning object to one of several predefined categories [1]. The universality of this task leads to many applications, from medicine through real time systems to searching engines. An important application of classification systems in medicine are Clinical Decision Support Systems (CDSS). Such systems can be described as "active knowledge systems which use two or more patient data to generate case-specific advice" [2]. A more specific class are Medical Diagnosis Systems (MDS). Those systems take as an input patient data, e.g. list of symptoms, and produce a set of diseases that can cause given symptoms. The core mechanism used to produce an output is usually a classification mechanism, which can be both simple similarity measure and complex ensemble of classifiers. Over the years, many MDS with various classification methods were developed. Some of them are: GIDEON [3], HEPAR II [4], Isabel [5] or SWD [6]. The latter was analyzed in this paper.

The purpose of this paper is to study the effects of adjusting rank threshold in multi-classifier systems, especially in applications on medical diagnosis. Moreover, we will show that specificity of medical data plays an important role in threshold adjustment. Section 2 of this paper contains definitions of basic concepts used in classification theory. In Section 3 we introduce a problem of thresholding the classifier output. In Section 4 the test results are presented. They are then discussed in Section 5.

2. Classification and classifiers

In the previous section, we provided a rough definition of the classification problem. It can be formulated more formally as follows: a classification problem is to find a function f that maps a set of objects X to a set of classes Y:

$$f: X \to Y \tag{1}$$

Function f is also called *classification model* or *classifier* [1]. Objects from X are typically represented by a *feature vector*:

$$x = (x_1, x_2, \dots, x_n), x \in X \tag{2}$$

where x_i is the value of i-th feature.

The case with two available classes ($|Y| = 2$) is called *binary classification*. If there are more classes available, we deal with *multiclass classification*.

An important characteristic of classifier is its *linearity*. A linear classifier chooses a class based on linear combination of the feature vector:

$$f(x) = g(w \cdot x) = g(\textstyle\sum_j w_j x_j) \tag{3}$$

where w is a real vector of weights.

There are many types of linear classifiers, such as Perceptron [7], Naive Bayes Classifier [8], Support Vector Machine [8, 9] and others. They performs well against many real-world tasks, such as document classification, while being relatively fast to train and use [10].

As it turns out, there are some cases which cannot be effectively solved using linear classification models. This led to the creation of non-linear classifiers. These models are often extensions of previously known linear methods, allowing them to operate on non-linear feature spaces. Some examples of non-linear classifiers are Multi-Layer Neural Networks [12], Bayesian Networks [12, 13] or similarity-based methods [14].

Single classifiers can be combined to form an *ensemble of classifiers*. There are two main approaches to combining classifiers: *fusion* and *selection* [15]. In classifiers fusion, outputs from base classifiers are treated as an input for some second-level classifier operating on intermediate feature space. In classifier selection, on the other hand, for given input, one of the base classifiers is being selected to give the output.

Combining the classifiers proved to be useful technique to improve a quality of classification model, even when using simple base classifiers. Several explanations for this phenomenon were given by Dietterich [16].

In order to compare classifiers, both single and combined, we need to measure their performance. There are three commonly used metrics: *sensitivity*, *specificity* and *ROC curve* [17].

Each classification result can be assigned to one of the following cases:

- True Positive (TP) – classifier correctly determined that object belongs to a class;
- False Positive (FP) – classifier incorrectly determined that object belongs to a class;
- True Negative (TN) – classifier correctly determined that object doesn't belong to a class;
- False Negative (FN) – classifier incorrectly determined that object doesn't belong to a class.

Using cardinalities of these sets, we can define sensitivity and specificity:

$$sensitivity = \frac{|TP|}{|TP| + |FN|} \tag{4}$$

$$specificity = \frac{|TN|}{|FP| + |TN|} \tag{5}$$

Sensitivity measures classifier's ability to correctly detect belonging to a class. On the other hand, specificity relates to classifier's ability to rule out classes. Using these two metrics we can create a *Receiver Operating Characteristic* (*ROC*) curve. It is a 2D plot where Y axis determines sensitivity and X is 1-specificity. Each point on this plot determines a single classifier; points closer to upper left corner have bigger sensitivity and specificity.

Above metrics were originally defined for binary classifiers returning one of two available classes. However, they can be easily extended for classifiers returning any number of classes. Then we consider those metrics for each class. For example, for given class y:

- TP_y are instances of y that are classified as y;
- FP_y are instances of non-y that are classified as y;
- TN_y are instances of non-y that are not classified as y;
- FN_y are instances of y that are not classified as y.

Then, the sensitivity and specificity *for class* y can be calculated by the formulas (4) and (5).

3. Threshold adjustment

The output of multiclass classification model can be categorized into one of three types [18]:

- Type I (abstract level) – classifier returns an unordered set of classes $\{y_1, \ldots, y_n\}$;
- Type II (rank level) – classifier returns an ordered sequence of classes $(y_1, \ldots, y_n)$;
- Type III (measurement level) – classifier returns an ordered sequence of classes along with their scores $((y_1, z_1), \ldots, (y_n, z_n))$.

In order to minimize the size of the classifier output, a final "filtering" is required. In case of type I classifier this can be achieved by random sampling. For type II, filtering can be done using *rank threshold*. For type III classification models, this is usually realized by using a *decision threshold*.

Let us take a classifier which returns a sequence of pairs (y, z), where each class y is associated with level z of classifier's belief that object belongs to this class. We can treat this as a *score function*:

$$s(x, y) = z \tag{6}$$

We can reduce the size of the output set by introducing a *decision threshold*: only classes with score greater than or equal to the threshold τ will be accepted:

$$Y_x = \{y \in Y : s(x, y) \geq \tau\} \tag{7}$$

where Y_x is a final set of classes for given input vector x.

Following problem arises: how to adjust the decision threshold in order to achieve the best classification performance? To investigate this problem, we need to measure specificity and sensitivity. Let us take a classifier which returns normalized score, i.e. $z \in < 0; 1 >$. Consider two extreme cases:

- $\tau = 0$: we choose all the classes,
- $\tau = 1$: we choose only the class with the highest available score.

In the first case we will have the highest sensitivity – simply because proper class will always be in the output set. However, the specificity of classifier will be very low. In the second case the situation will be reversed – the sensitivity will be the highest, and the specificity will be the lowest. As we can see, there is no "silver bullet". This is known as a problem of *decision threshold adjustment* [19].

One can distinguish two approaches to adjusting the decision threshold: static and dynamic. In the static approach, the threshold is set before classification and is a constant value. A two trivial examples of this approach were presented above. Mohammadi and van de Geer proposed a method of estimating threshold for classifier based on maximum likelihood [20]. In the dynamic approach, the threshold is estimated separately for each case. It was shown by Koford and Groner that one can dynamically calculate optimal threshold for linear pattern classifier [21]. Threshold estimation can be also used for ensembles of classifiers. G. Levitin analyzed both static and dynamic threshold for classifier ensemble method called *Weighted Voting Classifier* (*WVC*) [22].

In all the previously mentioned methods, the threshold is applied to numerical values returned by classifier. One can take a slightly different approach and apply the threshold to size of output class set. This can be applied to type II classifiers as well as type III by ordering the output set using score function. As a result we get a class ranking:

$$R_x = (y_1, \dots, y_n) : s(x, y_{i-1}) \leq s(x, y_i) \quad i = \overline{2, n}, y_i \in Y, x \in X \tag{8}$$

With a *class rank threshold* equal to k, a final output set of classifier will consist of k first classes from Y_x. In next section we study how adjusting this threshold affects sensitivity and specificity of the classifier ensemble.

4. Tests

The tests were performed in order to examine the effects of the following factors:

- combining classifiers using ranking-based ensemble methods,
- adjusting the class rank threshold in classifier ensemble.

Tests were performed on medical diagnosis system SWD [6]. It is a complex system developed for holistic support of medical diagnosis. A certain novelty which distinguishes this system is a usage of several classifiers rather than a single one.

We tested two classifier ensemble methods implemented in SWD: Highest Rank Method and Borda Count Method. Both of these methods operate on a class ranking, i.e. they are taking into account class ranks from base classifiers rather than class scores [23].

The *Highest Rank* is a simple fusion method where the final rank of class is chosen as the highest rank given by any of base classifiers to this class. If two classes have the same final rank, then the final ordering between them can be chosen arbitrarily.

The *Borda Count* was originally designed as an election system [24]. We define $B_i(y_j)$function, returning amount of classes with ranks lower than y_j for i-th base classifier. Based on this function, we create a final scoring function:

$$B(y_j) = \sum_i B_i(y_j) \tag{9}$$

Then the final ranking is created by sorting classes using above function in descending order.

Both of these ensemble methods use the same set of three base classifiers:

- Multilayer Feedforward Neural Network, with 878 input neurons, 344 neurons in hidden layer and 91 neurons in output layer,
- Jaccard Similarity Index:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{10}$$

- Lennon Similarity Index:

$$L(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)} \tag{11}$$

For similarity indices, A denotes input feature set, and B is a reference feature set associated with a single class. The classifier checks similarity of input set to reference set corresponding to each class.

We used two distinct datasets for training and testing. They were taken from SWD database of medical data for respiratory and skin diseases. Data model is based on a parametrized disease model proposed by Walczak and Paczkowski [25]. There are 91 diseases, 766 symptoms and 112 symptom values which form

a set 14086 triplets <disease, symptom, symptom value>. The test set consisted of 81 test vectors. Each test vector defined a single disease and a set of symptoms along with their values.

The first test investigated how combining the classifiers using ranking-based methods affects the classification performance. Each classifier and ensemble method were tested against a set of 81 test cases. The sensitivity and specificity of each model were measured for each test case (counted as shown in section 2). Then, the averages of these metrics were calculated. In this test the fixed rank threshold was used, equal to 10.

Tab. 1. Sensitivity and specificity of classifiers

Classifier	Sensitivity	Specificity
Neural Network	0,593	0,883
Jaccard Index	0,679	0,884
Lennon Index	0,420	0,892
Highest Rank	0,741	0,884
Borda Count	0,716	0,884

Results can be presented on ROC plot:

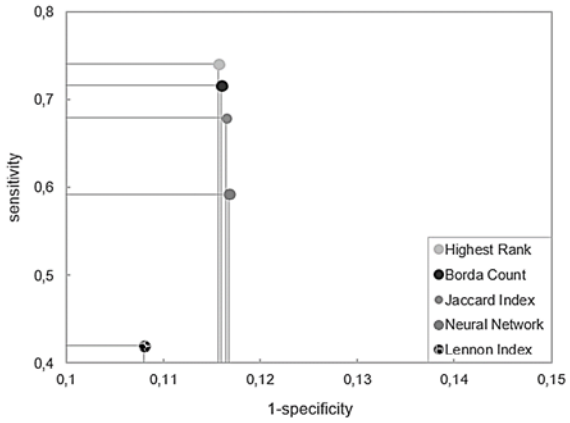

Fig. 1. ROC plot for classifiers

Both of ensemble methods had a higher sensitivity than any of base classifiers – 0,741 for Highest Rank and 0,716 for Borda Count. The specificity for all the classifiers was similar, ranging from 0,883 for Neural Network to 0,892 achieved by Lennon Index. The Jaccard Index proved to have the highest sensitivity from all the base classifiers (0,679).

In the second test we studied an effect of applying various class rank threshold for Highest Rank and Borda Count classifiers. Threshold values ranged from 0 (reject all classes) to 10 (accept 10 first classes). As before, the sensitivity and specificity were measured.

Tab. 2. Sensitivity and specificity for various rank threshold levels

Threshold	Classifier			
	Highest Rank		Borda Count	
	Sensitivity	Specificity	Sensitivity	Specificity
0	0,000	1,000	0,000	1,000
1	0,469	0,993	0,420	0,993
2	0,642	0,983	0,420	0,980
3	0,654	0,971	0,481	0,969
4	0,667	0,958	0,543	0,957
5	0,667	0,946	0,605	0,945
6	0,716	0,934	0,630	0,933
7	0,716	0,921	0,679	0,921
8	0,716	0,909	0,716	0,909
9	0,716	0,896	0,716	0,896
10	0,728	0,884	0,716	0,884

Corresponding ROC plots:

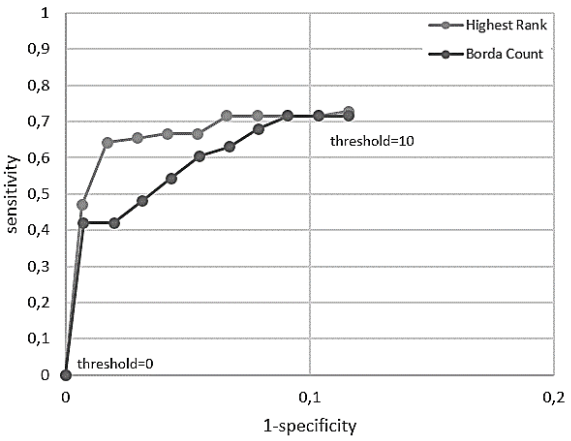

Fig. 2. ROC plots for various rank threshold levels

It turns out that increasing the threshold results in the higher sensitivity at the cost of specificity. Moreover, Highest Rank method proved to be more resistant to lower threshold values than Borda Count.

5. Discussion

Test results proved that using rank-based classifier combination method can improve classification performance. The main factor was improving the sensitivity. As of specificity, only Lennon Index had a higher score, but at the cost of having the lowest sensitivity. This *synergy effect* in ranking-based ensembles is caused by reasons of statistical nature, as explained by

Dietterich [16]. In our case, the two main factors are:

- error averaging,
- promoting common indications.

The *error averaging* occurs when some of the classifiers returns ranking with wrong class in the first place. If there are more "right" classifiers than "wrong" ones, the correct classifications will be outvoted and the final result will be correct. Below is an example of this phenomenon for Borda Count method. The correct class is marked "**A**":

Classifier 1: [**A**, B, C] \
Classifier 2: [B, **A**, C] – Borda Count: [**A**, B, C]
Classifier 3: [**A**, C, B] /

Another synergy factor is a result of *promoting common indication*. In this case, the correct class doesn't even have to be on the first place in any ranking. However, if it will constantly achieve high ranks, it can be first in the final ranking:

Classifier 1: [C, **A**, B, D] \
Classifier 2: [B, **A**, C, D] – Borda Count:
[**A**, B, C, D]
Classifier 3: [D, **A**, C, B] /

The Jaccard Index had the best performance of all of the base classifiers. This is an interesting behavior, due to this algorithm's relative simplicity in comparison to, for example, neural networks. This phenomenon can be explained by specificity of the medical datasets used for training and tests. It turns out that the Jaccard Index is a good approximation of *reasoning scheme* used by diagnosticians: they perform diagnosis by comparing the set of patient's symptoms with a set of typical symptoms for a disease, in a similar way to Jaccard Index. This result shows how important it is to choose the right classification model for a specific domain.

We showed that adjusting rank threshold affects performance of classifier ensemble. The higher rank threshold result in higher sensitivity and lower specificity. Below plot shows how often the correct class was ranked as n-th in Highest Rank and Borda Count methods:

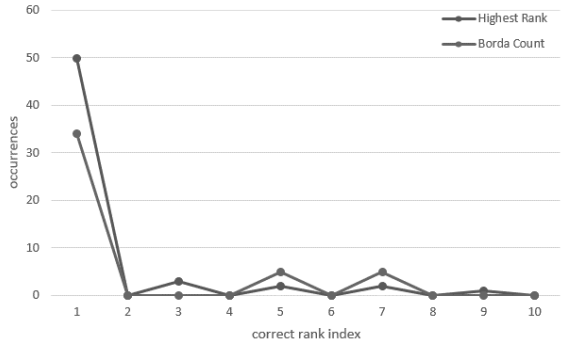

Fig. 3. Amounts of various correct class indices

As one can see, correct classes were mostly ranked as the first. This shows, why applying the ranking can improve the quality of classification.

Xu and others listed a several new problems regarding classifiers ensembles to be studied. One of them was: "How to adjust these thresholds in the combination phase such that the best combination can be achieved?" [18]. Our paper gives a partial answer for rank-based classifiers. In fact, the problem of adjusting rank threshold can be viewed as *multi-objective optimization problem* with specificity and sensitivity as objectives to maximize.

Chen and others studied the effects of adjusting decision threshold on sensitivity, specificity and concordance for several classifiers (logistic regression, classification tree, Fisher's linear discriminant analysis and a weighted k-nearest neighbor) [19]. Their results were similar to ours: the sensitivity and specificity increased and decreased, respectively, with the decision threshold.

Kam Ho and others analyzed several ranking-based classifier combination methods: Highest Rank, Borda Count and Logistic Regression [23]. These method were tested in a field of character recognition. They showed that the performance of a multiple classifier system can be better than those of each individual. They also raised an interesting question: "Is it possible to systematically create a multiple classifier system for a given problem, so that for each possible input pattern there exists one or a combination of several classifiers that can correctly identify its true class?". Occurrence of synergy effect in classifiers ensemble hints that this indeed can be possible.

6. Bibliography

[1] Tan P-N., Steinbach M., Kumar V., "Chapter 4. Classification: Basic Concepts, Decision Trees and Model Evaluation", in: *Introduction to Data Mining*, 145–205, Pearson Education, 2006.

[2] Wyatt J., Spiegelhalter D., "Field trials of medical decision-aids: potential problems and solutions", *Proceedings Annual Symposium Computer Applications in Medical Care*, 3–7, AMIA, 1991.

[3] Berger S.A., "GIDEON: A Computer Program for Diagnosis, Simulation, and Informatics in the Fields of Geographic Medicine and Emerging Diseases", *Emerging Infectious Diseases*, Vol. 7, No. 7 (2001).

[4] Oniśko A., Druzdzel M.J., Wasyluk H,. "Extension of the Hepar II Model to Multiple-Disorder Diagnosis", in: *Intelligent Information Systems*, 303–313, Physica-Verlag HD, 2000.

[5] Ramnarayan P., Tomlinson A., Kulkarni G., Rao A., Britto J., "A novel diagnostic aid (ISABEL): development and preliminary evaluation of clinical performance", in: *MEDINFO 2004*, Marius Fieschi, Enrico Coiera, Yu-Chan Jack Li (Eds.), of the series: Studies in Health Technology and Informatics, Vol. 107, 1091–1095, IOS Press, 2004.

[6] Walczak A., Gaj A., Jahnz-Różyk K., *Wybrane zagadnienia informatycznego wspomagania decyzji medycznych*, Andrzej Walczak (Ed.), Warszawa, Wojskowa Akademia Techniczna, 2013.

[7] Rosenblatt F., *The Perceptron: A Perceiving and Recognizing Automaton* (*Project PARA*), New York, 1957.

[8] Rish I., "An empirical study of the naive Bayes classifier", *Proceedings of IJCAI 2001 workshop on Empirical Methods in Artificial Intelligence*, 3: 41–46, IBM New York, 2001.

[9] Cortes C., Corinna C., Vladimir V., "Support-Vector Networks", *Machine Learning*, 20, 273–297 (1995).

[10] Yuan G-X., Guo-Xun Y., Chia-Hua H., Chih-Jen L., "Recent Advances of Large-Scale Linear Classification", *Proceeding of the IEEE*, Vol. 100, 2584–2603, 2012.

[11] Bland R., *Learning XOR: Exploring the Space of a Classic Problem*, University of Stirling, Department of Computing Sciennce and Mathematics, 1998.

[12] Rosenblatt F., *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, Washington, 1962.

[13] Neapolitan R.E., "Probabilistic Reasoning in Expert Systems: Theory and Algorithms", *Technometrics*, Vol. 34, No. 1, 99–100 (1992).

[14] Subasi M., Subasi E., Anthony M., Hammer P.L., "Using a similarity measure for credible classification", *Discrete Applied Mathematics*, 157(5), 1104–1112. (2009).

[15] Kuncheva L.I., "Switching between selection and fusion in combining classifiers: An experiment", *IEEE Transactions on Systems, Man, and Cybernetics – Part B*: *Cybernetics*, Vol. 32, No. 2, 146–156 (2002).

[16] Dietterich T.G., "Ensemble Methods in Machine Learning", in: *Multiple Classifier Systems 2000*, J. Kittler and F. Roli (Eds.), of series Lecture Notes in Computer Science, Vol. 1857, 1–15, Springer-Verlag, Berlin, 2000.

[17] Bradley A.P., "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognition*, Vol. 30, No. 7, 1145–1159 (1997).

[18] Xu L., Krzyzak A., Suen C.Y., "Methods of combining multiple classifiers and their applications to handwriting recognition", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 22, No. 3, 418–435 (1992).

[19] Chen J.J., Tsai C-A., Moon H., Ahn H., Young J.J., Chen C-H., "Decision threshold adjustment in class prediction", *SAR QSAR Environmental Research*, Vol. 17, No. 3, 337–352 (2006).

[20] Mohammadi L., van de Geer S., "On threshold-based classification rules", in: *Mathematical Statistics and Applications*: *Festschrift for Constance van Eeden*, Marc Moore, Sorana Froda and Christian Léger (Eds.), IMS Lecture Notes – Monograph Series, Vol. 42, 261–280, IMS, USA, 2003.

[21] Koford J., Groner G., "The use of an adaptive threshold element to design a linear optimal pattern classifier", *IEEE Transaction on Information Theory*, Vol. 12, 42–50 (1966).

[22] Levitin G., "Threshold optimization for weighted voting classifiers", *Naval Research Logistics*, Vol. 50(4), 322–344 (2003).

[23] Tin Kam Ho., Ho T.K., Hull J.J., Srihari S.N., "Decision combination in multiple classifier systems", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, 66–75 (1994).

[24] McLean I.S., McMillan A., Monroe B.L. (Eds.), "Partial Justification of the Borda Count", in: *The Theory of Committees and Elections by Duncan Black and Committee Decisions with Complementary Valuation by Duncan Black and RA Newing*, 369–385, Springer Netherlands, 1998.

[25] Walczak A., Paczkowski M., "Medical data preprocessing for increased selectivity of diagnosis", in: *Bio-Algorithms and Med-Systems*, Vol. 12(1), 39–43 (2016).

Progowanie rang dla zespołów klasyfikatorów w diagnostyce medycznej

K. ANTCZAK

Metody klasyfikacji mają wiele zastosowań, z których jednym z częściej spotykanych jest diagnostyka medyczna. Jakość klasyfikacji można w znaczący sposób podnieść, tworząc zespoły klasyfikatorów. Jedną z metod tworzenia takich zespołów jest łączenie rankingów generowanych przez klasyfikatory bazowe. W niniejszej pracy przeanalizowano dwie metody łączenia klasyfikatorów bazujące na rankingach: Najwyższej Rangi oraz Głosowanie Bordy. Dodatkowo zbadano wpływ progowania rankingu na jakość klasyfikacji. Testy przeprowadzono z użyciem rzeczywistych danych medycznych. Wykazano przy tym, że specyfika danych medycznych może wpłynąć na jakość klasyfikacji w zależności od typu klasyfikatora.

Słowa kluczowe: diagnostyka medyczna, łączenie klasyfikatorów, progowanie rankingu.