MULTI-CLASS CLASSIFICATION

Ardo Allik

DIFFERENCES BETWEEN BINARY AND MULTI-CLASS CLASSIFICATION

Classification tasks with more than two classes are very common.

Certain classification concepts are fundamentally binary and do not easily generalize for more than two classes:

- Some classification methods require multi-class models to be built out of binary models (such as linear classifiers).
- Other classification methods handle any number of classes naturally (such as decision trees).

TAL TECH

MULTI-CLASS CLASSIFICATION

Classification tasks with more than two classes are very common. For instance, once a patient has been diagnosed as suffering from a rheumatic disease, the doctor will want to classify him or her further into one of several variants. If we have k classes, performance of a classifier can be assessed using a k-by-k contingency table.

Assessing performance is easy if we are interested in the classifier's accuracy, which is still the sum of the descending diagonal of the contingency table, divided by the number of test instances. However, this can obscure differences in performance on different classes, and other quantities may be more meaningful.

	Predicted			
	15	**2**	**3**	20
Actual	**7**	**15**	**8**	30
	2	**3**	**45**	50
	24	20	56	100

Flach, P., Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge: Cambridge University Press, 2012

TAL TECH

MULTICLASS CLASSIFICATION VALIDATION

If we have k classes, performance of a classifier can be assesse using a k-by-k contingency table (confusion matrix).

Accuracy of this classifier is:

(15+15+45)/100 = 0.75

We can calculate precision and recall separately for each class:

Precision: 15/24 (I class), 15/20 (II), 45/56 (III)

Recall: 15/20 (I class), 15/30 (II), 45/50 (III)

	Predicted			
	15	**2**	**3**	20
Actual	**7**	**15**	**8**	30
	2	**3**	**45**	50
	24	20	56	100

Flach, P., Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge: Cambridge University Press, 2012

TAL TECH

MULTICLASS CLASSIFICATION VALIDATION

We could average the recall and precision to obtain single precision and recall numbers for the whole classifier, or we could take a weighted average taking the proportion of each class into account:

Average precision:

(0.63 + 0.75 + 0.80) / 3 = 0.73

Weighted average precision:

0.20·0.63 + 0.30·0.75 + 0.50·0.80 = 0.75

Another possibility is to perform a more detailed analysis, by looking at the characteristics for separate pair of classes.

Distinguishing the first class from the third:

Precision: 15/17 = 0.88

Recall: 15/18 = 0.83

Distinguishing the third class from the first:

Precision: 45/48 = 0.94

Recall: 45/47 = 0.96

	Predicted			
	15	**2**	**3**	20
Actual	**7**	**15**	**8**	30
	2	**3**	**45**	50
	24	20	56	100

Flach, P., Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge: Cambridge University Press, 2012

(Why are these numbers higher in the latter direction?)

ALSO USED: MACRO-AVERAGING AND MICRO-AVERAGING

Macro-averaging is straight forward. We just take the average of the precision and recall of the system on different sets

Macro-averaging precision:

(Prec.1 + Prec.2 + Prec.3) / 3 =

(0.63 + 0.75 + 0.80) / 3 = 0.73

In *micro-averaging* all TPs, TNs, FPs and FNs for each class are summed up and then the average is taken

Micro-averaging precision:

(TP1 + TP2 + TP3) /

(TP1 + FP1 + TP2 + FP2 + TP3 + FP3) = ?

		Predicted		
	15	**2**	**3**	20
Actual	**7**	**15**	**8**	30
	2	**3**	**45**	50
	24	20	56	100

Flach, P., Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge: Cambridge University Press, 2012

TAL TECH

ALSO USED: MACRO-AVERAGING AND MICRO-AVERAGING

Macro-averaging is straight forward. We just take the average of the precision and recall of the system on different sets

Macro-averaging precision:

(Prec.1 + Prec.2 + Prec.3) / 3 =

(0.63 + 0.75 + 0.80) / 3 = 0.73

In *micro-averaging* all TPs, TNs, FPs and FNs for each class are summed up and then the average is taken

Micro-averaging precision:

(TP1 + TP2 + TP3) /

(TP1 + FP1 + TP2 + FP2 + TP3 + FP3) =

(15 + 15 + 45) / (15 + 9 + 15 + 5 + 45 + 11) =

75 / 100 = 0.75

		Predicted		
	15	**2**	**3**	20
Actual	**7**	**15**	**8**	30
	2	**3**	**45**	50
	24	20	56	100

Flach, P., Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge: Cambridge University Press, 2012

TAL TECH

ACCURACY

Accuracy of this classifier is:

(15+15+45)/100 = 0.75

We can calculate accuracy individually for each class:

Acc1: (TP1 + TN1) / (TP1 + TN1 + FP1 + FN1) =

(15 + (15 + 45 + 3 + 8)) / 100 = 0.86

Acc2:

(15 + (15 + 45 + 3 + 2)) / 100 = 0.80

Acc3:

(45 + (15 + 15 + 2 + 7)) / 100 = 0.84

Accuracy averaged over individual accuracies:

(Acc1 + Acc2 + Acc3) / 3 =

(0.86 + 0.80 + 0.84) / 3 = 0.83

		Predicted		
	15	**2**	**3**	20
Actual	**7**	**15**	**8**	30
	2	**3**	**45**	50
	24	20	56	100

Flach, P., Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge: Cambridge University Press, 2012

TAL TECH

COMBINING TWO-CLASS CLASSIFIERS INTO A MULTICLASS CLASSIFIER

If we want to construct multi-class classifier, but we only have ability to train two-class models (such as lineaar classifiers), then we can combine the binary classifiers.

There are different ways to combine several two-class classifiers into a single k-class classifier

TAL TECH

COMBINING TWO-CLASS CLASSIFIERS INTO A MULTICLASS CLASSIFIER: ONE-VERSUS-REST, ONE-VERSUS-ONE

One-versus-rest:

1. We train n binary classifiers, the first separates class C_1 from $C_2 \ldots C_n$, the second separates C_2 from all other classes and so on.
2. When training the i-th classifier we treat all instances of class C_i as positive examples, and the remaining instances as negative examples.
3. Sometimes the classes are learned in a fixed order, in which case we learn $n-1$ models, the i-th one separating C_i from $C_i+1, \ldots, C_n$ with $1 \leq i < n$.

One-versus-one:

1. In this scheme, we train $n(n-1)/2$ binary classifiers, one for each pair of different classes.
2. If a binary classifier treats the classes asymmetrically, as happens with certain models, it makes more sense to train two classifiers for each pair, leading to a total of $n(n-1)$ classifiers.

TAL TECH

CLASSIFICATION ALGORITHMS THAT SUPPORT MULTI-CLASS CLASSIFICATION:

Some binary classification algorithms have been adapted to solve multi-class classification problems without requiring problem transformations.

These examples include:

- Neural networks
- Decision trees
- k-nearest neighbors
- Naive Bayes
- Support vector machines

TAL TECH

MULTICLASS CLASSIFICATION IN HEALTHCARE

F1-Scores for Single and All Possible Combinations of Accelerometer Sensors in the MHEALTH Dataset

	Ankle	Chest	Wrist	A+C	A+W	C+W	A+C+W
Lying	94.74	89.09	90.85	100.0	96.49	100.0	100.0
Sitting	45.00	44.41	79.23	26.12	81.65	84.93	67.67
Standing	61.02	60.78	86.59	56.54	85.26	86.59	75.95
Walking	95.40	85.94	83.59	97.58	98.70	89.11	94.08
Running	88.62	88.70	76.14	88.78	87.35	86.90	87.71
Cycling	99.36	93.98	94.92	97.32	96.56	99.84	100.0
C. Stairs	98.24	85.58	84.92	98.08	98.72	89.45	94.47
Jogging	86.78	88.47	76.56	89.20	88.41	87.60	88.82
Mean	83.64	79.62	84.10	81.70	91.64	90.55	88.59

Confusion Matrix for Ankle and Wrist Combination (A+W) in the MHEALTH Dataset

	1	2	3	4	5	6	7	8
1 Lying	289	0	0	0	0	21	0	0
2 Sitting	0	227	83	0	0	0	0	0
3 Standing	0	18	292	0	0	0	0	0
4 Walking	0	0	0	303	0	0	7	0
5 Running	0	0	0	0	259	0	0	51
6 Cycling	0	1	0	0	0	309	0	0
7 C. Stairs	0	0	0	1	0	0	309	0
8 Jogging	0	0	0	0	24	0	0	286

Chowdhury AK, Tjondronegoro D, Chandran V, Trost SG (2018) Physical Activity Recognition Using Posterior-Adapted Class-Based Fusion of Multiaccelerometer Data. IEEE J. Biomed. Health. Inform. 22:678-685

TAL TECH

REFERENCES AND USED MATERIALS

EMC Education Services, Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, John Wiley @ Sons, 2015

Flach, P., Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge: Cambridge University Press, 2012

TAL TECH

TALLINNA TEHNIKAÜLIKOOL

Ehitajate tee 5, 19086 Tallinn, Tel 620 2002 (E-R 8.30–17.00)

taltech.ee

MULTI-CLASS SCORES AND PROBABILITIES (EXTRA INFORMATION)

If we want to calculate multi-class scores and probabilities from binary classifiers, we have a number of different options:

1. We can use the distances obtained by loss-based decoding and turn them into scores by means of some appropriate transformation. This method is applicable if the binary classifiers output calibrated scores on a single scale.
2. Alternatively, we can use the output of each binary classifier as features (real-valued if we use the scores, binary if we only use the predicted class) and train a model that can produce multi-class scores, such as naive Bayes or tree models. This method is generally applicable but requires additional training.
3. A simple alternative that is also generally applicable and often produces satisfactory results is to derive scores from coverage counts: the number of examples of each class that are classified as positive by the binary classifer.

TAL TECH

MULTI-CLASS ROC CURVE AND AUC (EXTRA INFORMATION)

Once we have multi-class scores, we can ask the familiar question of how good these are. An important performance index of a binary scoring classifier is the area under the ROC curve or AUC, which is the proportion of correctly ranked positive–negative pairs.

Tanking does not have a direct multi-class analogue, and so the most obvious thing to do is to calculate the average AUC over binary classification tasks, either in a one-versus-rest or one-versus-one fashion.

For instance, the one-versus-rest average AUC estimates the probability that, taking a uniformly drawn class as positive, a uniformly drawn example from that class gets a higher score than a uniformly drawn example over all other classes.

The ‘negative’ is more likely to come from the more prevalent classes; for that reason the positive class is sometimes also drawn from a non-uniform distribution in which each class is weighted with its prevalence in the test set.

TAL TECH