Decision rules for ensembled probabilistic classifier chain for multilabel classification

Ostapets A.A.

MSU, Faculty of Computational Mathematics and Cybernetics

14-10-2016

Preliminaries

Let $\mathcal{X}$ denote the domain of instances and let $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ be the finite set of labels. Let $\mathcal{Y} = \{0, 1\}^k$ - the set of all binary vectors of length k.

Given a training set $S = (\mathbf{x}_i, \mathbf{Y}_i)$, $(\mathbf{x}_i \in \mathcal{X}, \mathbf{Y}_i \in \mathcal{Y}, 1 \leq i \leq M)$, i.i.d. drawn from an unknown distribution $\mathcal{D}$.

The goal of the learning system is to output a multilabel classifier $h : \mathcal{X} \to \mathcal{Y}$, which optimizes some specific evaluation metric [1].

In most cases however, instead of outputting a multilabel classifier, the learning system will produce a real-valued function of the form $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$.

An Algebraic Approach

Yu.I. Zhuravlev showed that an arbitrary algorithm could be represented as a product (successive execution) of two algorithms [2]:

- A recognition operator. The recognition algorithm converts original information and descriptions of objects to be recognized into a number matrix.
- A decision rule. The decision rule converts the number matrix into a binary matrix of final answers.

Problem Transformation Methods

There exists a number of very simple problem transformation methods which actually transform multilabel data in such a way so that existing classification algorithms (i.e. binary classifiers) can be applied.

- Label Powerset (LP).
- Binary Relevance (BR).

Label Powerset

Label Powerset is a straight forward method that considers each unique set of labels in a multilabel training data as one class in the new transformed data. Therefore, the new transformed problem is a single label classification task.
For a new instance, LP outputs the most probable class which actually is a set of classes in the original data.

Binary Relevance

Binary Relevance is one of the most popular approaches as a transformation method that actually creates k datasets ($k = |\mathcal{L}|$), each for one class label and trains a classifier on each of these datasets.
Each of these datasets contains the same number of instances as the original data, but each dataset $D_{\lambda_j}, 1 \leq j \leq k$ positively labels instances that belong to class λ_j and negative otherwise.
While BR has been used in many practical applications, it has been widely criticized for its implicit assumption of **label independence** which might not hold in the data.

Probabilistic Classifier Chains

Given a query instance $\mathbf{x}$, the (conditional) probability of each label combination $\mathbf{Y} = (y_1, \dots, y_k) \in \mathcal{Y}$ can be computed using the product rule of probability:

$$\mathbf{P_x}(\mathbf{y}) = \mathbf{P_x}(y_1) \times \prod_{i=2}^{k} \mathbf{P_x}(y_i | y_1, \dots, y_{i-1})$$

Thus, to estimate the joint distribution of labels, one possibility is to learn k functions f_i n an augmented input space $\mathcal{X} \times \{0,1\}^{i-1}$, taking $y_1, \dots, y_{i-1}$ as additional attributes:

$$f_i : \mathcal{X} \times \{0,1\}^{i-1} \to [0,1]$$

$$(\mathbf{x}, y_1, y_2, \dots, y_{i-1}) \to P(y_i = 1 | \mathbf{x}, y_1, y_2, \dots, y_{i-1}),$$

Decision rules

With a vector $(g_1, \ \ldots, g_k)$ of class scores obtained, the final class prediction $(a_1, \ \ldots, a_k)$ is made using one of the possible decision rules:

1. S-cut: $a_i(\mathbf{x}) = \mathbb{I}[g_i(\mathbf{x}) \geq t], \forall i \in \mathcal{L}$
2. R-cut: $a_i(\mathbf{x}) = \mathbb{I}[rank(i) \leq r], \forall i \in \mathcal{L}$
3. DS-cut: $a_i(\mathbf{x}) = \mathbb{I}[g_i(\mathbf{x}) \geq t_{rank(i)}], \forall i \in \mathcal{L}$
4. DSS-cut: $a_i(\mathbf{x}) = \mathbb{I}[\frac{g_i(\mathbf{x})}{g_{max}} \geq t_{rank(i)}], \forall i \in \mathcal{L}$

Dataset

To compare performance of different recognition operators and of the decision rules evaluation tests were done on a real task dataset.
The WISE-2014 dataset presents the task of multilabel classification of articles coming from Greek print media. Data was collected by scanning a number of Greek print media from May 2013 to September 2013.
The text of the articles is represented using the bag-of-words model and for each token encountered inside the text of all articles, the tf-idf statistic is computed and unit normalization is applied to the tf-idf values of each article.
There are therefore 301561 numerical attributes corresponding to the tokens encountered inside the text of the collected articles.
Articles were manually annotated by a human expert with one or more out of 203 labels.

Evaluation metrics

The evaluation metrics were:

- Mean F_1 score, also known as example-based F_1 score.
- Classification accuracy.

$$F_{score} = \frac{1}{M} \sum_{i=1}^{M} f^{i}_{score},$$

$$f^{i}_{score} = 2\frac{pr}{p+r}, \text{ where } p = \frac{tp}{tp+fp}, \quad r = \frac{tp}{tp+fn},$$

Classification accuracy

Classification accuracy or subset accuracy is defined as follows:

$$Accuracy = \frac{1}{M}\sum_{i=1}^{M} acc(Y_i^{pred}, Y_i^{true}),$$

$$acc(Y_i^{pred}, Y_i^{true}) = \begin{cases} 1, & Y_i^{pred} \text{ to be an exact match of } Y_i^{true}; \\ 0, & \text{otherwise.} \end{cases}$$

Recognition operators

The recognition operators were:

- Logistic Regression (from scikit-learn with parameters (penalty='l1', C=6.0, tol=0.001))
- Linear classifier with SGD training (from scikit-learn: SGDClassifier(loss="modified_huber")).

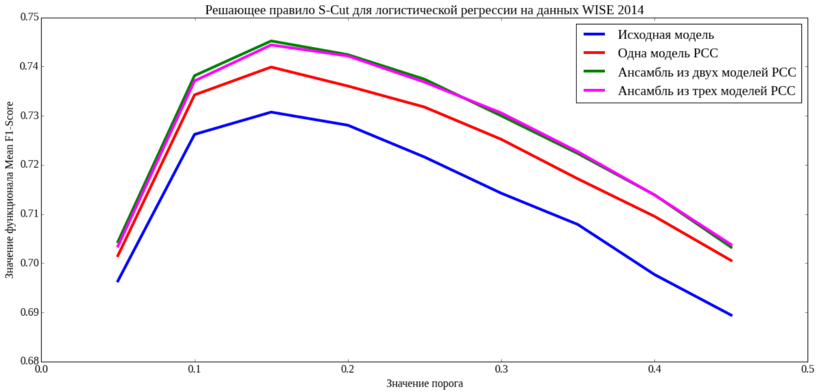
For each of these models 4 recognition operators were trained:

1. Original model with «Binary Relevance».
2. Probabilistic Classifier Chain based on the original model.
3. Ensemble of 2 PCCs.
4. Ensemble of 3 PCCs.

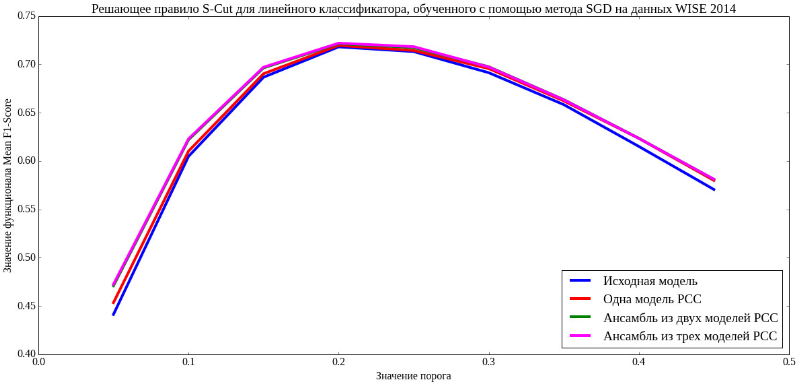
Mean F_1 score for different decision rules

Algorithm	S-cut	R-cut	DS-cut	DSS-cut
LR	73.07	73.58	76.36	78.28
1 PPC on LR	73.99	73.40	76.27	78.24
2 PPCs on LR	**74.52**	73.68	76.68	78.32
3 PPCs on LR	74.48	**73.73**	**76.74**	**78.41**
LC (SGD)	71.80	71.53	71.12	75.52
1 PPC on LC	71.96	71.46	71.06	75.41
2 PPCs on LC	72.13	71.66	71.41	75.55
3 PPCs on LC	**72.18**	**71.78**	**71.50**	**75.67**

Mean F1-Score, Logistic Regression, S-cut



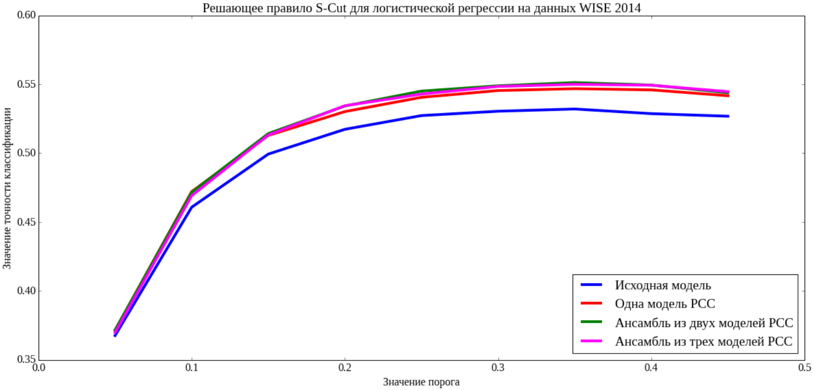
Mean F1-Score, Linear Classifier, S-cut



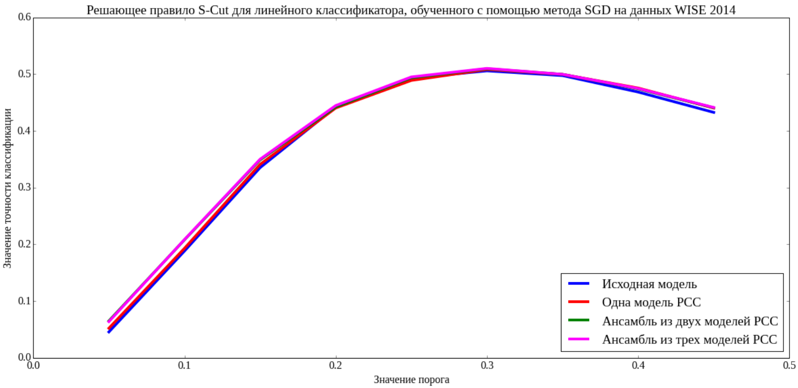
Subset accuracy for different decision rules

Algorithm	S-cut	R-cut	DS-cut	DSS-cut
LR	52.73	58.29	53.77	59.93
1 PPC on LR	54.68	58.17	54.00	59.85
2 PPCs on LR	55.13	58.42	54.19	60.15
3 PPCs on LR	**55.20**	**58.50**	**54.25**	**60.21**
LC (SGD)	50.58	56.77	53.40	53.20
1 PPC on LC	50.82	56.62	53.32	53.18
2 PPCs on LC	50.94	56.89	53.51	53.55
3 PPCs on LC	**51.00**	**56.96**	**53.64**	**53.73**

Subset accuracy, Logistic Regression, S-cut



Subset accuracy, Linear Classifier, S-cut



Conclusion

It is experimentally demonstrated that the quality of the forecast of the proposed composition exceeds the quality of the original models. It should be emphasized that a single probabilistic classifier chain does not improve the quality of the original model.
The noticeable growth can be achieved by using an ensemble of two or more probabilistic classifier chains.

References

1. Min L. Zhang and Zhi H. Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition, 40(7):2038–2048
2. Zhuravlev Yu.I. 1979. An Algebraic Approach to Recognition and Classification Problems. Problems of Cybernetics 33 P. 5–68

Questions

Thank you! Any questions?