Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
AAAI Technical Report CR-13-01

Aggregating Human-Expert Opinions for Multi-Label Classification

Evgueni Smirnov, Hua Zhang, Ralf Peeters
DKE, Maastricht University, Maastricht, The Netherlands
{smirnov,hua.zhang,ralf.peeters}@maastrichtuniversity.nl

Nikolay Nikolaev
University of London, London, UK
n.nikolaev@gold.ac.uk

Maike Imkamp
UCM, Maastricht University, Maastricht, The Netherlands
m.imkamp@student.maastrichtuniversity.nl

Abstract

This paper introduces a multi-label classification problem to the field of human computation. The problem involves training data such that each instance belongs to a set of classes. The true class sets of all the instances are provided together with their estimations presented by m human experts. Given the training data and the class-set estimates of the m experts for a new instance, the multi-label classification problem is to estimate the true class set of that instance. To solve the problem we propose an ensemble approach. Experiments show that the approach can outperform the best expert and the majority vote of the experts.

Introduction

The multi-label classification problem proposed in this paper is formalized as follows. Let X be an instance space, Y be a class set, and p be an unknown probability distribution over the set-labeled space $X \times 2^Y$. We assume m human experts estimating the true class set of instances $x \in X$ according to p. We draw n instances $(x, Y_x) \in X \times 2^Y$ from p. Each expert $i \in 1..m$ provides an estimate $Y^{(i)} \subseteq Y$ of the true class set Y_x of each instance x without observing Y_x. Thus, we consider each instance as a $m+2$-tuple $(x, Y^{(1)}, ..., Y^{(m)}, Y_x)$. The set of these n instances formed in this way results in training data D. In this context the multi-label classification problem is to estimate the true class set for new instance $x \in X$ according to p, given the training data D and the class-set estimates $Y^{(1)}, ..., Y^{(m)} \in Y$ given by the m experts for x.

The problem introduced can be compared with other classification problems considered in human computation (Yan et al. 2010; Raykar et al. 2010). In this field the emphasis is on single-label classification problems where the data is labeled by the experts and the true instance classes are not given. We note that our problem is simpler but it has not been considered so far and it has many applications. Consider for example meteorologists predicting for the next day whether there will be clouds, rain, wind, sun etc. The true set of classes arrives in 24 hours. We can record the meteorologist predictions and the true class set over a time period to form our data. Then using our new meta-classifier ensemble approach (given below) we can solve this multi-label classification problem. However we note that from a human-computation view point (Quinn and Bederson 2011) the solution is practical if we can predict better than the best meteorologist and the majority vote of the meteorologists.

Meta-Classifier Ensemble Approach

We propose a meta-classifier ensemble approach to our multi-label classification problem. The approach first transforms the problem into a set of single-label binary classification problems (Read et al. 2011). Then, it trains a single-label meta classifier for each binary classification problem. Finally, the approach combines the meta classifiers to form the multi-label classification solution.

The problem transformation we propose is a modification of the binary-relevance transformation (Read et al. 2011). It decomposes the multi-class classification problem into set of binary classification problems BP_y, one for each class $y \in Y$. The training data D_y for BP_y is formed from the training data D as follows: any instance $(x, Y^{(1)}, ..., Y^{(m)}, Y_x) \in D$ is transformed into an instance $(x, 1_{Y^{(1)}}(y), ..., 1_{Y^{(m)}}(y), 1_{Y_x}(y)) \in D_y$ where $1_Y(y)$ is the indicator function. Thus, (1) the class set $Y^{(i)}$ for expert i is substituted by a new binary feature that equals 1 iff the class y of the binary classification problem BP_y is in the set $Y^{(i)}$, and (2) the output class set Y_x is substituted by a new output binary feature that equals 1 iff the class y of the binary classification problem BP_y is in the set Y_x. In this context the single-label binary classification problem BP_y for any class y is to estimate the indicator $1_{Y_x}(y)$ for new instance $x \in X$, given the training data D_y and indicators $1_{Y^{(1)}}(y), ..., 1_{Y^{(m)}}$ provided by the m experts for x.

To solve the single-label binary classification problem BP_y for any class $y \in Y$ we propose to employ stacked generalization (Wolpert 1992). We consider each human expert $i \in 1..m$ as a base classifier and learn a single-label meta classifier m_y using the training data D_y that aggregates the opinions of the experts into final opinion. The classifier m_y is a function that can have two possible forms: $h : X, \{0,1\}^m \rightarrow \{0,1\}$ and $h : \{0,1\}^m \rightarrow \{0,1\}$ with and without the instance space X in the input (Wolpert 1992). The output is the estimate of the indicator $1_{Y_x}(y)$.

Meta classifiers m_y trained for all the classes $y \in Y$

form the meta-classifier ensemble. The ensemble estimates the class set of new instance $x \in X$ as follows. First, for each class $y \in Y$ meta classifier m_y estimates the indicator $1_{Y_x}(y)$ for that class w.r.t. x. Then, y is added to the estimated class set $\widehat{Y}_x$ of x iff the indicator estimation is 1.

The generalization performance of the meta-classifier ensembles is estimated using the multi-label accuracy rate a_m (Read et al. 2011): $a_m = \frac{1}{|D|} \sum_{(x,Y_x)\in D} \frac{|Y_x \cap \widehat{Y}_x|}{|Y_x \cup \widehat{Y}_x|}$.

We note that meta-classifier training can employ feature selection. Thus, we can identify a combination of the experts resulting into good generalization performance.

Experiments

We experimented with multi-label classification problem with 200 news from the BBC news website. Each news was given by heading and abstract, and belonged to one or more out of 12 categories: UK, Africa, Asia, Europe, Latin America, Mid-East, US&Canada, Business, Health, SciEnvironment, Tech, Entertainment. For example : {*Tech, Business*}*: "Blackstone pulls out of Dell bid. Blackstone has decided not to submit a bid for computer company Dell, citing falling sales and fears over the company's finances."*

The 200 news were labeled by 15 experts that did not know the true class sets of those news. The multi-label accuracy rate of the experts varied in [0.692, 0.910]. The multi-label accuracy rate of the best expert and experts' majority vote were 0.908 and 0.910, respectively.

To apply our multi-classifier ensemble approach we formed our data D as follows. First, the text of all the news was transformed to the bag-of-word representation. Then, each news with the class sets estimated by the 15 experts and its true class set was added to D. For each category y we generated the data D_y and then trained two meta classifiers $h^b : X, \{0,1\}^{15} \to \{0,1\}$ and $h^{\bar{b}} : \{0,1\}^{15} \to \{0,1\}$, where b ($\bar{b}$) indicates (non-) presence of the bag-of-word representation in the input. The output of both meta classifiers is the estimate of the indicator $1_{Y_x}(y)$.

The meta classifiers h^b and $h^{\bar{b}}$ were trained for any category y with and without a wrapper feature-selection procedure based on greedy stepwise search (Guyon et al. 2006). Thus, at the end we trained four types of meta classifiers: h^{bw}, $h^{\bar{b}w}$, $h^{b\bar{w}}$, and $h^{\bar{b}\bar{w}}$ for each category, where w ($\bar{w}$) indicates (non-) use of the wrapper method.

Meta-classifier ensembles were designed for each of the four meta-classifier types: h^{bw}, $h^{\bar{b}w}$, $h^{b\bar{w}}$, and $h^{\bar{b}\bar{w}}$. They consisted of the best meta classifiers for each category.

The multi-label accuracy rates of the ensembles were estimated using 10-fold cross-validation and are provided in Table 1. Two observations can be derived from the table:

- **(O1)** the ensembles have multi-label accuracy rates significantly greater than those of the best expert (0.908) and the experts' majority vote (0.910);
- **(O2)** the ensembles achieves the best multi-label classification accuracy rates when **(O2a)** the instances are represented by the expert class-set estimates only, and **(O2b)** the wrapper-based feature selection is employed.

$\bar{b}\bar{w}$	$b\bar{w}$	$\bar{b}w$	bw
0.928	**0.929**	**0.941**	**0.933**

Table 1: Multi-label accuracy rates of ensembles: h^{bw}, $h^{\bar{b}w}$, $h^{b\bar{w}}$, and $h^{\bar{b}\bar{w}}$. The rates in bold are significantly greater than that of the experts' majority vote (0.910) and that of the best expert (0.908) on 0.05 significance level.

For the wrapper ensembles bw and $\bar{b}w$ we observed that:

- **(O3)** different experts were selected for each of the 12 categories (meta classifiers). The number of experts selected varied in [1, 9], one expert was not selected at all.

Conclusion

This section analyzes observations (O1)-(O3). Observation (O1) implies that our meta-classifier ensembles can outperform the best expert and the majority vote of the experts. Thus, our multi-label classification problem and meta-classifier ensemble approach are useful. Observation (O2a) is a well-known fact in stacked generalization (Wolpert 1992). In the paper context however it states that for our multi-label classification problem we do have to know the class-set estimates of the experts only to receive good generalization performance. The input from the application domain (in our case English text of the news) is less important. Observation (O2b) is an expected result in feature selection (Guyon et al. 2006). However it also has a practical implication for our multi-label classification problem, namely it allows to choose combination of the most adequate experts (see observation O3). This means that we can reduce the number of human experts and thus the overall financial cost.

References

Guyon, I.; Gunn, S.; Nikravesh, M.; and Zadeh, L. 2006. *Feature Extraction, Foundations and Applications*. Springer.

Quinn, A., and Bederson, B. 2011. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011*, 1403–1412. ACM.

Raykar, V.; Yu, S.; Zhao, L.; Valadez, G.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from crowds. *Journal of Machine Learning Research* 11:1297–1322.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2011. Classifier chains for multi-label classification. *Machine Learning* 85(3):333–359.

Wolpert, D. 1992. Stacked generalization. *Neural Networks* 5(2):241–259.

Yan, Y.; Rosales, R.; Fung, G.; Schmidt, M.; Valadez, G. H.; Bogoni, L.; Moy, L.; and Dy, J. 2010. Modeling annotator expertise: Learning when everybody knows a bit of something. *Journal of Machine Learning Research* 9:932–939.