Leveraging Crowdsourcing to Make Models in Multi-label Domains Interoperable

Lei Duan* Oyama Satoshi Haruhiko Sato Masahito Kurihara
(Graduate School of Information Science and Technology, Hokkaido University)†

Abstract

Recently, the issue of learning from multi-label data has attracted significant attention. Due to different aspects that multi-label classifiers want to capture, personal preferences of researchers, or just inconsistency in terminology usage, the employed models may differ from each other. Therefore, model interoperability is a big concern in multi-label domains. Our study focuses on exploiting effective interoperation between two different models in a multi-label domain through the application of harmonised mapping established in a crowdsourced setting.

Keywords
Multi-label domain, Interoperability, Harmonised mapping, Crowdsourcing

1 Introduction

Traditional multi-class classification aims at categorizing instances into a set of candidate labels, in which each instance is associated with a single label. Multi-label classification is a generalization of multi-class classification where each instance is associated with a subset of candidate labels. Recently, the issue of learning from multi-label data has attracted significant attention, mainly motivated from applications such as topic categorization of news article [1] and web page [2], affect analysis in narrative [3] and music [4], and semantic annotation of image [5] and video [6]. A good survey on multi-label classification is presented by Tsoumakas *et al.* [7].

The first step towards solving a problem in a multi-label domain is to adopt or create an appropriate model. Simply put, a model can be represented by the candidate labels (also referred to as classes, categories, terms or tags) applied to the collected instances. Due to different aspects that multi-label classifiers want to capture, personal preferences of researchers, or just inconsistency in terminology usage, the employed models may differ from each other. A noteworthy example is affect analysis. Even though the model of Ekman's six basic emotions (*happiness*, *fear*, *anger*, *surprise*, *disgust* and *sadness* [8]) has been used very broadly to cover a wide range of affect analysis research, other emotion models also exist. For example, Trohidis *et al.* employed other six emotions (*amazed-surprised*, *happy-pleased*, *relaxing-calm*, *quiet-still*, *sad-lonely* and *angry-fearful*) based on the Tellegen-Watson-Clark model [9] to conduct the automated detection of emotion in music; the model that the *manifold* sentiment analyser [10] was developed on consists of a collection of 32 emotions; the WordNet-Affect [11] even hierarchically organized a collection of 288 emotions. Moreover, given the complexity of human thinking, social and cultural background plays an significant part in emotion interpretation. A noteworthy example is that being different from the English-oriented affect analysis research listed above, a lot of Japanese-oriented research (e.g., [3, 12]) prefers to employ the ten emotions (喜, 好, 安, 怒, 哀, 怖, 恥, 厭, 昂 and 驚) based on Nakamura's *Emotive Expression Dictionary* [13]. A complete discussion on the models for affect analysis is out of the scope of this paper, and can be found in Calvo *et al.* [14]. However, it is certain that unfortunately, as of date, no general model of emotions has yet been agreed on [15].

Although different emotion models are founded on different psychological theories and fit specific purposes of particular affect analysis research in various fields, complications still arise when they are employed. First, it is hard to integrate an affect analyser and an affect application to allow them work together. One typical example is that a text-oriented affect analyser classifies a linguistic unit (often a sentence) into relative emotions from the Nakamura's model, but a text-to-speech synthesis requires a linguistic unit with its relative emotions from the Ekman's model as the input for affective pronunciation. The output of the affect analyser cannot be used as the input of the affect application since the model followed by the output does not match the one followed by the input. Second, training data cannot be shared among supervised affect analysers employing different emotion models, which results in waste of resources. Third, the lack of harmonisation among different emotion models poses barriers to comparative experiments, so it is hard to evaluate that of two affect analysers (or applications) employing different emotion models, which one performs better. Thus, regardless of

*duan@ec.hokudai.ac.jp
†Kita 14, Nishi 9, Kita-ku, Sapporo, Japan

the models chosen to represent the emotions, how to make them interoperable is a necessary and important problem.

Interoperability, which is a big concern in the computer world, is the ability to make systems and organizations work together (inter-operate). The term was initially defined for information technology or systems engineering services to allow for information exchange. For most problems in multi-label domains, a large number of class terms have been employed, and of course, not everyone will agree on what a "standard" list of terms should be – such as the affect research illustrated above and film genre classification (taking the list of genres from IMDB[1] or Netflix[2]). Therefore, interoperability is actually a pretty big concern in multi-label domains as well.

On-line crowdsourcing services provide an inexpensive means for outsourcing various kinds of tasks to hundreds of thousands of people, and it is being used more frequently in the annotation community. We propose leveraging crowdsourcing to make different models in a certain multi-label domain interoperable. Suppose that there is a large collection of multi-label instances with model X, but information of model Y is considered more important for the goal being pursued. Therefore, we can first (randomly or deliberately) select a part of all instances, and ask crowdsourcing workers to assign the relative labels from model Y to each of the selected instances. The optimum mapping from model X to model Y are then established according to the obtained triplets {(instance, relative label set in model X, assigned label set in model Y)}. Using the established mapping, the rest of the multi-label instances with relative label set in source model X can directly obtain their respective relative label set in object model Y.

Although data can be obtained from a crowdsourcing service at very low cost (time and expense), crowdsourcing workers are rarely trained and generally do not have the abilities needed to accurately perform the offered task. Therefore, ensuring the quality of the results submitted by workers is one of the biggest challenges in crowdsourcing. In addition to the exploration of regulatory mechanism such as giving monetary bonuses to high-performance workers and denying payments to low-performance ones, and injecting a collection of tasks with known correct answers into tasks to measure a worker's performance automatically, crowdsourcing service researchers have also explored sophisticated statistical strategies. A commonly used approach is by aggregating the responses produced by multiple workers to produce a consensus result. The problem is how to construct reliable results with a minimum of human effort. In the multi-label domain, Duan *et al.* [12] has proposed an effective method for estimating multiple relative labels for each repeatedly crowdsourced multi-labeled instance. Our study focuses on exploiting effective interoperation between two different models in a multi-label domain through the application of harmonised mapping established in a crowdsourced setting.

[1] http://www.imdb.com/genre
[2] http://www2.netflix.com/allgenreslist

2 Problem Definition and Proposed Methods

2.1 Problem Definition

Let I be the set of annotated instances, X be the source model, and Y be the object model. $\mathcal{X}_i \subseteq X$ $(i \in I)$ denotes the gold (or estimated beforehand) labels in X of instance i. Let K be the set of crowdsourcing workers, and $\mathcal{K}_i \subseteq K$ $(i \in I)$ be the set of workers who annotated instance i. (Note that it is not necessary to ask every worker to annotate all the instances.) $\mathcal{Y}_i^k \subseteq Y$ $(k \in \mathcal{K}_i, i \in I)$ denotes the assigned labels in Y by worker k of instance i. Let $T = \left\{\mathcal{X}_i, \mathcal{Y}_i^k : k \in \mathcal{K}_i, i \in I\right\} \subseteq 2^X \times 2^Y$ be the set of obtained examples, where 2^X (or 2^Y) is the power set of X (or Y). The goal is to learn a mapping f: $2^X \to 2^Y$ from T, where f is chosen from a hypothesis class F, such that a loss function: $F \times 2^X \times 2^Y \to \mathbb{R}_0^+$ is minimized.

2.2 Proposed Algorithms

2.2.1 Maximum Likelihood Estimation (*MLE*)

The simplest way to learn a harmonised mapping between two models is the maximum likelihood estimation. For each instance i, two sets of indicator variables are defined as follows:

$$m_{i\mathcal{X}}\left(i \in I, \mathcal{X} \subseteq \mathcal{X}\right) = \begin{cases} 1, & \mathcal{X} = \mathcal{X}_i \\ 0, & \mathcal{X} \neq \mathcal{X}_i \end{cases}$$

$$n_{i\mathcal{Y}}^k\left(k \in \mathcal{K}_i, i \in I, \mathcal{Y} \subseteq Y\right) = \begin{cases} 1, & \mathcal{Y} = \mathcal{Y}_i^k \\ 0, & \mathcal{Y} \neq \mathcal{Y}_i^k \end{cases} .$$

The maximum likelihood that $\mathcal{Y}$ is the object label set of $\mathcal{X}$ is estimated as:

$$P\left(\mathcal{Y} \mid \mathcal{X}\right) = \frac{\sum_{i \in I} \sum_{k \in \mathcal{K}_i} m_{i\mathcal{X}} n_{i\mathcal{Y}}^k}{\sum_{i \in I} m_{i\mathcal{X}} |\mathcal{K}_i|}, \qquad (1)$$

and the optimum object label set in Y for $\mathcal{X}$ is the one that achieves the maximum likelihood:

$$\arg\max_{\mathcal{Y}\subseteq Y} P\left(\mathcal{Y} \mid \mathcal{X}\right).$$

Equation (1) just equally treats annotations given by different workers. However, the fact is that the ability of workers is varying. Another method is to estimate in advance the relative label set in Y for each instance using quality control methods (such as the one proposed by Duan *et al.* [12]) given the obtained annotations $\left\{n_{i\mathcal{Y}}^k \in \{0,1\} : k \in \mathcal{K}_i, i \in I, \mathcal{Y} \subseteq Y\right\}$. Let $t_{i\mathcal{Y}} \in \{0,1\}$ $(i \in I, \mathcal{Y} \subseteq \mathcal{Y})$ be the indicator variable: if the estimated label set of instance i is $\mathcal{Y}$, then $t_{i\mathcal{Y}} = 1$ and $t_{i\mathcal{Y}'} = 0$ for $\mathcal{Y} \neq \mathcal{Y}'$. At this time, Equation (1) is transformed into:

$$P\left(\mathcal{Y} \mid \mathcal{X}\right) = \frac{\sum_{i\in I} m_{i\mathcal{X}} t_{i\mathcal{Y}}}{\sum_{i\in I} m_{i\mathcal{X}}}.$$

Because the states of labels in both source model X and object model Y are binary-valued, the MLE method needs to estimate an object label set $\mathcal{Y}$ for each of the $2^{|X|}$ different source label sets in X. This means that it is necessary to at least select $2^{|X|}$ instances to cover all possible label sets in X. But at the practical level, it is too expensive and nearly impossible to select a sufficient number of instances for every perspective and ask workers to annotate them.

2.2.2 Optimum Transformation-matrix Estimation (*OTE*)

To overcome the shortage of the *MLE* method mentioned in Section 2.2.1, we propose a more robust method. In linear algebra, the problem of learning a mapping $2^X \to 2^Y$ can be solved by constructing a linear transformation mapping $f : \{0,1\}^{|X|} \to \{0,1\}^{|Y|}$:

$$\vec{y} = f\left(\vec{x}\right) = \vec{x}\mathbf{A}, \tag{2}$$

where $\mathbf{A}$ is a $|X| \times |Y|$ transformation matrix of mapping f. Establishing the optimal mapping f is then equivalent to minimizing the sum of the distances between the two vectors of all annotated instances:

$$\arg\min_{\mathbf{A}} \left\{ \sum_{i\in I} \sum_{k\in\mathcal{K}_i} \mathrm{dis}\left(\vec{x}_i\mathbf{A}, \vec{y}_{ki}\right) \right\},$$

where $\mathrm{dis}\left(\cdot,\cdot\right)$ denotes the distance between two vectors, and $\vec{x}_i$ (or $\vec{y}_{ki}$) is the corresponding binary vector: if an element's corresponding label exists in $\mathcal{X}_i$ (or $\mathcal{Y}_i^k$), its value is 1, and 0 otherwise.

Here we give a simple method to approximately calculate $\mathbf{A}$. Let $\mathbf{A}_i^k$ $(k \in \mathcal{K}_i, i \in I)$ be the transformation matrix, which is defined as:

$$\vec{y}_{ki} = \vec{x}_i\mathbf{A}_i^k,$$

each transformation matrix in $\left\{\mathbf{A}_i^k : k \in \mathcal{K}_i, i \in I\right\}$ is the single unique solution of the system of linear equations if $|X| = |Y|$, or the optimal solution of the underdetermined or overdetermined system if $|X| > |Y|$ or $|X| < |Y|$. And then $\mathbf{A}$ can be estimated as the average of $\left\{\mathbf{A}_i^k : k \in \mathcal{K}_i, i \in I\right\}$:

$$\mathbf{A} = \frac{\sum_{i\in I} \sum_{k\in\mathcal{K}_i} \mathbf{A}_i^k}{\sum_{i\in I} |\mathcal{K}_i|}. \tag{3}$$

Because the result of the object vector $\vec{y}$ calculated using Equation (2) is not necessarily a binary vector, we have to determine a threshold (normally being 0.5) so that an element's value is transformed to 1 if it is above the threshold, and 0 otherwise.

Similar to Equation (1), Equation (3) also treats assignments given by different workers equally. Therefore, we propose giving more weight to the annotations given by high-performance workers and less weight to those given by low-performance workers. We introduce a temporary vector $\vec{z}_i$ $(i \in I)$, whose elements' values are the assigning probabilities of their corresponding labels:

$$z_i^\imath = \frac{\sum_{k\in\mathcal{K}_i} y_{ki}^\imath}{|\mathcal{K}_i|},$$

where $z_i^\imath$ (or $y_{ki}^\imath$) $(\imath \in Y)$ is the value of element $\imath$ in $\vec{z}_i$ (or $\vec{y}_{ki}$). (Note that $\vec{z}_i$ can also be estimated using the method in Duan *et al.* [12] to be a binary vector.) Let $w_k \in [0,1]$ $(k \in K)$ denote the *individual contribution-rate* of worker k, which is defined as

$$w_k = \frac{\sum_{i\in\mathcal{I}_k} \mathrm{sim}\left(\vec{y}_{ki}, \vec{z}_i\right)}{|\mathcal{I}_k|},$$

where $\mathcal{I}_k \subseteq I$ $(k \in K)$ is the set of instances annotated by worker k and $\mathrm{sim}\left(\cdot,\cdot\right)$ is the similarity between two vectors. A simple similarity metric can be defined as

$$\mathrm{sim}\left(\vec{y}_{ki}, \vec{z}_i\right) = \frac{\sum_{\imath\in Y} \left(1 - |y_{ki}^\imath - z_i^\imath|\right)}{|Y|},$$

since $y_{ki}^\imath \in \{0,1\}$ and $z_i^\imath \in [0,1]$. We can also employ other metrics such as *cosine coefficient*, *correlation coefficient*, etc. The *individual contribution-rates* can be viewed as the probability that worker k assigns the correct label set to an instance. After incorporating the *individual contribution-rates*, Equation (3) is transformed to

$$\mathbf{A} = \frac{\sum_{i\in I} \sum_{k\in\mathcal{K}_i} w_k\mathbf{A}_i^k}{\sum_{i\in I} \sum_{k\in\mathcal{K}_i} w_k}.$$

3 Conclusion

In multi-label domains such as affect analysis, different models are employed on the basis of what information is considered important for the goals being pursued, researchers' personal preferences, or just inconsistency in terminology usage. We focused on leveraging crowdsourcing to making different models interoperable. This can benefit in a few ways. First, an application can have the advantage of using a classifier that has already been vetted, and one that may also come with an annotated corpus, which can be used to train other classifiers or just argument the original dataset if the usage restrictions on the corpus allow for that. Finally, it makes interrelated classifiers and applications comparable. We proposed two algorithms, *MLE* and *OTE*. The *MLE* algorithm has a high requirement on the coverage on source model, while the *OTE* algorithm is more robust in solving the data sparsity problem. In order to test the efficiency of these algorithms, in future work we will concentrate on conducting experiments on real crowdsourcing datasets to see whether prospective results can be obtained.

参考文献

[1] Schapire, R. E., & Singer, Y. (2000). BoosTexter: A boosting-based system for text categorization. Machine learning, 39(2-3), 135-168.

[2] Ueda, N., & Saito, K. (2002). Parametric mixture models for multi-labeled text. In Advances in neural information processing systems (pp. 721-728).

[3] Ptaszynski, M., Dokoshi, H., Oyama, S., Rzepka, R., Kurihara, M., Araki, K., & Momouchi, Y. (2013). Affect analysis in context of characters in narratives. Expert Systems With Applications, 40(1), 168-176.

[4] Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. P. (2008, September). Multi-Label Classification of Music into Emotions. In ISMIR (Vol. 8, pp. 325-330).

[5] Yang, S., Kim, S. K., & Ro, Y. M. (2007). Semantic home photo categorization. Circuits and Systems for Video Technology, IEEE Transactions on, 17(3), 324-335.

[6] Qi, G. J., Hua, X. S., Rui, Y., Tang, J., Mei, T., & Zhang, H. J. (2007, September). Correlative multi-label video annotation. In Proceedings of the 15th international conference on Multimedia (pp. 17-26). ACM.

[7] Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining multi-label data. In Data mining and knowledge discovery handbook (pp. 667-685). Springer US.

[8] Ekman, P. (1992). An argument for basic emotions. Cognition & Emotion, 6(3-4), 169-200.

[9] Tellegen, A., Watson, D., & Clark, L. A. (1999). On the dimensional and hierarchical structure of affect. Psychological Science, 10(4), 297-303.

[10] Kim, S., Li, F., Lebanon, G., & Essa, I. (2012). Beyond sentiment: The manifold of human emotions. arXiv preprint arXiv:1202.1568.

[11] Strapparava, C., & Valitutti, A. (2004, May). WordNet Affect: an Affective Extension of WordNet. In LREC (Vol. 4, pp. 1083-1086).

[12] Duan, L., Oyama, S., Sato, H., & Kurihara, M. (2014). Separate or joint? Estimation of multiple labels from crowdsourced annotations. Expert Systems with Applications, 41(13), 5723-5732.

[13] Nakamura, A. (1993). Kanjo hyogen jiten [Dictionary of Emotive Expressions] (in Japanese). Tokyodo.

[14] Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. Affective Computing, IEEE Transactions on, 1(1), 18-37.

[15] Schröder, M. (2004). Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. Institut für Photetik, Universität des Saarlandes.