1L4-OS-09a-5in1

クラウドソーシングによるマルチラベル分類のための RAkEL を用いた品質管理法

Quality Control Methods Using RAkEL for Crowdsourced Multi-Label Classification

吉村 皐亮 Kosuke YOSHIMURA　馬場 雪乃 Yukino BABA　鹿島 久嗣 Hisashi KASHIMA

京都大学大学院情報学研究科知能情報学専攻
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

The creation of an enormous data with annotations is usually expensive and time-consuming. By using crowdsourcing services, we can annotate large datasets at low cost. However, quality of products from crowdsourcing cannot be guaranteed because crowdsourcing workers are non-experts, therefore, quality control is essential. In this paper, we propose two quality control methods for crowdsourced multi-label classification tasks. Existing methods are not able to take a relation of each label into account or are computationally expensive. To solve this problem, we propose the methods using RAkEL, a method which captures a correlation between labels and is not computationally expensive. We demonstrate that our methods estimate labels more effectively than existing methods on crowdsourced multi-label classification tasks in the case that there is a correlation between labels. We also show that the proposed method is robust against spam workers.

1. 序論

大量のラベル付けされたデータを安価に得る方法としてクラウドソーシングを用いる方法が考えられる．クラウドソーシングとは，不特定多数の作業者 (ワーカ) に作業依頼を行い依頼主が必要とするサービスやコンテンツを得るためのプロセスのことである．従来，膨大な量のデータに対してのラベル付け作業は少数の専門家が長い時間をかけて行っていたため，コストが高いという問題があった．こうしたコスト面での問題をクラウドソーシングで解決するのだが，ラベル付けを行うワーカは専門家でないため能力や意欲に大きな幅があり，ワーカから得られる成果物は必ずしも品質の良いものばかりではないという別の問題が出てくる．そこで，成果物の品質を保証するために品質管理が重要になる．一般的な品質管理法として，一つのタスクに対して複数のワーカから回答を募り，得られた回答を統合する方法がある．素朴な統合手法として多数決を採ることが考えられるが，ワーカの能力を考慮することで精度向上を図ることができる [Dawid 79, Whitehill 09].

本研究では，マルチラベル分類タスクをクラウドソーシングを用いて解決する際の品質管理法を提案する．マルチラベル分類とは，一つの対象に対して複数のラベルが付与される分類問題を指す．一方，一つの対象に対して単一のラベルが付与される分類問題をシングルラベル分類問題と呼ぶ．マルチラベル分類はシングルラベル分類と異なり，ラベル間の共起関係の存在が仮定できるため，ラベル間の共起関係をうまく捉えた推定を行うことで精度の向上が見込まれる．しかし，既存研究ではラベル間の共起関係を捉えた推定ができるが，計算量が指数関数的に増大してしまうという問題をもっている [Duan 14]．一方で，計算量を抑えようとすればラベル間の共起関係は捉えた推定ができない．つまり，ラベル間の共起関係を捉えた推定を行うことと計算量を抑えた推定を行うことの間のトレードオフ関係がマルチラベル分類においては問題となる．この問題を解決するために，我々は小タスクを解き、結果を統合させる手法である RAkEL(RAndom k-labELsets) [Tsoumakas 07b, Tsoumakas 11] を用いて，既存のクラウドソーシングにおけるシングルラベル分類のための品質管理法である Dawid & Skene モデル [Dawid 79] と GLAD モデル [Whitehill 09] に帰着させてマルチラベル分類問題を解く．これら二つの品質管理法をそれぞれ RAkEL DS, RAkEL GLAD と名付ける．これらの手法により，ラベル間の共起関係を捉えた推定を現実的な時間で行うことが可能になる．

提案した二つの手法に対して既存の品質管理法との精度比較実験を行うことで提案手法が精度改善を達成することを示す．加えて，品質管理を行う上でラベルをランダムに付けたり，全て同じラベルを付けるようなワーカ (スパムワーカ) の存在も問題となるため，提案手法である RAkEL GLAD が既存手法よりもスパムワーカに対する耐性があることを実験により示す．

連絡先: 吉村皐亮, ykosuke@ml.ist.i.kyoto-u.ac.jp

2. 問題設定

m 個の対象に対して，n 人のワーカにマルチラベル分類タスクをクラウドソーシング上で依頼する．マルチラベル分類タスクとは，一つの対象が複数のラベル (ラベル集合) に分類されうる分類タスクで，ワーカ i が対象 j に付けたラベル集合を $\boldsymbol{l}_{ij} \subseteq \mathbf{L}_{\mathrm{T}}$ で表す．ここで，$\mathbf{L}_{\mathrm{T}}$ は候補ラベル集合とする．ただし，n 人のワーカ全てが m 個全ての対象にラベル付けするとは限らない．なお，タスクとはワーカが対象にラベルを付けることと定義する．本研究では，クラウドソーシングで得られた回答集合 $\mathbf{L}=\{\boldsymbol{l}_{ij}\}$ から，各対象 j の真のラベル集合を要素にもつ集合 $\mathbf{Z} = \{\boldsymbol{z}_j\}_{j=1,\cdots,m}$ を推定する．このとき，出力となる推定結果を $\hat{\mathbf{Z}} =\{\hat{\boldsymbol{z}}_j\}$ とする．ただし，$\boldsymbol{l}_{ij}$，$\boldsymbol{z}_j$，$\hat{\boldsymbol{z}}_j$ はいずれもラベル集合を表す．

3. シングルラベル分類タスクの品質管理法

シングルラベル分類タスクでよく用いられる品質管理法として Dawid と Skene による Dawid & Skene モデル [Dawid 79] と Whitehill らによる GLAD モデル [Whitehill 09] がある．本研究ではマルチラベル分類タスクを RAkEL を用いることで，これらシングルラベル分類のための品質管理法に帰着する．Dawid & Skene モデルは臨床医らが患者らに対して行っ

た診察の結果から，各患者の真の疾患を EM アルゴリズムに基づいて推定するモデルである．このモデルにおける臨床医・患者・疾患を，ワーカ・対象・ラベルと置き換えることでシングルラベル分類タスクの品質管理法として用いることができる．GLAD モデルは，ワーカの能力とタスクの難易度を考慮し，EM アルゴリズムに基づき真のラベルを推定するシングルラベル分類タスクの品質管理法である．

4. RAkEL を用いた提案手法

マルチラベル分類タスクをシングルラベル分類に対する品質管理手法へ帰着させる典型的な手法には，Binary Relavance (BR) と Label Powersets (LP) がある [Tsoumakas 07a]．BR は各ラベルごとにシングルラベル分類のための品質管理法を適用し，各ラベルの有無を推定する手法であり．計算量こそ小さいがラベル間の関係を捉えることができない．また，LP は候補ラベル集合の冪集合にシングルラベル分類のための品質管理法を適用し，一つの冪集合を選択的に推定する手法であり，ラベル間の関係を捉えることができるものの，計算量が指数関数的に増大してしまう．つまり，ラベル間の関係を捉えることと計算量の大きさとの間にトレードオフが存在する．

このトレードオフを解決するための手法として RAkEL を利用した帰着法を提案する．つまり，我々はクラウドソーシングにおけるマルチラベル分類問題を RAkEL を用いて Dawid & Skene モデルと GLAD モデルに帰着して解く品質管理法を提案する．

4.1 RAkEL

RAkEL は，小さなラベルのまとまり (サイズ k のラベル集合) ごとにラベル間の関係を考慮した計算をして，各計算結果を統合することで，ラベル間の関係を捉えた上で計算量を抑えた推定が可能となる手法である．

RAkEL では，まず，候補ラベル集合 $\mathbf{L}_{\mathrm{T}}$ の部分集合となるサイズ k のラベルセットを全 ${}_{|\mathbf{L}_{\mathrm{T}}|}C_k$ 通り作成する．この中から重複を許さずにラベルセットを一様ランダムに M 個選び出し，これら M 個の選出されたラベルセットそれぞれに対して LP を適用しシングルラベル分類タスクの品質管理法を実行する．最後に，こうして得られたラベルセットごとの各ラベルが真のラベルである確率を統合し，各タスクのラベルを決定する．

候補ラベル集合が $\{A, B, C\}$，サイズが $k = 2$ で $M = 3$ の場合を考える．まず，サイズ $k = 2$ の部分集合 $\{A, B\}$，$\{B, C\}$，$\{A, C\}$ を作成し，これらから $M = 3$ 個をランダムに選出し，LP を適用する (ただし，今回は全ての部分集合が選出される)．例えば，$\{A, B\}$ に LP を適用する場合を考える．$\{A, B\}$ のべき集合 $\{\{\}, \{A\}, \{B\}, \{A, B\}\}$ を作り，これを小さなラベルのまとまりに対する候補ラベル集合と見なして A と B それぞれが付与される確率をシングルラベル分類のための品質管理法を用いて計算する．これをサイズ $k = 2$ の部分集合のうち選出された M 個について行い，最終的に各ラベルの付与される確率の平均を求め，閾値を上回ればそのラベルを付与すべきであるという推定を行う．

RAkEL の統合手法として，disjoint (RAkEL$_d$) と overlap (RAkEL$_o$) の二つが提案されている [Tsoumakas 11]．RAkEL$_d$ は各ラベルに対する推定を一度ずつしか行わない．一方，RAkEL$_o$ は任意の数 M 個のラベルセットを選び出し，それに従い各ラベルの予測確率の平均値を計算し，閾値 $\theta \in [0, 1]$ を超えればそのラベルが付くと推定し，それ以外の場合にはそのラベルが付かないという推定する．

複数の推定結果を統合することで推定の精度が向上するため，一般的には RAkEL$_d$ より RAkEL$_o$ の方が高い精度を示す [Tsoumakas 11]．したがって，本論文においては特に断らない限りは RAkEL とは RAkEL$_o$ のことを指すこととする．

4.2 提案手法

RAkEL DS は RAkEL で帰着させるシングルラベル分類のための品質管理法として Dawid & Skene (DS) モデルを用いた手法であり，RAkEL GLAD は RAkEL で帰着させるシングルラベル分類タスクの品質管理法として GLAD モデルを用いた手法である．

ここで，帰着とはマルチラベル分類タスクを変形し、シングルラベル分類のための品質管理法を適用してラベル推定を行うことを表す．なお，RAkEL について，$k = 1, M = |\mathbf{L}_{\mathrm{T}}|$ なら BR への帰着と等価であり，$k = |\mathbf{L}_{\mathrm{T}}|, M = 2^{|\mathbf{L}_{\mathrm{T}}|}$ なら LP への帰着と等価であるため，RAkEL を用いた提案手法は，BR に帰着する手法と LP に帰着する手法の一般化となっている。

5. 実験

提案手法が既存手法や多数決よりも高い性能を達成することを確認するために，クラウドソーシングで集めた一つのデータセットと先行研究で用いられた四つのデータセットを用いて比較実験を行った．また，多数決よりも高いスパムワーカ耐性を示すことを確認するために，上述の五つのデータセット全てでスパムワーカを加えた場合の精度低下の比較実験も行った．ただし，推定時にはその他や Neutral と行ったラベルは，ラベルなしと見なした．

5.1 実験の設定

5.1.1 提案手法

本研究の提案手法は RAkEL DS と RAkEL GLAD の二つである．いずれの提案手法も候補ラベル集合の部分集合のサイズは $k = 2$ とし，作成された部分集合を全て推定に用いた．つまり，$M = {}_{|\mathbf{L}_{\mathrm{T}}|}C_2$ とした．ただし，$\mathbf{L}_{\mathrm{T}}$ はその他や Neutral のラベルを除いた候補ラベル集合とする。

5.1.2 比較手法

比較対象には下記の五つを用いた．(1) MV：ワーカが答えたラベルの組み合わせを一つの候補として多数決を採る手法．(2) binary MV：各ラベルごとに多数決を採り，統合する手法．(3) DS (Dawid & Skene) モデル：Dawid と Skene が提案した手法 [Dawid 79] を BR でシングルラベル分類問題に帰着させた手法．(4) D-DS (dependent-DS) モデル：Duan らが提案した手法 [Duan 14]．(5) GLAD モデル：Whitehill らが提案した手法 [Whitehill 09] を BR でシングルラベル分類問題に帰着させた手法．

5.1.3 評価指標

各手法による推定結果を Micro accuracy [Tsoumakas 07b] で評価する．TP_λ，FP_λ，FN_λ，TN_λ をそれぞれラベル λ についての True Positive の件数，False Positive の件数，False Negative の件数，True Negative の件数とし，候補ラベル集合を $\mathbf{L}_{\mathrm{T}}$ とするとき，Micro accuracy は式 (1) で表される．

$$\text{Micro accuracy} = \frac{\sum_{\lambda=1}^{|\mathbf{L}_{\mathrm{T}}|}(\mathrm{TP}_\lambda + \mathrm{TN}_\lambda)}{\sum_{\lambda=1}^{|\mathbf{L}_{\mathrm{T}}|}(\mathrm{TP}_\lambda + \mathrm{FP}_\lambda + \mathrm{FN}_\lambda + \mathrm{TN}_\lambda)} \quad (1)$$

5.2 データセット

5.2.1 映画のカテゴリ分類タスク

このタスクの目標は，与えられた映画をその内容と合致する適切なカテゴリを 1 個ないし複数個選択することである．

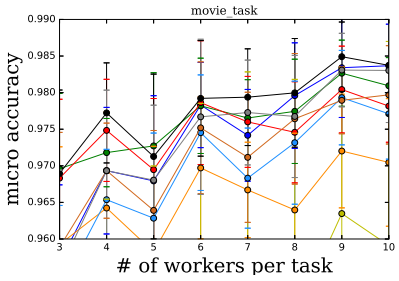

図 1: 評価実験 (映画のカテゴリ分類)

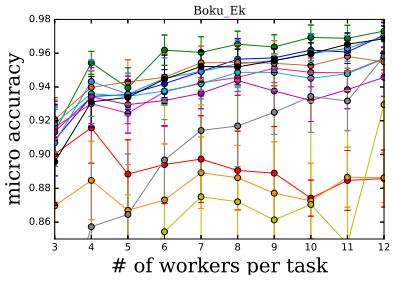

図 2: 評価実験 (Boku_Ek)

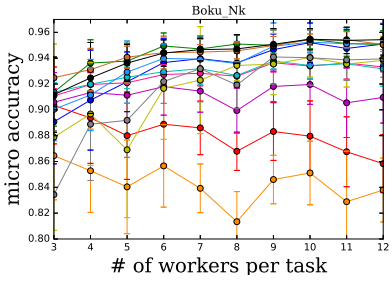

図 3: 評価実験 (Boku_Nk)

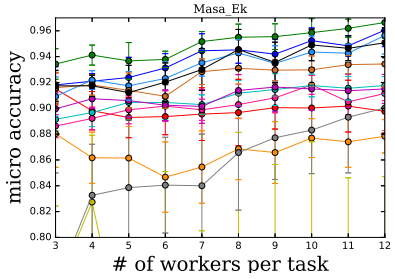

図 4: 評価実験 (Masa_Ek)

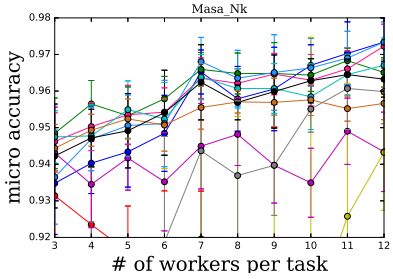

図 5: 評価実験 (Masa_Nk)

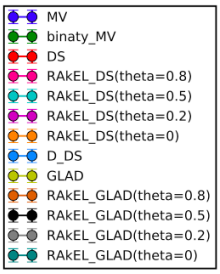

図 6: 図 1〜 図 5 の凡例

分類対象となる映画として映画ランキングドットコム [*1] に掲載されている日本国内歴代総合興行収入ランキングトップ 1000 にランクインしているものから 100 作品を選び，日本のクラウドソーシングサービスであるランサーズ [*2] 上で作業依頼を行った．ワーカは映画名のみを見て，当てはまると思うカテゴリを複数個選択する作業を行う．候補ラベル数は，その他を含めた 20 個とした．一映画あたり 35 人ずつに対象映画のカテゴリ分類作業を依頼した．

5.2.2 小説内の感情分類タスク

このタスクの目標は，小説内の一文から読み取れる登場人物の感情として適切なものを 1 個ないし複数個選択することである．感情を表すラベル集合として，Ekman により提案されたもの [Ekman 92] と中村により提案されたもの [中村 93](以下それぞれ Ek，Nk と表す) を用いる．Ek の候補ラベル数は Neutral を含めて 7 個，Nk の候補ラベル数は Neutral を含めて 11 個である。

本データセットは，Duan らの論文 [Duan 14] で用いられたもので，ラベル付けの対象となる小説は小川未明の「僕たちは愛するけれど」と「政ちゃんの赤いりんご」(以下それぞれ Boku，Masa と表す) である．

5.3 精度評価実験の結果

提案手法である RAkEL DS と RAkEL GLAD はいずれも，$\theta = 0, 0.2, 0.5, 0.8$ の 4 通りの閾値で実験を行った．

5.3.1 映画のカテゴリ分類タスク

映画のカテゴリ分類タスクにおける精度評価実験の結果を図 1 に示す. 提案手法である RAkEL GLAD (閾値 $\theta = 0.5$) が概ね高い精度を達成した．

5.3.2 小説内の感情分類タスク

小説内の感情分類タスクにおける精度評価実験の結果を図 2, 3, 4, 5 に示す．Boku_Nk のデータセット (図 3) では，ワーカ数が 9 人を超えたあたりから提案手法である RAkEL GLAD (閾値 $\theta = 0.5$) の精度が高くなっている．

*1 映画ランキングドットコム (http://www.eiga-ranking.com)
*2 ランサーズ (http://www.lancers.jp)

表 1: 一対象あたりの真のラベル数の平均

データセット名	一対象あたりの真のラベル数の平均
映画のカテゴリ分類	1.9081
Boku_Ek	0.7096
Boku_Nk	1.0000
Masa_Ek	0.8857
Masa_Nk	0.8028

5.4 スパムワーカ耐性の評価実験の結果

元のデータセットから一対象に対して 7 人ずつ回答するようにサンプリングを行い，スパムワーカを追加した場合の精度低下を比較する実験を行った．

5.4.1 映画のカテゴリ分類タスクのスパムワーカ耐性評価

映画のカテゴリ分類タスクにおいて，スパムワーカを加えた場合の Micro accuracy の低下傾向を確認する実験の結果を図 7 に示す．ここで，スパムワーカは全ての対象にラベル付けを行い，その他を含む候補ラベルの中からランダムに一つずつ各対象にラベル付けを行う．提案手法である RAkEL GLAD (閾値 $\theta = 0.2$) は高いスパムワーカ耐性を達成した．

5.4.2 小説内の感情分類タスクのスパムワーカ耐性評価

小説内の感情分類タスクにおいて，スパムワーカを加えた場合の Micro accuracy の低下傾向を確認する実験の結果を図 8，9，10，11 に示す．Boku_Nk では，RAkEL GLAD (閾値 $\theta = 0.2$) のみ高いスパムワーカ耐性を達成した．スパムワーカの数が 5 人程度までは RAkEL-GLAD (閾値 $\theta = 0.5$) のスパムワーカに対する耐性が保たれていることがわかる．

5.5 分析

精度の評価実験では，提案手法である RAkEL GLAD が映画のカテゴリ分類データセットにおいて概ね高い精度を示した．

RAkEL GLAD がほかの手法に比べて精度が高くなるのは，一対象あたりの真のラベル数の平均 (表 1) が多い場合である．映画のカテゴリ分類タスクでは，一対象あたりの真のラベル

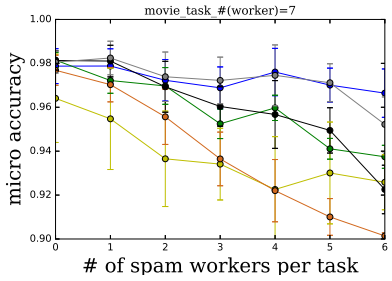

図 7: 耐性比較 (映画のカテゴリ分類)

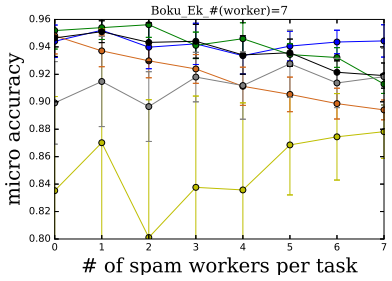

図 8: 耐性比較 (Boku_Ek)

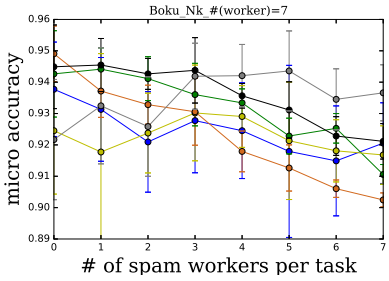

図 9: 耐性比較 (Boku_Nk)

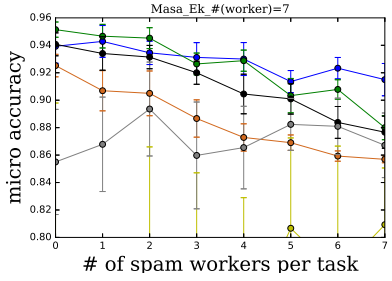

図 10: 耐性比較 (Masa_Ek)

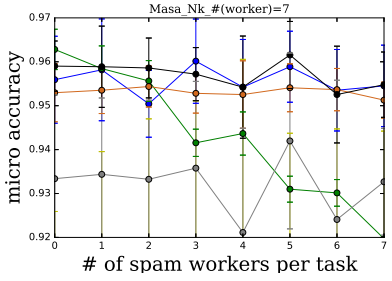

図 11: 耐性比較 (Masa_Nk)

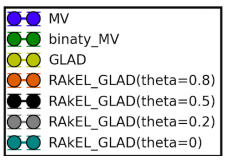

図 12: 図 7〜 図 11 の凡例

数の平均が約 1.9 個であるため，ラベル間の関係の存在を仮定できる．したがって、ラベル間の関係を捉えることで精度向上を図った RAkEL GLAD で高い精度を達成した．一方で，小説内の感情分類タスクでは，いずれのデータセットでも真のラベル数の平均が 1.0 以下であり，ラベル間の関係の存在を仮定できないために，RAkEL GLAD の精度があまり高くならなかったと推測される．

映画のカテゴリ分類タスクにおける RAkEL GLAD のスパムワーカ耐性は閾値によって大きく左右されている．各ワーカの回答数は最大で 100 回，最小で 10 回であり，回答数が少ないワーカの能力をうまく推定できない．その上，スパムワーカが各回答の信頼性を下げてしまう．つまり，真のラベルが選ばれてもスパムワーカによりその信頼性が下げられ，真のラベルである推定確率が低く見積もられてしまうため，スパムワーカ耐性が閾値によって大きく左右されていると考えられる．

また，RAkEL DS に関しては精度改善は見られなかったが，閾値の設定が大きく精度に影響することがわかった．

6. 結論

本研究では，クラウドソーシングを用いて集めたマルチラベル分類タスクに対する回答から，各対象の真のラベル集合を推定する手法を二つ提案した．既存の品質管理法にはラベル間の共起関係が掴めない，又は，ラベル間の共起関係は掴めるが計算量が候補ラベル数に対して指数関数的に増大するという問題点があった．そこで，本研究ではマルチラベル分類問題を RAkEL を用いて既存のシングルラベル分類に対する品質管理法に帰着させることでこの問題を解決した．

既存手法との比較実験によって，提案手法である RAkEL GLAD が真のラベル数が複数である場合に精度改善することを示した．スパムワーカを追加した場合の精度低下比較実験から，RAkEL GLAD が高いスパムワーカ耐性を達成することを示した．提案手法の RAkEL GLAD を用いることで，真のラベル数が複数である場合に従来手法より品質の良い結果が得られ，数人のスパムワーカの存在が仮定される場合にもより信頼性の高い結果が得られることが期待できる．

参考文献

[Dawid 79] Dawid, A. P. and Skene, A. M.: Maximum likelihood estimation of observer error-rates using the EM algorithm, *Applied Statistics*, Vol. 28, pp. 20–28 (1979)

[Duan 14] Duan, L., Oyama, S., Sato, H., and Kurihara, M.: Separate or joint? Estimation of multiple labels from crowdsourced annotations, *Expert Systems with Applications*, Vol. 41(13), pp. 5723–5732 (2014)

[Ekman 92] Ekman, P.: An argument for basic emotions, *Cognition & Emotion*, Vol. 6(3-4), pp. 169–200 (1992)

[Tsoumakas 07a] Tsoumakas, G. and Katakis, I.: Multi-label classification: An overview, *International Journal of Data Warehousing and Mining*, Vol. 3(3), pp. 1–11 (2007)

[Tsoumakas 07b] Tsoumakas, G. and Vlahavas, I.: Random k-labelsets: An ensemble method for multilabel classification, in *Proceedings of the 18th European Conference on Machine Learning*, pp. 406–417 (2007)

[Tsoumakas 11] Tsoumakas, G., Katakis, I., and Vlahavas, I.: Random k-labelsets for multi-label classification, *IEEE Transactions on, Knowledge and Data Engineering*, Vol. 23(7), pp. 1079–1089 (2011)

[Whitehill 09] Whitehill, J., Wu, fan T., Bergsma, J., Movellan, J. R., and Ruvolo, P. L.: Whose vote should count more: Optimal integration of labels from labelers of unknown expertise, in *Advances in Neural Information Processing Systems 22*, pp. 2035–2043 (2009)

[中村 93] 中村 明：感情表現辞典, 東京堂出版 (1993)