ラベルノイズに頑健な Dual Encoder 学習: 大規模マルチラベル分類における自己推定型損失重み付け

牧野拓哉　馬春鵬
株式会社リクルート Megagon Labs, Tokyo, Japan
{makino,ma.chunpeng}@megagon.ai

概要

大規模マルチラベル分類においてラベルノイズは不可避な課題である．既存の Generalized Cross Entropy は，損失の大きさに応じてノイズを判定するため学習が難しい事例をノイズと誤認して学習への影響を低下させ，網羅性を損なう．本研究は意味的類似度に基づく自己推定型損失重み付け手法を提案する．提案手法はラベルの学習中のモデルによる事例とラベルの意味的類似性に応じて，事例–ラベル単位で重みを適応的に与える．具体的には，意味的類似性が低い正例への学習を抑制すると同時に，意味的に類似した負例への学習の寄与を緩和することで潜在的な正例を保護する．人工ノイズ環境下の実験において，提案手法は既存手法と比較して高頻度ラベルの精度 (P@1) および頻度バイアスを除去した精度 (PSP@1) を改善した．

1　はじめに

大規模なラベル空間を扱う大規模マルチラベル分類（Extreme Multi-label Classification; XMC）は，商品カテゴリ予測 [1] や法務・医療文書へのタグ付与 [2, 3] など広く応用されている．XMC におけるラベル分布は典型的なロングテール分布を示し，大多数のラベルは出現頻度が低い．これら低頻度ラベルの正確な予測は，医療分野の希少疾患の診断コード付与や，電子商取引のニッチ商品の推薦など，実応用上重要である．Gupta ら [4] は Decoupled Softmax と呼ばれる対照学習で dual encoder ベースの分類モデルを学習させることで高い分類精度を示すことを報告した．しかしながら，XMC は人手による完全なアノテーションが困難であり，一定のラベルノイズが含まれることは避けられない [5, 6]．誤ったラベルの付与や必要なラベルの欠損はモデルの学習を阻害する要因となる [7]．ノイズに対処するため，Generalized Cross Entropy (GCE) [8] は損失の大きさに依存してノイズを判定し，学習への影響を下げる．XMC には出現頻度が極端に低いラベルや抽象度が高いラベルなど，モデルが確信を持ちにくい学習困難な正例が少なからず存在する．そのため，GCE は「真のラベル」と「ノイズ」を区別できず，確信度が低い事例をノイズとして学習してしまう．

本研究は学習中のモデル自身の予測を用いてラベルノイズを自己推定し，Decoupled Softmax を頑健化する手法を提案する．本手法は「モデルの予測確信度」と「ラベル間の意味的類似性」の 2 つの観点から損失におけるラベルの重要度を適応的に制御する．具体的には，1) 正例ラベルに対して学習がある程度進んだモデルが低い確信度を示すほど，これを偽陽性の疑いが強いと判断し，損失の重みを小さくして学習への悪影響を抑える．2) 負例ラベルに対して正例ラベルと意味的に類似しているほど，偽陰性の疑いがあると判断する．こうしたラベルに対する重みを正例との類似度に応じて減衰することで類似ラベルの学習阻害を防ぐ．XMC で標準的に用いられる EURLEX-4K および LF-Amazon-131K に対して人工的なラベルノイズを注入したデータセットを用いた実験では，提案手法が比較手法に対して精度 (P@1) および頻度バイアスを除去した精度 (PSP@1) を改善させることを確認した． また，学習後の正例ラベルおよび負例ラベルに対する重み分析により，提案手法が正例・負例それぞれのノイズを適切に判別し，学習への影響を抑制していることを明らかにした．

2　Decoupled Softmax

Decoupled Softmax は XMC における dual encoder 学習のための損失関数である [4]．分母から正例ラベルを除外することで，負例ラベルの中に正例ラベルが含まれることによる影響を軽減している．事例

i に対する損失は次のように表される:

$$\ell_i = -\frac{1}{\sum_j^L y_{ij}} \sum_{j=1}^{L} \log \frac{y_{ij} e^{s_{ij}/\tau}}{y_{ij} e^{s_{ij}/\tau} + \sum_{r=1}^{L} (1 - y_{ir})\, e^{s_{ir}/\tau}}, \tag{1}$$

y_{ij} はラベル j $(1 \le j \le L)$ に対する二値であり，正例ラベルは 1，負例ラベルは 0 を取る．τ は温度パラメータであり，s_{ij} は入力テキスト i とラベル j の埋め込みの間のコサイン類似度である．

3 提案手法

提案手法は通常の Decoupled Softmax による warm up を経て，モデルがある程度の識別能力を獲得した段階でラベルノイズの自己推定に基づく損失の重み付け調整へと切り替える．具体的には学習中のモデル自身の予測スコアを信頼度の指標とし，以下の直感に基づいて式 (1) の各ラベルに対する重みを適応的に操作する．

- 偽陽性対策: データ上で正例とされていてもモデルが低い類似度を示すほど，「誤って付与されたラベル」の疑いがあるため，ラベルの重みを小さくして学習への寄与を抑える．
- 偽陰性対策: データ上で負例とされていても正例ラベルと意味的に類似しているほど，「本来付与されるべきラベルの欠損」の疑いがあるため，同様に重みを小さくする．

これらにより事例–ラベル単位で学習におけるノイズの影響を軽減させる．提案手法における事例 i に対する損失は以下のように表される:

$$\ell_i = -\frac{1}{\sum_j^L y_{ij}} \sum_{j=1}^{L} \log \frac{w_{ij}^{+} y_{ij} e^{s_{ij}/\tau}}{w_{ij}^{+} y_{ij} e^{s_{ij}/\tau} + \sum_{r=1}^{L} w_{ir}^{-} (1 - y_{ir})\, e^{s_{ir}/\tau}}. \tag{2}$$

正例ラベルに対する重み w_{ij}^{+} は入力テキスト i とラベル j の類似度 s_{ij} に基づいて計算される．ただし，あらかじめ付与されている正例ラベルに対してのみ重みを計算し，負例ラベルに対しては $w_{ij}^{+} = 0$ とする:

$$w_{ij}^{+} = \begin{cases} \sigma(s_{ij}/\tau) & \text{if } y_{ij} = 1, \\ 0 & \text{if } y_{ij} = 0, \end{cases} \tag{3}$$

σ はシグモイド関数 $1/(1 + e^{-x})$ を表す．正例は学習への影響が大きいため，確信度が著しく低い場合のみ重みを下げるべく，非線形な飽和特性を持つシグモイド関数を採用した．負例ラベルに対する重み

Algorithm 1 自己推定型ラベルノイズ損失重み付け．

Require: 学習データ $\mathcal{D}$, warm up エポック N_{warm}, 総エポック N_{total}, 温度パラメータ τ，warm up した dual encoder のパラメータ θ，類似ラベル数 m
Ensure: dual encoder のパラメータ θ

```
for エポック e = 1 to N_total − N_warm do
    for バッチ B ∈ 𝒟 do
        埋め込み表現から類似度 s_ij を計算
        バッチの各事例 i の正例ラベル j に対する w⁺_ij と負例ラベル r に対する w⁻_ir を計算
        Σ_{i=1}^{|B|} l_i/|B| に基づき θ を更新
    end for
    現在の θ を用いて全ラベルに対する top-m 類似ラベル集合 𝒩_j を再計算
end for
```

w_{ir}^{-} は，正例ラベルとの類似度に基づいて計算される．具体的には，事例 i の正例ラベル集合を $\mathcal{P}_i$ とし，その中で最も類似度が高いラベルとの類似度に基づいて重みを計算する．ここで，$\mathcal{N}_j$ は各ラベル j に対して top-m の類似ラベルの集合を表す:

$$w_{ir}^{-} = \begin{cases} 1 - \max(\bar{s}_{j^* r}, 0) & \text{if } r \in \mathcal{N}_{j^*} \setminus \mathcal{P}_i, \\ 1 & \text{otherwise}, \end{cases} \tag{4}$$

ただし $j^* = \arg\max_{j \in \mathcal{P}_i} \bar{s}_{jr}$ を表す．$\bar{s}$ はラベル埋め込み間の類似度を表し，コサイン類似度とした．類似度を非負とするため $max(\cdot, 0)$ を適用した．負例は類似度による急激な変化を避け，意味的な近さに応じて緩やかに罰則を与える線形関数を採用した．提案手法の学習手順をアルゴリズム 1 に示す．通常の Decoupled Softmax で N_{warm} エポック warm up したのち，提案手法を $N_{total} - N_{warm}$ エポック学習する．w_{ir} を計算するための top-m 類似ラベル集合を得るために，各エポックの終わりに各ラベルに対して top-m 類似ラベルを計算する．このデータは $L \times m$ $(m << L)$ であるためメモリ負荷は小さい．

4 実験

実験には大規模マルチラベル分類で標準的に用いられる EURLEX-4K [9] および LF-Amazon-131K [10] を使用した．ラベルノイズが存在する設定で実験するため，現実的なノイズ環境を想定した 2 種類の人工ノイズを加える．**FP の挿入**: 確率 p_{FP} で事例の埋め込み類似度の高い負例ラベルを正例ラベルに変更する．**FN の挿入**: 確率 p_{FN} で事

表 1　EURLEX-4K と LF-Amazon-131K における実験結果. 括弧内の値は改善度（pt）であり，太字はノイズあり設定における最高値を示す．灰色背景の行はノイズなし設定の結果を示す．

Dataset	Method	Precision		PSP		Recall
		P@1	P@5	PSP@1	PSP@5	R@50
EURLEX-4K 平均ラベル数 5.32 平均トークン数 127.77	Decoupled Softmax (w/o noise)	86.86	60.22	43.37	52.77	87.48
	Decoupled Softmax	84.55	58.42	42.40	51.44	86.18
	+ Label Smoothing	84.63 (+0.08)	**58.70 (+0.28)**	41.85 (-0.55)	51.01 (-0.43)	**86.28 (+0.10)**
	Decoupled GCE	84.27 (-0.28)	58.54 (+0.02)	42.16 (-0.24)	**51.47** (+0.03)	85.47 (-0.71)
	提案手法	**84.71 (+0.16)**	58.35 (-0.07)	**42.62 (+0.22)**	51.38 (-0.06)	86.25 (+0.07)
LF-Amazon-131K 平均ラベル数 2.29 平均トークン数 97.53	Decoupled Softmax (w/o noise)	45.93	22.03	38.15	49.74	71.04
	Decoupled Softmax	43.84	20.88	36.15	46.88	68.90
	+ Label Smoothing	43.58 (-0.26)	20.68 (-0.20)	35.98 (-0.17)	46.39 (-0.49)	69.09 (+0.19)
	Decoupled GCE	43.94 (+0.10)	20.84 (-0.04)	36.31 (+0.16)	46.83 (-0.05)	68.36 (-0.54)
	提案手法	**44.11 (+0.27)**	**21.17 (+0.29)**	**36.46 (+0.31)**	**47.59 (+0.71)**	**69.48 (+0.58)**

例の正例ラベルを削除する．ただし少なくとも 1 つの正例ラベルは残す．p_{FP} と p_{FN} は 0.1 とし，FP 挿入には `bert-base-uncased` を用いて埋め込みを計算した．エンコーダは `distilbert-base` を用いた．本研究ではノイズ環境下での性能を 3 つの観点で評価する．第一に上位予測の品質として，Precision@K (P@k) ($k = 1, 5$) を用いた．第二に低頻度ラベルへの予測精度である．高頻度ラベルへのバイアスを排除するため，逆傾向スコアで重み付けした Propensity scored P@K (PSP@K) ($K = 1, 5$) を採用した [11]．第三に候補生成としての網羅性を測るため，Recall@50 (R@50) を採用した．比較対象としてノイズを考慮しない通常の Decoupled Softmax と，ラベルノイズに対応するための Generalized Cross Entropy [8] を Decoupled Softmax に適用した手法 (Decoupled GCE)，および一般的な正則化手法である Label Smoothing [12] を適用した Decoupled Softmax を採用した．Decoupled GCE は $q = 0.1$ とした[1)]．Label Smoothing は一般的に用いられる $\epsilon = 0.1$ とした．総エポック数 N_{total} はいずれの手法も 100 とした．提案手法の warm up エポック数 N_{warm} は 40，$\tau = 0.05$，$m = 10$ とした．実験には NVIDIA A100-80GB GPU を 2 枚利用し，バッチサイズは EURLEX-4K で 1,024，LF-Amazon-131K で 2,048 とした．損失は勾配キャッシュ [13] を用いて全ラベルを対象として計算した．パラメータ更新には AdamW [14] を用い学習率は EURLEX-4K で 5e-5，LF-Amazon-131K で 3e-4 とした．トークン数は最大 128 に制限した．

1)　開発データを用いて 0.1 から 0.9 まで 0.1 刻みで評価して選択した

5　実験結果

表 1 に EURLEX-4K および LF-Amazon-131K データセットにおける実験結果を示す．まず，EURLEX-4K の結果に着目する．提案手法は P@1 において 84.71%を達成し，ベースライン (+0.16pt) および Label Smoothing (+0.08pt) を上回る精度を示した．一方，Decoupled GCE は P@1 や R@50 を低下させており，XMC のノイズ環境下では単純な損失の一般化だけでは不十分であることが示唆される．特筆すべきは提案手法における低頻度ラベル性能（PSP）との両立である．Label Smoothing は R@50 で僅かな改善 (+0.10pt) を示したものの，その副作用として PSP@1 を 0.55pt 悪化させている．これは，一律な平滑化が学習頻度の低い低頻度ラベルの識別を阻害したためと考えられる．対照的に，提案手法は PSP@1 を 0.22pt 向上させつつ，P@1 および Recall も高水準を維持している．これは，提案手法がモデルの確信度に応じて適応的に重みを制御し，ラベルの頻度に関わらずノイズの影響を低減できたためと考えられる．次に LF-Amazon-131K について述べる．本データセットにおいて，提案手法は R@50 で 69.48%を達成し，ベースライン (+0.58pt) や Label Smoothing (+0.39pt) に対して優位性を示した．Decoupled GCE は P@1 や PSP@1 で僅かな改善が見られるものの，R@50 を 0.54pt 悪化させている．これに対し，提案手法は全ての指標においてベースラインを上回り，特に PSP@5 (+0.71pt) や R@50 (+0.58pt) で顕著な改善を示した．また P@5 においても改善が見られる．

表 2　正例ラベルの重み付け (FP weighting) と負例ラベルの重み付け (FN weighting) の効果 (EURLEX-4K).

Configuration	P@1	PSP@1	R@50
Decoupled Softmax	84.55	42.40	86.18
+ FP weighting only	84.63	42.56	86.23
+ FN weighting only	84.61	42.51	**86.27**
提案手法 (Full)	**84.71**	**42.62**	86.25

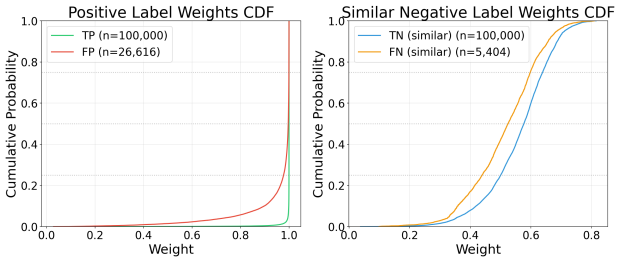

図 1　重み分布の累積分布関数．負例については top-m 類似ラベル（$\mathcal{N}_j$）に属するもののみを対象としている．

6　分析

提案手法における偽陽性対策 (FP weighting) と偽陰性対策 (FN weighting) のそれぞれの寄与を調査した (表 2)．まず正例ラベルの重み付けのみを適用した設定は Decoupled Softmax と比較して P@1 が 84.55 から 84.63 へと向上した．これはモデルが自信を持てない正例ラベルの重みを下げることでノイズラベルによる学習への悪影響を軽減できたためと考えられる．次に負例ラベルの重み付けのみを適用した設定は R@50 が全部設定の中で最も高い 86.27 を示した．これは正例ラベルと類似する負例ラベルに対するペナルティを緩和したことで意味的に妥当なラベルが候補として残りやすくなり，網羅性が向上したことを示している．最後に両者を組み合わせた提案手法は P@1 (84.71) および PSP@1 (42.62) において最高値を達成した．これは FP 対策による精度の向上と FN 対策による類似ラベルの保護が相乗的に機能しノイズ環境下においても頑健な学習が可能となったことを示唆する．提案手法がノイズ（FP, FN）と正例，負例（TP, TN）をどのように区別しているかを検証するため学習終了時点における重みの累積確率分布を可視化した (図 1)．図 1 左側に示す通り，TP (緑線) の重みはほぼ全ての事例で 1.0 付近に維持されている．これは提案手法がノイズ除去を行う過程で，必要な正例データを誤って抑制する副作用が生じていないことを示唆している．一方でノイズである FP (赤線) については，下位約 20%の事例で重みが顕著に抑制されている．XMC ではラベル間の意味的重複が多く，FP の中には文脈的に正例と区別が困難なものも含まれる．提案手法はこれらを一律に排除するのではなく，モデルが確信を持てないノイズをフィルタリングすることで P@1 を改善させている．図 1 の右側から，TN (青線) と FN (橙線) の挙動に差異が見られる．これは式 (4) で導入した類似度に基づく減衰が機能し，正例ラベルと意味的に類似したラベルを FN として保護していることを示唆している．このことが実験結果における R@50 や PSP@1 の維持・向上に寄与していると考えられる．

7　関連研究

XMC における dual encoder の学習の高度化 [4, 15] は盛んに進められているが，これらの手法はラベルノイズを明示的に考慮していない．対照学習の分野では，False Negative が学習を阻害することが理論的に示されている [6]．検索タスク等では教師モデルを用いてこれを除去する手法 [16] もあるが，推論コストが高い cross-encoder を必要とする．マルチラベル分類におけるノイズ対策として ILaCo [17] があるが，$O(L^2)$ メモリを必要とするため大規模マルチラベル分類への適用に制約がある一方で，提案手法は $O(L \times m)$ のメモリで効率的に動作する．ラベルノイズ対策の方法として事例単位でノイズを推定する方法 [18, 19, 20] が提案されているが，本稿が想定する事例–ラベル単位のノイズには対応していない．学習中のモデルの確率に基づきノイズを補正する方法 [21] は提案されているが，本稿は XMC に特化して，学習中のモデルを利用して正解ラベルと類似するラベルの学習への影響を低下させる点で異なる．

8　おわりに

本研究は XMC における dual encoder 学習のラベルノイズに頑健な学習手法を提案した．提案手法は予測確信度と意味的類似性に基づいて，事例–ラベル単位でノイズを推定して学習における重みを適応的に調整する．EURLEX-4K と LF-Amazon-131K に人工的に False Negative および False Positive を加えた実験において，提案手法は，比較手法よりも高い P@1 および PSP@1 を示した．重み分析により提案手法がラベルノイズの学習への影響を部分的に軽減していることも確認された．

参考文献

[1] Hsiang-Fu Yu, Kai Zhong, Jiong Zhang, Wei-Cheng Chang, and Inderjit S Dhillon. PECOS: Prediction for Enormous and Correlated Output Spaces. **Journal of Machine Learning Research (JMLR)**, 2022.

[2] Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Extreme Multi-Label Legal Text Classification: A Case Study in EU Legislation. In **Proceedings of the Natural Legal Language Processing Workshop (NLLP)**, pp. 78–87, 2019.

[3] Yang Liu, Hua Cheng, Russell Klopfer, Matthew R. Gormley, and Thomas Schaaf. Effective Convolutional Attention Network for Multi-Label Clinical Document Classification. In **Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 5941–5953, 2021.

[4] Nilesh Gupta, Devvrit Khatri, Ankit S Rawat, Srinadh Bhojanapalli, Prateek Jain, and Inderjit Dhillon. Dual-Encoders for Extreme Multi-Label Classification. In **Proceedings of the International Conference on Learning Representations (ICLR)**, 2024.

[5] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit S. Dhillon. Large-Scale Multi-Label Learning with Missing Labels. In **Proceedings of the International Conference on Machine Learning (ICML)**, pp. 593–601, 2014.

[6] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased Contrastive Learning. In **Advances in Neural Information Processing Systems (NeurIPS)**, pp. 8765–8775, 2020.

[7] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding Deep Learning (Still) Requires Rethinking Generalization. **Communications of the ACM**, pp. 107–115, 2021.

[8] Zhilu Zhang and Mert Sabuncu. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In **Advances in Neural Information Processing Systems (NeurIPS)**, 2018.

[9] Eneldo Loza Mencía and Johannes Fürnkranz. Efficient Pairwise Multilabel Classification for Large-Scale Problems in the Legal Domain. In **Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)**, 2008.

[10] Julian McAuley and Jure Leskovec. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In **Proceedings of the ACM Conference on Recommender Systems (RecSys)**, pp. 165–172, 2013.

[11] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme Multi-Label Loss Functions for Recommendation, Tagging, Ranking & Other Missing Label Applications. In **Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)**, pp. 935–944, 2016.

[12] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 2818–2826, 2016.

[13] Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. Scaling Deep Contrastive Learning Batch Size under Memory Limited Setup. In **Proceedings of the Workshop on Representation Learning for NLP (RepL4NLP)**, pp. 316–321, 2021.

[14] Ilya Loshchilov and Frank Hutter. Fixing Weight Decay Regularization in Adam. **CoRR**, Vol. abs/1711.05101, , 2017.

[15] Siddhant Kharbanda, Devaansh Gupta, Gururaj K, Pankaj Malhotra, Amit Singh, Cho-Jui Hsieh, and Rohit Babbar. UniDEC: Unified Dual Encoder and Classifier Training for Extreme Multi-Label Classification. In **Proceedings of the ACM Web Conference (WWW)**, pp. 4124–4133, 2025.

[16] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In **Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)**, pp. 5835–5847, 2021.

[17] Pengyu Xu, Mingyang Song, Linkaida Liu, Bing Liu, Hongjian Sun, Liping Jing, and Jian Yu. Noisy Multi-Label Text Classification via Instance-Label Pair Correction. In **Findings of the Association for Computational Linguistics: NAACL**, pp. 1446–1458, 2024.

[18] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. In **Advances in Neural Information Processing Systems (NeurIPS)**, 2018.

[19] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. In **Proceedings of the International Conference on Learning Representations (ICLR)**, 2022.

[20] Dan Qiao, Chenchen Dai, Yuyang Ding, Juntao Li, Qiang Chen, Wenliang Chen, and Min Zhang. SelfMix: Robust Learning against Textual Label Noise with Self-Mixup Training. In **Proceedings of the International Conference on Computational Linguistics (COLING)**, pp. 960–970, 2022.

[21] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin Mcguinness. Unsupervised Label Noise Modeling and Loss Correction. In **Proceedings of the International Conference on Machine Learning (ICML)**, pp. 312–321, 2019.