Wang, Dongsheng; Tiwari, Prayag; Shorfuzzaman, Mohammad; Schmitt, Ingo

Deep neural learning on weighted datasets utilizing label disagreement from crowdsourcing

Published in:
Computer Networks

DOI:
10.1016/j.comnet.2021.108227

Published: 04/09/2021

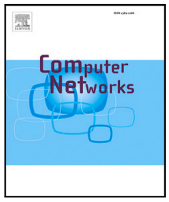
Document Version
Publisher's PDF, also known as Version of record

Computer Networks 196 (2021) 108227

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet



Deep neural learning on weighted datasets utilizing label disagreement from crowdsourcing

Dongsheng Wang [a], Prayag Tiwari [b,c], Mohammad Shorfuzzaman [d], Ingo Schmitt [e,*]

[a] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
[b] Department of Computer Science, Aalto University, Espoo, Finland
[c] Department of Information Engineering, University of Padova, Padova, Italy
[d] Department of Computer Science, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia
[e] Institute of Computer Science and Information and Media Technology, Brandenburg University of Technology Cottbus - Senftenberg, Germany

ARTICLE INFO

Keywords:
Neural networks
Assessed datasets
Instance weight

ABSTRACT

Experts and crowds can work together to generate high-quality datasets, but such collaboration is limited to a large-scale pool of data. In other words, training on a large-scale dataset depends more on crowdsourced datasets with aggregated labels than expert intensively checked labels. However, the limited amount of high-quality dataset can be used as an objective test dataset to build a connection between disagreement and aggregated labels. In this paper, we claim that the disagreement behind an aggregated label indicates more semantics (e.g. ambiguity or difficulty) of an instance than just spam or error assessment. We attempt to take advantage of the informativeness of disagreement to assist learning neural networks by computing a series of disagreement measurements and incorporating disagreement with distinct mechanisms. Experiments on two datasets demonstrate that the consideration of disagreement, treating training instances differently, can promisingly result in improved performance.

1. Introduction

A well labeled dataset is essential for training neural networks where all instances are usually considered correctly labeled. However, some researchers have noticed that dataset collections more or less have inevitable wrong labeling problems, that can mislead the training. Our goal is to take the labeling process into account to reduce the impact of wrong labeling problem. A few works try to use selective attention, e.g., [1], over multiple instances to automatically learn to reduce the weights of noise instances. They achieve this by applying a conditional probabilistic prediction with an accumulative probabilistic optimization function. But this leads to an absence of sound explanation and an absence of insight of labeling process.

Recently, mostly large-scale datasets are assessed on crowd-sourcing platforms where a labeling task can be set up with each instance being assessed by, e.g., three assessors. Then, a deterministic label is aggregated for each instance through majority voting or averaging, sometimes with experts involved. As a result, one instance with one single label dataset is presented. In other words, the disagreement involved in human assessments is hidden in the dataset.

Though crowd and expert work together in order to give rise to a better quality of data, it is not realistic to obtain such labels for a large pool of data. In other words, training on a large dataset more depends on crowdsourced data with aggregated labels. Hereby, we claim that the small high quality dataset can be used as an objective test dataset that enables us to gain an insight how to take full advantage of the crowdsourced dataset together with the disagreement information.

In this paper, instead of dropping these conflicting human assessments, we take advantage of the disagreement during aggregation as a complementary weighting strategy. The underlying assumption is that the disagreement is an important indicator of the label quality that can be used to improve the learning for a better prediction. Specifically, the labeling of certain instances with high disagreement can be caused by

- The instance itself is wrong or with error information (Error),
- The assessors produce spams or have different background knowledge (Spam),
- The instance itself is ambiguous among several labels (Ambiguity), or
- Correct labeling requires too much effort (Difficulty).

Hence, the instances with conflicting labeling would mislead the learning of neural networks. Assuming that these raw assessments are

* Corresponding author.
E-mail addresses: wang@di.ku.dk (D. Wang), prayag.tiwari@aalto.fi (P. Tiwari), m.shorf@tu.edu.sa (M. Shorfuzzaman), ingo.schmitt@b-tu.de (I. Schmitt).

https://doi.org/10.1016/j.comnet.2021.108227
Received 11 November 2020; Received in revised form 4 May 2021; Accepted 4 June 2021
Available online 9 June 2021

available before aggregation, we propose to take advantage of the disagreement of the labeling as a weight of importance.

1.1. Contribution

The main contribution of this paper is follows:

- We first measure the disagreement with entropy and Gini index, then normalize them with Gaussian or rank based weight decision mechanisms.
- Consequently, we define an adapted loss function where the weight decision is adopted in a neural network model, as shown in Fig. 1.
- The experimental results on two scenarios, ideal and realistic, show a promising improvement with our method. By this work, we encourage that the dataset provider can keep the raw assessments before aggregation since they contain more semantics than only spam or noise, from which the dataset consuming community may benefit.

1.2. Organization

This paper is organized as follows: Section 2 discusses the works done on label disagreement. Section 3 provides brief details about the proposed methodology consist of problem formulation, disagreement computation, label distance, entropy-based disagreement, GINI-based disagreement, along other approaches. Experimental results can be found in Section 4 followed with conclusion in Section 5.

2. Related works

Imbalanced datasets have been investigated on class level imbalance [2–6]. For instance, in [3], class distribution is measured and was taken into account to weight the imbalanced datasets, and a modified kNN algorithm was proposed to demonstrate the improvement of the adoption. Same with kNN, [4] employs k-means to compute the weights for training samples. [2] focused on reducing the minority class influence, by using a mixed-kernel based weighted extreme learning machine (MK-WELM).

Some researchers proposed alternative decision tree model [7] to improve the classification accuracy by assigning weights to each training instance using naive Bayes classifier.

[8] introduced an active learning method for classification that handles label noise without relying on crowdsourcing. The basic idea is to select those instances of high influence and eliminate noisy labels to assist the classification. This is the opposite of our method while somehow complementary to each other. All of the above-mentioned works rely on datasets with aggregated labels (one instance one label).

The closest work to us is the CrowdTruth measures for language ambiguity [9] on instance level. It shows the benefit that a classifier can gain from ambiguity measures of weighted label. However, the adoption is normalized and re-scaled in a simple linear way, and the adoption of the threshold requires elaborate manipulation on the crowd-sourced raw data. More importantly, they only focus on improving the quality of dataset instead of on the classification kernels while our work focuses on the formalization of disagreement and its adoption on neural learning models.

Supervised learning models are solely dependent upon the ground truth which are annotated by humans. Perhaps, these ground truths are very noisy and also comes from noisy platforms like Amazon Turk.[1] Generally, multiple labels are collected for each example and then combine the outputs to alleviate the noises. In this way, unnecessary annotation takes place at the cost of insufficient labeled instances. Two very basic questions arise; how to learn from the noisy workers in the best way, and how to manage the annotation budget to enlarge the classifier performance?. [10] proposed a novel algorithm for modeling the worker's quality and labels from the noisy crowd sourced data. The proposed model uses the current model to evaluate worker quality from disagreement, and then the model is updated by optimizing the loss function responsible for the current evaluation of worker quality.

The essence of repeated labeling is also analyzed in several papers. For example, [11] did deep research for analyzing the consequences of repeated labeling and demonstrated that it is likely to be dependent upon the cost of labeling as well as the respective cost of obtaining an unlabeled instance. [12] demonstrates that repeated labeling is very essential if the work quality is below the threshold value. [13,14] mentioned that the expressiveness of any classifier, as well as several factors assessed by others, also play an important role.

3. Methodology

3.1. Problem formulation

Assume there is a set of N labeled instances $X = \{(x_i, l_i, A_i)\}$ where x_i is the instance, l_i is its deterministic label that is aggregated from assessment set A_i where $A_i = \{a_k\}$, indicating for each instance there are k workers giving their assessment $a_k \in L$ from the label set $L = \{l_1, \ldots, l_M\}$. We also introduce the disagreement values of $X^d = \{x_1^d, x_2^d, \ldots, x_N^d\}$, computed from assessments A in different ways as discussed in Section 3.2.

Given a neural network with parameters θ, the softmax layer that outputs the probability distribution $P(l_i|x_i, \theta)$ of a neural network can be expressed as below,

$$P(l_i|x_i, \theta) = softmax(M(x_i h) + b) \tag{1}$$

where $softmax(o_i) = \frac{e^{o_i}}{\sum_{l \in L} e^{o_l}}$ is the output of the neural network which corresponds to the probabilities associated with all labels L. M as function reeturns the vector of the labels; h is the hidden layer, e.g., CNNs, RNNs, or Transformers; and b is a bias. We have the predicted label with the maximum probability, indicated as l_p. Then, we define the loss function as below,

$$loss = \sum_{i=1}^{N} weight(x_i^d) * dist(l_p - l_i) \tag{2}$$

where we incorporate the weight decision $weight(x_i^d)$ for each instance in this formula, i.e., the bigger the weight the more influence does the instance have. Subsequently, in our assumption, the labels of training data l_i is a probability for each label instead of a single aggregated one. Therefore, we define the new version of loss function as Eq. (3) where we replace l_i with l_m of the label set L.

$$loss' = \sum_{i=1}^{N} \sum_{m=1}^{M} weight(x_i^d) * dist(l_p - l_m) \tag{3}$$

As a result, we focus on two factors, the computation of disagreement x_i^d (discussed in Section 3.2) and then the weight decision $weight(x_i^d)$ based on disagreement (introduced in Section 3.3).

3.2. Disagreement computation

We employ a variety of variants to compute the disagreement X^d, including entropy, GINI, deviation, etc. as shown in Table 1. Ordinal or graded labels are widely used in different datasets, therefore, we take label distances into account when computing disagreement. We achieve this by bringing alternative versions of disagreement formulas with or without considering label distance. For theory competency, we discuss the ordinal type with label distance as well, but our experiment only covers binary and category type; while we leave that for ordinal type as future work.

[1] https://www.mturk.com/.

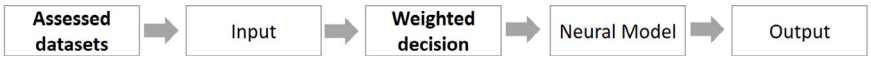


Fig. 1. The scheme of the training on a weighted network.

Table 1
Label types that different disagreement measurements support.

Label type	Example	Entropy	Gini	Deviation
Binary	{0,1}	Yes	Yes	No
Category	{cat,dog,mouse}	Yes	Yes	No
Ordinal	{1,2,3,4,5}	Yes	Yes	Yes

3.2.1. Label distance

For ordinal type, we first define the average labeling distance for A_i as,

$$Dist(A_i, l_i) = \sqrt{\frac{\sum_{k=1}^{K}(a_k - l_i)^2}{K}} \tag{4}$$

where $Dist(A_i)$ computes the distance between each assessor's label a_k and the deterministic label l_i, similar to deviation formula. Then we normalize the distance indicator as $Norm(A_i) = \frac{Dist(A_i, l_i)}{max_j(Dist(A_j, l_j))}$, with value between $[0, 1]$.

3.2.2. Entropy based disagreement

For categorical labels, the disagreement is defined as the entropy among assessments A_i as Eq. (5),

$$x_i^d = H(A_i) = -\sum_{m=1}^{M} fr(l_m) log(fr(l_m)) \tag{5}$$

where $fr(l_m)$ is the relative frequency of the label l_m accumulated from K workers.

For datasets that have ordinal type labels, we have,

$$x_i^d = H'(A_i) = H(A_i) * Norm(A_i) \tag{6}$$

where $Norm(A_i)$ is the normalized distance indicator, meaning the more similar labeling among assessors the less distance and the less disagreement. For example, if two instances with label 1 aggregated from three assessments $\{1, 2, 1\}$ and $\{1, 5, 1\}$ respectively, the entropy values will be the same but we can observe the former case has smaller disagreement than the latter. Thus, the ordinal version penalizes the latter case with larger label distance.

3.2.3. GINI based disagreement

For categorical labels, we have

$$x_i^d = Gini(A_i) = 1 - \sum_{m=1}^{M} fr(l_m)^2 \tag{7}$$

For labels that are ordinal or graded, we have, $x_i^d = Gini'(A_i) = Gini(A_i) * Norm(A_i)$. It is noted that entropy based disagreement has smoother range than Gini based version.

3.2.4. Deviation based disagreement

For completeness, we introduce deviation based disagreement as well, but leave the experimental validation for the future work. Standard deviation is only applicable for ordinal (graded) or numeric labels. It is defined as below,

$$x_i^d = std(A_i) = \sqrt{\frac{\sum_{m=1}^{M}(fr(l_m) - \overline{fr})^2}{M - 1}} \tag{8}$$

where the $\overline{fr}$ is the average frequency of the M labels, N is the total number of the instances.

3.2.5. Confidence based disagreement

Confidence is calculated by some crowd-sourcing websites, e.g. Figure Eight[2]. The confidence is similar to the reversed version of disagreement but calculated with distinct parameters with the internal worker trust calculated based on their historical performance. There are three steps to calculate confidence with Figure Eight. First, $trust(l)$ is defined as the sum of trust of workers who assessed an instance as l as Eq. (9),

$$trust(l) = \sum_{k \in K} t_k * \delta[a_k, l] \tag{9}$$

where t_k is the internal trust of Figure Eight for worker k, and $\delta[a_k, l]$ is a indicator of whether $a_k == l$. Then, the sum of all trust is defined as Eq. (10).

$$trust(L) = \sum_{k \in K} t_k \tag{10}$$

Consequently, the disagreement is defined as Eq. (11),

$$x_i^d = 1.0 - \frac{trust(l)}{trust(L)} \tag{11}$$

where we use 1.0 minus confidence to gain the disagreement.

3.3. Weight decision based on disagreement

In this section, we introduce approaches to compute the weight $weight(x_i^d)$, including normalized, Gaussian distribution and ranked list. We introduce them next.

3.3.1. Normalized weighting

One simple way is to normalize the disagreement values as the weight,

$$weight(x_i^d) = 1.0 - \frac{x_i^d}{max(X^d)} \tag{12}$$

where $\frac{x_i^d}{max(X^d)}$ is the normalized disagreement value. The 1.0 minus the normalized value is to turn the disagreement into "agreement" value since it works as a influence weight in Eq. (2).

3.3.2. Gaussian distribution weighting

Assuming X^d satisfies a Gaussian distribution (normal distribution). Then, we consider awarding some instances while penalizing the other by defining the weight as below,

$$weight(x_i^d) = e^{-x_i^d} \tag{13}$$

where $e^{-x_i^d} \in (0, 1]$. e is the natural constant. The formula makes the instance the more disagreement it contains the less weight it gains.

3.3.3. Rank based weighting

As a complementary, we propose another alternative weighting formula based on rank position. The underlying assumption is that disagreement values vary case by case, especially for those datasets that have few assessors; while the rank of them reflect simple but stable pattern. Therefore, we normalize the weight only based on rank position of disagreement value. We turn X^d into a ranked disagreement values as permutation $X^r = \{x_{(1)}^1, x_{(2)}^2, \ldots, x_{(N)}^N\}$ in a descending order from most disagreement (worst quality) to least disagreement (best quality) instances. It is noted that we take tiers into account, for

[2] https://success.figure-eight.com/hc/en-us/articles/201855939-How-to-Calculate-a-Confidence-Score.

example, values of [0.8, 0.7, 0.7, 0.5] are turned into rank [1, 2, 2, 3]. Then, the weight value is defined as below,

$$weight'(x_i) = log_N(r_i + 1) \tag{14}$$

where we have the root as N so that the weight falls on the range of $[0, 1]$. For instance, the first instance has $log_N 1.0 = 0$ weight, while the last instance has value $log_N N = 1$.

3.4. Models

We choose the most representative deep neural networks, i.e., convolutional neural networks (CNNs) and Transformers (DistilBERT) in our experiment to conduct an apples-to-apples comparison.

For CNNs, we use the multi-scale CNNs in line with the work [15], which concatenates a set of convolutional kernels with different kernel sizes. For Transformer, we use DistilBERT, which is a fast and lightweight version of BERT.

3.4.1. RNN, CNN and transformers

RNN [16–18] and CNN [19–21] are two well-studied successful frameworks of neural networks that perform effectively on various tasks of different datasets. RNN has the advantage of capturing long-term dependencies on a series of data in tasks that can be shaped as "what will happen next given…". In NLP scenario, words are often treated as a series of data in a given sentence. In contrary, CNNs perform significantly successful on image classification tasks, then being adapted to text classification tasks as well with competitive performance [22–27].

A more recent and more powerful deep learning framework is the transformers, such as BERT (Bidirectional Encoder Representations from Transformers) [28]. The core component of the BERT is the Transformer's encoder representation, which practically pre-trains the bidirectional encoder representation on unlabeled texts with masks. Therefore, BERT is also called the masked language model.

3.5. Parameters

We train our model with the following parameter sets: drop rate [0.1,0.2,0.3,0.4,0.5], learning rate [0.01, 0.001, 0.0001], and batch size [32,64,96]. For Transformers, we default and fix with 8 attention heads and 6 layers.

4. Experiment

Taking reality into consideration, we conduct our experiments on two scenarios, the first one is on an ideal dataset that has both the raw information of crowd-sourcing and the deterministic labels after aggregation; while the other is on a realistic dataset with only aggregated labels and confidence for the labels. In real-world scenarios, the latter case is more commonly existing. We discuss both within our framework. Though we introduce an ordinal version disagreement for the completeness of our method, we do not include the validation of them in this section, but leave it as the future work.

4.1. Dataset description

For the ideal scenario, we use the GrowdTruth medical relation extraction dataset [9] where they aggregate the label and the weights of disagreement with different strategies. There is a total of 3984 instances.[3] They include 1043 instances containing treatment relations and 1787 containing causal relations. Agreement of crowd and expert in sentences for negative and positive threshold for Cause, and treat can be seen in Figs. 2 and 3. We use the same partitions as provided with train, valid and test. In particular, the labels we use in train and validate dataset are expected to be "relatively" correct since they are aggregated from crowd, while in the test dataset the labels are expected to be "absolute" correct, by selecting those that have more than 75% agreement between crowd and expert.

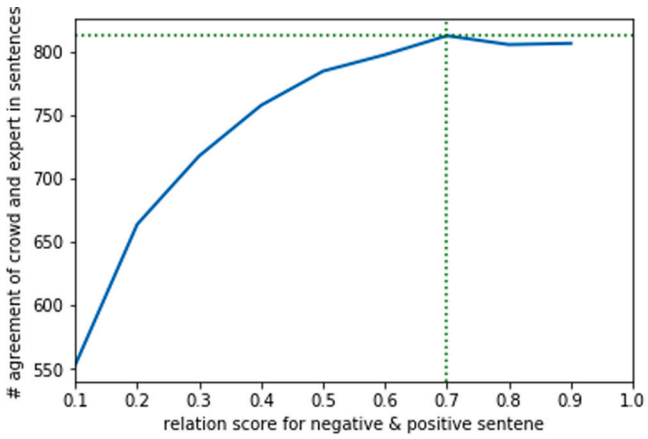


Fig. 2. Agreement of crowd and expert in sentences for negative and positive threshold for Cause.

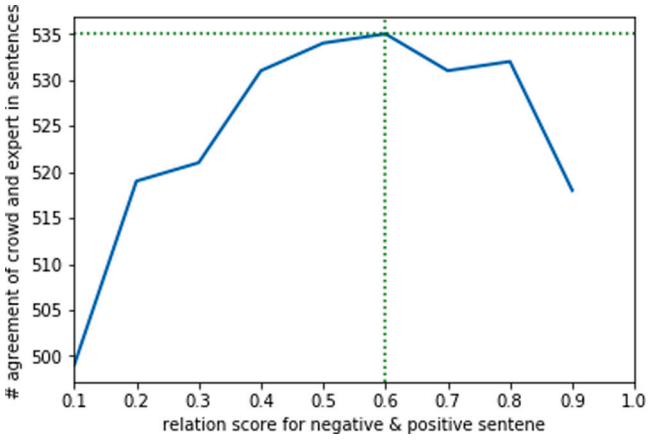


Fig. 3. Agreement of crowd and expert in sentences for negative and positive threshold for Treat.

For a more realistic scenario, we use "Sentiment Analysis – Global Warming/Climate Change" from Figure Eight.[4] Global warm dataset assess tweets for belief in the existence of global warming or climate change. The label is "Yes" if the tweet suggests global warming is occurring, "No" if the tweet suggests global warming is not occurring, and "I cannot tell" if the tweet is either unrelated or ambiguous to global warming. It also includes a confidence score for the classification of each tweet. There is a total of 6090 instances. Following a tradition, we randomly partition the dataset into 80% for training and 20% for testing.

4.2. Hardware setting

The hardware setting is listed below in Table 2. It is necessary to mention here that we used normal CPU server instead of using GPU server.

[3] https://www.figure-eight.com/dataset/medical-sentence-summary-and-relation-extraction/.

[4] https://www.figure-eight.com/data-for-everyone/.

Table 2
Hardware settings.

Property	Modes	CPU	Memory	System	Threads per core
Value	64 bits	40	125G	Ubuntu 14.04	2

Table 3
Performance of different methods on medical info dataset for *cause* relation with CNN encoder.

Weight decision	Precision	Recall	F1
Baseline	0.4363	0.8449	0.5753
Expert	**0.6720**	0.6046	0.6381
Single	0.4958	0.4731	0.4830
[29]	0.620	0.611	0.613
Gini+Gaussian	0.55	0.89	0.68
Gini+Rank	0.58	0.90	0.71
Entropy+Gaussian	0.57	0.91	0.70
Entropy+Rank	0.62	**0.92**	**0.73**

Table 4
Performance of different methods on medical info dataset for *cause* relation with DistilBERT encoder.

Weight decision	Precision	Recall	F1
Baseline	0.4363	0.8449	0.5753
Expert	**0.6720**	0.6046	0.6381
Single	0.4958	0.4731	0.4830
[29]	0.620	0.611	0.613
Gini+Gaussian	0.56	**0.94**	0.70
Gini+Rank	0.53	0.95	0.69
Entropy+Gaussian	0.59	0.90	0.69
Entropy+Rank	0.60	0.90	**0.72**

4.3. Evaluation measures

We used precision, recall and f1 score as a evaluation metric in our paper to check the performance of propose models and baselines. Precision, recall and f score can be defined as below,

	Positive	Negative
Positive	TP(True Positive)	FP(False Positive)
Negative	FN(False Negative)	TN(True Negative)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$F1 = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{18}$$

4.4. Results and discussion

For the ideal datasets, the result is shown in Table 3. "Baseline" is the performance for labels (positive or negative) aggregated for each instance by the distant supervision method, based on whether the relation is expressed between the two terms in the sentence; "Expert" is the performance for labels based on an expert's judgment as to whether the baseline label is correct; and "Crowd" is that for the score used to train the relation extraction classifier [29] with crowd data.

We can observe that for the "cause" relation dataset with the CNN encoder, we achieve an F1 of 0.73 with an absolute 9.2% improvement, while an F1 of 0.858, with a slight improvement of 0.2% in treat relation. For the first one, we have high recall and satisfactory precision, while for the latter, the recall is lower, but the accuracy is satisfactory. In order to check the different encoder, we show the results with DistilBERT on the same datasets, as shown in Table 4. We can observe the best improvement is almost the same or similar with an F1 of 0.72. However, the consistent fact is that the adoption of disagreement will lead to improvements.

For our methods, we found that the combination of entropy-based disagreement and rank gain better performance. This can be that the entropy has a more smooth range for disagreement measurement while the rank based weighting is in a controlled range for distinct datasets.

For realistic dataset of three-class prediction, the result is demonstrated in Table 6. In this case, we do not have a disagreement computation but only a weight decision. The baseline is the result without any weighting, treating each instance equally important. We can observe that rank based measuring achieves the best performance with an absolute improvement of 2.39% accuracy. This brings us the hint that (1) even with the lack of raw data, we still can use the confidence to assist learning (2) different weighting mechanisms of confidence value lead to different performance (see Table 5).

Table 5
Performance of different methods on medical info dataset for *treat* relation.

Weighted version	Precision	Recall	F1
Baseline	0.767	**0.968**	0.856
Expert	0.834	0.60460.833	0.832
Single	0.774	0.856	0.811
[29]	0.823	0.891	0.854
Gini+Gaussian	0.80	0.71	0.802
Gini+Rank	0.79	0.77	0.790
Entropy+Gaussian	**0.846**	0.71	**0.858**
Entropy+Rank	0.832	0.79	0.845

Table 6
Performance of different methods on global warm dataset.

	Accu (adopt confidence)
Baseline	56.62%
Normalized	56.96%
Gaussian	57.86%
Ranked list	**59.01%**

Relation	Train	Validate	Test
Cause	3190	400	400
Treat	3190	400	400

5. Conclusion

We claim that the disagreement behind an aggregated label of an instance contains more semantics than singly spam or noise, which can be employed to assist the learning of neural networks. Therefore, we propose to incorporate the disagreement as instance weight into an adapted loss function in deep neural networks. To achieve this, we measure the disagreement with distinct mechanisms, including entropy and Gini index, followed by a normalization of Gaussian or rank based weighting decision.

The design has the advantage of avoiding threshold analysis from the raw annotating data. We validate our method on two scenarios where one is on an ideal dataset with information from crowdsourcing to aggregation (medical information); while the other is a close to reality dataset with only aggregated label and confidence.

The experiments demonstrate that the weighted decision improves the performance by an absolute improvement of 7.19% (F1) for ideal dataset and an absolute 2.39% (accu) in realistic dataset.

One direction for future work is to combine the label disagreement with instance semantics. The underlying assumption is that an instance weight can be decided by both instance content itself and its labeling conflicts. This can be achieved with the consideration of quantum entropy [30,31].

CRediT authorship contribution statement

Dongsheng Wang: Proposed an idea, Experiment, Writing - original draft. **Prayag Tiwari:** Experiment, Writing - original draft. **Mohammad Shorfuzzaman:** Experiment. **Ingo Schmitt:** Idea, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

M. Shorfuzzaman is grateful to the Taif University Researchers Supporting Project Number (TURSP-2020/79), Taif University, Taif, Saudi Arabia for funding this work. This work was also supported by the Academy of Finland (grants 336033, 315896), Business Finland (grant 884/31/2018), and EU H2020 (grant 101016775).

References

[1] Y. Lin, S. Shen, Z. Liu, H. Luan, M. Sun, Neural relation extraction with selective attention over instances, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 2124–2133.

[2] D. Wu, Z. Wang, Y. Chen, H. Zhao, Mixed-kernel based weighted extreme learning machine for inertial sensor based human activity recognition with imbalanced dataset, Neurocomputing 190 (2016) 35–49.

[3] H. Dubey, V. Pudi, Class based weighted k-nearest neighbor over imbalance dataset, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2013, pp. 305–316.

[4] C. Wang, Z. Pan, C.-S. Ma, L.-L. Dong, T. Zhang, Classification for imbalanced dataset of improved weighted KNN algorithm, Comput. Eng. 20 (2012).

[5] D. Wang, P. Tiwari, S. Garg, H. Zhu, P. Bruza, Structural block driven enhanced convolutional neural representation for relation extraction, Appl. Soft Comput. 86 (2020) 105913.

[6] P. Tiwari, H. Zhu, H.M. Pandey, DAPath: Distance-aware knowledge graph reasoning based on deep reinforcement learning, Neural Netw. 135 (2021) 1–12.

[7] D.M. Farid, C.M. Rahman, Assigning weights to training instances increases classification accuracy, Int. J. Data Min. Knowl. Manage. Process 3 (1) (2013) 13.

[8] M.-R. Bouguelia, S. Nowaczyk, K. Santosh, A. Verikas, Agreeing to disagree: Active learning with noisy labels without crowdsourcing, Int. J. Mach. Learn. Cybern. 9 (8) (2018) 1307–1319.

[9] A. Dumitrache, L. Aroyo, C. Welty, Crowdtruth measures for language ambiguity, in: Proc. of LD4IE Workshop, ISWC, 2015.

[10] A. Khetan, Z.C. Lipton, A. Anandkumar, Learning from noisy singly-labeled data, 2018.

[11] V.S. Sheng, F. Provost, P.G. Ipeirotis, Get another label? improving data quality and data mining using multiple, noisy labelers, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 614–622.

[12] P.G. Ipeirotis, F. Provost, V.S. Sheng, J. Wang, Repeated labeling using multiple noisy labelers, Data Min. Knowl. Discov. 28 (2) (2014) 402–441.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[14] C.H. Lin, M. Mausam, D.S. Weld, Re-active learning: Active learning with relabeling, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[15] J.G. Simonsen, B. Larsen, C. Lioma, The copenhagen team participation in the factuality task of the competition of automatic identification and verification of claims in political debates of the CLEF-2018 fact checking lab, in: CLEF (Working Notes), 2018.

[16] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, S. Khudanpur, Recurrent neural network based language model, in: Eleventh Annual Conference of the International Speech Communication Association, 2010.

[17] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (1997) 2673–2681.

[18] W. Zaremba, I. Sutskever, O. Vinyals, Recurrent neural network regularization, 2014, arxiv preprint arXiv:1409.2329.

[19] Y. Kim, Convolutional neural networks for sentence classification, 2014, arxiv preprint arXiv:1408.5882.

[20] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[21] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 655–665.

[22] S. Lai, L. Xu, K. Liu, J. Zhao, Recurrent convolutional neural networks for text classification, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[23] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, in: Advances in Neural Information Processing Systems, 2015, pp. 649–657.

[24] R. Johnson, T. Zhang, Semi-supervised convolutional neural networks for text categorization via region embedding, in: Advances in Neural Information Processing Systems, 2015, pp. 919–927.

[25] P. Wang, J. Xu, B. Xu, C. Liu, H. Zhang, F. Wang, H. Hao, Semantic clustering and convolutional neural network for short text categorization, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 352–357.

[26] L. Yao, C. Mao, Y. Luo, Graph convolutional networks for text classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 7370–7377.

[27] J.Y. Lee, F. Dernoncourt, Sequential short-text classification with recurrent and convolutional neural networks, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 515–520.

[28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arxiv preprint arXiv:1810.04805.

[29] C. Wang, J. Fan, Medical relation extraction with manifold models, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 828–838.

[30] A. Khrennikov, I. Basieva, Possibility to agree on disagree from quantum information and decision making, J. Math. Psych. 62 (2014) 1–15.

[31] A. Sen, Arithmetic of quantum entropy function, J. High Energy Phys. 2009 (08) (2009) 068.

Dongsheng Wang received his PhD degree from the Department of Computer Science, the University of Copenhagen in Jan. 2021, and his master degree from Korea University in Aug. 2013. He had worked as an engineer for more than two years for Chinese academy of sciences (CAS) where he attended to build up two well-known large-scale knowledge bases, i.e., CASIA-KB and Linked Brain Data (cooperate with EPFL, involving with Human Brain Project). Also, he had several months working in a startup company and Tencent AI department.

His research topics include NLP, Knowledge Graph, CNNs, Transformers, and information retrieval. He was awarded the first in the fact-checking task in CLEF2018 and second place in the "entity linking" task in NLP&CC2013. He has published more than fifteen papers in international conferences and journals and two patents of innovations. Now he is a research scientist in action.ai where he leverages the challenges of conversational AI.

Prayag Tiwari received his MS Degree from NUST MISIS, Moscow. He also worked as a Research Assistant at NUST MISIS, and he has had Teaching and Industrial work experience. He is working as a Postdoctoral Researcher at the Aalto University. Previously, he was working as a Marie Curie Researcher under the QUARTZ project at the University of Padova, Italy.

He has several publications in top journals and conferences. His research interests include Machine Learning, Deep Learning, Quantum-Inspired Machine Learning, Information Retrieval, Health Informatics, and IoT.

Mohammad Shorfuzzaman is currently an Associate Professor with the Department of Computer Science, College of Computers and Information Technology (CCIT), Taif University, Taif, Saudi Arabia. He is also a member of the Big Data Analytics and Applications (BDAAG) Research Group, CCIT. His current research interests include applied artificial intelligence in the areas of computer vision, natural language processing, big data, and cloud computing.

Ingo Schmitt is Professor at department of computer science of the Brandenburg University of Technology. The chair has existed since 1993 and Prof. Ingo Schmitt has been the head of the chair since 2006. His current research interests include quantum logic in different intelligence.