PAPER *Special Section on Data Engineering and Information Management*

A Novel Label Aggregation with Attenuated Scores for Ground-Truth Identification of Dataset Annotation with Crowdsourcing

Ratchainant THAMMASUDJARIT[†a)], Anon PLANGPRASOPCHOK[††b)], *Members*, *and* Charnyote PLUEMPITIWIRIYAWEJ[†c)], *Nonmember*

SUMMARY Ground-truth identification – the process, which infers the most probable labels, for a certain dataset, from crowdsourcing annotations – is a crucial task to make the dataset usable, e.g., for a supervised learning problem. Nevertheless, the process is challenging because annotations from multiple annotators are inconsistent and noisy. Existing methods require a set of data sample with corresponding ground-truth labels to precisely estimate annotator performance but such samples are difficult to obtain in practice. Moreover, the process requires a post-editing step to validate indefinite labels, which are generally unidentifiable without thoroughly inspecting the whole annotated data. To address the challenges, this paper introduces: 1) Attenuated score (A-score) – an indicator that locally measures annotator performance for segments of annotation sequences, and 2) label aggregation method that applies A-score for ground-truth identification. The experimental results demonstrate that A-score label aggregation outperforms majority vote in all datasets by accurately recovering more labels. It also achieves higher F1 scores than those of the strong baselines in all multi-class data. Additionally, the results suggest that A-score is a promising indicator that helps identifying indefinite labels for the post-editing procedure.

key words: *ground-truth identification, crowdsourcing, label aggregation, attenuation scoring*

1. Introduction

Recently, crowdsourcing is a promising cost-effective and time saving solution for dataset annotations by leveraging collective opinions from multiple non-expert annotators [1]. An annotated dataset with crowdsourcing needs another step, namely *ground-truth identification*, to infer the most probable labels, which should ideally be identical to the ground-truth labels, from annotations contributed by many annotators. Such procedure is crucial because annotated labels are inconsistent and noisy due to several factors, including, e.g., different expertise levels, labeling inconsistency, and spamming behaviors of annotators. In the past, a majority voting method is commonly applied since it is intuitive and easy to use. However, the method is prone to noisy labels. Moreover, given an instance annotated by several annotators, the majority vote may produce uncertain aggregation if multiple majority labels occur.

Existing ground-truth identification methods [2]–[13] perform well in many crowdsourcing datasets but they generally do not address an important issue about the post-editing procedure, which requires some human experts to finalize the most probable labels from crowdsourcing annotations. Suppose the majority vote yields 70% aggregation accuracy on a given dataset, which, in turn, implies that each instance in the dataset shares the same label accuracy at 70%. Unfortunately, it is not informative enough for the validation procedure. Specifically, if the validation is needed, all instances have to be investigated anyway, which is especially impractical for a large dataset. This scenario indicates the need for a certain indicator that describes annotation quality at the instance level. Another issue is the impractical assumption about an availability of ground-truth labels for estimating annotator performance. Several previous methods require the good set of data sample that truly represents annotator performance which is difficult to obtain in practice.

This paper presents a novel and practical label aggregation method with similar time complexity to the majority voting method that offers: 1) an indicator manifesting annotation quality at the instance level without using ground-truth labels, and 2) a label recovery technique for uncertain aggregations. The unique part in this work is the attenuated score (*A-score*) which expresses annotator's local performance in a sequence of annotations. The A-score is extended from its original version, proposed in our previous work [14] that does not take annotation orders into account. We also propose the novel label aggregation using A-score, called *A-score aggregation*. Experiments on real crowdsourcing datasets show that the A-score aggregation does not only yield the better aggregation quality and quantity than the majority voting but also obtain higher F1 scores for all multi-class data sets, as compared to many strong baselines.

By using the A-score aggregation, dataset annotation project can be accomplished without using true labels. Dataset owners can obtain larger number of annotated instances, compared to those inferred from the baseline approach. Moreover, the post-editing procedure is more ro-

Manuscript received June 29, 2016.
Manuscript revised November 5, 2016.
Manuscript publicized January 17, 2017.

[†]The authors are with Faculty of Information and Communication Technology, Mahidol University, Thailand.
[††]The author is with National Electronics and Computer Technology Center, Thailand.
a) E-mail: ratchainant.tha@student.mahidol.ac.th
b) E-mail: anon.plangplasopchok@nectec.or.th
c) E-mail: charnyote.plu@mahidol.ac.th

DOI: 10.1587/transinf.2016DAP0024

bust than using inter-rater agreement that could be unstable with missing values. Our A-score aggregation also makes the post-editing easier with the A-score indicator. In particular, the score suggests experts, who finalize the data labels, to pay more attention on low A-score instances, whose inferred labels are unlikely precise due to annotators' disagreement and poor recent performance.

2. Ground-Truth Identification in Crowdsourcing

For dataset annotation with crowdsourcing, the ground-truth identification is the problem of inferring the most probable label from a collection of annotated labels contributed by multiple annotators on a given instance. For example, in the music genre classification task, three annotators give annotated labels after listening a given pop music as follows: {*pop*, *rock*, *pop*}. The ground-truth label is either pop or rock depending on ground-truth identification method.

Formally, the ground-truth identification is defined as follows: Let m be the number of instances, n be the number of annotators, $x_i \in \mathrm{x}^{m\times 1}$ be the observed instance i, $y_{i,j} \in \mathrm{Y}^{m\times n}$ be the annotated label on the instance i by the annotator j, $\mathrm{y}_{\bullet,j}$ be a collection of labels given by the annotator j, and $z_i \in \mathrm{z}^{m\times 1}$ be the identified ground-truth label of the instance i.

Table 1 shows the formal representation of ground-truth identification. There are two main approaches: *label discrimination* and *label aggregation*. For label discrimination, the ground-truth label is identified from an annotated label that has the highest probability among a given collection of annotated labels [3], [5]–[8], [10]. This approach estimates model parameters using EM algorithm [15] and uses those parameters to determine the posterior probability of ground-truth labels. One potential limitation is the number of annotated data might not enough to learn all parameters in order to obtain a robust model. For label aggregation, the ground truth label is obtained from weight aggregation with respect to a particular scheme.

In label aggregation, weighting scheme represents reliability of annotators or degree of expertise. A simple method is majority vote that assumes all annotators are equally reliable with static weight. Previous works improve from majority vote by adjusting the weights more dynamically such as weight majority [2] or worker history of agreement [9] with binary classification. The more advance methods assume a latent confusion matrix bounded with true labels [16] or two latent confusion matrices of instances and annotators [17] in order to estimate annotator reliability. The further extensions take prior distribution of worker confusion matrix into account and apply to multi-classification [4], [18], [19] or ordinal labels [20].

However, label aggregation could generate uncertain labels and impractical. For the uncertainty issue, if there are two or more labels which have the highest weight, the label aggregation method has to choose predicted label randomly from such candidates. This circumstance is called *uncertain aggregation* that makes correct prediction by chance, which is considered as undesirable for ground-truth identification.

Table 1 Formalization of ground-truth identification

x	$\mathrm{y}_{\bullet,1}$	$\mathrm{y}_{\bullet,2}$	...	$\mathrm{y}_{\bullet,j}$	...	$\mathrm{y}_{\bullet,n}$	z
x_1	$y_{1,1}$	$y_{1,2}$	...	$y_{1,j}$	...	$y_{1,n}$	z_1
...	...	...	...	...	...	...	...
x_i	$y_{i,1}$	$y_{i,2}$	...	$y_{i,j}$	...	$y_{i,n}$	z_i
...	...	...	...	...	...	...	...
x_m	$y_{m,1}$	$y_{m,2}$	...	$y_{m,j}$	...	$y_{m,n}$	z_m

For the impractical problem, most of the existing methods still need some ground-truth labels for weight adjustment or parameter estimation. Even though prior probability can be assumed, it is obtained by the known ground-truth labels from a subset of dataset. The difficulty of these methods is to prepare the good data samples with ground-truth labels that represent the distribution of annotator performance.

Furthermore, post-editing is required to finalize the annotated dataset [21]. With existing methods, targeting instances to verify is difficult because they report only overall annotation accuracy. This process is normally performed by analyzing of inter-rater agreement such as Kohen's kappa [22] but it is generally unstable due to the nature of crowdsourcing data that contains very large number of missing values.

3. Attenuated Score

Intuitively, for human, an annotation performed with concentration is likely to be a correct label if annotator's prior knowledge is sufficient for a given task. However, human annotators might not be able to concentrate throughout the whole task because fatigue, boredom, or interruptions are difficult to control. Since crowdsourcing data is hard to identify annotator's concentration, we introduce the A-score that can be observed from an annotated dataset directly.

The A-score is based on the annotator's recent performance. In this paper, the annotator performance on a given instance is the binary outcome whether his annotated label is a member of majority labels or not. Therefore, annotator's recent performance is a sequence of binary numbers within a given range of annotations. The range of annotations, k-gram, starts from the current annotated instance to $k-1$ previously annotated instances. The A-score takes into account the annotation order that makes the annotated instance i more important than its $k-1$ previously annotated instances. This characteristic can discriminate the same pattern size but different order. As an example, consider that 1 indicates the annotated label to be a member of majority labels and 0 otherwise, the annotator's bi-gram recent performance of $\langle 1, 0\rangle$ will be better than that of $\langle 0, 1\rangle$. The A-score is obtained by transforming a sequence of recent performance represented with binary numbers into a decimal number.

Figure 1 shows an overview of A-score calculation from crowdsourcing data. The calculation of the A-score starts with transforming the *crowdsourcing data* into an *an-*

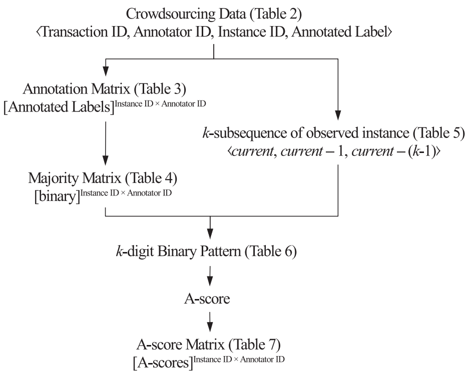
Crowdsourcing Data (Table 2)
⟨Transaction ID, Annotator ID, Instance ID, Annotated Label⟩
Annotation Matrix (Table 3)
[Annotated Labels]^Instance ID × Annotator ID
k-subsequence of observed instance (Table 5)
⟨current, current – 1, current – (k-1)⟩
Majority Matrix (Table 4)
[binary]^Instance ID × Annotator ID
k-digit Binary Pattern (Table 6)
A-score
A-score Matrix (Table 7)
[A-scores]^Instance ID × Annotator ID

Fig. 1 Processes of A-score calculation

Table 2 Example of crowdsourcing data

Transaction ID	Annotator ID	Instance ID	Annotated Labels
1	A0	classic01	pop
2	A1	rock01	classic
3	A2	pop01	rock
4	A0	pop01	pop
5	A1	classic01	classic
6	A2	rock01	rock
7	A0	rock01	rock
8	A1	pop01	pop
9	A2	classic01	classic

Table 3 Annotation matrix

	A0	A1	A2
classic01	pop	classic	classic
pop01	pop	pop	rock
rock01	rock	classic	rock

notation matrix and a *majority matrix*, respectively. Then, for each transaction of crowdsourcing data, a *k-subsequence of observed instances* is extracted and transformed to be a *k-digit binary pattern* with respect to the majority matrix. Next, the k-digit binary pattern is converted into an *A-score* of a label annotated by a particular annotator. Finally, the set of A-score is represented in the form of *A-score matrix*.

The A-score calculation can be described through the following example: Suppose that there are nine transactions of annotated data for music genre classification returned from crowdsourcing, as listed in the Table 2. Each transaction consists of a transaction ID, an annotator ID, an instance ID, and an annotated label. In this example, the crowdsourcing data are annotated by three annotators (A0, A1, and A2) on three unique instances (classic01, pop01, and rock01).

Table 3 shows the annotation matrix corresponding to the crowdsourcing data. The rows and the columns of the matrix represent instances and annotators, respectively. The values in the matrix are the annotated labels given by a particular annotator on his observed instances. This example does not have missing values. However, in practice, the missing values are ubiquitous because annotators generally contribute to some portions of a given dataset.

Table 4 Majority matrix

	A0	A1	A2
classic01	0	1	1
pop01	1	1	0
rock01	1	0	1

Table 5 k-subsequence of observed instances

Transaction ID	Annotator ID	k-subseq of observed instances
1	A0	⟨classic01, N/A⟩
2	A1	⟨rock01, N/A⟩
3	A2	⟨pop01, N/A⟩
4	A0	⟨pop01, classic01⟩
5	A1	⟨classic01, rock01⟩
6	A2	⟨rock01, pop01⟩
7	A0	⟨rock01, pop01⟩
8	A1	⟨pop01, classic01⟩
9	A2	⟨classic01, rock01⟩

Table 6 k-digit binary patterns

Transaction ID	Annotator ID	k-digit binary patterns
1	A0	⟨0, 0⟩
2	A1	⟨0, 0⟩
3	A2	⟨0, 0⟩
4	A0	⟨1, 0⟩
5	A1	⟨1, 0⟩
6	A2	⟨1, 0⟩
7	A0	⟨1, 1⟩
8	A1	⟨1, 1⟩
9	A2	⟨1, 1⟩

Note: the patterns are organized as ⟨current, previous⟩.

Table 4 shows the majority matrix. The value of 1 represents the majority label and 0 otherwise. It is derived from the annotation matrix with respect to the majority vote excluding missing values. For example, at the row of classic01, the annotated labels are {pop, classic, classic}. The result from majority vote is classic; thus, the values in the majority matrix at this row will be $\langle 0, 1, 1\rangle$.

Table 5 shows k-subsequence of observed instances extracted from the crowdsourcing data. Each sequence consists of a current instance and $k - 1$ previous instances. In this case, a sequence of bi-gram ($k = 2$) is observed. For example, in case of annotator A0, the transactions 1, 4, and 7 are extracted and the bi-gram sequences are observed as ⟨classic01, N/A⟩, ⟨pop01, classic01⟩, and ⟨rock01, pop01⟩, respectively.

Table 6 shows k-digit binary patterns which are derived from the sequence of k observed instances and the majority matrix. In our continuing example, in the forth transaction performed by the annotator A0, the pattern ⟨pop01, classic01⟩ is observed. By consulting the majority matrix, the majority score of the instance pop01 annotated by A0 is 1 and that of the instance classic01 annotated by A0 is 0; thus, the corresponding k-digit binary pattern is $\langle 1, 0\rangle$.

Table 7 A-score matrix

	A0	A1	A2
classic01	0	2	3
pop01	2	3	0
rock01	3	0	2

Table 7 shows the A-score matrix whose layout is similar to the majority matrix but its values are A-scores each of which is derived from a binary-to-decimal conversion of a corresponding k-digit binary pattern. For example, the A-score of the k-digit binary pattern of the forth transaction, $\langle 1, 0\rangle$, is $(1 \times 2^1) + (0 \times 2^0) = 2$. Note that each value in the A-score matrix represents the A-score of the current instance annotated by a particular annotator with consideration of $k - 1$ previous annotated instances.

4. Label Aggregation with A-Score

In this section, a formal description of applying A-score to label aggregation, called *A-score aggregation*, is described. Let m be the total number of unique instances and n be the total number of annotators. The A-score aggregation can be defined as follows: Given an annotation matrix $\mathrm{Y}^{m\times n}$, identify a list of ground-truth labels z^m.

Let $y_{i,j} \in \mathrm{Y}^{m\times n}, 1 \leq i \leq m, 1 \leq j \leq n$ be an annotated label on the instance i by the annotator j; thus, $y_{i,\bullet}$ is a collection of annotated labels on the i^{th} instance. The instance i has r_i annotated labels where $1 \leq r_i \leq n$ because annotators do not necessarily contribute to all instances.

The majority matrix $\mathrm{U}^{m\times n}$ can be derived from $\mathrm{Y}^{m\times n}$ as follows: for each $u_{i,j} \in \mathrm{U}^{m\times n}$, $1 \leq i \leq m, 1 \leq j, \leq n$

$$u_{i,j} = \begin{cases} 1 & ,\text{if } y_{i,j} = mode(\mathrm{y}_{i,\bullet}) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $mode(\mathrm{y}_{i,\bullet})$ returns the statistical mode of the collection $\mathrm{y}_{i,\bullet}$. Assume that, for each annotator j, the number of observed instances is c_j, where $1 \leq c_j \leq m$, and the sequence of observed instances is $\mathrm{v}_j = \langle v_j^1, v_j^2, \ldots, v_j^{c_j}\rangle$. The k-subsequence of instances observed at the current position p by the annotator j, where $1 \leq k \leq c_j$, is defined as follows:

$$\mathrm{v}_j(p,k) = \langle v_j^p, v_j^{p-1}, \ldots, v_j^{p-(k-1)}\rangle \tag{2}$$

where v_j^q, $p-(k-1) \leq q \leq p$ represents the row index of the majority matrix. We use it to come up with the corresponding k-digit binary pattern which is defined as follows:

$$\mathrm{w}_j(p,k) = \langle w_j^p, w_j^{p-1}, \ldots, w_j^{p-(k-1)}\rangle \tag{3}$$

where

$$w_j^q = \begin{cases} u_{v_j^q,j} & ,\text{if } q > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Now, an A-score matrix $\mathrm{A}^{m\times n}, 1 \leq i \leq m, 1 \leq j \leq n$ can be obtained. Let $a_{i,j} \in \mathrm{A}^{m\times n}$ be an A-score of the label annotated on the instance i by the annotator j. The $a_{i,j}$ is defined as follows:

$$\begin{aligned} a_{i,j} &= a_{f(p,j),j} \\ &= \sum_q w_j^q \cdot 2^{q-1} \quad ,p-(k-1) \leq q \leq p \end{aligned} \tag{5}$$

where $f(p, j)$ returns the A-score matrix's row index that represents the current instance in crowdsourcing data at the position p observed by the annotator j.

Finally, with the A-score matrix, a list of identified ground-truth labels $\mathrm{z}^m = \langle z_1, z_2, \ldots, z_m\rangle$ can be determined as follows:

$$z_i = y_{i,j} \quad \text{for } 1 \leq i \leq m \text{ and } j = \underset{t}{\arg\max}\,(a_{i,t}) \tag{6}$$

The Eq. (6) is interpreted as follows: *The identified ground-truth label of the instance i contributed by r_i annotators where $1 \leq r_i \leq n$ is the label that has highest A-score*. It is possible that two or more unique labels have the maximum A-score. In this case, the aggregated label need to be sampled from the list of candidates.

For a given annotated dataset of m unique instances by n annotators, the complexity of the majority voting method is defined as $O(mn)$. For the pattern size of k, the complexity of A-score aggregation is defined as $O(kmn)$.

5. Experiment

The experiments have been conducted to evaluate the A-score aggregation on four datasets: music genre (MG), sentiment polarity (SP), dog breed (DB) and adult contents (AC). The first two datasets, available at https://eden.dei.uc.pt/~fmpr/malr/, have been used in the study of logistic regression modeling from multiple annotators [6]. For the dog breed classification dataset, which is a part of the Stanford dog dataset [23], has been used in the study of a crowdsourcing technique using minimax entropy [11]. For the adult contents classification dataset, available at http://ir.ischool.utexas.edu/square/data.html, has been used in SQUARE [24] – a benchmarking framework for ground-truth identification methods. Statistically, the label distributions of MG, SP, DB are relatively uniform while AC is rather skewed. We summarized the dataset characteristics in terms of distributions between annotator accuracies and contributions are summarized in Fig. 2, where dash lines are the average values.

The experiments are comparative studies between the A-score aggregation and other four baselines: majority vote, Raykar's model [4], GLAD [12], and Zencrowd [13] with the following objectives: 1) to evaluate prediction performance 2) to explore the relationship between the A-score values of annotated labels and the probability of ground-truth label and 3) to compare the number of post-editing workloads.

For the first objective, the A-score aggregation with k gram pattern is compared with four methods by precision, recall, F1, and accuracy. Then, we report the comparison of label recovery of A-score aggregation for each size of k. Since the A-score has 2^k possible values, the larger k could have patterns which are unobservable in annotated datasets based on their annotator contribution. Therefore, the k is set to be 2, 3, and 4 which produce 4, 8, and 16 possible patterns, respectively. With these settings, the number of possible patterns are suitable for the average contribution in the four datasets.

For the second objective, we measure the probability of ground-truth label given the A-score of the annotated label. This experiment extends our previous study [14] with the real-world dataset instead of the simulated dataset. Let k be a pattern size, A = $\{0, 1, \ldots 2^k\}$ be a random variable of A-score associated with the identified ground-truth label for the instance i and T = $\{True, False\}$ be a random variable of the comparison between the identified ground-truth label and the real ground-truth label where the *True* represents the label-matched and *False* otherwise. This experiment measures the probability of ground-truth label given the A-score using Eq. (7), (8) and (9).

$$P(T = True|A = \alpha) = \frac{\sum_i f(T, \alpha)}{\sum_i g(T, \alpha)} \tag{7}$$

$$f(T, \alpha) = \begin{cases} 1 & \max(a_{i,\bullet}) = \alpha \text{ and } T = True \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$g(T, \alpha) = \begin{cases} 1 & \max(a_{i,\bullet}) = \alpha \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

For the third objective, the post-editing is assumed to produce the true labels. This experiment explores to what extent the A-score can speed up the post-editing task to achieve the target quality. The post-editing of A-score aggregation re-annotates instances from the lowest to the highest A-score. We compare the growth of ground-truth labels obtained from the A-score aggregation and the majority vote. Raykar, Zencrowd, and GLAD are not included in this experiment because their post-editing procedures are similar to the majority vote although each method has different starting points with respect to their label aggregation quality. For the majority vote, the post-editing re-annotates instances randomly because we assume that only overall aggregation accuracy is known.

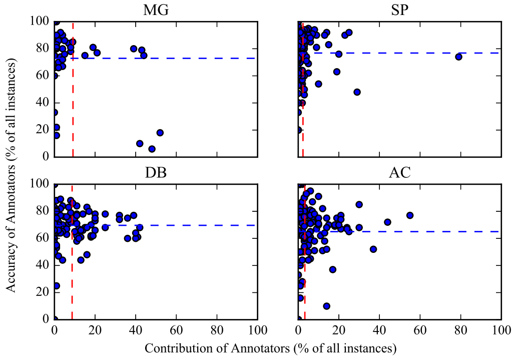

Fig. 2 Dataset characteristics

6. Findings and Discussions

This section reports experimental results. For the first objective, our A-score aggregation with k-gram pattern (ASAGGk) is compared with majority vote (MV), Zencrowd (ZC), Raykar (RY), and GLAD. Their performances are reported in Table 8, comprising of precision, recall, F1 and accuracy values in all datasets. For ASAGGk, almost all numbers are raised when increasing k, which confirms our assumption that the larger k would yield the better performance.

For multi-class datasets, all ASAGGk with $k >= 2$ clearly show superior recall, F1, and accuracy to all other methods in MG, which has the highest proportion of uncertain aggregations at 26.86% (detected by the majority vote). For DB and AC, which have relatively low rates of uncertain aggregations (6.32% and 3.00% respectively), ASAGGk also perform better than the majority vote and achieve higher F1, compared to ZC and GLAD. For SP, a

Table 8 Performance

Datasets	Method	Evaluation			
		Precision	Recall	F1	Accuracy
MG	MV	0.720	0.694	0.695	0.939
	ASAGG2	0.756	0.731	0.734	0.946
	ASAGG3	0.785	0.753	0.756	0.951
	ASAGG4	0.812	**0.774**	**0.776**	**0.955**
	ZC	0.833	0.741	0.749	0.949
	RY	N/A	N/A	N/A	N/A
	GLAD	**0.921**	0.457	0.558	0.938
SP	MV	0.887	0.886	0.886	0.886
	ASAGG2	0.889	0.888	0.888	0.888
	ASAGG3	0.890	0.889	0.889	0.889
	ASAGG4	0.889	0.888	0.888	0.899
	ZC	0.915	0.915	0.915	0.915
	RY	0.911	0.911	0.911	0.911
	GLAD	**0.917**	**0.917**	**0.917**	**0.917**
DB	MV	0.824	0.820	0.820	0.909
	ASAGG2	0.825	0.822	0.821	0.910
	ASAGG3	0.830	**0.827**	**0.826**	**0.912**
	ASAGG4	0.826	0.823	0.822	0.911
	ZC	0.832	0.823	0.821	0.909
	RY	N/A	N/A	N/A	N/A
	GLAD	**0.869**	0.781	0.817	**0.912**
AC	MV	0.724	0.757	0.726	0.860
	ASAGG2	0.725	0.757	0.724	0.858
	ASAGG3	0.729	**0.760**	**0.729**	0.861
	ASAGG4	0.725	0.757	0.725	0.858
	ZC	0.685	0.718	0.672	0.824
	RY	N/A	N/A	N/A	N/A
	GLAD	**0.821**	0.679	0.683	**0.871**

binary-class dataset with a relatively low rate of uncertain aggregations (3.62%), although ASAGGk outperform the majority vote, they are not superior to ZC, RY, and GLAD.

The key factor influencing ASAGGk performance is its ability to capture annotator's recent performance. When $k = 1$, only the current instance will be considered and it is equivalent to majority vote. At k greater than 1, such information is richer than the majority voting case, which helps turning uncertain aggregations to be definitive, or reducing a number of majority labels in some difficult instances. For example, the ground-truth of a given music instance is *classic*. Suppose its annotated labels associate with A-scores are ⟨(*pop*,1),(*rock*,2),(*classic*,3),(*country*,3),(*jazz*,2)⟩. Obviously, all labels are the majority labels. The probability of correct aggregation for the MV becomes 0.2 because it samples one from all majority labels while the probability of correct aggregation for the ASAGGk becomes 0.5 because it samples one from all labels with maximum A-score.

Since the A-score improves label aggregation by recovering labels from uncertain aggregations, the majority vote is only compared because the number of uncertain aggregations in other methods are unobservable. From Table 9, the recovered label quantity is increasing when the size of k is increased. Obviously, each dataset has its appropriate size of k. For example, in Table 9, the A-score 2-gram could not recover label for the DB dataset but increasing the pattern size from 2 to 3 and 4 shows that the A-score aggregation yield more quantity and quality of label recovery. Although increasing the size of k yields higher recovery rate, the behaviors of A-score could be unstable. We suggest to trade-off between quality and the quantity of label recovery by choosing the appropriate k.

For the second objective, behaviors of A-score are studied from the probability to be the ground-truth label of a given annotated label and its A-score, P(ground-truth|A-score). Figure 3 shows the behaviors of A-scores in all datasets. The probability to be ground-truth is proportional to the values of A-score within a certain range of k for each dataset. We observe that the growth of probability to be ground-truth become unstable when k is too large, e.g. $k = 4$ for all datasets. These unstable growth are caused by the rare patterns as k increases.

We also explore the relationship between the annotator accuracy and the average A-score shown in Fig. 4, where each annotator is represented in the scatter dot. The larger dot size indicates that annotator has the greater contribution to the dataset than the smaller dot size. All datasets show the positive relationship between annotator accuracy and average A-score.

For the third objective, dataset annotation projects have the expected target quality for the annotation correctness. To achieve the target quality, the post-editing procedure is required. The post-editing workload is indicated by the number of instances to be validated until the target quality is achieved. Figure 5 shows the growth of ground-truth labels after applying post-editing to annotated instances from the majority vote and A-score aggregations. The projection on the x-axis of the interception between the target quality and the growth of ground-truth labels indicates the post-editing workloads.

Figure 5 demonstrates the stability of the A-score for the post-editing task. Specifically, as validation progressing according to the A-score ranks, the cumulative accuracy has risen constantly. Such ranking is desirable because for each

Table 9 Label recovery quantity

Datasets	Unc Agg. Rate	Methods		
		ASAGG2	ASAGG3	ASAGG4
MG	26.86%	38.89%	65.95%	77.66%
SP	3.62%	2.21%	6.63%	15.47%
DB	6.32%	0.00%	1.96%	5.88%
AC	3.00%	10.00%	20.00%	20.00%

Table 10 Label recovery quality

Datasets	Methods		
	ASAGG2	ASAGG3	ASAGG4
MG	69.33%	63.71%	65.06%
SP	50.00%	66.67%	75.00%
DB	N/A	0.00%	33.33%
AC	0.00%	50.00%	50.00%

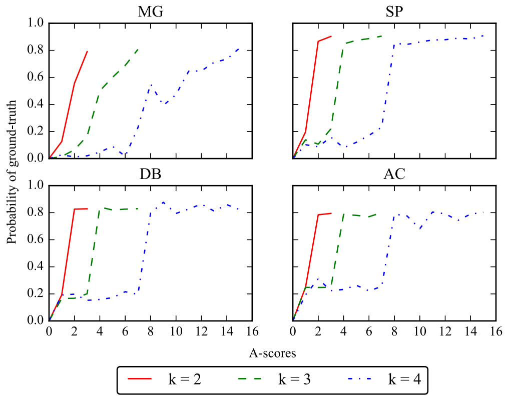

Fig. 3 Behaviors of A-scores

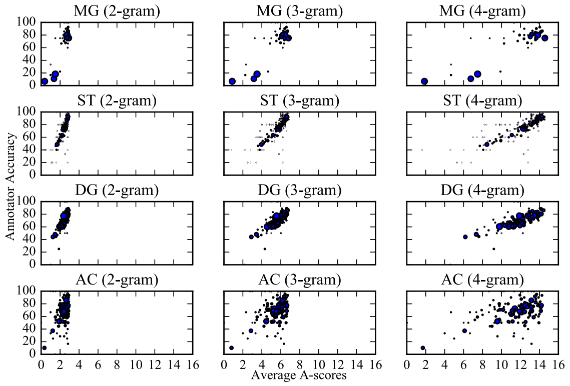

Fig. 4 Relationship between annotator's accuracy and individual's average A-score

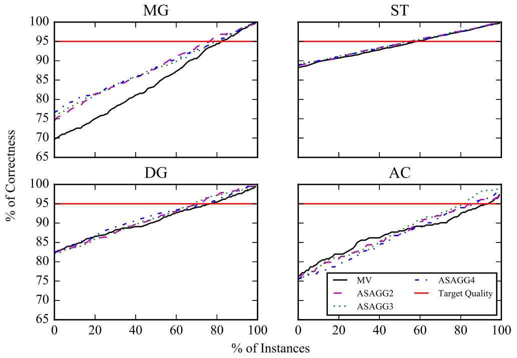

Fig. 5 Post-editing workload

inspection, especially at the very beginning, it keeps correcting the misclassified labels. A method with poor ranking, on the other hand, might rank correctly classified labels at the beginning, which would make small or even no improvement at the beginning of the verification. Unfortunately, it is difficult to perceive in ST, DG and AC datasets as both methods (ASAGGk & MV) roughly yielded the same pattern in terms of their slopes and interceptions. This is firstly because the majority vote and A-score approaches achieves about the same accuracies, which, in turn, yield similar interception values on y-axis. Secondly, the distribution of aggregated A-score in those datasets is heavily skewed to the highest value (approximately 95% of instances for all datasets), implying that annotators in these datasets are somewhat consistent. Consequently, similar improvements appear as the post-editing progresses. For MG dataset, nevertheless, A-score noticeably gave a relatively less steep slope, since fewer incorrect instances remain to be corrected for the A-score case, as comparing to the majority vote case.

7. Conclusions and Future Works

We have presented the A-score and the novel label aggregation for practical ground-truth identification of dataset annotation with crowdsourcing. Our A-score aggregation does not require ground-truth labels of partial dataset for learning and can be performed with similar computational complexity to majority vote that makes it achieves satisfactory processing time. The experimental findings reveal that the A-score aggregation can improve the overall accuracy by recovering labels from uncertain aggregations without deteriorating certain aggregations. The recovery performances depend on the size of annotated pattern and the size of dataset. With our A-score aggregation, new datasets can be created easily and effectively. Although the post-editing workload does not show obvious improvement due to the heavily skewed distribution of aggregated A-score, the lack of instance prioritization is still one of the interesting research directions.

In the future works, the A-score can be further extended by taking instance priority into account. The priority of instances could be very useful for particular types of datasets such as text data that the words are not equally important in term of lexical meaning. Validating the important words which are likely to get wrong annotations could be desirable than correcting functional words those have little lexical meaning. Moreover, the A-score can be used as a random variable for a probabilistic model. Finally, comparative studies between the A-score aggregation, the probabilistic model using A-score, and other existing methods are our research directions.

Acknowledgments

We would like to thank Dr. Jacob Whitehill to share the source code of GLAD implementation.

References

[1] R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng, "Cheap and fast - but is it good? Evaluation non-expert annotiations for natural language tasks," Proc. Conference on Empirical Methods in Natural Language Processing, pp.254–263, 2008.

[2] N. Littlestone and M. Warmuth, "The Weighted Majority Algorithm," Information and Computation, vol.108, pp.212–261, 1994.

[3] P.-Y. Hsueh, P. Melville, and V. Sindhwani, "Data quality from crowdsourcing: a study of annotation selection criteria," Proc. NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, pp.27–35, 2009.

[4] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning From Crowds," J. Mach. Learn. Res., vol.11, pp.1297–1322, 2010.

[5] X. Wu, W. Fan, and Y. Yu, "Sembler: Ensembling Crowd Sequential Labeling for Improved Quality," Proc. 26th Association for the Advancement of Artificial Intelligence, pp.1713–1719, 2012.

[6] F. Rodrigues, F. Pereira, and B. Ribeiro, "Learning from multiple annotators: Distinguishing good from random labelers," Pattern Recognit. Lett., vol.34, no.12, pp.1428–1436, 2013.

[7] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, "Learning Whom to Trust with MACE," Naacl-Hlt '13, vol.3, no.June, pp.1120–1130, 2013.

[8] D. Hovy, B. Plank, and A. Søgaard, "Experiments with crowdsourced re-annotation of a POS tagging data set," Proc. 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp.377–382, 2014.

[9] M. Georgescu, "Aggregation of Crowdsourced Labels Based on Worker History Categories and Subject Descriptors," Proc. 4th International Conference on Web Intelligence, Mining, and Semantics, 2014.

[10] M.J.y. Chung, M. Forbes, M. Cakmak, and R.P.N. Rao, "Accelerating Imitation Learning through Crowdsourcing," Proc. IEEE International Conference on Robotics and Automation, pp.4777–4784, 2014.

[11] D. Zhou, J. Platt, S. Basu, and Y. Mao, "Learning from the wisdom of crowds by minimax entropy," Advances in Neural Information Processing Systems 25, pp.2204–2212, 2012.

[12] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, "Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise," Advances in Neural Information Processing Systems, pp.1–9, 2009.

[13] G. Demartini, D.E. Difallah, and P. Cudré-Mauroux, "Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking," Proc. World Wide Web Conference Committee (IW3C2), pp.469–478, 2012.

[14] R. Thammasudjarit, A. Plangprasopchok, and C. Pluempitiwiriyawej, "Exploring Consensus Sequential Pattern For Ground Truth Identifi-

cation in Corpus Annotation With Crowdsourcing," Symposium on Natural Language Processing, 2016.

[15] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Statistical Society Series B Methodological, vol.39, no.1, pp.1–38, 1977.

[16] A.P. Dawid and A.M. Skene, "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm," Applied Statistics, vol.28, no.1, p.20, 1979.

[17] Y. Tian and J. Zhu, "Learning from crowds in the presence of schools of thought," Proc. 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12, New York, New York, USA, p.226, ACM Press, 2012.

[18] Q. Liu, J. Peng, and A. Ihler, "Variational Inference for Crowdsourcing," Proc. Advances in Neural Information Processing Systems, pp.701–709, 2012.

[19] X. Chen, Q. Lin, and D. Zhou, "Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing," Proc. 30th International Conference on Machine Learning (ICML-13), pp.64–72, 2013.

[20] D. Zhou, Q. Liu, J.C. Platt, and C. Meek, "Aggregating Ordinal Labels from Crowds by Minimax Conditional Entropy," Proc. 31st International Conference on Machine Learning, pp.262–270, 2014.

[21] G. Leech, "Adding Linguistic Annotation," in Developing Linguistic Corpora: a Guide to Good Practice, ed. M. Wynne, pp.17–29, 2005.

[22] A.J. Viera and J.M. Garrett, "Understanding interobserver agreement: The kappa statistic," Family Medicine, vol.37, no.5, pp.360–363, 2005.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Comput. Vis. Pattern Recognit., pp.2–9, 2009.

[24] A. Sheshadri and M. Lease, "SQUARE: A Benchmark for Research on Computing Crowd Consensus," Proc. AAAI Conference on Human Computation and Crowdsourcing, pp.156–164, 2013.

Ratchainant Thammasudjarit is a Ph.D. candidate at Faculty of Information and Communication Technology, Mahidol University, Thailand. He received his B.Eng. degree in Electrical Engineering from Kasetsart University in 2000, M.S. in Management of Information Systems from the King Mongkut's University of Technology North Bangkok in 2008. His research lies in the area of machine learning and natural language processing.

Anon Plangprasopchok is a research scientist at National Electronics and Computer Technology Center (Thailand). He obtained a PhD in the Computer Science Department at the University of Southern California in 2010. His research lies in the area of pattern recognition and machine learning techniques.

Charnyote Pluempitiwiriyawej is an Assistant Professor in the Faculty of Information and Communication Technology at the Mahidol University. He received his B.Eng. degree in Computer Engineering (2nd Class Honors) from the King Mongkut's University of Technology Thonburi in 1994, M.S. in Computer Science from the University of Maryland in 1997, and Ph.D. in Computer Engineering from the University of Florida in 2001. He was a visiting faculty in the Computer Science and Engineering Department at the York University in 2008. His areas of research are data warehousing, data mining, big data analytics, information retrieval, ontology and lexical database development, data management for natural language processing, and data science.