Identifying and Mitigating Labelling Errors in Active Learning

Mohamed-Rafik Bouguelia(✉), Yolande Belaïd, and Abdel Belaïd

Université de Lorraine - LORIA, UMR 7503, 54506 Vandoeuvre-les-Nancy, France
{mohamed.bouguelia,yolande.belaid,abdel.belaid}@loria.fr

Abstract. Most existing active learning methods for classification, assume that the observed labels (i.e. given by a human labeller) are perfectly correct. However, in real world applications, the labeller is usually subject to labelling errors that reduce the classification accuracy of the learned model. In this paper, we address this issue for active learning in the streaming setting and we try to answer the following questions: (1) which labelled instances are most likely to be mislabelled? (2) is it always good to abstain from learning when data is suspected to be mislabelled? (3) which mislabelled instances require relabelling? We propose a hybrid active learning strategy based on two measures. The first measure allows to filter the potentially mislabelled instances, based on the degree of disagreement among the manually given label and the predicted class label. The second measure allows to select (for relabelling) only the most informative instances that deserve to be corrected. An instance is worth relabelling if it shows highly conflicting information among the predicted and the queried labels. Experiments on several real world data show that filtering mislabelled instances according to the first measure and relabelling few instances selected according to the second measure, greatly improves the classification accuracy of the stream-based active learning.

Keywords: Label noise · Active learning · Classification · Data stream

1 Introduction

In usual supervised learning methods, a classification model is built by performing several passes over a static dataset with sufficiently many labelled data. Firstly, this is not possible in the case of data streams where data is massively and continuously arriving from an infinite-length stream. Secondly, manual labelling is costly and time consuming. Active learning reduces the manual labelling cost, by querying from a human labeller only the class labels of data which are informative (usually uncertain instances). Active learning methods [5] are convenient for data stream classification. Several active learning methods [2–5] and stream-based active learning methods [1,6,7] have been proposed. Most of these methods assume that the queried label is perfect. However, in real world scenarios this

A. Fred et al. (Eds.): ICPRAM 2015, LNCS 9493, pp. 35–51, 2015.
DOI: 10.1007/978-3-319-27677-9_3

assumption is often not satisfied. Indeed, it is difficult to obtain completely reliable labels, because the labeller is prone to mislabelling errors. Mislabelling may occur for several reasons: inattention or accidental labelling errors, uncertain labelling knowledge, subjectivity of classes, etc.

Usually, the active learner queries labels of instances that are uncertain. These instances are likely to improve the classification model if we assume that their queried class label is correct. Under such assumption, the active learner aims to search for instances that reduce its uncertainty. However, when the labeller is noisy, mislabelling errors cause the learner to incorrectly focus the search on poor regions of the feature space. This represents an additional difficulty for active learning to reduce the label complexity [8]. If the potential labelling errors are not detected and mitigated, the active learner can easily fail to converge to a good model. Therefore, label noise is harmful for active learning and dealing with it is an important issue.

Detecting label noise is not trivial in stream-based active learning, mainly for two reasons. Firstly, in a data stream setting, the decision to filter or not a potentially mislabelled instance should be taken immediately. Secondly, because the learning is active, the mislabelled instances are necessarily among those that the classifier is uncertain about their class.

Usual methods to deal with label noise like those surveyed in [9], assume that a static dataset is available beforehand and try to clean it before training occurs by repeatedly removing the most likely mislabelled instances among all instances of the dataset. A method proposed in [10] is designed for cleansing noisy data streams by removing the potentially mislabelled instances from a stream of labelled data. However, the method does not consider an active learning setting where the mislabelling errors concern uncertain instances. Moreover, they divide the data stream into large chunks and try to detect mislabelled instances in each chunk, which makes the method partially online and reduces the importance of its streaming nature. The method in [11] considers an active label correction, but the learning itself is not active. Rather, the method iteratively selects the top k likely mislabelled instances from a labelled dataset and presents them to an expert for correction rather than discarding them. Some other methods like [12,13] are designed for active learning with label noise and are intended only for label noise whose source is the uncertain labelling knowledge of the labeller. Generally speaking, they try to model the knowledge of the labeller and avoid asking for the label of an instance if it belongs to the uncertain knowledge set of the labeller. However, this may lead to discarding many informative data. Moreover, the method implicitly assumes that the labeller is always the same (since his knowledge is modelled). Methods like [14,15] can be applied to active learning but they try to mitigate the effect of label noise differently: rather than trying to detect the possibly mislabelled instances, they repeatedly ask for the label of an instance from noisy labellers using crowd-sourcing techniques [16]. However, all these methods require multiple labellers that can provide redundant labels for each queried instance and are not intended to be used with a single alternative labeller.

The method we propose is different. We consider a stream-based active learning with label noise. The main question we address is whether some possibly mislabelled instances among the queried ones are worth relabelling more than others. A potentially mislabelled instance is filtered as soon as it is received according to a mislabelling likelihood. An alternative expert labeller can be used to correct the filtered instance. The method is able to select (for relabelling) only those instances that deserve correction according to an informativeness measure.

This paper is organized as follows. In Sect. 2 we give background on stream-based active learning with uncertainty and its sensibility to label noise. In Sect. 3 we firstly propose two measures to characterize the mislabelled instances. Then, we derives an informativeness measure that determines to which extent a possibly mislabelled instance would be useful if corrected. In Sect. 4 we present different strategies to mitigate label noise using the proposed measures. In Sect. 5 we present the experiments. Finally, we conclude and present some future work in Sect. 6.

2 Background

2.1 Active Learning with Uncertainty

Let X be the input space of instances and Y the output space. We consider a stream-based active learning where at each time step t, the learner receives a new unlabelled instance $x_t \in X$ from an infinite-length data stream and has to make the decision (at time t) of whether or not to query the corresponding class label $y_t \in Y$ from a labeller. Each $x \in X$ is presented in a p-dimensional space as a feature vector $x \stackrel{\text{def}}{=} (x_{f_1}, x_{f_2}, ..., x_{f_p})$, where $x_{f_i} \in \mathbb{R}$ and f_i is the i'th feature.

If the label y_t of x_t is queried, the labelled instance (x_t, y_t) is used to update a classification model h. Otherwise, the classifier outputs the predicted label $y_t = h(x_t)$. In this way, the goal is to learn an efficient classification model $h : X \rightarrow Y$ using a minimal number of queried labels. In order to decide whether or not to query the label of an instance, many active learning strategies have been studied [5]. The most common ones are the uncertainty sampling based strategies. The instances that are selected for manual labelling are typically those for which the model h is uncertain about their class. If an uncertain instance x is labelled manually and correctly, two objectives are met: (i) the classifier avoids to output a probable prediction error, and (ii) knowing the true class label of

Algorithm 1. Stream-based active learning (δ).

1: **Input:** uncertainty threshold δ, underlying classification model h
2: **for** each new data point x from the stream **do**
3: **if** the uncertainty $\Delta_x^h > \delta$ **then**
4: query y the label of x from a labeller
5: train h using (x, y)
6: **end if**
7: **else** output the predicted label $y = h(x)$
8: **end for**

x would be useful to improve h and reduce its overall uncertainty (x is said to be informative). A simple uncertainty measure that selects instances with a low prediction probability can be defined as $\Delta_x^h = 1 - \underset{y \in Y}{max}\, P_h(y|x)$, where $P_h(y|x)$ is the probability that x belongs to class y. A general stream-based active learning process is described by Algorithm 1. Any base classifier can be used to learn the model h. The algorithm queries labels of instances with an uncertainty beyond a given threshold δ.

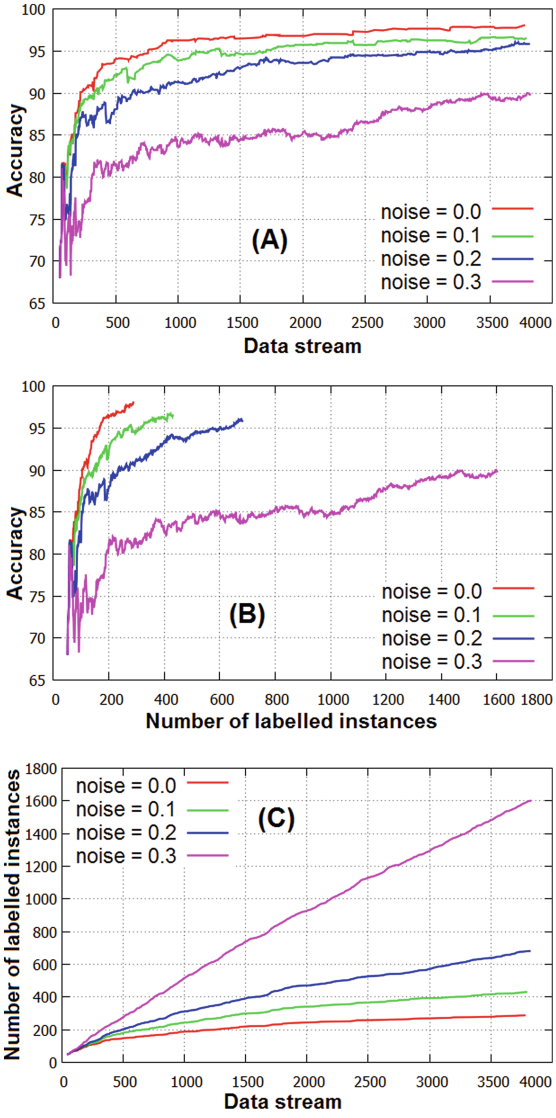

Fig. 1. A stream-based active learning with different levels of label noise (σ). The data set used is optdigits (UCI repository). SVM is used as a base classifier.

2.2 Impact of Label Noise

As mentioned previously, we are not always guaranteed to obtain a perfectly reliable label when querying it from a human labeller. We consider a random label noise process where the noisy labeller has a probability σ for giving a wrong answer and $1 - \sigma$ for giving the correct answer, each time a label is queried.

Figure 1 shows the results obtained using Algorithm 1 in the presence of label noise with different intensities σ and compared to the noise-free setting $\sigma = 0$. Figure 1(A) shows the accuracy of the model h on a test set, according to the number of instances from the stream. As for the usual supervised learning, it is not surprising to see that in active learning, label noise also reduces the overall classification accuracy. Figure 1(B) shows the accuracy according to the number of queried labels (manually labelled instances). We can see that in addition to achieving a lower accuracy, more label noise also causes the active learner to make more queries. This is confirmed in Fig. 1(C) that shows the number of instances whose label is queried, according to the number of instances seen from the stream. This is explained by the fact that the most uncertain instance can be informative if we obtain its true class label, but may easily become the most misleading one if it is mislabelled. Therefore, mislabelled instances causes the active learner to incorrectly focus the query on poor regions of the feature space, and deviates it from querying the truly informative instances.

In summary, stream-based active learning is very sensitive to label noise since it not only impacts the predictive capabilities of the learned model but also leads it to query labels of instances which are not necessarily informative. This results in more queried instances and represents a bottleneck for minimizing the label complexity of active learning [8].

3 Characterizing Mislabelled Instances

In this section we propose measures for characterizing mislabelled instances and their importance. First, we present in Sect. 3.1 a disagreement coefficient that reflects the mislabelling likelihood of instances. Then, we derive in Sect. 3.2 an informativeness measure that reflects the importance of the instance, which is later used to decide if the instance merits to be relabelled.

Let x be a data point whose label is queried. The class label given by the labeller is noted y_g. Let $y_p = \underset{y \in Y}{argmax}\, P_h(y|x)$ be the class label of x which is predicted by the classifier. If $y_p = y_g$ then the label given by the labeller is trusted and we consider that it is not a mislabelling error. Otherwise, a mislabelling error may have occurred.

3.1 Mislabelling Likelihood

Assume that $y_p \neq y_g$. We express how likely x is mislabelled by estimating the degree of disagreement among the predicted class y_p and the observed class y_g, which is proportional to the difference in probabilities of y_p and y_g.

Let $p_p = P(y_p|x)$ and $p_g = P(y_g|x)$. As in the silhouette coefficient [17] and given that $p_p \geq p_g$, we define the degree of disagreement $D_1(x) \in [0, 1]$ as:

$$D_1(x) = \frac{p_p - p_g}{\max(p_p, p_g)} = \frac{p_p - p_g}{p_p} = 1 - \frac{P(y_g|x)}{P(y_p|x)}$$

The higher the value of $D_1(x)$, the more likely that x has been incorrectly labelled with y_g, because the probability that x belongs to y_g would be small relatively to y_p.

Inspired by multi-view learning [18], we present a second measure to estimate the degree of disagreement. In multi-view learning, classifiers are learned on different views of data using different feature subsets. Data points of one view (using a feature f_1) are scattered differently in a second view (using a different feature f_2). Therefore, it is possible for some instance x that we are uncertain if its label is y_p or y_g in one view, to be less uncertain in another view.

Let us take features separately. Each feature value f_i of an instance x has a contribution $q_y^{f_i}$ for classifying x into a class y. As an analogy, a textual document contains terms (features) that attracts towards a given class more strongly than another one. $q_y^{f_i}$ can be considered as any score that shows how much the feature value f_i attracts x towards class y. For example, let x_{f_i} be the instance x restricted to the feature f_i. Let $d_y^{f_i}$ be the mean distance from x_{f_i} to its k nearest neighbours belonging to class y, restricted to feature f_i. Then, $q_y^{f_i}$ can be defined as inversely proportional to the distance $d_y^{f_i}$, e.g. $q_y^{f_i} = \frac{1}{d_y^{f_i}}$.

Considering the predicted class y_p and the given class y_g, $q_{y_p}^{f_i}$ and $q_{y_g}^{f_i}$ represent how much the feature f_i is likely to contribute at classifying x into y_p and y_g respectively. Let F_p be the set of features that contributes at classifying x in the predicted class more than the given class, and inversely for F_g:

$$F_p = \{f_i | q_{y_p}^{f_i} > q_{y_g}^{f_i}\} \qquad F_g = \{f_i | q_{y_p}^{f_i} \leq q_{y_g}^{f_i}\}$$

The amount of information reflecting the membership of x to y_p (resp. y_g) is:

$$q_p = \sum_{f_i \in F_p} (q_{y_p}^{f_i} - q_{y_g}^{f_i}) \qquad q_g = \sum_{f_i \in F_g} (q_{y_g}^{f_i} - q_{y_p}^{f_i})$$

Note that $q_p \in [0, +\infty)$ and $q_g \in [0, +\infty)$. Again, by applying the silhouette coefficient, a degree of disagreement $D_2'(x) \in [-1, 1]$ among y_p and y_g can be expressed as:

$$D_2'(x) = \frac{q_p - q_g}{\max(q_p, q_g)}$$

Note that D_2' can be normalized to be in $[0, 1]$ rather than $[-1, 1]$ simply as $D_2 = \frac{D_2'+1}{2}$.

Instances distributed according to D_1 and D_2 are shown on Fig. 2. A final mislabelling score D can be expressed either by D_1 or D_2 or by using possible combinations of both including: the average $\frac{D_1+D_2}{2}$, $\max(D_1, D_2)$, or $\min(D_1, D_2)$. In order to decide whether an instance x is potentially mislabelled, a usual way is to define a threshold (that we denote t_D) on D.

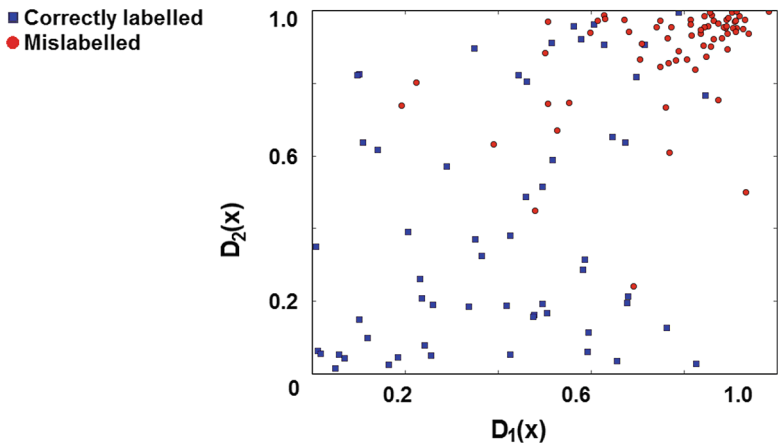

Fig. 2. Mislabelled and correctly labelled instances distributed according to D_1 and D_2.

In summary, the presented disagreement measure only expresses how likely x has been incorrectly labelled. Strong information reflecting the predicted class label and low information reflecting the given class label, indicates a mislabelling error. For example, most terms in a textual document strongly attract towards a class, but the other words weakly attracts towards the class given by the labeller. Nonetheless, the mislabelling score does not give information about the importance of an instance or how much its queried label deserves to be reviewed. This is discussed in the next section.

3.2 Informativeness of Possibly Mislabelled Instances

In active learning with uncertainty sampling, instances for which the model is uncertain on how to label, are designated as informative and their label is queried. In this section we are not referring to informativeness in terms of uncertainty (the considered instances are already uncertain). Rather, we are trying to determine to which extent a possibly mislabelled instance would be useful if corrected.

It is possible for the mislabelling likelihood using D_1 and/or D_2 to be uncertain if an instance x is mislabelled or not. This appears on Fig. 2 as the overlapped region of the mislabelled and the correctly labelled instances. This happens essentially in the presence of either strong or weak conflicting informations in x with respect to y_p and y_g, which leads to $P(y_p|x) \simeq P(y_g|x)$ and $q_p \simeq q_g$. Let us again consider the example of a textual document:

- *Strongly conflicting information*: some terms strongly attract the document towards y_p and other terms attract it with the same strength towards y_g. In this case q_p and q_g are both high and close to each other.
- *Weakly conflicting information*: terms equally but weakly attract the document towards y_p and y_g, that is, there is no persuasive information for y_p or y_g. In this case q_p and q_g are both low and close to each other.

In both the above cases $q_p - q_g$ would be low. However, instances showing strongly conflicting information are more informative if their true class label is available, and deserve to be reviewed and corrected more than the other instances. Therefore, in addition to the mislabelling likelihood (Sect. 3.1), we define the informativeness measure $I \in [0, +\infty)$ as

$$I = \sqrt{q_p \times q_g}$$

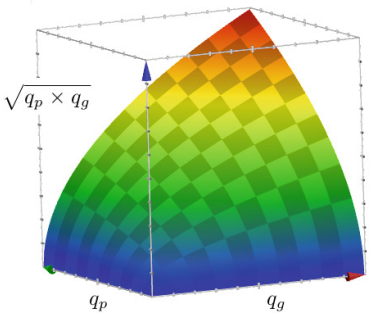

Fig. 3. $\sqrt{q_p \times q_g}$ is high when both q_{y_p} and q_{y_g} are high and close to each other.

The higher I, the more its queried label deserves to be reviewed (and eventually corrected). Figure 3 justifies the choice of $I = \sqrt{q_p \times q_g}$ since it gives a high value when both q_{y_p} and q_{y_g} are high and close to each other. The measure I is unbounded but can be normalized, for example by dividing it on a maximum value of I which can be computed on a validation set. This way, a threshold (that we denote t_I) can be defined on the informativeness measure I.

4 Mitigating Label Noise

When an instance x is detected as potentially mislabelled, the next step is to decide what to do with this instance. In usual label noise cleansing methods, the dataset is available beforehand and the methods just discard or remove the instances that are likely to be mislabelled or predict their labels. In order to mitigate the impact of label noise on the stream-based active learning, we study three strategies including discarding, weighting and relabelling by an expert labeller, and we show a hybrid approach.

4.1 Discard, Weight and Relabel

What not to do. In a stream-based active learning, correcting a mislabelled instance by predicting its labels is not the right way to go. Indeed, updating the model h using the predicted label of x (i.e. $y = h(x)$) rather than the queried label, is usually more harmful to active learning than the label noise itself. This is due to the fact that the mislabelled instances are those instances for which the model h was primarily uncertain how to classify. Therefore, predicting their labels will more likely result in an error that the model would be unable to detect (otherwise it would have avoided that error).

Discarding. When $D(x) > t_D$, x is considered as mislabelled, otherwise it is considered as correctly labelled. This way, if x is identified as being mislabelled, it can be just discarded, that is, we do not update the classification model h with x.

Weighting. Depending on the base classifier, the instances can be weighted so that the classifier learns more from instances with a higher weight. Therefore, a possible alternative to mitigate label noise without defining a threshold on D, is to update the model h using every instance x with its queried label y_g weighted by $w = 1 - D(x)$ which is inversely proportional to its mislabelling likelihood D. Indeed, instances with a high mislabelling likelihood will have a weight closer to 0 and will not affect the classification model h too much, unlike an instance with a low mislabelling likelihood.

Relabelling. If an alternative reliable labeller is available, the label of the potentially mislabelled instance (having a mislabelling likelihood $D > t_D$) can be verified and eventually corrected. Then, the model h is updated using the instance and its corrected label.

Figure 4 shows the results obtained using the different strategies. It is obvious that the best alternative is to relabel correctly the instances that are identified as mislabelled. However, this is done under a cost by an expert labeller which is assumed to be reliable. Discarding the possible mislabelled instances may improve the classification accuracy. However, informative instances that were correctly labelled may be lost, especially if many instances are wrongly discarded (depending on threshold t_D). Rather than discarding possible mislabelled instances, weighting all the instances with an importance which is inversely proportional to their mislabelling likelihood may improve the performances of the active learner without loosing informative instances (at the risk of underweighting some informative instances). Finally, Fig. 4 confirms that it is very harmful to predict the label of a filtered instance x and updating the model using the instance x with its predicted label. Indeed, some dataset cleansing methods for supervised learning propose to predict the label of the potentially mislabelled instances. However, in an active learning configuration, this becomes very harmful, because the queried labels are precisely those of uncertain instances (about which the classifier is uncertain about its prediction).

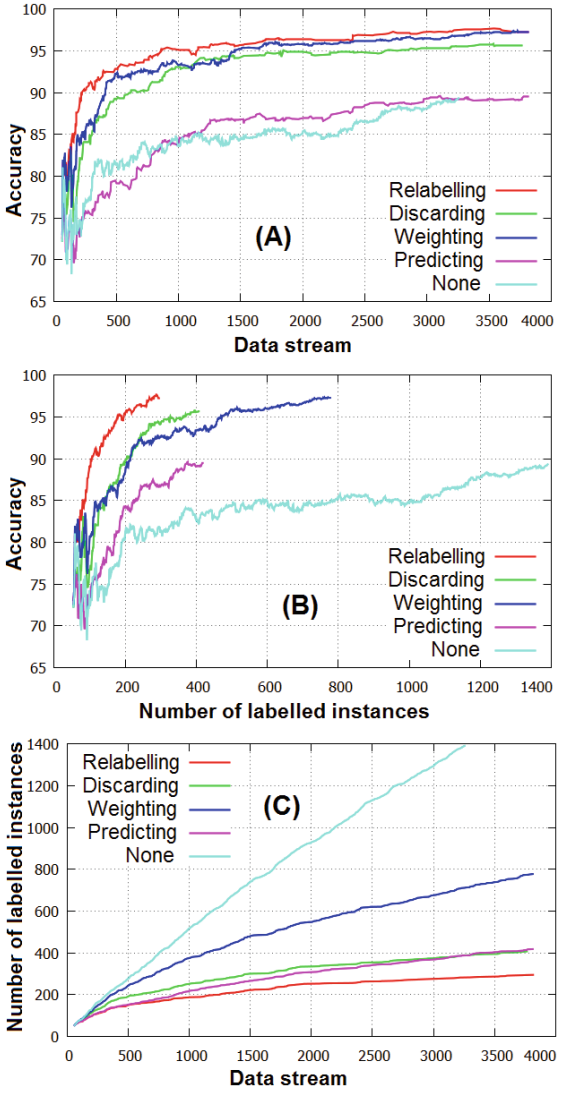

Fig. 4. Active learning from a data stream with label noise intensity $\sigma = 0.3$ and $t_D = 0.6$. The dataset used is optdigits (UCI repository).

Hybrid Strategy. Correcting mislabelled instances using their true class label gives the best results. However, this requires an expert labeller which implies a high cost since it is assumed to be a reliable labeller. We present a hybrid approach that minimizes the cost required by using the alternative expert labeller. Since relabelling is costly, we assume that we have a limited budget B for relabelling, that is, the expert can review and relabel no more than B instances. Given the budget B, the problem can be stated as which instances are worth

to be relabelled. Actually, relabelling instances that are informative according to the measure I (see Sect. 3.2) is more likely to improve the classification accuracy. Therefore, if an instance x is identified as being mislabelled and has a high informativeness $I(x) > t_I$, then it is relabelled. Otherwise, either the discarding or the weighting strategy can be used.

5 Experiments

We use for our experiments different public datasets obtained from the UCI machine learning repository[1]. We also consider a real administrative documents dataset provided by the ITESOFT[2] company. Each document was processed by an OCR and represented as a bag-of-words which is a sparse feature-vector containing the occurrence counts of words in the document. Without loss of generality, the label noise intensity is set to $\sigma = 0.3$. An SVM is used as a base classifier (we use the python implementation available on scikit-learn [21]). Threshold t_D is fixed to 0.6 for all datasets to allow reproducibility of the experiments, although a better threshold can be found for each dataset.

5.1 Mislabelling Likelihood

This experiment figures-out the ability of the proposed mislabelling likelihood to correctly characterize mislabelled instances as proposed in Sect. 3.1, without considering a stream-based active learning. We corrupt the labels of each dataset such that instances with a low prediction probability are more likely to be mislabelled. Then, instances of each dataset are ranked according to degrees of disagreement defined in terms of D_1 and D_2. Results are compared with an entropy measure E which is commonly used in active learning. Instances with a low entropy implies a confident classification, thus, instances for which the classifier disagrees with their queried label are more likely to be mislabelled when they have a low entropy $E = -\sum_{y \in Y} P_h(y|x) \times \log P_h(y|x)$

We select the top n ranked instances of each dataset and we compute 2 types of errors: e_1 represents the correctly labelled instances that are erroneously selected, and e_2 the mislabelled instances that are not selected:

$$e_1 = \frac{\text{correct\_selected}}{\text{correct}} \quad e_2 = \frac{\text{mislabelled\_unselected}}{\text{mislabelled}}$$

where "correct_selected" is the number of correctly labelled instances that are selected as being potentially mislabelled. "mislabelled_unselected" is the number of mislabelled instances that are not selected. "correct" and "mislabelled" are the total number of correctly labelled and mislabelled instances respectively.

[1] http://archive.ics.uci.edu/ml/.
[2] http://www.itesoft.com/.

We also compute the percentage of selected instances that are actually mislabelled *prec*, which represents the noise detection precision:

$$prec = \frac{\text{mislabelled\_selected}}{\text{selected}}$$

where "mislabelled_selected" is the number of mislabelled instances that are selected, and "selected" is the number of selected instances.

Table 1 shows the obtained results. We can see that for *documents* and *pendigits* datasets, D_2 achieves better results than D_1 and inversely for *optdigits* and *letter-recognition* datasets. However combining D_1 and D_2 using $D = \max(D_1, D_2)$ may yield better results than both D_1 and D_2, and always achieves better results than the entropy measure E. This is confirmed by the average results obtained over all the datasets. For the reminder of the experiments we use the D which represents a convenient disagreement measure for almost all datasets.

Table 1. Mislabelling likelihood measures.

	D_1	D_2	D	E
Optdigits dataset				
e_1	**0.67**%	1.49%	1.12%	1.27%
e_2	**1.22**%	3.14%	2.26%	2.61%
prec	**98.44**%	96.53%	97.4%	97.05%
Documents dataset				
e_1	3.51	3.29%	**2.96**%	4.50%
e_2	3.07	2.56%	**1.79**%	5.38%
prec	92.2	92.69%	**93.42**%	90.0%
Pendigits dataset				
e_1	2.34%	2.19%	**2.07**%	2.61%
e_2	1.37%	1.02%	**0.75**%	2.00%
prec	94.75%	95.10%	**95.35**%	94.15%
Letter-recognition dataset				
e_1	**1.82**%	3.2%	1.93%	2.29%
e_2	**21.1**%	24.3%	21.35%	22.18%
prec	**94.87**%	91.03%	94.6%	93.57%
Average results over all datasets				
e_1	2.08%	2.54%	**2.02**%	2.66%
e_2	6.69%	7.75%	**6.53**%	8.04%
prec	95.07%	93.84%	**95.20**%	93.70%

5.2 Label Noise Mitigation

In this experiment we consider a stream-based active learning where the label noise is mitigated according to different strategies: relabelling, discarding and weighting. A hybrid strategy is also considered where only a small number of instances are manually relabelled according to a relabelling budget B. For the hybrid strategy, without any loss of generality, we used in our experiments a budget of $B = 20$ instances allowed to be relabelled. Results with others values of B lead to similar conclusions but are not reported due to the space limitation. The considered strategies are listed below:

- Full relabelling: relabelling every instance x that is identified as mislabelled (i.e. if $D(x) > t_D$, then x is relabelled)
- Full discarding: discarding every instance x that is identified as mislabelled
- Full weighting: using every instance and its queried label (x, y_g) weighted by $w = 1 - D(x)$ to update the classification model.
- Hybrid discarding and relabelling: consists in relabelling an instance that is identified as mislabelled only if it shows a high informativeness ($I(x) > t_I$) and the budget B is not yet exhausted. Otherwise, the instance is discarded.
- Hybrid weighting and relabelling: same as the above hybrid strategy but using weighting instead of discarding.

The results obtained for each strategy are illustrated in Fig. 5 and Table 2. The classification accuracy obtained on a test set is shown on Fig. 5 according to the number of labelled instances. Table 2 shows the final classification accuracy, the final number of instances N_1 whose label was queried from the first (unreliable) labeller, and the number of instances N_2 that are relabelled by the alternative expert labeller (fixed to $N_2 = B = 20$ for the hybrid strategies). Let c_1 and c_2 respectively be the cost required by the first labeller and the expert labeller to label a single instance. It is assumed that $c_2 > c_1$ since the expert labeller is supposed to be reliable (or much more reliable than the first labeller). Then, the labelling cost is $c_1 \times N_1$, the relabelling cost is $c_2 \times N_2$, and the overall cost is $C = c_1 \times N_1 + c_2 \times N_2$.

Firstly, the results on Fig. 5 confirm that the "full relabelling" is obviously the most effective strategy in terms of classification accuracy, since all instances that are identified as being mislabelled are relabelled by the expert. Secondly, the results obtained on all datasets show that discarding the possibly mislabelled instances is not better than the weighting strategy. Actually, it has been observed in many works on label noise cleansing [9,19,20] that learning with mislabelled instances harms more than removing too many correctly labelled instances, but this is not true in the active learning setting, as it is confirmed by the obtained results. This is due to the fact that in the active learning setting, the discarded instances are more likely to improve the classification model if they are correctly labelled, thus, discarding them may negatively impact the performances of the active learning. Fore the same reason, we can see that the "hybrid weighting and relabelling" strategy performs better than the "hybrid discarding and relabelling" strategy. Also, Fig. 5 shows for the "hybrid weighting and relabelling"

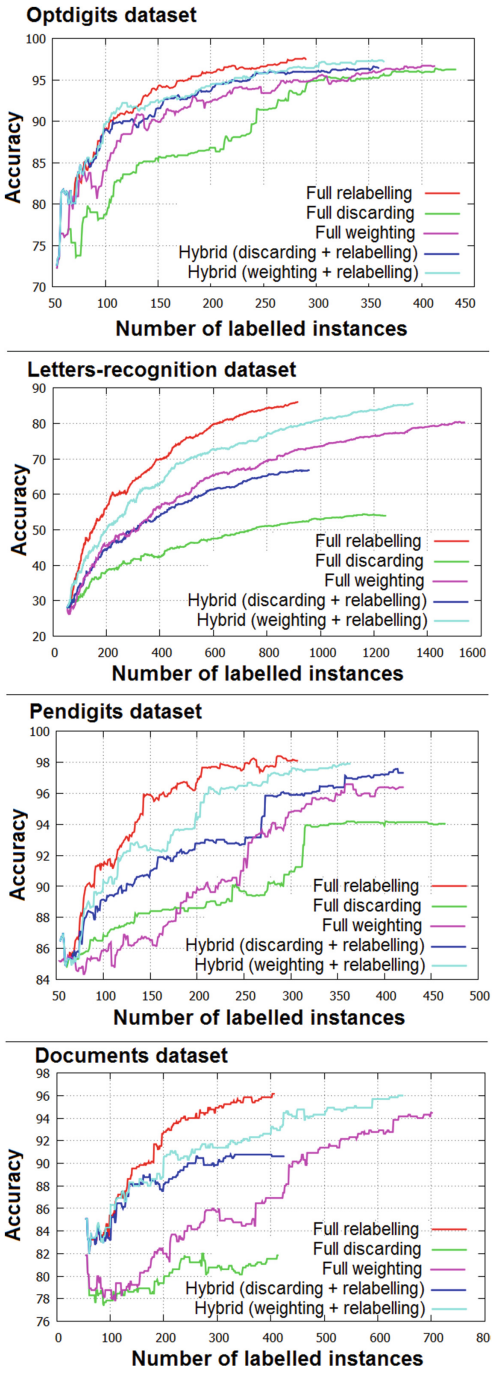

Fig. 5. Classification accuracy according to the number of actively queried labels using different strategies.

strategy that relabelling only $B = 20$ instances with a high value of I, greatly improves the accuracy compared to the "discarding" or the "weighting" strategy. We can see on Table 2 that the "hybrid weighting and relabelling" strategy achieves a final classification accuracy which is pretty close to that of the "full relabelling" strategy. For example, the final accuracy achieved by the "hybrid weighting and relabelling" strategy for the optdigits dataset is 97.21 %, whereas the one achieved by the "full relabelling" strategy is 97.49 %. As explained in Sect. 2.2, mislabelling errors causes the active learner to ask for more labelled instances. This explains why N_1 in Table 2 is smaller in the "full relabelling" strategy. However, by taking into consideration the cost induced by relabelling the mislabelled instances, the "full relabelling" strategy will have a higher overall cost than the other strategies, since all the instances that are identified as being mislabelled are relabelled.

Finally, we can conclude that although the "hybrid weighting and relabelling" strategy has a low relabelling cost, it achieves a final classification accuracy which is close to the one achieved by the "full relabelling" strategy. Therefore, if a limited relabelling budget is available, then this budget should be devoted to relabelling instances with a high informativeness I.

Table 2. Final accuracy and cost according to different strategies.

	Relabel	Discard	Discard and relabel	Weight	Weight and relabel
Optdigits dataset					
Accuracy	97.49 %	96.21 %	96.43 %	96.66 %	97.21 %
N_1	290	432	359	412	364
N_2	92	0	20	0	20
Pendigits dataset					
Accuracy	98.11 %	94.02 %	97.31 %	96.36 %	97.88 %
N_1	307	464	420	420	363
N_2	105	0	20	0	20
Letters-recognition dataset					
Accuracy	85.98 %	53.95 %	66.79 %	80.26 %	85.52 %
N_1	914	1242	956	1537	1344
N_2	211	0	20	0	20
Documents dataset					
Accuracy	96.1 %5	81.84 %	90.61 %	94.46 %	96.0 %
N_1	406	413	424	701	646
N_2	104	0	20	0	20

6 Conclusion and Future Work

In this paper we addressed the label noise detection and mitigation problem in stream-based active learning for classification. In order to identify the potentially

mislabelled instances, we proposed a mislabelling likelihood based on the disagreement among the probabilities and the quantity of information that the instance carries for the predicted and the queried class labels. Then, we derived an informativeness measure that reflects how much a queried label would be useful if it is corrected. Our experiments on real datasets show that the proposed mislabelling likelihood is more efficient in characterizing label noise compared to the commonly used entropy measure. The experimental evaluation also shows that the potentially mislabelled instances with high conflicting information are worth relabelling.

Nonetheless, one limitation of the current hybrid label noise mitigation strategy is that it requires a threshold on the informativeness measure I which depends on the data and its automatic adaptation constitute one of our perspectives. As future work, we want to minimize the correction cost by defining and optimizing a multi-objective function that combines together (i) the mislabelling likelihood, (ii) the informativeness, and (iii) the cost of relabelling instances. Also, in the current work we observed that manually relabelling few instances chosen according to their informativeness I can improve results, but figuring out the number of labelled instances that are required to achieve closer accuracy to the case where all instances are relabelled still constitute one of our future work.

References

1. Zliobaite, I., Bifet, A., Pfahringer, B., Holmes, G.: Active learning with drifting streaming data. IEEE Trans. Neural Netw. Learn. Syst. **25**(1), 27–39 (2014)
2. Kremer, J., Steenstrup Pedersen, K., Igel, C.: Active learning with support vector machines. In: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, pp. 313–326 (2014)
3. Huang, L., Liu, Y., Liu, X., Wang, X., Lang, B.: Graph-based active semi-supervised learning: a new perspective for relieving multi-class annotation labor. In: IEEE International Conference on Multimedia and Expo, pp. 1–6 (2014)
4. Kushnir, D.: Active-transductive learning with label-adapted kernels. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 462–471 (2014)
5. Settles, B.: Active learning. In: Synthesis Lectures on Artificial Intelligence and Machine Learning, pp. 1–114 (2012)
6. Bouguelia, M-R., Belaïd, Y., Belaïd, A.: A stream-based semi-supervised active learning approach for document classification. In: IEEE International Conference on Document Analysis and Recognition, pp. 611–615 (2013)
7. Goldberg, A., Zhu, X., Furger, A., Xu, J.M.: OASIS: online active semi-supervised learning. In: AAAI Conference on Artificial Intelligence, pp. 1–6 (2011)
8. Dasgupta, S.: Coarse sample complexity bounds for active learning. In: Neural Information Processing Systems (NIPS), pp. 235–242 (2005)
9. Frénay, B., Verleysen, M.: Classification in the presence of label noise: a survey. IEEE Trans. Neural Netw. Learn. Syst. **25**(5), 845–869 (2013)
10. Zhu, X., Zhang, P., Wu, X., He, D., Zhang, C., Shi, Y.: Cleansing noisy data streams. In: IEEE International Conference on Data Mining, pp. 1139–1144 (2008)

11. Rebbapragada, U., Brodley, C.E., Sulla-Menashe, D., Friedl, M.A.: Active label correction. In: IEEE International Conference on Data Mining, pp. 1080–1085 (2012)
12. Fang, M., Zhu, X.: Active learning with uncertain labeling knowledge. Pattern Recogn. Lett. **43**, 98–108 (2013)
13. Tuia, D., Munoz-Mari, J.: Learning user's confidence for active learning. IEEE Trans. Geosci. Remote Sens. **51**(2), 872–880 (2013)
14. Sheng, V.S., Provost, F., Ipeirotis, P.G.: Get another label? improving data quality and data mining using multiple noisy labelers. In: ACM Conference on Knowledge Discovery and Data Mining, pp. 614–622 (2008)
15. Ipeirotis, P.G., Provost, F., Sheng, V.S., Wang, J.: Repeated labeling using multiple noisy labelers. In: ACM Conference on Knowledge Discovery and Data Mining, pp. 402–441 (2014)
16. Yan, Y., Fung, G.M., Rosales, R., Dy, J.G.: Active learning from crowds. In: International Conference on Machine Learning, pp. 1161–1168 (2011)
17. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput. Appl. Math. **20**, 53–65 (1987)
18. Sun, S.: A survey of multi-view machine learning. Neural Comput. Appl. **23**(7–8), 2031–2038 (2013)
19. Gamberger, D., Lavrac, N., Dzeroski, S.: Noise elimination in inductive concept learning: a case study in medical diagnosis. In: Arikawa, Setsuo, Sharma, A.K. (eds.) ALT 1996. LNCS, vol. 1160, pp. 199–212. Springer, Heidelberg (1996)
20. Brodley, C.E., Friedl, M.A.: Identifying mislabeled training data. J. Artif. Intell. Res. **11**, 131–167 (1999)
21. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)